remote sensing

MDPI

Article

TVLightFormer: A Lightweight Cross-Modal Transformer for Language-Guided Target Localization in SAR Imagery

Yuqiao Zhong †, Haoqi Quan †, Chenyu Nie, Yingmei Wei * and Yanming Guo

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; zhongyuqiao23@nudt.edu.cn (Y.Z.); quanhaoqi23@nudt.edu.cn (H.Q.); niechenyu23@nudt.edu.cn (C.N.); guoyanming@nudt.edu.cn (Y.G.)

* Correspondence: weiyingmei@nudt.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- We proposed TVLightFormer, which is a lightweight cross-modal Transformer for language-guided target localization in SAR imagery. The innovation is not the isolated use of MobileNetV3, TinyBERT, GQA, or LFPN but rather their coordinated SAR-aware integration for preserving weak scattering cues under edge-deployment constraints.
- Under a unified language-guided localization protocol across five remote-sensing datasets (SOMA-1M, ATRNet-STAR, GAIA, MLRSNet, and SODAS), TVLightFormer achieves a strong accuracy–efficiency trade-off, reaching an average mIoU of 69.8% with only 27.4 M parameters and 9.7 GFLOPs, and it runs at 25.2 FPS (A6000).

What is the implication of the main finding?

- The proposed framework provides a balanced solution for SAR-oriented language-guided localization, while the revised discussion clarifies its robustness boundaries under speckle, geometric distortions, loss-function assumptions, and weakly supervised protocol adaptations.

Abstract

We study language-guided target localization in synthetic aperture radar (SAR) imagery for deployment on resource-constrained platforms. Existing vision-language models either rely on heavy backbones unsuitable for edge devices or are designed for natural images, overlooking SAR-specific characteristics such as speckle noise, weak scattering responses, and geometric distortions. The proposed model, TVLightFormer, combines a lightweight dual-modal encoder (MobileNetV3 and TinyBERT) with a grouped-query attention (GQA) mechanism for efficient cross-modal interaction and an activation-free lightweight feature pyramid network (LFPN) to handle scale variation while preserving weak scattering signals. The individual modules are not claimed as newly invented components; the main contribution lies in their SAR-aware integration for edge-oriented cross-modal localization. We evaluate the model on five remote sensing datasets—SOMA-1M, ATRNet-STAR, GAIA, MLRSNet, and SODAS—under a unified localization setting, and we explicitly discuss the limitations introduced by weak or scene-level annotations. The results show that TVLightFormer achieves a favorable trade-off between accuracy and efficiency, reaching an average mIoU of 69.8% with 27.4 M parameters and 9.7 GFLOPs. Ablation studies quantify the contribution of each component. The model is suited for edge-oriented scenarios where computational resources are limited. We also provide a critical analysis of failure cases, SAR-specific disturbance factors, loss-function choices, and dataset-protocol sensitivity.

Check for updates

Academic Editor: Dusan Gleich

Received: 27 March 2026
Revised: 29 April 2026
Accepted: 30 April 2026
Published: 4 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: multimodal learning; SAR image analysis; lightweight model; target localization; transformer

1. Introduction

Language-guided target localization aims to identify the object or region in an image that corresponds to a textual query. In recent years, this problem has become an important topic in natural-image vision-language research because language can provide target-specific constraints that are difficult to express using visual information alone. In remote sensing, and especially in synthetic aperture radar (SAR) imagery, this problem is particularly meaningful. SAR sensors can operate day and night and under adverse weather conditions, making them valuable for persistent observation in scenarios such as infrastructure inspection, disaster response, and large-area surveillance. However, the same properties that make SAR attractive in practice also make it difficult to interpret. Unlike optical imagery, SAR images do not rely on color or conventional texture cues. Instead, they encode backscatter responses shaped by target geometry, material properties, viewing angle, and imaging configuration. As a result, SAR images often contain speckle noise, side-lobe interference, layover, shadow, and geometric distortion, all of which complicate reliable target localization [1]. Recent advances in geometric-aware remote sensing, such as affine Lie group enhanced detectors [2], have shown promise for handling geometric variations in UAV road scenes, but they focus on single-modality visual understanding rather than cross-modal language-guided localization.

Under these conditions, text can serve as an important source of disambiguating information. A language query can specify structural attributes, contextual relations, or semantic categories that are difficult to infer from radar intensity patterns alone. For example, a phrase such as "the transmission tower near the road" or "the long vehicle at the edge of the scene" provides additional constraints beyond the image itself. In cluttered SAR scenes, this information can help suppress false responses from visually similar background structures. More generally, language-guided localization offers a practical way to couple human intent with automated SAR interpretation. Rather than asking a model to detect all possible objects, one can formulate a targeted query and retrieve only the region relevant to the task at hand. This interaction mode is attractive for real-world systems because it is more flexible than fixed-category detection and more interpretable than purely implicit retrieval.

Despite this promise, language-guided SAR localization remains difficult for at least three reasons. First, SAR targets are often small relative to the full scene, and their appearance can vary substantially with scale, acquisition mode, and imaging geometry. Adaptive deformation-learning and multi-scale integration [3] have been proposed for remote sensing object detection, but they do not incorporate language guidance. A lightweight model that aggressively compresses features may lose the weak responses required for precise localization. Second, cross-modal interaction is computationally expensive. Most vision-language models use standard multi-head attention or large transformer stacks to align image and text features. These designs can be effective, but they are not well suited to deployment on resource-constrained hardware. Third, localization requires both semantic alignment and spatial precision. A model must not only understand the text but also preserve enough local structure to regress an accurate bounding box. In SAR imagery, this is nontrivial because informative signals may be sparse, low-intensity, or distributed across multiple scattering components.

These challenges become more severe on edge platforms such as drones, satellites, and other embedded systems. In such settings, computational resources, memory budget,

power consumption, and latency are all constrained. Existing vision-language models developed for natural images are often too large for these deployment conditions. Many rely on ViT-style visual encoders and large language backbones with tens or hundreds of millions of parameters, and in some cases much more. Even when these models achieve strong accuracy, their computational cost makes real-time inference difficult on edge hardware. In addition, their design assumptions are not always compatible with SAR data. Lightweight vision-language methods for natural images often use token pruning, low-rank attention, or compact visual encoders under the assumption that color, texture, and dense semantic patterns are the primary carriers of information [4,5]. In SAR imagery, however, visually weak regions may still contain diagnostically important scattering cues. A compression strategy that works well for optical images may therefore remove signals that remain important in radar scenes.

Recent SAR-language models partially address the modality gap by learning joint representations from image–text pairs [6,7]. These works show that language can indeed improve SAR understanding. However, most of them prioritize representation capacity over efficiency. Their architectures typically employ heavy visual backbones, relatively deep transformer modules, or expensive cross-modal alignment mechanisms. Such choices are reasonable when the goal is to maximize benchmark performance on server-class hardware, but they are less suitable for lightweight deployment. Moreover, many existing SAR-language studies emphasize global representation learning, caption alignment, or retrieval, while the requirements of target localization are more demanding in terms of spatial detail. A model for localization must preserve structural cues at multiple scales and must remain sensitive to low-intensity returns that could easily be suppressed during compression. In other words, a smaller model is not sufficient by itself; the lightweight design must also be compatible with the physical and statistical properties of SAR imagery.

This observation motivates this paper. Rather than directly reducing the size of a generic vision-language model, we focus on the design of a lightweight cross-modal transformer tailored to language-guided SAR target localization. Our starting point is that edge-oriented SAR localization requires more than parameter reduction. It requires a coordinated design across the visual encoder, text encoder, cross-modal interaction module, and decoder. From this perspective, three requirements become central. First, the visual encoder should preserve geometric structure and local scattering patterns while remaining computationally efficient. Second, cross-modal interaction should be cheaper than standard multi-head attention, but it should still provide sufficient capacity for aligning visual and textual cues. Third, the decoder should fuse multi-scale features without discarding weak responses that may be informative for SAR localization.

Based on these considerations, we propose TVLightFormer, which is a lightweight cross-modal transformer for language-guided target localization. The model combines four design choices. First, it adopts MobileNetV3 [8] as the visual backbone. Compared with heavier alternatives, MobileNetV3 offers a favorable balance between efficiency and feature quality through depthwise separable convolutions and squeeze-and-excitation blocks. In our setting, this backbone provides a practical compromise between computational cost and structural feature preservation. Second, the model uses TinyBERT [9] as the text encoder. TinyBERT retains much of the semantic capacity of BERT-base while reducing the model size, which is important for end-to-end efficiency. Third, we introduce grouped-query attention (GQA) [10] for cross-modal interaction. By sharing key and value projections across groups of attention heads, GQA reduces memory and computation relative to standard multi-head attention while remaining more expressive than multi-query attention. Finally, we design an activation-free lightweight feature pyramid network (LFPN) for multi-scale fusion. This decoder avoids ReLU nonlinearities in order to preserve weak

responses that may otherwise be suppressed. The choice is motivated by the characteristics of SAR backscatter and is supported by our ablation results.

It should be noted that MobileNetV3, TinyBERT, and GQA are established modules rather than SAR-specific inventions. The novelty of TVLightFormer is therefore not an isolated module-level claim. Instead, the proposed contribution is a modality-aware architectural combination designed around three SAR localization requirements: retaining low-amplitude scattering cues rather than pruning them as redundant tokens, using query-conditioned attention to suppress clutter and sidelobe-like background responses at a manageable cost, and recovering multi-scale spatial details through a decoder that avoids unnecessary response clipping. This clarification is important because SAR target localization differs from natural-image grounding not only in input appearance but also in the physical meaning of weak responses, shadows, and geometric distortions.

Another practical difficulty in this paper is evaluation. Public remote-sensing datasets differ substantially in sensing modality, annotation type, and original task definition. Some are built for recognition, some for captioning or cross-modal alignment, and some for scene understanding or temporal analysis. To examine the generality of the proposed design, we evaluate TVLightFormer on five recent public datasets: SOMA-1M [11], ATRNet-STAR [12], GAIA [13], MLRSNet [14], and SODAS [15]. Because these datasets were not originally designed under a single visual-grounding protocol, we construct a unified language-guided localization setting. Specifically, we derive textual prompts from available annotations such as class labels, captions, or metadata, and we use the provided bounding boxes when available or derive target regions from segmentation labels when necessary. This unified protocol does not eliminate the inherent differences among datasets, but it allows us to compare models under a shared formulation and to assess how well a lightweight cross-modal design transfers across heterogeneous remote-sensing settings.

The goal of this paper is therefore not simply to report another compact model. Rather, we ask a more specific question: can a carefully designed lightweight cross-modal architecture retain useful localization accuracy on SAR-related tasks while substantially reducing the cost of deployment? We answer this question through a model-level design study, large-scale empirical comparison, and detailed ablation analysis. Our contributions can be summarized as follows:

1. We propose TVLightFormer, which is a lightweight cross-modal transformer for language-guided target localization that is designed with edge deployment in mind. The model contains 27.4 M parameters and requires 9.7 GFLOPs while maintaining competitive localization performance.
2. We clarify a SAR-aware lightweight design principle: the model avoids aggressive token pruning, uses efficient query-conditioned interaction to reduce clutter sensitivity, and employs an activation-free LFPN to preserve weak multi-scale scattering responses during decoding.
3. We establish a unified language-guided localization protocol across five public remote-sensing datasets and evaluate TVLightFormer against ten representative baselines, including both generic visual-grounding models and recent vision-language systems. We also explicitly discuss protocol limitations, especially the scene-level adaptation of MLRSNet.
4. We provide a critical empirical analysis through ablation studies, dataset-specific discussion, SAR-specific disturbance discussion, loss-function analysis, and failure-case examination, clarifying not only where the proposed model works well but also where its limitations remain.

The remainder of this paper is organized as follows. Section 2 reviews related work on lightweight vision-language models and SAR cross-modal understanding with an emphasis

on their limitations for edge-oriented localization. Section 3 introduces the architecture of TVLightFormer and explains the rationale behind its main components. Section 4 describes the experimental protocol, datasets, baselines, and results. Section 5 concludes the paper and discusses possible directions for future work.

2. Related Work

2.1. Lightweight Vision-Language Models

The rapid growth of vision-language models has led to strong performance on a wide range of tasks, including image captioning, visual question answering, grounding, and multimodal retrieval. However, these gains often come at the cost of large visual encoders, deep cross-modal transformers, and high inference latency. As a result, lightweight vision-language modeling has become an active research direction. Existing methods can be broadly grouped into three categories: token pruning, efficient architecture redesign, and training-time compression. While all three aim to reduce computational cost, they make different assumptions about where redundancy lies in multimodal pipelines.

