

RAH-VLA: Resolution-Adaptive Hierarchical Vision–Language Alignment for Multimodal Remote Sensing Understanding

Siyu Zhang¹, Lianlei Shan², Runhe Qiu¹, and Wenxin Zhong¹

Abstract—Multimodal vision–language modeling has emerged as a promising paradigm for remote sensing (RS) image understanding. However, existing methods are limited by fixed-resolution visual processing and single-scale vision–language alignment, making it difficult to simultaneously preserve fine-grained details and maintain semantic consistency across different spatial granularities. To address these challenges, we propose RAH-VLA, a resolution-adaptive hierarchical vision–language alignment framework for multimodal RS understanding. Specifically, a dynamic resolution input strategy (DRIS) is developed to enable resolution-adaptive visual representations, while a multiscale vision–language alignment mechanism (MS-VLAM) is introduced to establish hierarchical semantic correspondence across object-level, region-level, and global-level representations. Extensive experiments on multiple RS benchmarks demonstrate that RAH-VLA consistently improves image captioning, visual grounding, and cross-modal reasoning performance while reducing computational redundancy. Qualitative analyses further illustrate the effectiveness of the proposed resolution-adaptive perception and hierarchical vision–language alignment mechanisms, as well as the cross-modality generalization capability of the proposed framework. Overall, RAH-VLA provides an effective and scalable solution for multimodal RS interpretation, offering a practical pathway toward efficient and semantically robust RS vision–language models.

Index Terms—Cross-modal semantic understanding, hierarchical alignment, multimodal remote sensing (RS), resolution adaptation, vision–language modeling.

NOMENCLATURE

$R(x, y)$	Resolution assignment function (1).
$R_{\text{high}}, R_{\text{low}}$	High-/low-resolution modes (1).
$s(x, y)$	Regional saliency score (1).
τ	Adaptive threshold value (1) and (2).
λ	Threshold sensitivity coefficient (2).
k	Number of selected ROIs (6).
Ω	Selected region of interest (3)–(5).
F_{coarse}	Coarse-resolution feature map (3)–(5).
F_{fine}	Fine-resolution feature map (3)–(5).
F_{fused}	Fused multiresolution features (3)–(5).

Received 15 June 2026; revised 10 July 2026; accepted 11 August 2026. Date of publication 14 August 2026; date of current version 27 August 2026. This work was supported by the Sanda University Special Research Fund Project under Grant 2024BSZX05. (Corresponding author: Wenxin Zhong.)

Siyu Zhang, Runhe Qiu, and Wenxin Zhong are with Sanda University, Shanghai 201209, China (e-mail: zhangsiyu0066@gmail.com; qiuhr@sanda.edu.cn; wxzhong@sanda.edu.cn).

Lianlei Shan is with Tsinghua University, Beijing 100084, China (e-mail: shanlianlei18@mails.ucas.edu.cn).

Digital Object Identifier 10.1109/TGRS.2026.3723919

P_{ROI}	Selected ROI set (6).
V	Visual feature representation (7) and (17).
g	Global visual representation (17) and (18).
t_{CLS}	Global textual representation (18).
L_{obj}	Object-level alignment loss (10).
L_{reg}	Region-level alignment loss (16).
L_{glob}	Global-level alignment loss (18).
α, β, γ	Alignment loss weights (19).
L_{align}	Hierarchical alignment loss (19).
L_{caption}	Caption generation loss (20).
δ	Loss balancing coefficient (21).
L	Overall training objective (21).

I. INTRODUCTION

REMOTE sensing (RS) imagery, acquired from satellite-borne and airborne platforms, provides a fundamental means for large-scale Earth observation and dynamic environment monitoring [1]. Benefiting from wide spatial coverage, high spatiotemporal resolution, and rich spectral diversity, modern RS data support a broad range of applications, including land-cover analysis, urban management, precision agriculture, and disaster monitoring [2], [3], [4]. With the rapid advancement of sensing technologies and Earth observation systems, remote sensing imagery is evolving toward increasingly large-scale and heterogeneous data distributions, driving growing demands for intelligent semantic understanding [5].

Despite the rapid advancement of sensing technologies and Earth observation systems, achieving reliable semantic understanding of RS imagery remains highly challenging, particularly in scenarios requiring the joint perception of large-scale spatial context and fine-grained object details [6], [7]. Traditional unimodal analysis is inherently insufficient to capture the complementary and heterogeneous information distributed across modern RS data [e.g., optical, synthetic aperture radar (SAR), and LiDAR modalities], motivating the transition toward multimodal semantic understanding paradigms [8], [9]. Existing multimodal approaches mainly rely on implicit feature interaction and feature-level fusion, making it difficult to establish consistent semantic correspondence across heterogeneous modalities and multiple semantic granularities [10], [11]. Consequently, these methods often struggle to achieve robust high-level semantic understanding in complex RS scenarios.