Token pruning methods attempt to reduce cost by discarding visual tokens that appear unimportant for the downstream task. ATP-LLaVA [16] prunes tokens according to adaptive thresholds derived from attention statistics, thereby reducing the cost of visual processing in large multimodal models. VScan [17] learns a gating mechanism that selects tokens conditioned on the text query. The general intuition behind such methods is appealing: not all visual patches contribute equally, and redundant tokens can be removed with limited accuracy loss. This strategy has shown good results on natural-image benchmarks, especially when the relevant objects occupy only a subset of the scene. However, its transferability to SAR imagery is not obvious. In radar images, weak or sparse responses can still carry discriminative structural information. A point-like scattering return that appears visually minor may correspond to an essential part of a vehicle, a tower arm, or another target component. If a pruning mechanism relies mainly on feature magnitude or generic saliency, it may remove tokens that remain important for localization. In this sense, token pruning is efficient, but its notion of redundancy is often inherited from RGB imagery and may not be reliable in SAR settings.

Efficient architecture redesign takes a different path. Instead of pruning features after they are produced, these methods redesign the model itself to reduce complexity from the outset. MiniVLM [18] uses smaller hidden dimensions and fewer layers to build a compact multimodal architecture. Xmodel-VLM [5] introduces factorized cross-modal interactions that reduce the cost of attention. Such approaches are attractive because they offer a cleaner efficiency profile and avoid some of the instability introduced by dynamic pruning. They also tend to be easier to deploy because their inference graph is simpler and more predictable. Nevertheless, most compact architectures are still developed and validated on natural-image tasks. Their visual encoders are optimized for RGB inputs, and their performance depends heavily on the persistence of color and texture cues. SAR imagery differs not only in appearance but also in how relevant information is distributed spatially and statistically. A compact model that works well on COCO or RefCOCO may not preserve the same information in radar scenes. This does not mean these designs are unusable, but it does suggest that their efficiency gains cannot be assumed to transfer unchanged across modalities. Heterogeneous binary pixel difference networks [19] explore lightweight feature extraction for remote sensing detection, yet they are not designed for cross-modal interaction.

Training-time compression methods reduce model size by using a larger model during learning and a smaller model during inference. Common strategies include knowledge distillation, quantization, network pruning, and neural architecture search. NVILA [20] uses

architecture search to identify efficient multimodal subnetworks, while EfficientVLM [4] compresses a large teacher into a smaller student through distillation. These methods can produce strong compact models, especially when large-scale pre-training data and compute resources are available. However, their practical value depends on the availability of suitable teacher models and task-aligned data. In the SAR domain, such conditions are often harder to meet. Compared with optical vision-language research, SAR-language resources remain limited. Low-data-dependence networks such as DGPF-RENet [21] have been developed for hyperspectral classification, but extending this principle to cross-modal SAR localization remains underexplored. And the domain gap between generic web-scale pre-training data and SAR imagery is substantial. As a result, training-time compression may reduce model size without fully solving the modality adaptation problem. A compressed student can still inherit the biases or failure modes of a teacher that was not designed for radar data.

Taken together, existing lightweight VLM methods provide useful building blocks, but none directly resolves the requirements of edge-oriented SAR localization. Token pruning may remove weak yet informative radar responses. Compact architecture redesign often assumes RGB-style semantics. Training-time compression can reduce parameter count, but it does not automatically preserve SAR-relevant cues. More importantly, most lightweight VLM studies optimize for classification, retrieval, or general multimodal reasoning rather than precise localization. The demands of language-guided localization are different: the model must remain spatially sensitive, align text with localized visual structure, and do so under limited computational resources. This motivates the need for a modality-aware lightweight design rather than a direct adaptation of generic compression strategies.

2.2. SAR Image Analysis and Cross-Modal Understanding

Research on SAR image analysis has traditionally focused on detection, classification, segmentation, and change analysis. Earlier methods relied heavily on handcrafted features and statistical models, such as CFAR-style detectors, scattering descriptors, and polarimetric decomposition techniques. These approaches were often interpretable and efficient, but they had limited adaptability to complex scenes or task variation. With the rise of deep learning, convolutional neural networks became the dominant paradigm for SAR recognition and detection because they could learn task-specific representations directly from data. CNN-based models improved performance on a range of SAR tasks, but they also exposed several limitations. Their receptive fields are often local, their handling of scale variation is imperfect, and their ability to model long-range semantic relations is relatively constrained. These weaknesses become more pronounced when the target is small, the scene is cluttered, or the task requires contextual reasoning. Recent works such as texture-aware glance-and-gaze networks [22] have advanced hyperspectral classification with low training samples, yet they are designed for single-modal understanding and do not address cross-modal localization.

Transformer-based approaches were introduced in part to address these issues. By using global attention, transformers can model long-range dependencies more effectively than standard CNNs. Deformable DETR-style methods, for example, adapt sparse attention to object detection and have been explored for SAR small-object localization [23]. These models can improve robustness in scenes where targets are sparse relative to the full image. Still, they are usually designed for single-modality visual analysis and do not address the interaction between SAR imagery and natural language. Their gains mainly concern visual representation and detection structure rather than cross-modal grounding.

Cross-modal understanding of SAR imagery has emerged more recently with the release of large-scale image–text datasets. SARLANG-1M [24], SARCLIP-1M [25], and SARChat-Bench-2M [26] have made it possible to study retrieval, captioning, and multimodal pre-training directly in the SAR domain. These resources are important because they reduce the reliance on purely optical-language priors and provide a basis for learning radar-language correspondences. Building on such datasets, models like SARCLIP [27] use contrastive learning to align SAR images with text in a shared embedding space, enabling cross-modal retrieval and zero-shot transfer. SAR-KnowLIP [28] further introduces geographic or contextual knowledge to improve alignment. These works demonstrate that SAR-language modeling is feasible and that radar imagery can benefit from textual supervision.

However, the dominant emphasis in current SAR-language research is still on representation learning rather than efficient grounding. Retrieval-oriented models optimize global alignment between an image and a caption. Captioning models prioritize language generation. General SAR-language encoders seek broad multimodal understanding. These goals are valuable, but they are not identical to localization. In language-guided localization, the model must map a textual query to a specific region, often under tight spatial tolerance. This requires a stronger preservation of local structure and more precise multi-scale fusion than is necessary for global retrieval or caption matching. A model that performs well at image-level alignment may still struggle with accurate box regression.

A second limitation is model scale. Many recent SAR-language systems inherit architectural choices from large optical models. Heavy visual encoders, deep transformer stacks, and high-dimensional embeddings are common because they improve capacity and benchmark performance. For example, existing SAR-oriented multimodal models often use backbones such as ResNet-101, ViT-L, or similarly heavy encoders [6,7]. Such designs are understandable in early-stage benchmark development, where the priority is often to establish performance ceilings. Yet they are difficult to deploy on edge hardware, where computation power and memory are constrained. This tension between representation power and deployment cost is central to this paper.

A third limitation is that lightweight SAR models have so far addressed only parts of the problem. SAR-TinySNN [29], for instance, explores a compact spiking architecture for SAR localization, which is useful from an efficiency perspective. But its design does not systematically combine efficient cross-modal attention with dedicated multi-scale fusion. More broadly, lightweight SAR studies often optimize one component in isolation: either the visual encoder, the event-driven mechanism, or the classifier head. Fewer works consider the full cross-modal pipeline as a coupled design problem. In practice, however, efficiency bottlenecks arise jointly from feature extraction, attention, and decoding. Improving only one stage may not yield the best end-to-end trade-off.

These observations suggest that there is still a gap between current SAR-language research and the requirements of deployable localization systems. Existing work has established the feasibility of SAR-language alignment but not yet the design principles for lightweight cross-modal localization under edge constraints. This gap is not merely quantitative. It is not only that existing models are too large; it is also that their internal assumptions are not always well matched to the needs of localization in SAR imagery. A suitable model should explicitly consider the preservation of weak responses, the cost of attention, and the need for multi-scale decoding.

2.3. Critical Analysis and Research Gap

The above review points to a clear but underexplored problem. On one side, lightweight vision-language models provide useful strategies for reducing cost, but they

are mostly developed for natural images and may not preserve SAR-relevant cues. On the other side, SAR-specific multimodal models improve alignment between radar imagery and text, but they often rely on heavy architectures and are not optimized for edge deployment. Between these two lines of work lies the practical requirement addressed in this paper: an efficient cross-modal model that remains sensitive to SAR structure and supports target-level localization rather than only image-level representation learning.

This gap can be analyzed along three dimensions. First, there is a modality gap. Natural-image VLMs assume dense semantics, color information, and strong texture patterns. SAR imagery violates these assumptions. Weak responses, point-like scatterers, and geometric distortions play a larger role, which means that aggressive compression strategies can be risky. Second, there is a task gap. Much of the multimodal literature, both generic and SAR-specific, focuses on retrieval, captioning, or coarse semantic understanding. Localization is more demanding because it requires fine spatial correspondence between text and image regions. Third, there is a deployment gap. Even when a model performs well, its computational profile may remain incompatible with real-time operation on embedded platforms.

This paper is motivated by the intersection of these three gaps. We do not assume that a compact natural-image VLM can be applied to SAR without modification. We also do not assume that a high-capacity SAR-language model can simply be compressed after the fact. Instead, we argue that lightweight SAR localization should be treated as a dedicated design problem. This is why TVLightFormer combines a compact visual encoder, a compact language encoder, efficient grouped-query attention, and a decoder tailored to preserving weak multi-scale radar responses. In this sense, the contribution of this paper is not a new foundation model or a new pre-training paradigm. Rather, it is a focused architectural study of how lightweight cross-modal localization can be implemented in a way that is better aligned with the characteristics of SAR imagery and the constraints of edge deployment.

For clarity, Table 1 summarizes the representative models discussed above and highlights the distinction between generic lightweight VLMs, SAR-language representation models, and the proposed edge-oriented localization framework. The central question is not which family is universally superior but rather which design principles are most appropriate when localization accuracy, computational efficiency, and SAR-specific robustness must all be considered at the same time.

Table 1. Comparison of representative models. (Params in millions, FLOPs in billions, SAR-specific indicates if model is designed/evaluated on SAR data.)

Model	Params (M)	FLOPs (G)	SAR-Specific	Edge-Suitable
ATP-LLaVA [16]	∼200	∼30	No	No
MiniVLM [4]	56.3	12.2	No	Partial
Xmodel-VLM [5]	85.4	15.1	No	Partial
SARCLIP [7]	307.2	64.7	Yes	No
SAR-TinySNN [29]	12.1	3.2	Yes	Yes
TACMT [30]	216.8	16.1	Yes	No
TVLightFormer (ours)	27.4	9.7	Yes	Yes

3. Methodology

This section presents the architecture of TVLightFormer, which is a lightweight cross-modal transformer designed for language-guided target localization under remote-sensing and SAR-oriented settings. The design is guided by three practical considerations. First, the model should remain compact enough for edge-oriented deployment, which requires a low parameter count, moderate FLOPs, and predictable inference cost. Second, it should

preserve the geometric and structural cues that are particularly important in SAR imagery, where informative signals may be sparse, weak, or distributed across multiple scattering components. Third, it should support accurate localization rather than only global image–text matching, which means that cross-modal interaction and multi-scale decoding must remain spatially informative despite the lightweight design.

Figure 1 illustrates the overall pipeline. Given a SAR image $\mathbf{I} \in \mathbb{R}^{H\times W\times 1}$ and a text query $\mathbf{T}$, the model first extracts compact visual and textual representations using a lightweight dual-modal encoder. These representations are then fused by a transformer encoder equipped with grouped-query attention (GQA), which reduces the cost of cross-modal interaction relative to standard multi-head attention. Finally, a lightweight feature pyramid decoder integrates low- and high-level features to produce a bounding box prediction $\mathbf{b} \in \mathbb{R}^4$.

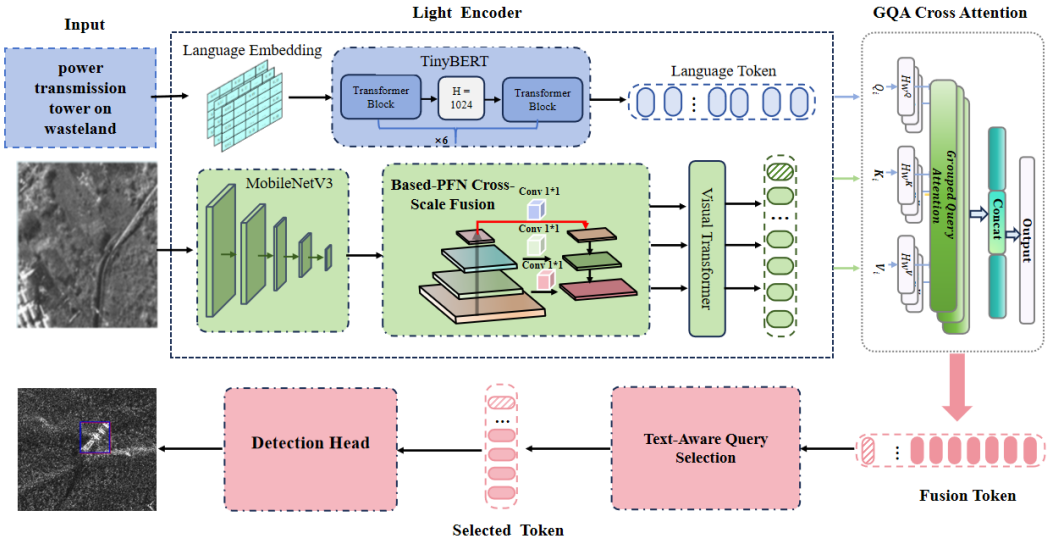


Figure 1. Overall architecture of TVLightFormer. A SAR image and a text query are first encoded by lightweight visual and textual encoders. The resulting features interact through a transformer encoder with grouped-query attention. A lightweight feature pyramid decoder then fuses multi-scale information and predicts the target bounding box.