To overcome the limitations of conventional multimodal fusion, vision–language models (VLMs) have recently emerged

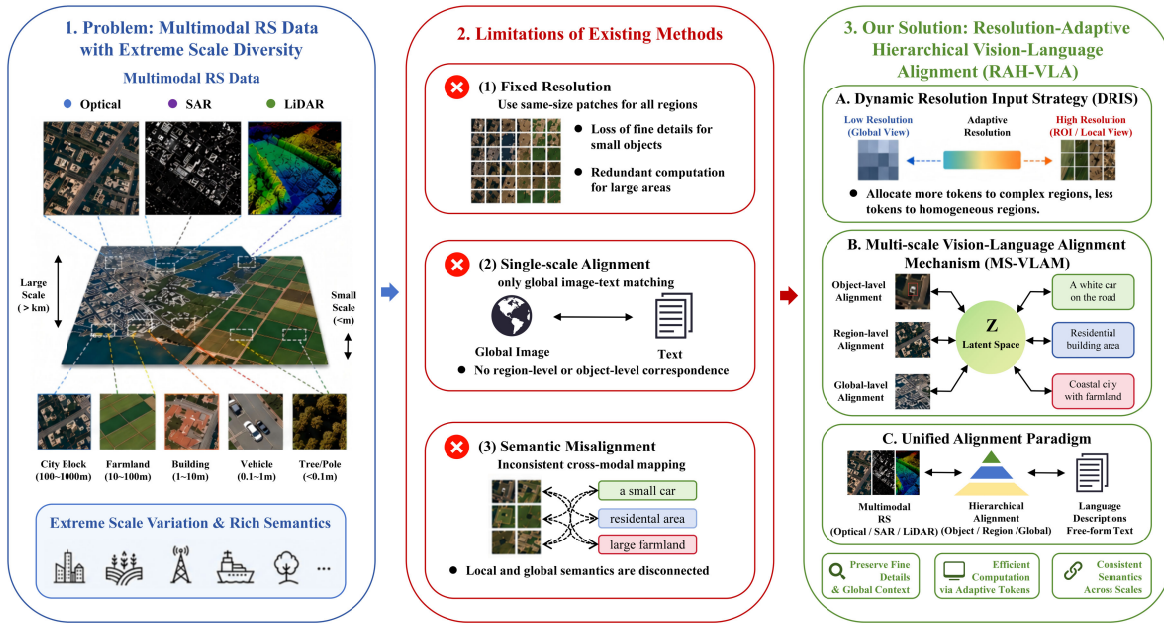


Fig. 1. *Motivation and Overview of the Proposed RAH-VLA Framework for RS Image Understanding.* RS imagery exhibits heterogeneous modalities, extreme scale variation, and hierarchically distributed semantics, posing challenges for multimodal understanding. Existing methods are limited by fixed-resolution processing and single-scale alignment. RAH-VLA provides a unified framework for jointly modeling resolution-adaptive visual perception and hierarchical vision-language semantic alignment.

as a promising paradigm for cross-modal semantic understanding by explicitly bridging visual representations and language-based semantic concepts [12], [13]. Compared with traditional multimodal frameworks, VLMs provide a unified framework for connecting visual content with semantic concepts, enabling high-level tasks such as image captioning, visual grounding, and multimodal reasoning [14], [15]. More importantly, VLMs enable explicit modeling of semantic relationships beyond low-level feature interaction, offering a principled way to align visual representations with language-guided semantics [16], [17].

Despite their success, existing VLMs are primarily designed for natural image domains and do not explicitly account for the intrinsic characteristics of RS imagery [7]. In particular, RS imagery exhibits extreme scale variation, where large homogeneous regions coexist with small but semantically critical objects, making fixed-resolution processing inefficient and prone to losing fine-grained semantic details [18]. Meanwhile, semantic concepts in RS imagery are inherently spatially distributed and hierarchically structured across multiple scales [19]. However, most existing VLMs mainly rely on single-scale alignment, limiting their ability to capture consistent cross-modal semantics across object-level, region-level, and global-level representations [20]. In addition, the weak and indirect correspondence between visual patterns and textual descriptions further increases the difficulty of robust cross-modal semantic alignment under heterogeneous sensing conditions [14].

These limitations reveal a fundamental mismatch between resolution-dependent visual representations and hierarchically structured semantic concepts in RS imagery, where fixed-resolution visual encoding fails to preserve consistent semantic correspondence across spatial scales and semantic granularities

[21]. As illustrated in Fig. 1, multimodal RS imagery exhibits extreme scale variation and hierarchically distributed semantic structures, while existing vision-language frameworks mainly rely on fixed-resolution processing and single-scale alignment, resulting in inefficient representation learning and inconsistent semantic correspondence across different spatial granularities. These observations highlight the necessity of a unified modeling paradigm that can simultaneously account for scale variation and hierarchical semantic structures in RS [22].

To address these challenges, we formulate multimodal RS understanding as a *resolution-adaptive hierarchical vision-language alignment* problem and propose resolution-adaptive hierarchical vision-language alignment (RAH-VLA), a unified framework for efficient and semantically consistent multimodal RS understanding. Specifically, we introduce a dynamic resolution input strategy (DRIS), which adaptively allocates computational resources and visual tokens according to scene complexity, enabling efficient visual perception while preserving fine-grained semantic details. Building upon these resolution-adaptive representations, we further develop a multi-scale vision-language alignment mechanism (MS-VLAM), which decomposes cross-modal alignment into object-level, region-level, and global-level semantic spaces to establish hierarchical semantic correspondence across different semantic granularities. By jointly integrating adaptive visual representation learning and hierarchical semantic alignment, RAH-VLA effectively bridges the gap between visual representations and semantic understanding in RS imagery, providing a unified framework for image captioning, visual grounding, and multimodal reasoning.

The main contributions of this work are summarized as follows.

- 1) We propose RAH-VLA, a unified framework for multimodal RS understanding that jointly models adaptive visual perception and hierarchical semantic alignment.
- 2) We develop a DRIS that adaptively allocates visual computation according to scene complexity, reducing computational redundancy while preserving fine-grained visual details.
- 3) We introduce an MS-VLAM to establish hierarchical semantic correspondence across object-level, region-level, and global-level representations.
- 4) Extensive experiments on multiple RS benchmarks demonstrate the effectiveness of RAH-VLA in image captioning, visual grounding, and multimodal reasoning tasks, while additional analyses validate its interpretability and cross-modality generalization capability.

II. RELATED WORK

A. Multimodal Semantic Understanding in RS

The rapid growth of heterogeneous Earth observation data has driven increasing interest in multimodal RS semantic understanding, where complementary information from optical imagery, SAR, LiDAR, multispectral, and hyperspectral modalities is jointly exploited [23], [24]. Compared with single-modality analysis, multimodal frameworks integrate complementary spatial, spectral, geometric, and structural information across different sensing mechanisms, significantly improving robustness and scene understanding capability under complex observation conditions, including noise interference, cloud contamination, and heterogeneous terrain distributions [10], [25]. Consequently, multimodal semantic understanding has gradually evolved into a fundamental research direction for intelligent RS analysis [26].

1) *Early Shallow Fusion Stage:* Early multimodal RS methods mainly relied on handcrafted features and shallow fusion strategies, where different modalities were combined at the pixel, feature, or decision levels [27], [28]. Traditional RS analysis heavily depended on manually designed texture, shape, and spectral priors to characterize heterogeneous observations. Beyond handcrafted feature extraction, traditional hyperspectral image processing further explored spectral information exploitation through techniques such as image compression and spectral unmixing, aiming to improve data representation and reveal intrinsic spectral characteristics of hyperspectral imagery [29], [30]. Representative handcrafted descriptors include SIFT [31], HOG [32], LBP [33], and spectral indices [34]. Although these methods improve multimodal information utilization to some extent, they implicitly assume that heterogeneous modalities can be directly projected into compatible feature spaces through shallow aggregation. However, substantial differences in sensing mechanisms, statistical distributions, and semantic characteristics among heterogeneous modalities often lead to severe modality gaps and inconsistent cross-modal semantics in RS scenarios [21]. Consequently, shallow fusion methods generally suffer from limited semantic interpretability, weak robustness to modality-specific noise, and poor generalization capability across complex RS environments.

The limitations of shallow fusion methods in modeling structured multimodal interaction have driven the development of deep learning-based architectures for more effective contextual understanding.

2) *CNN-Based Multimodal Interaction Paradigms:* In the early deep learning era, convolutional neural network (CNN)-based fusion networks became the dominant paradigm for multimodal RS analysis by replacing handcrafted features with automatically learned hierarchical features. Representative architectures typically adopted dual-stream or multibranch CNN structures to extract modality-specific features, followed by multimodal interaction through feature concatenation, attention weighting, or multiscale aggregation mechanisms [35]. For example, Guo et al. proposed a CNN-based spatial feature fusion framework for hyperspectral image classification [36], while Feng et al. developed ICIF-Net to model intrascale cross-modal interaction and interscale feature fusion for complex RS scenes [37].