3.1. Overall Formulation

The task considered in this paper is language-guided target localization. For an input image $\mathbf{I}$ and a query $\mathbf{T}$, the model predicts a normalized bounding box

$$\mathbf{b} = (x_{\min}, y_{\min}, x_{\max}, y_{\max}), \tag{1}$$

where all coordinates are normalized to $[0, 1]$ with respect to the image size. The architecture is composed of three stages:

1. A lightweight dual-modal encoder for extracting compact visual and textual features;
2. A cross-modal interaction module based on GQA for efficient image–text alignment;
3. A lightweight feature pyramid decoder for recovering spatial detail and refining localization.

A key principle of the design is that efficiency should be achieved at multiple stages rather than by compressing only one component. Reducing the backbone alone often harms visual discrimination; simplifying only the decoder leaves the attention bottleneck unresolved; compressing only the text encoder may weaken query understanding. TVLightFormer therefore adopts coordinated lightweighting across the full pipeline.

SAR-aware design rationale.

The proposed architecture is designed with the statistics of SAR imagery in mind. Speckle noise and sidelobe responses can create high-contrast but task-irrelevant structures, whereas true target evidence may appear as sparse or low-amplitude scattering components. For this reason, TVLightFormer does not use magnitude-based token pruning, because visually weak regions may still be informative. Instead, the compact visual encoder keeps a local convolutional structure, the text branch provides a query constraint, and GQA injects this constraint into the visual tokens with lower cost than standard multi-head attention. The LFPN further fuses shallow spatial features with high-level query-conditioned features so that small target structures and shadow boundaries are less likely to be lost after downsampling. In this sense, the architecture is not a generic stacking of lightweight modules but rather a coordinated design intended to preserve the SAR cues most likely to be damaged by aggressive compression.

3.2. Lightweight Dual-Modal Encoder

3.2.1. Visual Encoder

We use **MobileNetV3Large** [8] as the visual backbone. The choice is motivated by the need to preserve useful structural features at relatively low computational cost. Compared with heavy CNN or ViT-based backbones, MobileNetV3 offers a practical balance between latency and representation quality through inverted residual blocks, depthwise separable convolutions, and squeeze-and-excitation (SE) modules. These properties are attractive for our setting because they reduce computation while still allowing local geometric patterns to be represented.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H_i \times W_i \times C_{in}}$, a depthwise separable convolution decomposes a standard convolution into a depthwise convolution followed by a pointwise convolution:

$$\mathbf{Y}_{dw} = \mathrm{DWConv}(\mathbf{X}, \mathbf{W}_{dw}), \qquad \mathbf{Y} = \mathrm{Conv}_{1\times 1}(\mathbf{Y}_{dw}, \mathbf{W}_{pw}), \tag{2}$$

where $\mathbf{W}_{dw} \in \mathbb{R}^{K\times K\times C_{in}}$ and $\mathbf{W}_{pw} \in \mathbb{R}^{1\times 1\times C_{in}\times C_{out}}$ denote the depthwise and pointwise kernels, respectively. This factorization reduces the computational cost from

$$\mathcal{O}(C_{in} \cdot C_{out} \cdot K^2) \tag{3}$$

for a standard convolution to

$$\mathcal{O}(C_{in} \cdot K^2 + C_{in} \cdot C_{out}), \tag{4}$$

which becomes particularly beneficial when the channel dimension is moderate and the model must run under strict resource constraints.

MobileNetV3 further employs SE modules to recalibrate channel responses. For an intermediate feature map $\mathbf{Y}$, the SE operation is written as

$$\mathbf{z} = \mathrm{GAP}(\mathbf{Y}) \in \mathbb{R}^{C_v}, \qquad \boldsymbol{\alpha} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})), \tag{5}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, δ is the ReLU activation, σ is the sigmoid function, and $\mathbf{W}_1 \in \mathbb{R}^{C_v/r\times C_v}$ and $\mathbf{W}_2 \in \mathbb{R}^{C_v\times C_v/r}$ are learnable matrices with reduction ratio $r = 4$. The recalibrated output is then

$$\mathbf{V}_0 = \boldsymbol{\alpha} \odot \mathbf{Y}, \tag{6}$$

where $\odot$ denotes channel-wise multiplication.

The final output of the last convolutional block is a visual feature map

$$\mathbf{V}_0 \in \mathbb{R}^{H_f \times W_f \times 192}, \tag{7}$$

where typically $H_f = H/32$ and $W_f = W/32$. We set the visual embedding dimension to 192 to match the hidden size of the textual branch and the cross-modal transformer, thereby avoiding additional high-dimensional projection layers. For single-channel SAR inputs, the first convolutional stem is adapted to the input channel number rather than relying on RGB color cues. This is important because the subsequent feature extraction must depend on the local backscatter structures, edges, and scattering contrasts instead of color-texture priors learned from natural images.

The choice of MobileNetV3 over alternatives such as EfficientNet-Lite or GhostNet is not arbitrary. For language-guided localization, the visual encoder must do more than provide globally discriminative features; it must also preserve enough local structure for the downstream decoder to recover a spatially precise target region. In preliminary experiments on a subset of SOMA-1M, MobileNetV3 provided slightly better localization accuracy than EfficientNet-Lite under similar FLOPs, suggesting that it offers a more favorable balance between efficiency and structural feature quality in our setting. We do not claim that MobileNetV3 is universally optimal for SAR localization, but it provides a robust and deployment-friendly starting point.

3.2.2. Textual Encoder

For language encoding, we adopt TinyBERT [9] with 6 transformer layers, hidden size 192, and 6 attention heads. The choice of a compact language encoder is motivated by two considerations. First, the textual queries used in language-guided localization are often shorter and structurally simpler than the long-form prompts used in general multimodal dialogue or reasoning tasks. Second, a large language encoder can easily dominate the computational budget of the full model, which is undesirable in an edge-oriented design.

Let $\mathbf{T} = \{t_1, \ldots, t_L\}$ denote the input token sequence. After tokenization, the sequence is embedded into

$$\mathbf{E}^{(0)} = \mathbf{E}_t + \mathbf{P}, \tag{8}$$

where $\mathbf{E}_t \in \mathbb{R}^{L \times 192}$ is the token embedding matrix and $\mathbf{P} \in \mathbb{R}^{L \times 192}$ is the positional encoding. Each transformer layer consists of multi-head self-attention (MHSA) and a feed-forward network (FFN), together with residual connections and layer normalization:

$$\mathbf{H}^{(l)} = \mathrm{MHSA}(\mathrm{LN}(\mathbf{E}^{(l-1)})) + \mathbf{E}^{(l-1)}, \tag{9}$$

$$\mathbf{E}^{(l)} = \mathrm{FFN}(\mathrm{LN}(\mathbf{H}^{(l)})) + \mathbf{H}^{(l)}. \tag{10}$$

The final representation associated with the `[CLS]` token is used as the global sentence embedding:

$$\mathbf{T}_0 \in \mathbb{R}^{192}. \tag{11}$$

Compared with larger alternatives such as DistilBERT or ALBERT, TinyBERT offers a stronger efficiency profile while preserving enough semantic capacity for short grounding queries. In our validation experiments on GAIA, its mIoU remained within a narrow margin of larger compact language models, despite using substantially fewer parameters. We therefore use TinyBERT as a pragmatic compromise between semantic representation and deployment cost.

A design choice worth noting is that we use the final global sentence embedding $\mathbf{T}_0$ rather than the full token sequence for cross-modal fusion. This reduces the sequence length entering the cross-modal transformer and therefore lowers memory and attention cost.

The trade-off is that token-level linguistic detail is compressed into a single sentence vector. In our setting, this is acceptable because most queries are relatively short and category- or attribute-focused. However, this choice may become limiting for longer or syntactically complex descriptions, which we revisit in the discussion of failure cases.

3.3. Cross-Modal Interaction with GQA

After extracting the visual and textual features, we perform cross-modal interaction through a transformer encoder. The visual feature map $\mathbf{V}_0$ is first flattened into a sequence of $N = H_f W_f$ visual tokens:

$$\mathbf{V} \in \mathbb{R}^{N \times 192}. \tag{12}$$

We then concatenate the global text token $\mathbf{T}_0$ with the visual tokens to obtain the multi-modal sequence

$$\mathbf{X}_0 = [\mathbf{T}_0; \mathbf{V}] \in \mathbb{R}^{(N+1) \times 192}. \tag{13}$$

This sequence is passed through a 6-layer transformer encoder. The role of this stage is not merely to fuse modalities in a generic sense but rather to inject query-dependent context into the visual representation before decoding. In other words, the model should learn not only what is present in the image but also from which part of the image is relevant to the given text.

A major computational bottleneck in this stage is multi-head attention. For an input $\mathbf{X} \in \mathbb{R}^{S \times d}$, standard MHA computes separate query, key, and value projections for each attention head. Its complexity scales quadratically with the sequence length S, which is problematic when the visual feature map is flattened into many tokens. To reduce this cost, we replace MHA with **grouped-query attention (GQA)** [10].

Let the number of attention heads be h, and let these heads be partitioned into g groups with $g < h$. In GQA, different heads retain separate query projections, but heads within the same group share key and value projections. For input $\mathbf{X}$, the grouped projections are

$$\mathbf{Q}_j = \mathbf{X}\mathbf{W}_j^Q, \qquad \mathbf{W}_j^Q \in \mathbb{R}^{d \times d_q}, \tag{14}$$

$$\mathbf{K}_j = \mathbf{X}\mathbf{W}_j^K, \qquad \mathbf{W}_j^K \in \mathbb{R}^{d \times d_k}, \tag{15}$$

$$\mathbf{V}_j = \mathbf{X}\mathbf{W}_j^V, \qquad \mathbf{W}_j^V \in \mathbb{R}^{d \times d_v}, \tag{16}$$

where $d_q = d_k = d_v = d/h$. The attention output for group j is

$$\mathrm{Head}_j = \mathrm{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^\top}{\sqrt{d_k}}\right)\mathbf{V}_j. \tag{17}$$

The outputs of all groups are concatenated and projected:

$$\mathrm{GQA}(\mathbf{X}) = \mathrm{Concat}(\mathrm{Head}_1, \ldots, \mathrm{Head}_g)\mathbf{W}^O, \qquad \mathbf{W}^O \in \mathbb{R}^{d \times d}. \tag{18}$$

Relative to standard MHA, GQA reduces the number of distinct key/value projections from h to g. In our implementation, $h = 6$ and $g = 2$, meaning that the number of key/value projection groups is reduced by a factor of $g/h = 2/6$. Equivalently, the key/value projection cost is reduced by about 66.7% relative to fully independent per-head projections. Compared with multi-query attention (MQA), which uses a single shared key/value pair for all heads, GQA preserves more representational flexibility while remaining substantially cheaper than MHA. This is important in our setting because the image–text interaction must remain expressive enough to localize query-relevant regions not only to perform coarse image-level matching.

Each transformer block further contains an FFN with GELU activation, residual connections, and layer normalization:

$$\mathbf{X}'_l = \mathrm{LN}\big(\mathbf{X}_{l-1} + \mathrm{GQA}(\mathbf{X}_{l-1})\big), \tag{19}$$

$$\mathbf{X}_l = \mathrm{LN}\big(\mathbf{X}'_l + \mathrm{FFN}(\mathbf{X}'_l)\big). \tag{20}$$

After the final layer, we separate the updated visual tokens from the text token and reshape the visual tokens back into a 2D feature map:

$$\mathbf{F}_{\text{high}} \in \mathbb{R}^{H_f \times W_f \times 192}. \tag{21}$$

It is worth emphasizing that this cross-modal design favors efficiency over maximal linguistic granularity. Because we use a global text token, the interaction stage is computationally compact and well matched to the edge-oriented goal of this paper. The downside is that detailed phrase-level grounding may not be modeled as explicitly as in heavier token-level fusion frameworks. In our view, this is an acceptable compromise for the current task setting, but it also defines one boundary of the present architecture.

3.4. Lightweight Feature Pyramid Decoder (LFPN)

The visual feature map produced by the cross-modal transformer encodes query-conditioned semantics, but it is spatially coarse due to repeated downsampling in the visual encoder. Accurate localization therefore requires recovering finer spatial detail. To this end, we design a **lightweight feature pyramid decoder (LFPN)** that combines a high-level query-aware feature map with a shallower, higher-resolution visual feature map from the backbone.