Compared with shallow fusion methods, CNN-based architectures significantly improve local contextual perception and multimodal interaction through hierarchical feature extraction. However, due to the localized nature of convolutional operations, these methods mainly capture short-range neighborhood dependencies and struggle to model long-range spatial interaction in large-scale RS scenes [38]. Moreover, semantic interaction is typically performed through implicit fusion modules without explicitly constraining structured cross-modal consistency across heterogeneous modalities.

The limitations of CNN-based methods in modeling global contextual interaction have further motivated the introduction of Transformer-based architectures.

3) *Transformer-Based Multimodal Interaction Paradigms:* Transformer-based architectures introduce self-attention mechanisms to model long-range dependencies and global contextual interaction across heterogeneous modalities. Representative models such as ViT [39] and Swin Transformer [40] significantly improve global contextual modeling and multimodal interaction in RS scenarios. Building upon these architectures, recent studies further extend Transformer-based frameworks toward multimodal RS understanding through large-scale contextual modeling and cross-modal attention mechanisms, while hybrid CNN-Transformer architectures are also explored to jointly capture local spatial details and global semantic dependencies [41].

However, despite their stronger global modeling capability, most existing Transformer-based methods still mainly rely on implicit feature interaction and coarse feature-level fusion without explicitly modeling hierarchical semantic structures or structured cross-modal consistency under heterogeneous sensing conditions. Consequently, consistent semantic interaction across multiple spatial granularities remains insufficiently explored in complex RS scenarios.

These limitations have recently driven the development of scalable multimodal foundation paradigms for more generalized RS understanding.

4) *Recent Multimodal Foundation Paradigms:* Building upon Transformer architectures and large-scale self-supervised learning, recent research has gradually shifted toward scalable

multimodal foundation paradigms for increasingly large-scale and heterogeneous RS data [25]. Multimodal self-supervised learning enables transferable semantic interaction across heterogeneous sensing modalities by leveraging large-scale unlabeled RS data [23]. Representative studies further incorporate geospatial relations, contrastive learning, and multimodal semantic interaction into self-supervised RS frameworks [42], [43].

Recent RS foundation models further improve multimodal understanding capability through large-scale Earth observation pretraining, while semantic reasoning and semantic reconstruction mechanisms are increasingly introduced to strengthen multimodal interaction and contextual consistency [44], [45], [46], [47].

Nevertheless, most existing methods still rely on implicit feature interaction and coarse feature-level fusion without explicitly modeling hierarchical semantic correspondence across heterogeneous modalities [48]. Consequently, robust multimodal understanding across multiple spatial granularities remains insufficiently explored, particularly under extreme scale variation and spatially distributed semantics. These challenges have motivated the development of more adaptive and structured multimodal interaction paradigms for RS understanding.

B. Vision-Language Modeling for RS

Vision-language modeling has emerged as a dominant paradigm for unifying visual perception and natural language understanding through aligned visual-textual representations. In natural image domains, large-scale VLMs such as contrastive language-image pre-training (CLIP), foundational language and vision alignment (FLAVA), and CoCa demonstrate that pretraining on web-scale image-text data can produce highly transferable multimodal representations, enabling strong performance in image captioning, visual question answering, and cross-modal retrieval [12], [49], [50]. Such success largely benefits from the intrinsic characteristics of natural images, where semantic concepts are typically object-centric, spatially compact, and strongly correlated with textual descriptions.

1) *General VLM Paradigm and Downstream Extension*: Driven by the increasing demand for intelligent RS interpretation, recent studies have gradually extended general vision-language paradigms toward RS scenarios [7]. Unlike conventional visual recognition tasks that mainly focus on category-level recognition, RS-oriented VLMs require joint understanding of complex spatial layouts, structural relationships, and contextual scene information for downstream multimodal reasoning tasks [51]. Consequently, RS VLMs have been increasingly applied to image captioning, visual grounding, and multimodal reasoning tasks [52].

Recent RS-oriented VLM frameworks further enhance multimodal interaction by incorporating geospatial grounding, large-scale domain-adaptive pretraining, and RS structural priors into vision-language architectures. Representative examples such as GeoChat and remote sensing GPT (RSGPT) demonstrate the feasibility of adapting large-scale

vision-language models to complex RS scenarios through grounded semantic interaction and domain-specific multimodal pretraining [13], [53], [54]. The rapid emergence of these RS-oriented VLM frameworks has also driven increasing efforts toward systematic reviews of datasets, architectures, and multimodal reasoning paradigms for RS vision-language modeling [15], [25].

2) *Paradigm Mismatch Between Conventional VLMs and RS Imagery*: Despite their promising capability, conventional VLMs are primarily developed under natural-image assumptions, where visual semantics are usually concentrated on a few salient foreground objects with relatively compact spatial distributions. However, such assumptions are fundamentally violated in RS scenarios due to the intrinsic properties of RS imagery.

Unlike natural images, RS scenes exhibit extreme scale variation, where large homogeneous land-cover regions coexist with small but visually critical objects, making uniform feature extraction pipelines ineffective for simultaneously preserving global contextual information and fine-grained local details [18]. Moreover, semantic information in RS imagery is inherently spatially distributed, where meaningful scene understanding often depends on complex interactions among multiple regions rather than isolated object-centric semantics [55]. In addition, the correspondence between visual patterns and textual descriptions is frequently weak and implicitly correlated, further increasing the difficulty of precise visual-textual interaction and semantic correspondence modeling [56].

Consequently, directly transferring conventional VLM paradigms to RS scenarios often leads to insufficient visual-textual correspondence and unstable multimodal interaction under complex spatial distributions. These limitations reveal a fundamental mismatch between resolution-dependent visual encoding and hierarchically distributed RS semantics.

3) *Limitations of Existing RS VLM Methods*: To alleviate the above challenges, existing RS VLM approaches mainly focus on enhancing visual representation capability and improving multimodal interaction mechanisms [25]. Nevertheless, most existing methods still rely on fixed-resolution visual encoding pipelines and single-level global interaction strategies, which are inherently incompatible with the hierarchical semantic structures and multiscale spatial characteristics of RS scenes [57]. Fixed-resolution visual encoding often introduces a trade-off between computational efficiency and fine-grained detail preservation, while global-only interaction mechanisms fail to explicitly model semantic correspondence across different spatial granularities [22].

Moreover, semantic interaction in most existing RS VLMs is predominantly performed at a single semantic granularity, making it difficult to establish consistent correspondence across object-level, regional-level, and scene-level semantics. Consequently, current RS VLMs struggle to simultaneously preserve local fine-grained details and maintain globally coherent contextual understanding, resulting in sub-optimal multimodal interaction performance under complex RS scenarios.

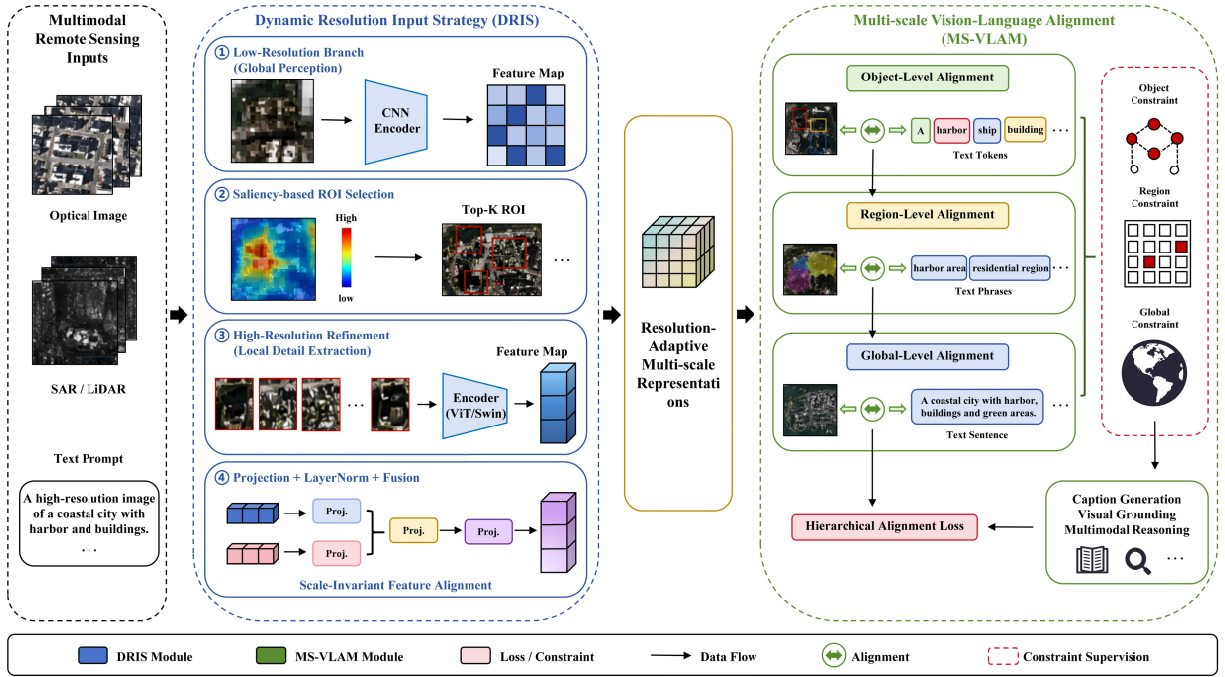


Fig. 2. Overview of the Proposed RAH-VLA Framework for Multimodal RS Interpretation. DRIS combines low-resolution global perception and high-resolution local refinement for adaptive resource allocation. MS-VLAM performs hierarchical cross-modal alignment at object, region, and global levels. The framework is jointly optimized by hierarchical alignment loss.