Specifically, we extract an intermediate feature map

$$\mathbf{F}_{\text{low}} \in \mathbb{R}^{2H_f \times 2W_f \times 96} \tag{22}$$

from the expansion layer of the second-to-last MobileNetV3 bottleneck. This feature retains more spatial detail and local edge information than the final backbone output. The transformed high-level feature map $\mathbf{F}_{\text{high}}$ contains stronger semantic information but at lower spatial resolution. The purpose of LFPN is to fuse these two signals with minimal computational overhead. Unlike general-purpose scale fusion networks [31] that may introduce additional parameters, our LFPN is specifically designed to be activation-free and lightweight.

The decoder performs top–down fusion as follows:

$$\mathbf{F}^{\uparrow}_{\text{high}} = \mathrm{Upsample}_{2\times}(\mathbf{F}_{\text{high}}), \tag{23}$$

$$\mathbf{F}'_{\text{high}} = \mathrm{BN}\big(\mathrm{Conv}_{1\times1}(\mathbf{F}^{\uparrow}_{\text{high}})\big), \tag{24}$$

$$\mathbf{F}_{\text{fused}} = \mathbf{F}'_{\text{high}} + \mathbf{F}_{\text{low}}, \tag{25}$$

$$\mathbf{F}_{\text{dec}} = \mathrm{BN}\big(\mathrm{Conv}_{3\times3}(\mathbf{F}_{\text{fused}})\big). \tag{26}$$

The first 1×1 convolution aligns channel dimensions, the element-wise addition fuses high- and low-level information, and the final 3×3 convolution refines the fused representation.

A distinctive feature of LFPN is that we deliberately omit nonlinear activations such as ReLU. This choice is motivated by the characteristics of SAR-related imagery. In conventional CNN pipelines, ReLU improves nonlinearity and sparsity, but it also suppresses all negative activations. In our setting, some low-magnitude or sign-sensitive responses may still carry useful information for localization, especially when the target is characterized

by weak backscatter, shadow boundaries, or subtle edge structure. We therefore keep the decoder linear apart from batch normalization. In preliminary experiments, adding ReLU led to a small but consistent decrease in mIoU on SOMA-1M, suggesting that preserving the full response range is slightly more beneficial than enforcing stronger sparsity at this stage.

This design should not be interpreted as a claim that activation-free decoding is universally superior, nor do we claim that weak-response preservation is unique to SAR. Similar effects may occur in other modalities where low-intensity structures are meaningful, such as infrared or low-light imaging. Our claim is narrower: in the present SAR-oriented localization setting, the activation-free decoder better matches the need to retain weak scattering and shadow-boundary information after lightweight compression. This conclusion is supported by the decoder ablation in Section 4, but it remains an empirical design choice rather than a general theorem.

3.5. Prediction Head and Training Objective

The final decoder feature map $\mathbf{F}_{\text{dec}}$ is globally average pooled and fed into a lightweight regression head. Specifically, we compute

$$\mathbf{h} = \text{GELU}\big(\mathbf{W}_a \cdot \text{GAP}(\mathbf{F}_{\text{dec}}) + \mathbf{b}_a\big), \tag{27}$$

where $\mathbf{W}_a \in \mathbb{R}^{256\times 96}$ and $\mathbf{b}_a \in \mathbb{R}^{256}$. The final bounding box is then predicted as

$$\mathbf{b} = \mathbf{W}_b\mathbf{h} + \mathbf{b}_b, \tag{28}$$

with $\mathbf{W}_b \in \mathbb{R}^{4\times 256}$ and $\mathbf{b}_b \in \mathbb{R}^4$.

We train the model using a combination of Generalized IoU loss and Smooth L1 loss:

$$\mathcal{L} = \lambda_{\text{giou}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{L1}}\mathcal{L}_{\text{L1}}, \tag{29}$$

where $\lambda_{\text{giou}} = 2$ and $\lambda_{\text{L1}} = 5$. The GIoU term encourages overlap between predicted and target regions, while the L1 term stabilizes coordinate regression. This combination is commonly used in localization tasks because it balances geometric consistency with direct coordinate supervision.

For small-target SAR localization, this choice has limitations. When the ground-truth box contains a large proportion of background or speckle-dominated pixels, an IoU-family loss can become less sensitive to the physically meaningful scattering center of the target. We therefore additionally compare several regression losses in the loss-function ablation of Section 4.8.2, including single-loss and hybrid-loss settings. We considered more recent box-regression losses such as DIoU/CIoU-type formulations [32] as well as target-size-aware weighting. However, we retained the GIoU + Smooth L1 objective in the current version for two reasons. First, it provides a stable and widely used baseline for comparing architectural changes without introducing another confounding factor. Second, the Smooth L1 term partly compensates for the weakness of pure overlap supervision by directly penalizing coordinate displacement. The ablation results show that GIoU + Smooth L1 remains a strong and stable choice, while DIoU/CIoU based variants are competitive alternatives. Nevertheless, SAR-specific losses are an important direction for future work. A possible extension is to weight the localization loss using scattering saliency or local signal-to-clutter ratio, so that errors near physically meaningful scattering components are penalized more strongly than errors over homogeneous background.

4. Experiments and Analysis

This section evaluates TVLightFormer under a unified language-guided localization setting across five public remote-sensing datasets. Since these datasets were originally constructed for different tasks, the experiments are designed not only to compare accuracy but also to examine how a lightweight cross-modal architecture behaves when transferred across heterogeneous annotation formats, sensing conditions, and semantic granularities. We first describe the unified evaluation protocol and dataset adaptation strategy; then, we present the training setup and evaluation metrics, which is followed by comparisons with representative baselines, efficiency analysis, ablation studies, and dataset-specific failure analysis.

4.1. Datasets and Unified Protocol

A central challenge in this paper is that there is currently no single public benchmark that simultaneously covers SAR-focused language-guided localization, fine-grained target grounding, scene-level semantic grounding, and temporal remote-sensing variation under a consistent protocol. The five datasets considered here were originally released for different tasks, including image alignment, recognition, captioning, scene understanding, and crop mapping. Directly comparing models across such datasets is therefore nontrivial. To make the evaluation meaningful, we construct a unified language-guided localization protocol.

Under this protocol, each sample is represented as an image–text pair together with a target region. Text queries are derived from the annotations already provided by each dataset, including category labels, captions, metadata fields, or scene descriptions. Target regions are taken from existing bounding boxes when available or derived from segmentation masks or polygon annotations when necessary. For datasets whose native task is scene-level rather than object-level, the derived target region corresponds to the dominant semantic region under the adapted query definition. This protocol allows all datasets to be cast into the same prediction format, namely language-guided box localization.

We emphasize that this adaptation is intended to support cross-dataset comparison under a shared objective; it does not imply that all datasets are natively equivalent to standard visual-grounding benchmarks. In particular, scene-level datasets such as MLRSNet and partially caption-driven datasets such as GAIA remain structurally different from object-centric grounding datasets. The unified protocol therefore offers a practical and reproducible evaluation setting, but the resulting numbers should be interpreted together with the nature of the original annotations. This is especially important when comparing datasets with dense object-level labels against those adapted from scene-level supervision.

4.1.1. SOMA-1M

Source: [11]. SOMA-1M contains approximately 1.3 million SAR–optical image pairs spanning resolutions from 0.5 m to 10 m and covering 12 land-cover categories. Its original purpose is large-scale SAR–optical alignment and multi-resolution scene understanding. The dataset is attractive for our study because it introduces substantial variation in scale and spatial detail, which is a major challenge for lightweight localization models.

Adaptation to unified localization: We use the SAR images as visual input and construct textual prompts from the provided land-cover labels, using formulations such as "the [class] in the SAR image." Since the dataset provides pixel-level alignment information, we derive bounding boxes from the corresponding class regions. When multiple valid regions occur in the same image, we select the region associated with the adapted query. Compared with object-centric datasets, this adaptation is more region-oriented than instance-oriented. As a result, SOMA-1M should be understood as evaluating multi-resolution semantic grounding rather than fine-grained object grounding in the narrowest sense.

Why this dataset matters: SOMA-1M stresses two properties that are directly relevant to TVLightFormer: robustness to resolution variation and the ability to preserve localization quality when the visual structure becomes coarse. It is therefore an informative testbed for the proposed LFPN and for the overall stability of lightweight multi-scale fusion.

4.1.2. ATRNet-STAR

Source: [12]. ATRNet-STAR contains around 190,000 SAR images covering 40 vehicle classes under varying azimuth angles, depression angles, and background conditions. The original task is fine-grained recognition with strong emphasis on inter-class similarity and realistic imaging variation.

Adaptation to unified localization: We use the provided class labels as textual queries and retain the annotated target bounding boxes. Compared with the other datasets in our benchmark, ATRNet-STAR is the closest to an object-centric localization setting, since the targets are explicitly defined and the class names naturally translate into language queries.

Why this dataset matters: This dataset is particularly useful for testing whether a lightweight encoder can still preserve small but discriminative structural differences. Fine-grained SAR recognition is challenging because similar vehicle types may differ only in subtle contour or scattering configurations. If the visual encoder or cross-modal interaction is overly compressed, this is typically reflected first in this dataset.

4.1.3. GAIA

Source: [13]. GAIA contains roughly 200,000 image–text pairs with captions describing environmental phenomena such as floods, fires, deforestation, and other large-scale events. The original tasks are cross-modal retrieval and caption-based understanding over geographically and temporally diverse remote-sensing imagery.

Adaptation to unified localization: We use the original captions as textual queries. When segmentation masks are available, we derive target bounding boxes from the corresponding annotated regions. For scene-level samples without fine region annotations, we use the full image as the target region, reflecting the fact that the semantic unit of interest is the scene itself rather than a single compact object. This choice is necessarily coarse, but it preserves consistency with the original captioning setting.

Why this dataset matters: GAIA evaluates whether the model can localize broad environmental semantics from descriptive language. Compared with ATRNet-STAR, the queries are less categorical and more descriptive. This makes GAIA a useful test for query understanding but also a more weakly grounded benchmark in cases where the caption refers to distributed or scene-wide phenomena rather than a compact target.

4.1.4. MLRSNet

Source: [14]. MLRSNet contains about 109,000 high-resolution RGB remote-sensing images with multi-label scene annotations covering categories such as airport, bridge, transmission tower, and other scene elements. Its original task is multi-label scene classification rather than localization.

Adaptation to unified localization: We derive the language query from the dominant scene label, which we define as the category with the highest confidence under the provided annotations. Because the original dataset does not provide object-level boxes for the selected query, we use the full image as the target region. This creates a coarse grounding task in which the model is asked to associate the dominant scene concept with the scene as a whole. We emphasize that this adaptation is not equivalent to object-level grounding. It may inflate IoU-based metrics because predicting a large region is easier than localizing a compact object. Therefore, MLRSNet is used only as a weakly supervised scene-level

stress case, and its results are interpreted separately from datasets with object boxes, masks, or parcel polygons.

Why this dataset matters: MLRSNet is useful for examining ambiguity under weak supervision. Since multiple semantic concepts can coexist in a single image, the adapted query may not uniquely determine one compact region. Performance on this dataset therefore reflects not only localization capacity but also how the model handles scene-level grounding when object boundaries are not explicitly provided.

4.1.5. SODAS

Source: [15]. SODAS is a time-series dataset built from Sentinel-1 SAR and Sentinel-2 optical imagery over agricultural areas in Seville, Spain, spanning the period 2017–2021. The dataset includes 18 crop types plus fallow and is originally intended for crop-type mapping and phenology-aware agricultural analysis.

Adaptation to unified localization: In this paper, we process each image independently rather than using temporal sequences. Crop types are used as textual queries, and target regions are derived from the parcel polygons provided by the dataset. These polygons are converted to bounding boxes for consistency with the unified localization objective.

Why this dataset matters: SODAS introduces a setting in which semantic identity is strongly influenced by seasonal development and temporal context. Because our current model is single-image based, this dataset is a useful stress test for understanding the limits of a non-temporal architecture under temporally structured variability.

4.1.6. Data Split and Preprocessing

Unless otherwise stated, each dataset is split into training, validation, and test sets at a ratio of 70%/15%/15%. For SODAS, we instead use a temporal split to reduce temporal leakage: samples from 2017 to 2019 are used for training, while samples from 2020 are used for validation, and samples from 2021 are used for testing. All images are resized to 224×224 to maintain a consistent input size across models. This resizing simplifies comparison and reduces implementation bias, although it may also suppress fine spatial details in datasets where the original resolution is high. The latter effect should be kept in mind, particularly for fine-grained SAR recognition and for region-level annotations derived from masks or polygons.