These observations suggest that effective RS vision-language modeling requires moving beyond fixed-resolution encoding and coarse global interaction toward more adaptive and structured multimodal semantic modeling paradigms, which remain insufficiently explored in existing research. These limitations motivate the development of RAH-VLA, which provides a unified framework for jointly modeling resolution-adaptive visual perception and hierarchical vision-language semantic alignment in multimodal RS understanding.

III. METHOD

A. Overview of the Proposed Framework

This study proposes RAH-VLA, a resolution-adaptive hierarchical vision-language alignment framework for multimodal RS understanding, as illustrated in Fig. 2. The framework is designed to address two key challenges in RS vision-language modeling: scale inconsistency caused by fixed-resolution visual processing, and semantic mismatch caused by single-level vision-language alignment. For ease of reference, the major mathematical symbols and variables used throughout this section are summarized in Nomenclature.

As shown in the figure, the framework consists of two closely coupled core mechanisms, which implement the four basic components of a typical vision-language model: the DRIS serves as the visual encoder, and the MS-VLAM integrates the connector, text encoder, and alignment functions.

Given an RS image and its corresponding textual prompt, the DRIS visual encoder first processes the image in a coarse-to-fine manner: it performs low-resolution global perception to estimate regional saliency, selects semantically

important regions for high-resolution refinement, and then generates resolution-adaptive visual representations through cross-resolution feature fusion. Meanwhile, the text prompt is embedded into semantic vectors by the text encoder. The connector module (implemented within MS-VLAM) maps visual features into a language-compatible space, enabling the framework to perform hierarchical cross-modal alignment at object-level, region-level, and global-level granularities. Finally, the aligned vision-language representations are fed into the large language model (LLM) to support downstream multimodal understanding tasks.

B. Relationship Between DRIS and MS-VLAM

Unlike conventional vision-language pipelines, where adaptive visual processing and cross-modal alignment are optimized independently, the proposed RAH-VLA establishes a unified framework that tightly couples the DRIS and the MS-VLAM. The motivation behind this design is that RS images simultaneously exhibit large-scale scene diversity and fine-grained object variations, making efficient visual perception and hierarchical semantic alignment inherently interdependent rather than isolated optimization objectives.

To better illustrate this design philosophy, Table I compares representative vision-language frameworks with the proposed RAH-VLA. Existing adaptive visual processing approaches, such as DynamicViT and Token Merging, primarily improve computational efficiency through token selection or adaptive computation while lacking explicit cross-modal semantic modeling. In contrast, representative RS vision-language models, including GeoChat and RSGPT, generally rely on fixed-resolution visual representations and perform semantic

TABLE I
COMPARISON OF REPRESENTATIVE VISION–LANGUAGE FRAMEWORKS IN TERMS OF ADAPTIVE VISUAL PERCEPTION,
SEMANTIC ALIGNMENT, AND OPTIMIZATION STRATEGY

Design Aspect	DynamicViT	Token Merging	GeoChat	RSGPT	RAH-VLA
Adaptive Resolution	×	×	×	×	✓
Token Selection	✓	✓	×	×	✓
Object-level Alignment	×	×	×	×	✓
Region-level Alignment	×	×	×	×	✓
Global-level Alignment	×	×	✓	✓	✓
Joint Optimization	×	×	×	×	✓

alignment mainly at the global level. Consequently, adaptive visual perception and hierarchical vision–language alignment are usually designed and optimized as separate components.

The proposed framework bridges this gap by enabling mutual interaction between DRIS and MS-VLAM. Specifically, DRIS is not merely introduced to reduce computational redundancy. By allocating high-resolution computation to semantically informative regions while preserving low-resolution global context, DRIS generates resolution-adaptive multiscale visual representations with enhanced object boundaries, richer regional structures, and more coherent scene-level semantics. These structured representations provide informative visual inputs for subsequent hierarchical semantic alignment.

Building upon these adaptive visual representations, MS-VLAM establishes semantic correspondence at the object-, region-, and global-levels through hierarchical cross-modal supervision. Rather than treating visual encoding and semantic alignment as two independent stages, the hierarchical alignment objective continuously refines the visual representations produced by DRIS during training. Consequently, DRIS improves the quality of visual features for semantic alignment, whereas MS-VLAM further enhances their semantic consistency through multilevel supervision.

The two modules are jointly optimized in an end-to-end manner, allowing adaptive visual perception and hierarchical semantic alignment to mutually reinforce each other throughout the learning process. This unified optimization strategy enables the proposed framework to simultaneously improve computational efficiency, preserve fine-grained visual details, and achieve more consistent cross-modal semantic understanding for multimodal RS tasks.

C. Dynamic Resolution Input Strategy

RS images usually cover large spatial areas and contain complex land-cover structures. Directly using fixed high-resolution inputs leads to excessive graphics processing unit (GPU) memory consumption and computational cost, whereas fixed low-resolution inputs may lose small objects, boundary information, and texture details. Consequently, existing fixed-resolution processing strategies often suffer from an inherent trade-off between computational efficiency and fine-grained semantic perception.

To address this limitation, we propose a DRIS, which dynamically allocates computational resources according to scene complexity and semantic importance. As illustrated

in Fig. 3, DRIS adopts a coarse-to-fine resolution allocation paradigm. Specifically, a low-resolution branch is first employed to obtain global scene understanding and estimate regional saliency. Subsequently, semantically important regions are selected for high-resolution refinement, and cross-resolution feature fusion is performed to generate resolution-adaptive visual representations.

The dynamic resolution allocation function is defined as

$$R(x, y) = \begin{cases} R_{\text{high}}, & \text{if } s(x, y) \geq \tau \\ R_{\text{low}}, & \text{otherwise} \end{cases} \quad (1)$$

where $R(x, y)$ denotes the resolution mode assigned to spatial position (x, y) ; R_{high} and R_{low} denote the high-resolution and low-resolution processing modes, respectively; $s(x, y)$ denotes the regional significance score estimated from the low-resolution global analysis; and τ is the threshold used to determine whether a region should be processed at high resolution. When $s(x, y) \geq \tau$, the corresponding region is regarded as semantically salient or structurally complex and is processed using R_{high} ; otherwise, it is processed using R_{low} .

To make the resolution allocation adaptive to different RS scenes, the threshold τ is further defined as

$$\tau = \mu_s + \lambda \cdot \sigma_s \quad (2)$$

where μ_s and σ_s denote the mean and standard deviation of the saliency scores in the coarse-resolution attention map, respectively, and λ is a sensitivity coefficient controlling the strictness of high-resolution region selection. A larger λ selects fewer but more salient regions, while a smaller λ allows more regions to enter the high-resolution branch. In all experiments, λ is set to 1.0 unless otherwise stated. This adaptive threshold design enables DRIS to dynamically adjust resolution allocation according to scene complexity.

In practice, the adaptive threshold first identifies candidate salient regions according to scene-dependent saliency statistics. To maintain a fixed computational budget across different scenes, a Top- k selection strategy is subsequently applied to the candidate set, and the k highest-scoring regions are selected for high-resolution refinement. Therefore, the adaptive threshold is responsible for candidate region generation, whereas the Top- k operation determines the final regions of interests (ROIs) used for local detail extraction. In all experiments, k is fixed to 5 to ensure consistent computational complexity, stable training behavior, and fair comparison across different evaluation settings.