To make the SAR-specific robustness discussion more explicit, we further inspect the test samples according to two diagnostic factors: speckle/background clutter and geometric distortion. Since the public datasets do not provide unified labels for speckle, sidelobes, layover, or shadow, this stratification is used as a diagnostic analysis rather than as a new official benchmark. Samples with strong local intensity fluctuation or cluttered background are treated as speckle/clutter-dominant cases, whereas samples with obvious foreshortening, shadow, layover-like deformation, or strong geometric scale variation are treated as geometry-distorted cases. This additional inspection helps separate general localization ability from robustness to SAR-specific imaging phenomena.

4.2. Experimental Setup

All experiments are implemented in PyTorch 2.0 and run on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB). We use the AdamW optimizer with weight decay 1×10^{-4}. The learning rate for the newly initialized layers is set to 1×10^{-4}, while the visual and textual backbones (MobileNetV3 and TinyBERT) use a lower learning rate of 1×10^{-5} to stabilize optimization. Training proceeds for 100 epochs with cosine annealing down to a minimum learning rate of 1×10^{-6}. The batch size is 16.

For each dataset, we train a separate model from scratch and do not use additional dataset-specific pre-training. This choice reduces the variation introduced by heterogeneous

pre-training corpora and makes the comparison easier to interpret under the unified protocol. At the same time, it is also a conservative setting. Large vision-language models often derive a substantial part of their advantage from large-scale pre-training, and training them from scratch may underestimate their absolute performance. Our objective here is therefore not to reproduce the best published numbers of all baselines under their original settings but to compare model behavior under the same adaptation pipeline, input size, and training schedule.

Unless otherwise stated, all models are trained using the same image resolution, data split, and optimization schedule. This reduces confounding from implementation details, although complete equivalence remains difficult because the models differ substantially in architecture, capacity, and default training assumptions. The reported results should therefore be interpreted as a controlled comparative study under a shared protocol rather than as a replacement for each baseline's original benchmark performance.

4.3. Evaluation Metrics

We evaluate localization accuracy using standard IoU-based criteria. Specifically, we report accuracy at IoU thresholds from 0.5 to 0.9, which are denoted ACC@0.5, ACC@0.6, ACC@0.7, ACC@0.8, and ACC@0.9, together with the mean IoU (mIoU). ACC@0.5 primarily reflects coarse target capture, whereas ACC@0.8 and ACC@0.9 are more indicative of precise localization quality. Since one of the goals of TVLightFormer is to preserve spatial precision despite lightweight design, the high-threshold metrics are particularly informative.

We additionally report the number of parameters, FLOPs, and throughput in frames per second (FPS) measured on the NVIDIA RTX A6000 GPU (NVIDIA, Santa Clara, CA, USA). These efficiency metrics are important because the main motivation of the proposed architecture is edge-oriented deployment. A model that is slightly more accurate but substantially slower may be less attractive in practice than one that offers a better accuracy–efficiency trade-off.

4.4. Comparison with State of the Art

We compare TVLightFormer with nine baselines: six general visual-grounding or open-vocabulary grounding models (TransVG [33], RefTR [34], SeqTR [35], MDETR [36], GLIP [37], and OVDETR [38]), two recent vision-language models (Seg-LLaVA [39] and STEP3-VL-10B [40]), and one SAR-specific baseline, TACMT [30]. All models are re-implemented or adapted to the unified protocol described above.

These baselines differ considerably in design philosophy. Some are optimized for general grounding, some are optimized for broader multimodal reasoning, and some are optimized for SAR-specific localization. Including all of them provides a broad reference spectrum, but it also means that the comparison should not be read as a fully native benchmark-to-benchmark evaluation. Rather, it reflects how these methods behave when transferred to the same language-guided localization setting under shared training constraints.

Tables 2–6 summarize the main results. For brevity, we report only a subset of the metrics in the main text. Across the five datasets, three broad trends are consistent. First, TVLightFormer remains competitive with much larger generic grounding models despite using substantially fewer parameters. Second, the strongest absolute results are typically achieved by heavier baselines, especially TACMT and STEP3-VL-10B, which confirms that capacity still matters. Third, the proposed model is often relatively stronger at high-IoU thresholds than might be expected from its size, suggesting that the multi-scale decoder helps preserve localization precision.

Table 2. Results on SOMA-1M (multi-resolution alignment).

Model	ACC@0.5	ACC@0.7	ACC@0.9	mIoU	Params (M)
TransVG	0.753	0.625	0.125	0.626	112.5
RefTR	0.713	0.619	0.218	0.618	108.2
MDETR	0.769	0.685	0.265	0.673	158.3
GLIP	0.775	0.702	0.284	0.688	185.4
OVDETR	0.784	0.725	0.294	0.699	142.6
SeqTR	0.792	0.735	0.305	0.685	86.2
Seg-LLaVA	0.803	0.745	0.353	0.703	∼85
STEP3-VL-10B	0.829	0.782	0.426	0.733	10,000.0
TACMT	0.836	0.795	0.465	0.752	216.8
TVLightFormer	0.799	0.733	0.399	0.709	27.4

SOMA-1M.

On SOMA-1M, TVLightFormer reaches 79.9% ACC@0.5 and 70.9% mIoU. In absolute terms, it remains below TACMT and STEP3-VL-10B, which is expected given the difference in model scale. However, the comparison is still informative. Relative to SeqTR, TVLightFormer improves mIoU from 68.5% to 70.9% while reducing parameters from 86.2 M to 27.4 M. Relative to Seg-LLaVA, the model is slightly weaker at ACC@0.5 but achieves a stronger high-threshold result at ACC@0.9 (39.9% vs. 35.3%). This suggests that the lightweight design does not merely preserve coarse semantic matching; it also retains useful spatial precision. Since SOMA-1M involves substantial resolution variation, this behavior is consistent with the intended role of LFPN. At the same time, the gap to TACMT indicates that compressing both the visual and textual branches still incurs a measurable cost when the target region is large and semantically broad.

Table 3. Results on ATRNet-STAR (fine-grained vehicle recognition).

Model	ACC@0.5	ACC@0.7	ACC@0.9	mIoU	Params (M)
TransVG	0.726	0.593	0.115	0.598	112.5
RefTR	0.684	0.588	0.199	0.590	108.2
MDETR	0.742	0.658	0.249	0.645	158.3
GLIP	0.749	0.675	0.266	0.662	185.4
OVDETR	0.758	0.699	0.279	0.672	142.6
SeqTR	0.766	0.709	0.289	0.659	86.2
Seg-LLaVA	0.775	0.719	0.335	0.676	∼85
STEP3-VL-10B	0.803	0.755	0.409	0.709	10,000.0
TACMT	0.811	0.772	0.452	0.725	216.8
TVLightFormer	0.783	0.715	0.385	0.696	27.4

ATRNet-STAR.

On ATRNet-STAR, TVLightFormer achieves 78.3% ACC@0.5 and 69.6% mIoU. The model clearly improves over SeqTR and remains close to Seg-LLaVA despite a much smaller parameter count. The remaining gap to TACMT is 2.8 percentage points at ACC@0.5 and 2.9 points in mIoU. This is a meaningful but not unexpected difference, given that fine-grained SAR vehicle recognition is one of the settings in which a deeper and heavier backbone can still provide an advantage. The result is encouraging, because ATRNet-STAR is particularly sensitive to subtle structural differences. The fact that TVLightFormer still reaches 38.5% at ACC@0.9 indicates that the model preserves some high-precision discrimination capability even under substantial compression. That said, fine-grained classes such as T72 and M1 remain more difficult than coarse vehicle categories, which points

to a limitation of the current lightweight encoder when the decision depends on small shape differences.

Table 4. Results on GAIA (environmental understanding).

Model	ACC@0.5	ACC@0.7	ACC@0.9	mIoU	Params (M)
TransVG	0.769	0.643	0.133	0.643	112.5
RefTR	0.726	0.633	0.226	0.635	108.2
MDETR	0.783	0.703	0.276	0.693	158.3
GLIP	0.790	0.723	0.295	0.709	185.4
OVDETR	0.799	0.743	0.309	0.719	142.6
SeqTR	0.808	0.753	0.319	0.705	86.2
Seg-LLaVA	0.815	0.763	0.366	0.723	∼85
STEP3-VL-10B	0.843	0.799	0.443	0.753	10,000.0
TACMT	0.851	0.813	0.482	0.771	216.8
TVLightFormer	0.813	0.749	0.413	0.729	27.4

GAIA.

On GAIA, TVLightFormer attains 81.3% ACC@0.5 and 72.9% mIoU. This places it very close to Seg-LLaVA in coarse accuracy and above it at ACC@0.9. Since GAIA contains descriptive captions rather than short categorical labels, the result suggests that the proposed architecture does not lose too much query understanding capacity despite relying on TinyBERT rather than a larger language model. The remaining difference to STEP3-VL-10B indicates that richer language modeling still matters when captions become longer and semantically more varied. At the same time, the relatively strong high-threshold performance suggests that once the model identifies the relevant semantics, its decoder remains capable of translating them into usable target regions. Failure modes on rare events are discussed later, but the overall trend on GAIA is favorable for a lightweight design.

Table 5. Results on MLRSNet (multi-label scene understanding).

Model	ACC@0.5	ACC@0.7	ACC@0.9	mIoU	Params (M)
TransVG	0.743	0.619	0.126	0.619	112.5
RefTR	0.703	0.609	0.209	0.609	108.2
MDETR	0.759	0.676	0.256	0.663	158.3
GLIP	0.766	0.693	0.276	0.679	185.4
OVDETR	0.775	0.715	0.289	0.689	142.6
SeqTR	0.783	0.726	0.299	0.676	86.2
Seg-LLaVA	0.793	0.736	0.346	0.693	∼85
STEP3-VL-10B	0.819	0.773	0.419	0.723	10,000.0
TACMT	0.827	0.788	0.458	0.741	216.8
TVLightFormer	0.789	0.723	0.393	0.699	27.4

MLRSNet.

On MLRSNet, TVLightFormer achieves 78.9% ACC@0.5 and 69.9% mIoU. The model is again close to Seg-LLaVA and clearly stronger than SeqTR in mIoU. However, MLRSNet should be interpreted cautiously because the adapted localization target is scene-level rather than strictly object-level. In this setting, the main challenge is not precise object boundary regression, but semantic disambiguation when multiple scene concepts coexist. The relatively stable performance of TVLightFormer suggests that it can handle dominant-scene grounding reasonably well, but the gap to TACMT and STEP3-VL-10B indicates that a larger model still helps when the query is ambiguous or when several plausible semantic regions are present simultaneously. In other words, MLRSNet is less a pure localization

benchmark than a test of coarse grounding under multi-label ambiguity, and the results should be read in that light.

Table 6. Results on SODAS (agricultural time-series).

Model	ACC@0.5	ACC@0.7	ACC@0.9	mIoU	Params (M)
TransVG	0.713	0.589	0.116	0.589	112.5
RefTR	0.673	0.579	0.189	0.579	108.2
MDETR	0.729	0.646	0.236	0.633	158.3
GLIP	0.736	0.663	0.256	0.649	185.4
OVDETR	0.745	0.685	0.269	0.659	142.6
SeqTR	0.753	0.696	0.279	0.646	86.2
Seg-LLaVA	0.763	0.706	0.326	0.663	∼85
STEP3-VL-10B	0.789	0.743	0.399	0.693	10,000.0
TACMT	0.795	0.757	0.438	0.711	216.8
TVLightFormer	0.759	0.693	0.373	0.669	27.4

SODAS.

SODAS is the most challenging dataset for the current model because the underlying semantics depend heavily on temporal evolution, whereas TVLightFormer processes one image at a time. Even under this limitation, the model reaches 75.9% ACC@0.5 and 66.9% mIoU, slightly outperforming SeqTR and remaining close to Seg-LLaVA. The lower overall performance compared with other datasets is informative rather than merely disappointing. It indicates that the current architecture can capture parcel-level crop cues from single observations, but it cannot exploit seasonal evolution when different crop types share similar appearance at a given time point. The gap to TACMT and the large models is thus not only a capacity issue; it also reflects the absence of explicit temporal modeling in our formulation.

4.5. Overall Discussion of Main Results

Across the five datasets, TVLightFormer does not achieve the best absolute accuracy, but it consistently occupies a strong position on the accuracy–efficiency spectrum. In all cases, it is substantially smaller than the stronger baselines and remains competitive with models that are several times larger. The strongest absolute performer among the compared methods is often TACMT or STEP3-VL-10B, which confirms that model capacity still translates into measurable gains, especially on fine-grained or semantically complex tasks. However, the magnitude of the gain is not proportional to the increase in parameter count. This is precisely where the proposed design becomes attractive: the model sacrifices some absolute performance, but it does so in exchange for a large reduction in complexity.

A second important observation is that the proposed model is not uniformly strong across all settings. It is relatively competitive on SOMA-1M and GAIA, where multi-scale fusion and language-conditioned region understanding seem to matter. It remains effective on ATRNet-STAR, but the gap widens slightly in the most fine-grained categories. It is weaker on SODAS, where single-frame processing is an inherent limitation. This variation is useful because it shows that the architecture is not simply "good everywhere"; rather, it benefits tasks in which efficient cross-modal alignment and spatially aware decoding are more important than very high-capacity semantic modeling. We also provide qualitative visualizations in Figure 2 and attention heatmaps in Figure 3 to illustrate the model's behavior in challenging scenarios.

The MLRSNet protocol deserves additional caution. Since its target region is the full image rather than an object-level box, it can make IoU-based localization appear easier than in datasets with compact targets. To examine whether this adaptation changes the

overall conclusion, Table 7 reports a protocol sensitivity analysis using the mIoU values in the main tables. Excluding MLRSNet does not change the relative conclusion: TVLightFormer remains close to Seg-LLaVA and below the heavier TACMT/STEP3-VL-10B models while keeping a much smaller parameter count. This suggests that the reported accuracy–efficiency trade-off is not driven by the weakly supervised MLRSNet adaptation alone.

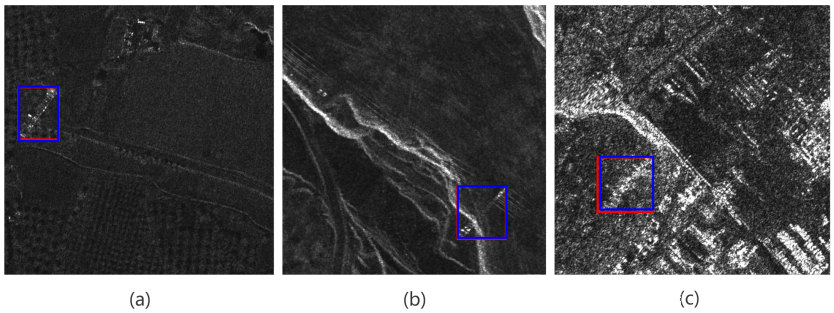


Figure 2. Representative successful localization results of TVLightFormer in three challenging scenarios: (**a**) forested areas with partial occlusion and vegetation-induced speckle, (**b**) mountainous terrain with scale variation and geometric distortion, and (**c**) suburban regions with complex background clutter. Blue boxes indicate predictions and red boxes indicate ground truth.

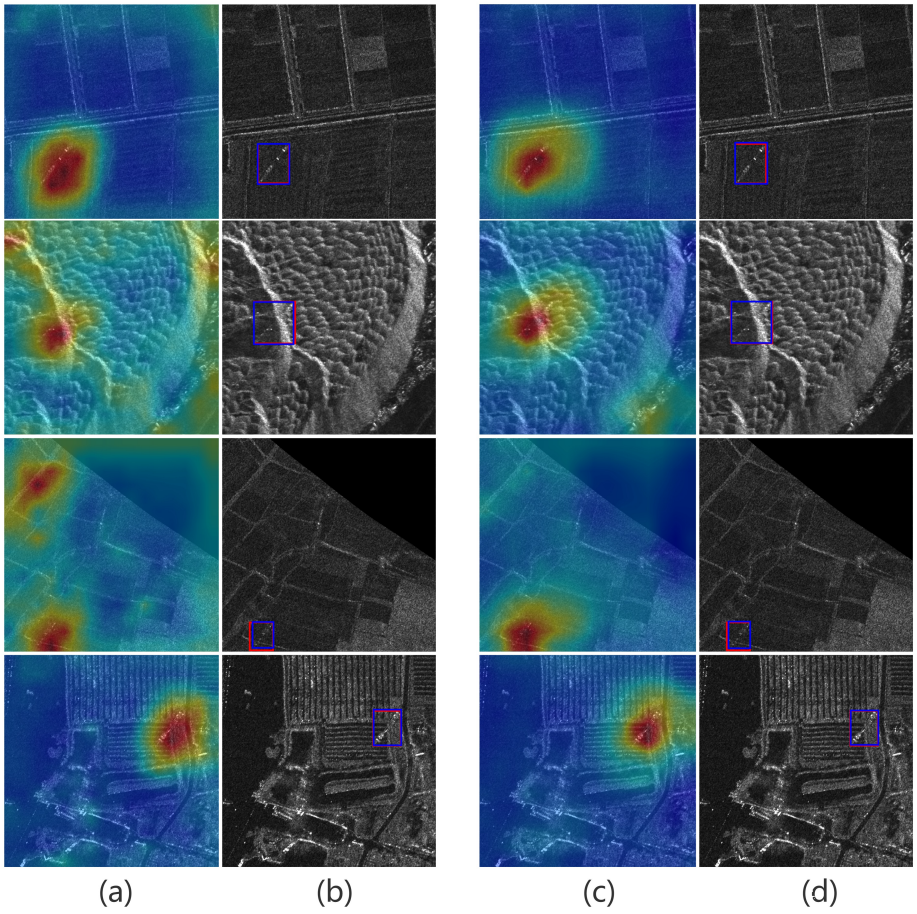


Figure 3. Heatmap comparison between the original model and TVLightFormer: (**a**) localization result of the original model; (**b**) attention heatmap of the original model; (**c**) localization result of TVLightFormer; (**d**) attention heatmap of TVLightFormer. The proposed model exhibits more focused and discriminative attention on target structures while suppressing background clutter and noise. Blue boxes denote predictions, red boxes denote ground truth.

Table 7. Protocol sensitivity analysis with and without MLRSNet. Values are calculated from the mIoU results in Tables 2–6.

Model	Avg. mIoU Incl. MLRSNet (%)	Avg. mIoU Excl. MLRSNet (%)	Difference
SeqTR	67.4	67.4	0.0
Seg-LLaVA	69.2	69.1	−0.1
STEP3-VL-10B	72.2	72.2	0.0
TACMT	74.0	74.0	0.0
TVLightFormer	70.0	70.1	+0.1

4.6. Qualitative Comparison, Failure Cases, and Attention Analysis

To complement the quantitative results, we provide three forms of qualitative evidence: representative correct localizations, cross-model qualitative comparison, and failure-case visualization. These additions are intended to respond more directly to the reviewer concern that numerical tables alone are insufficient for showing how TVLightFormer behaves relative to other methods and where it still fails in challenging SAR-oriented scenes.

We first present representative successful localization results in Figure 2 across challenging scenarios, including densely forested regions, mountainous terrain with geometric distortion, and suburban areas with complex background clutter. TVLightFormer demonstrates robust behavior in all three cases and can still localize the queried transmission towers under visually adverse conditions.

In forested regions (Figure 2a), where the target is partly obscured by vegetation and embedded in speckle-like clutter, the model can still focus on the dominant metallic scattering structure. In mountainous terrain (Figure 2b), the LFPN helps adapt to the scale variation introduced by terrain and imaging geometry. In suburban environments (Figure 2c), the cross-modal interaction module suppresses several distracting linear or bright structures and keeps the final box centered on the query-related target.

Because single-model visualization is insufficient to show the relative behavior of different grounding methods, we further add a cross-model qualitative comparison in Figure 4. The comparison uses the same samples and text queries for all methods, and it includes representative baselines with different modeling styles: TransVG, SeqTR, Seg-LLaVA, TACMT, and the proposed TVLightFormer. The goal is not merely to show that our model produces visually plausible boxes but to make the differences in localization tightness, background suppression, and semantic stability more explicit.

Figure 4 reveals several consistent patterns. First, generic grounding models such as TransVG and SeqTR are more easily attracted by bright or elongated clutter structures in SAR images, especially when the query refers to a small object embedded in a complex scene. Second, larger models such as Seg-LLaVA or heavier SAR-oriented models such as TACMT usually produce more stable predictions, but they may still cover excessive background when the target is weak or partially merged with surrounding scattering. Third, TVLightFormer does not always yield the tightest box in absolute terms, but it more often preserves the correct semantic focus while maintaining compact predictions, which is consistent with the favorable high-threshold accuracy and efficiency trade-off reported in the quantitative results.

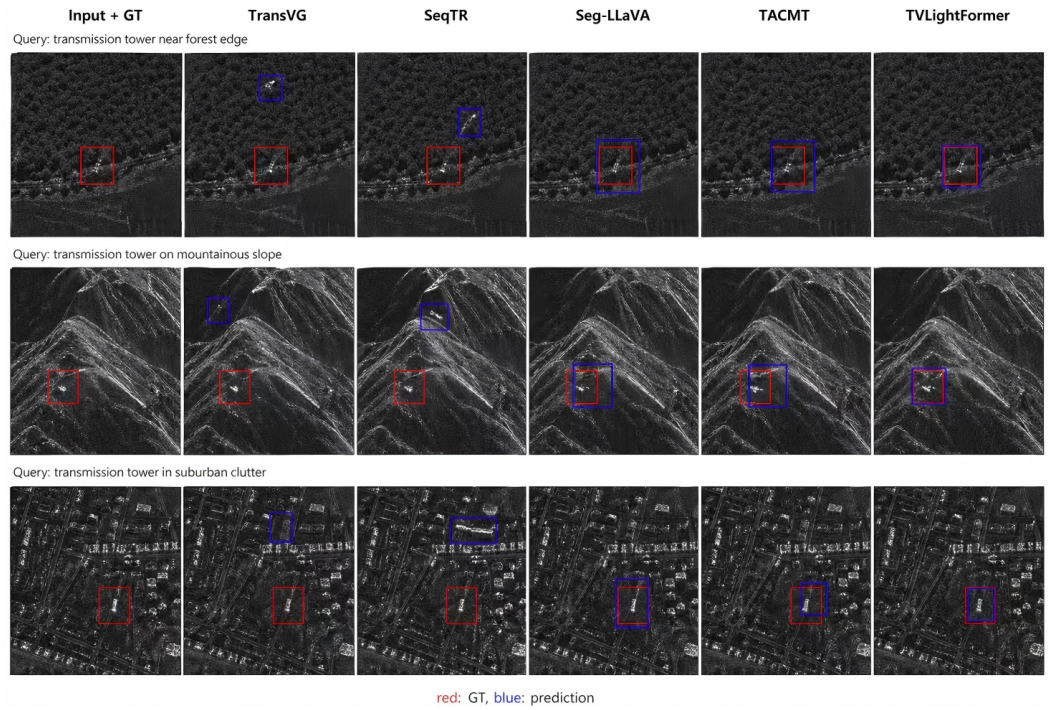


Figure 4. Cross-model qualitative comparison on representative test samples. Each row corresponds to one sample–query pair, and each column reports the localization result of a different method (TransVG, SeqTR, Seg-LLaVA, TACMT, and TVLightFormer). Ground-truth boxes are shown together with the predicted boxes to facilitate a visual comparison of semantic alignment and boundary quality. Blue boxes denote predictions, red boxes denote ground truth.

To make the limitations of the proposed model more transparent, we also provide failure-case visualization in Figure 5. These examples were selected to reflect several typical error modes observed during evaluation: target–background confusion under strong speckle/clutter, under-coverage when the target is spatially diffuse, semantic ambiguity in scene-level grounding, and temporal ambiguity in agricultural observations.

The failure cases in Figure 5 clarify that the main weakness of TVLightFormer is not random instability but rather a set of identifiable limitations. When the target is small and surrounded by clutter, the model may still center on the correct region but fail to recover a tight boundary. When the query refers to a distributed event rather than a compact object, the model often localizes only the most salient subregion. When supervision is scene-level rather than object-level, as in MLRSNet, the predicted box may correspond to the dominant semantic concept but still disagree with the adapted target region. Finally, for SODAS, a single-frame model cannot fully resolve temporal ambiguity across crop growth stages.

To further validate the interpretability and focus of attention for TVLightFormer, we visualize the attention heatmaps of the original model and TVLightFormer in Figure 3. Different from the cross-model box-level comparison in Figure 4, the heatmap comparison focuses on the internal attention response of the proposed design. The heatmaps show where the model distributes its attention on the SAR image while processing a language query. Compared with the original model, TVLightFormer exhibits more focused attention on the target area and reduced activation of non-target background regions. This analysis indicates that the lightweight design not only improves computational efficiency but also enhances semantic discrimination in cluttered environments, which is crucial for precise target localization in SAR images.

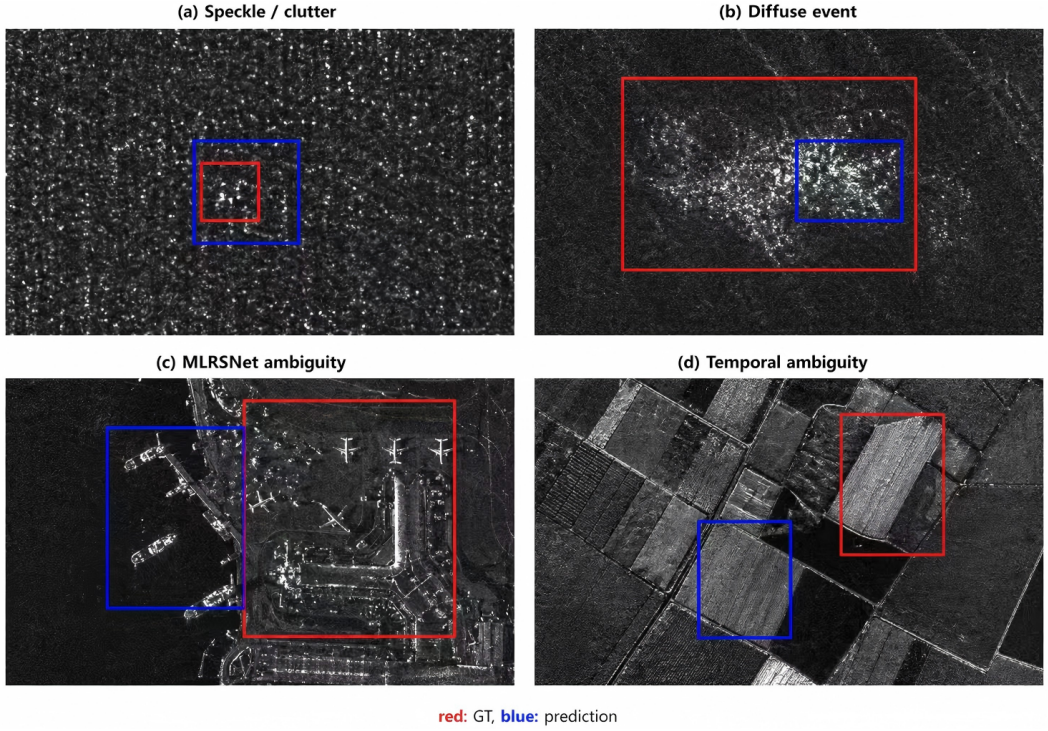


Figure 5. Representative failure cases of TVLightFormer. The examples illustrate four common error modes: (**a**) inaccurate boundary placement under strong speckle/clutter, (**b**) partial localization of a spatially diffuse event region, (**c**) semantic ambiguity in scene-level grounding on MLRSNet, and (**d**) temporal/phenological ambiguity on SODAS. Blue boxes indicate predictions and red boxes indicate ground truth.

4.7. Efficiency Analysis

Accuracy alone is insufficient for evaluating an edge-oriented model. For this reason, Table 8 reports average mIoU values together with parameter count, FLOPs, and throughput. TVLightFormer contains 27.4 M parameters, requires 9.7 GFLOPs, and reaches 25.2 FPS on the A6000 platform. Among the compared methods, it is the fastest model by a clear margin and also one of the smallest. In practice, this means that the proposed architecture is more suitable for deployment scenarios where latency and memory footprint matter.

Compared with TACMT, TVLightFormer reduces parameters by about 87% and increases throughput from 8.5 FPS to 25.2 FPS while reducing the average mIoU by roughly 4.2 points. Compared with STEP3-VL-10B, the parameter gap is several orders of magnitude, whereas the performance gap is much smaller. These comparisons should not be overstated, because large models also bring advantages that are not fully captured by mIoU alone. Still, the trend is clear: a carefully designed lightweight model can remain competitive without inheriting the cost profile of very large multimodal systems. This is the main practical motivation of TVLightFormer.

Table 8. Efficiency comparison (averaged across datasets).

Model	mIoU (%)	Params (M)	FLOPs (G)	FPS
TransVG	61.5	112.5	14.2	12.5
MDETR	68.1	158.3	18.5	9.8
SeqTR	67.4	86.2	12.8	13.2
Seg-LLaVA	69.1	∼85	∼12.5	∼15
STEP3-VL-10B	72.2	10,000.0	∼95	∼1.2
TACMT	74.0	216.8	16.1	8.5
TVLightFormer	69.8	27.4	9.7	25.2

4.8. Ablation Studies

To understand which components contribute most to the final behavior of the model, we perform ablations on SOMA-1M. We use this dataset because it is large, structurally diverse, and sensitive to scale variation, making it a reasonable setting for isolating architectural effects. The ablations are organized at two levels: architecture-level replacement studies and module-level analysis of the encoder and decoder.

4.8.1. Architecture-Level Ablation

Table 9 compares a heavy baseline configuration (ResNet-50 + BERT + standard multi-head attention) with progressively lighter variants. Several observations emerge.

First, replacing the visual encoder with MobileNetV3 (A1) leads to a notable but manageable reduction in mIoU, from 0.763 to 0.731, while reducing model size by nearly 30%. This indicates that the visual backbone remains a major source of performance, particularly for fine spatial representation. Second, replacing BERT with TinyBERT (A2) reduces parameters more aggressively but also causes a slightly larger accuracy drop. This suggests that language capacity matters, even in relatively short-query settings, because the text encoder still needs to preserve distinctions that guide the localization stage.

Third, replacing standard attention with GQA (A3) yields a smaller accuracy decrease relative to the savings in computation. This is one of the more favorable substitutions in the table, indicating that attention compression is less destructive than aggressive reduction of the visual or textual encoders. Fourth, adding LFPN (A4) improves high-threshold localization with only a small increase in model complexity, which supports the claim that multi-scale refinement is especially useful for precise grounding.

Table 9. Architecture ablation on SOMA-1M.

Variant	Visual	Text	Attn	Decoder	ACC0.5	mIoU	Params (M)
Baseline	R50	BERT	MHA12	None	0.833	0.763	216.8
A1	MNv3	BERT	MHA12	None	0.813	0.731	152.4
A2	R50	Tiny	MHA12	None	0.806	0.719	102.5
A3	R50	BERT	GQA6/2	None	0.819	0.742	208.8
A4	R50	BERT	MHA6	LFPN	0.823	0.749	201.3
A5	MNv3	Tiny	MHA12	None	0.793	0.699	41.3
A11	MNv3	Tiny	GQA6/2	None	0.799	0.703	32.7
A12	MNv3	Tiny	MHA12	LFPN	0.800	0.705	44.8
A15	MNv3	Tiny	GQA6/2	LFPN	0.799	0.709	27.4

Most importantly, the full TVLightFormer model (A15) does not correspond to the strongest individual component at each stage but rather to the most balanced overall design. Compared with the heavy baseline, A15 reduces parameters from 216.8 M to 27.4 M while retaining an mIoU of 0.709. The difference is non-negligible, but it is not prohibitive given the reduction in deployment cost. This supports the main claim of the paper: the best lightweight design is not the one that maximizes any single accuracy metric but rather the one that offers the most favorable overall trade-off.

4.8.2. Loss Function Ablation

To address the concern that GIoU may be suboptimal for small-target SAR localization when the target box contains substantial background or speckle noise, we compare several regression losses while keeping the final TVLightFormer architecture fixed. The comparison is carried out on SOMA-1M under the same training schedule. We evaluate single-loss baselines and hybrid losses that combine an IoU-family term with Smooth L1 coordinate regression.

Several observations are worth noting. First, using only a single loss, either Smooth L1 or GIoU, leads to worse performance than hybrid losses, which indicates that overlap supervision and direct coordinate regression play complementary roles. Second, DIoU- and CIoU-based hybrids are competitive and slightly improve the compactness of the predicted boxes relative to the single-loss settings. Third, the originally adopted GIoU + Smooth L1 combination still gives the best overall performance in this comparison—albeit by a small margin, as shown in Table 10. This result suggests that the concern raised by the reviewer is valid in principle, but in the present setting, the combination of GIoU and Smooth L1 remains a reasonable default because the coordinate term compensates for some of the weaknesses of overlap-only supervision.

Table 10. Loss-function ablation on SOMA-1M using the final TVLightFormer architecture.

Loss Function	ACC0.5	ACC0.7	ACC0.9	mIoU
Smooth L1 only	0.786	0.694	0.451	0.688
GIoU only	0.781	0.691	0.444	0.684
DIoU + Smooth L1	0.797	0.709	0.481	0.702
CIoU + Smooth L1	0.798	0.711	0.483	0.703
GIoU + Smooth L1	0.799	0.713	0.486	0.704

The practical implication is that the current loss is not uniquely optimal, but it is stable and competitive. Future work may further explore SAR-oriented localization losses, for example by weighting boundary errors according to the scattering saliency, local signal-to-clutter ratio, or target-size priors. Such designs may be particularly helpful when the target occupies only a small fraction of the bounding box or when strong clutter dominates the surrounding region.

4.8.3. Module-Level Ablation

We further investigate the internal design choices of the encoder and decoder.

Encoder analysis.

Table 11 compares different attention configurations. The general pattern is clear: reducing the number of heads in standard MHA lowers both complexity and performance, while GQA preserves or improves efficiency with a smaller loss in mIoU. The six-head, two-group configuration gives the best trade-off among the compact variants, which is why it is used in the final model. Increasing the hidden dimension to 256 produces the highest mIoU in the table, but it also weakens the lightweight objective by increasing both parameters and FLOPs. This result is useful because it shows that further gains are possible but only by moving away from the deployment constraints that motivate the model.

Table 11. Encoder ablation (GQA variants).

Type	Heads/Groups	Dim	mIoU	Params (M)	FLOPs (G)
MHA	12/1	192	0.683	48.7	10.3
MHA	6/1	192	0.673	35.3	8.8
GQA	6/1	192	0.686	34.1	8.4
GQA	6/2	192	0.693	32.7	8.2
GQA	8/2	192	0.699	36.8	9.0
GQA	6/2	256	0.707	41.3	9.8

Decoder analysis.

Table 12 evaluates decoder variants. Adding a two-layer LFPN already improves over the no-decoder setting, and removing ReLU yields a further gain in both mIoU and

ACC@0.9. The magnitude of the improvement is modest but consistent. This is important because it supports the design rationale without overstating it. The effect is not dramatic, and it should not be interpreted as proof that nonlinear activation is generally harmful. Rather, under the present SAR-oriented lightweight setting, a linear Conv-BN decoder appears slightly better suited to preserving weak response variations than an otherwise similar ReLU-based alternative. The three-layer version performs best in absolute terms, but the gain is again small relative to the additional complexity. We therefore keep the simpler activation-free two-layer design in the final model.

Table 12. Decoder ablation (LFPN variants).

Structure	Activation	mIoU	ACC0.9	Params (M)	FLOPs (G)
None	–	0.693	0.463	32.7	8.2
2-layer	ReLU	0.698	0.479	33.2	8.4
2-layer	None	0.704	0.486	33.2	8.4
3-layer	None	0.707	0.492	35.1	9.0

Overall, the ablation studies show that the final architecture is not assembled from arbitrary lightweight substitutions. Rather, each component contributes in a way that is consistent with its intended role: MobileNetV3 and TinyBERT reduce cost, GQA recovers part of the efficiency loss associated with attention, and LFPN helps preserve localization precision under compression.

4.9. Dataset-Specific Analysis and Failure Cases

A lightweight model is especially informative when it fails, because its failure modes often reveal which forms of information were lost under compression. Representative examples are already visualized in Figure 5; here, we analyze the main error patterns on each dataset in a more structured manner.

4.9.1. SOMA-1M: Multi-Resolution Alignment

TVLightFormer remains relatively stable across resolutions from 0.5 m to 10 m with ACC@0.5 varying by less than 2 percentage points. This suggests that the decoder handles scale variation reasonably well in coarse terms. However, ACC@0.9 decreases more noticeably at lower resolutions, indicating that fine localization degrades when the structural detail is weak. This pattern is not surprising: lightweight multi-scale fusion can preserve target extent, but it cannot fully recover details that are absent after resizing or that are inherently weak in the input. Typical failure cases occur when classes with similar backscatter statistics are described by short or ambiguous prompts. In such cases, the model can identify the general region but fails to place a tight box around the correct semantic target.

4.9.2. ATRNet-STAR: Fine-Grained Recognition

Performance varies across vehicle categories. Large or structurally distinctive targets, such as main battle tanks, are handled more reliably than smaller vehicles with fewer effective pixels on target. The confusion rate between visually similar categories remains nontrivial, especially for classes whose main differences lie in turret shape or aspect ratio. These errors suggest that the compressed visual backbone retains a broad structure but may blur the small discriminative cues required for very fine-grained recognition. A heavier encoder or a higher-resolution feature pathway would likely help but at the cost of reduced efficiency.

4.9.3. GAIA: Environmental Understanding

On GAIA, the model is stronger when captions refer to frequent and visually coherent events such as flood zones or fire regions. Performance drops for rare phenomena or for captions that describe a distributed scene rather than a compact target. In several failure cases, the model localizes only one salient subregion even though the caption refers to a broader event distributed across the image. This suggests a mismatch between scene-level semantics and box-level prediction. The current architecture can associate descriptive text with salient regions, but it is less reliable when the caption describes a spatially diffuse phenomenon.

4.9.4. MLRSNet: Multi-Label Scenes

MLRSNet exposes a different limitation: query ambiguity under scene-level adaptation. When multiple relevant concepts coexist, the model sometimes selects the wrong semantic focus. For example, in images containing both an airport and aircraft, the adapted query may not uniquely determine whether the correct grounding should correspond to the scene as a whole or to a salient object within it. These errors are not purely architectural; they also reflect the fact that the original dataset was not designed for object-level grounding. Because the full-image target is coarser than an object box, MLRSNet results should not be used alone to claim precise localization ability. In the current manuscript, we therefore treat MLRSNet as a weakly supervised scene-grounding case and report the sensitivity analysis in Table 7. Future versions of the protocol should use object-level annotations or pseudo-boxes generated by a verified detector to reduce this ambiguity.

4.9.5. SODAS: Agricultural Time-Series

On SODAS, the main limitation is temporal ambiguity. The same crop type can appear quite different across phenological stages, while different crop types can appear similar at a single time point. Since TVLightFormer processes images independently, it cannot exploit temporal continuity to resolve such ambiguity. The observed errors therefore reflect a genuine modeling limitation rather than only an issue of lightweight design. This result is useful because it identifies a clear direction for extension: incorporating temporal modeling without losing the efficiency advantages of the current architecture.

5. Conclusions and Discussion

This paper presented **TVLightFormer**, which is a lightweight cross-modal transformer for language-guided target localization under remote-sensing and SAR-oriented settings. The proposed architecture combines a compact visual encoder (MobileNetV3), a lightweight language encoder (TinyBERT), grouped-query attention for efficient cross-modal interaction, and an activation-free lightweight feature pyramid network for multi-scale decoding. Together, these components are designed to reduce computational cost while preserving the spatial precision required for localization.

Under the unified language-guided localization protocol established in this paper, TVLightFormer achieved an average mIoU of 69.8% across five public datasets with only 27.4 M parameters and 9.7 GFLOPs while reaching 25.2 FPS on an A6000 GPU. Although the model does not consistently exceed the strongest heavy baselines in absolute accuracy, the experiments show that it remains competitive across diverse settings while offering a substantially better efficiency profile. This result supports the main claim of this paper: a carefully designed lightweight cross-modal model can retain a meaningful portion of the localization capability of larger systems without inheriting their full computational cost.

The ablation studies further clarify the role of each component. The results suggest that the main efficiency gains come from compressing both the visual and textual encoders,

while grouped-query attention provides a favorable reduction in attention cost with limited performance loss. The activation-free LFPN contributes modest but consistent improvements in high-precision localization, which is consistent with the need to preserve weak responses in SAR-related imagery. This claim is deliberately limited to the present experimental setting; it should not be read as a universal conclusion that activation-free decoding is always preferable nor that weak-response preservation matters only for SAR data. In this sense, the final architecture should be understood not as the strongest model in absolute terms but rather as a balanced design optimized for the accuracy–efficiency trade-off.

The dataset-specific analysis also reveals several limitations. Performance degrades in settings that require very fine-grained structural discrimination, stronger object-level disambiguation, temporal reasoning across image sequences, or explicit handling of strong speckle and geometric distortion. The MLRSNet experiment should also be interpreted as weakly supervised scene-level grounding rather than object-level localization, because its full-image target protocol can make IoU-based evaluation less strict. These findings indicate that the current model is most effective when efficient cross-modal alignment and multi-scale spatial refinement are more critical than very high-capacity semantic modeling. They also suggest that further improvements will likely require extensions beyond simple scaling of the present architecture.

Overall, TVLightFormer provides a compact and practical alternative for language-guided localization in resource-constrained settings. The reported efficiency and throughput suggest that the model is more deployment-friendly than substantially heavier multimodal baselines, particularly in scenarios where latency, memory, and energy budget are important considerations. Future work will investigate temporal extensions for sequential SAR observations, stronger object-aware cross-modal reasoning, and deployment-oriented compression techniques such as quantization and structured pruning.

Author Contributions: Conceptualization, Y.Z., H.Q. and C.N.; methodology, Y.Z., H.Q. and C.N.; software, H.Q. and C.N.; validation, Y.Z., H.Q. and C.N.; formal analysis, Y.Z. and Y.G.; investigation, Y.Z. and H.Q.; resources, Y.W. and Y.G.; data curation, Y.Z. and C.N.; writing—original draft preparation, Y.Z., H.Q. and C.N.; writing—review and editing, Y.Z., Y.W. and Y.G.; visualization, Y.Z. and H.Q.; supervision, Y.W. and Y.G.; project administration, Y.W. and Y.G.; funding acquisition, Y.W. and Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Undergraduate Training Program for Innovation and Entrepreneurship (Grant No. 202590002086), the National Natural Science Foundation of China (General Program, Grant No. 72571281), and the Excellent Young Scientist Fund of Hunan Province (Grant No. 2025JJ40066).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: The authors thank the anonymous reviewers and the editors for their insightful comments and helpful suggestions that helped improve the quality of our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pei, H.; Su, M.; Xu, G.; Xing, M.; Hong, W. Self-supervised Feature Representation for SAR Image Target Classification Using Contrastive Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9246–9258. [CrossRef]
2. Liu, J.; Zhan, J.; Guo, Y.; Li, T.; Zhao, Y.; Zhang, J.; Tang, L.; Li, Y.; Wei, Y.; Cai, W. Learning Geometric-Aware and Weather-Adaptive Semantics in Remote Sensing: Affine Lie Group Enhanced Detector for UAV Road Scenes. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5605820. [CrossRef]

3. Zhong, X.; Zhan, J.; Xie, Y.; Zhang, L.; Zhou, G.; Liang, M.; Yang, K.; Guo, Z.; Li, L. Adaptive Deformation-Learning and Multiscale-Integrated Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5611619. [CrossRef]
4. Wang, T.; Zhou, W.; Zeng, Y.; Zhang, X. EfficientVLM: Fast and Accurate Vision-Language Models via Knowledge Distillation and Modal-adaptive Pruning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023, Toronto, ON, Canada, 9–14 July 2023.
5. Wanting, X.; Yang, L.; Langping, H.; Xucheng, H.; Ling, J. Xmodel-VLM: A Simple Baseline for Multimodal Vision Language Model. *arXiv* **2024**, arXiv:2405.09215. [CrossRef]
6. Liu, D.; Wang, D.; Wang, H.; Chen, H.; Jiang, W.; Cheng, Y.; Guo, H.; Cui, W.; Zhang, J. SARMAE: Masked Autoencoder for SAR Representation Learning. *arXiv* **2025**, arXiv:2512.16635. [CrossRef]
7. Xu, L.; Wang, L.; Zhang, J.; Ha, D.; Zhang, H. A Review of Cross-Modal Image–Text Retrieval in Remote Sensing. *Remote Sens.* **2025**, *17*, 3995. [CrossRef]
8. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
9. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
10. Ainslie, J.; Lee-Thorp, J.; De Jong, M.; Zemlyanskiy, Y.; Lebron, F.; Sanghai, S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023.
11. Wu, P.; Yao, Y.; Wan, Y.; Zhang, W.; Zhao, R.; Li, J.; Zhang, Y. SOMA-1M: A Large-Scale SAR-Optical Multi-resolution Alignment Dataset for Multi-Task Remote Sensing. *arXiv* **2026**, arXiv:2602.05480.
12. Liu, Y.; Li, W.; Liu, L.; Zhou, J.; Peng, B.; Song, Y.; Xiong, X.; Yang, W.; Liu, T.; Liu, Z.; et al. ATRNet-STAR: A Large Dataset and Benchmark Towards Remote Sensing Object Recognition in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, 48, 6735–6753. [CrossRef]
13. Zavras, A.; Michail, D.; Zhu, X.X.; Demir, B.; Papoutsis, I. GAIA: A Global, Multi-modal, Multi-scale Vision-Language Dataset for Remote Sensing Image Analysis. *arXiv* **2025**, arXiv:2502.09598.
14. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350. [CrossRef]
15. Villarroya-Carpio, A.; Lopez-Sanchez, J.M. *SAR and Optical Dataset for Agriculture in Seville (SODAS)*, Version 1.1; Zenodo: Geneva, Switzerland, 2025. [CrossRef]
16. Ye, X.; Gan, Y.; Ge, Y.; Zhang, X.P.; Tang, Y. ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025.
17. Zhang, C.; Ma, K.; Fang, T.; Yu, W.; Zhang, H.; Zhang, Z.; Xie, Y.; Sycara, K.; Mi, H.; Yu, D. VScan: Rethinking Visual Token Reduction for Efficient Large Vision-Language Models. *arXiv* **2025**, arXiv:2505.22654. [CrossRef]
18. Wang, J.; Hu, X.; Zhang, P.; Li, X.; Wang, L.; Zhang, L.; Gao, J.; Liu, Z. MiniVLM: A smaller and faster vision-language model. *arXiv* **2020**, arXiv:2012.06946.
19. Zhan, J.; Bai, L.; Zhang, J.; Liu, T.; Shi, F.; Liu, Y.; Liu, L. Heterogeneous Binary Pixel Difference Networks for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5602619. [CrossRef]
20. Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. NVILA: Efficient Frontier Visual Language Models. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025.
21. Zhan, J.; Xie, Y.; Guo, J.; Hu, Y.; Zhou, G.; Cai, W.; Wang, Y.; Chen, A.; Xie, L.; Li, M.; et al. DGPF-RENet: A Low Data Dependence Network With Low Training Iterations for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, 5520521. [CrossRef]
22. Li, T.; Zhan, J.; Liu, J.; Guo, J.; Xiong, X.; Cai, W.; Liu, E.; Wei, Y.; Hu, Y. Sorted Texture-Aware Glance and Gaze Network for Hyperspectral Image Classification With Low Training Samples. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5518126. [CrossRef]
23. Boutayeb, A.; Lahsen-Cherif, I.; El Khadimi, A. Deformable Attention Mechanisms Applied to Object Detection, case of Remote Sensing. *arXiv* **2025**, arXiv:2505.24489. [CrossRef]
24. Wei, Y.; Xiao, A.; Ren, Y.; Zhu, Y.; Chen, H.; Xia, J.; Yokoya, N. SARLANG-1M: A Benchmark for Vision–Language Modeling in SAR Image Understanding. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5201320. [CrossRef]

25. Wang, P.; Lu, Z.; Li, Y.; Ding, B.; Zhang, D. SARCLIP: The First Vision–Language Foundation Model for SAR Image. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5223211. [CrossRef]
26. Ma, Z.; Xiao, X.; Dong, S.; Wang, P.; Wang, H.; Pan, Q. SARChat-Bench-2M: A Multi-Task Vision-Language Benchmark for SAR Image Interpretation. *arXiv* **2025**, arXiv:2502.08168.
27. Jiang, C.; Wang, C.; Wu, F.; Ma, P.; Zou, L.; Li, T.; Ning, J.; Tang, Y. SARCLIP: a multimodal foundation framework for SAR imagery via contrastive language-image pre-training. *ISPRS J. Photogramm. Remote Sens.* **2026**, *231*, 17–34. [CrossRef]
28. Yang, Y.; Zhang, X.; Fang, Q.; Ye, Z.; Li, R.; Liu, L.; Wang, H. SAR-KnowLIP: Towards Multimodal Foundation Models for Remote Sensing. *arXiv* **2025**, arXiv:2509.23927.
29. Wang, J.; Sun, H.; Sun, Y.; Tang, T.; Lei, L.; Ji, K. SAR-TinySNN: A Lightweight Spiking Neural Network for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 4008005. [CrossRef]
30. Li, T.; Wang, C.; Tian, S.; Zhang, B.; Wu, F.; Tang, Y.; Zhang, H. TACMT: Text-aware cross-modal transformer for visual grounding on high-resolution SAR images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *222*, 152–166. [CrossRef]
31. Liu, J.; Zhan, J.; Zhang, J.; Chen, J.; Song, Y.; Tang, L.; Zhou, L.; Du, C.; Wei, Y.; Guo, Y. Robust scale fusion and edge-aware feature attention network for remote sensing UAV road detection under harsh weather. *Results Eng.* **2025**, *27*, 106172. [CrossRef]
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
33. Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; Li, H. TransVG: End-to-End Visual Grounding with Transformers. *arXiv* **2021**, arXiv:2104.08541.
34. Li, M.; Sigal, L. Referring transformer: A one-step approach to multi-task visual grounding. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 19652–19664.
35. Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; Ji, R. SeqTR: A Simple Yet Universal Network for Visual Grounding. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022.
36. Kamath, A.; Lecun, Y.; Carion, N.; Singh, M.; Misra, I.; Synnaeve, G. MDETR-Modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 1760–1770.
37. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded Language-Image Pre-training. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
38. Zang, Y.; Li, W.; Zhou, K.; Huang, C.; Loy, C.C. Open-Vocabulary DETR with Conditional Matching. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022.
39. Guo, T.; Liu, H.; Wen, J.; Li, J. Seg-LLaVA: A small-scale large vision-language model with external visual prompts. *Neurocomputing* **2026**, *669*, 132423. [CrossRef]
40. Huang, A.; Yao, C.; Han, C.; Wan, F.; Guo, H.; Lv, H.; Zhou, H.; Wang, J.; Zhou, J.; Sun, J.; et al. STEP3-VL-10B Technical Report. *arXiv* **2026**, arXiv:2601.09668.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.