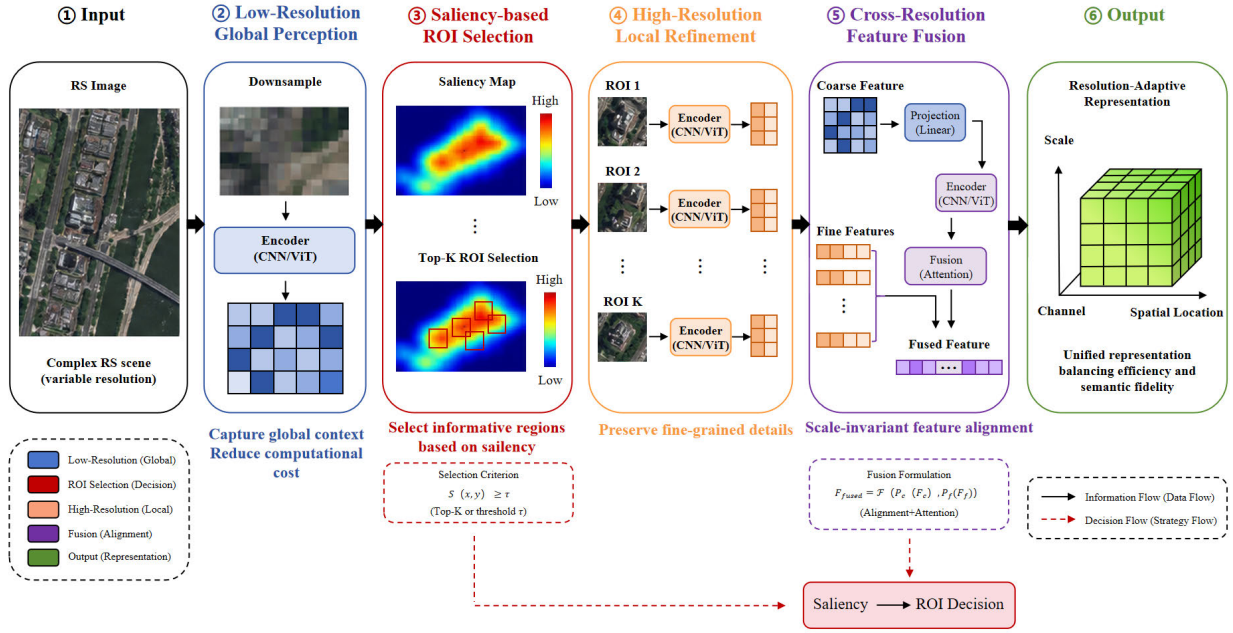


Fig. 3. Illustration of the Proposed DRIS. The DRIS adopts a coarse-to-fine, resolution-adaptive paradigm. It first captures global context via low-resolution perception and generates saliency maps for ROI selection. Then, selected regions are refined at high resolution to preserve fine details. Cross-resolution feature fusion finally produces resolution-adaptive representations balancing semantic fidelity and computational efficiency.

The mechanism first performs low-resolution processing on a large-format RS image to obtain global scene understanding, coarse localization, and regional saliency estimation. If the downsampling factor is n , the computational complexity of low-resolution global processing is approximately $1/n^2$ of full-resolution processing. Subsequently, high-resolution image blocks are extracted only from selected ROIs, enabling fine-grained analysis of small targets, boundary structures, and complex textures.

To integrate features from different resolutions, a multi-resolution feature fusion strategy is adopted. Let F_{coarse} denote the low-resolution feature map extracted from the global view, and let F_{fine} denote the high-resolution localized feature map extracted from selected ROIs. The initial fusion process can be expressed as

$$F_{\text{fused}} = \text{Upsample}(F_{\text{coarse}}) + \text{Crop}(F_{\text{fine}}, \Omega) \quad (3)$$

where F_{fused} denotes the fused multiresolution feature map; Ω denotes the spatial range of the selected ROI; $\text{Upsample}(\cdot)$ denotes the operation that resizes the coarse feature map to match the target spatial scale; and $\text{Crop}(\cdot, \Omega)$ denotes the operation that extracts the high-resolution localized feature region corresponding to Ω . This operation aims to combine global contextual information from the low-resolution branch with fine-grained structural details from the high-resolution branch.

However, F_{coarse} and F_{fine} originate from different resolutions and receptive fields, and therefore may exhibit different statistical distributions. Directly adding them may introduce feature inconsistency across scales. To address this issue, both feature maps are first projected into a shared latent space and normalized before fusion

$$\tilde{F}_{\text{coarse}} = \text{LN}(W_c F_{\text{coarse}}), \quad \tilde{F}_{\text{fine}} = \text{LN}(W_f F_{\text{fine}}) \quad (4)$$

where $\tilde{F}_{\text{coarse}}$ and $\tilde{F}_{\text{fine}}$ denote the normalized coarse and fine feature maps after projection; W_c and W_f are learnable projection matrices for coarse and fine features, respectively; and $\text{LN}(\cdot)$ denotes layer normalization. The final aligned feature fusion is then formulated as

$$F_{\text{fused}} = \text{Upsample}(\tilde{F}_{\text{coarse}}) + \text{Crop}(\tilde{F}_{\text{fine}}, \Omega). \quad (5)$$

Before fusion, the projected coarse- and fine-resolution features are resized to the same spatial and channel dimensions, ensuring compatibility for cross-resolution aggregation. This design can be interpreted as a projection-based cross-resolution feature alignment process. By mapping coarse and fine features into a shared latent space before fusion, the model preserves global semantic context from low-resolution inputs while incorporating fine-grained details from high-resolution regions, thereby alleviating semantic discontinuity caused by direct cross-scale aggregation.

Furthermore, the projection layers implicitly learn a scale-invariant embedding space, in which features from different resolutions are aligned according to semantic similarity rather than spatial scale. This enables the fused representation to maintain semantic consistency while integrating multiscale information, thus mitigating potential discontinuities caused by heterogeneous receptive fields.

For ROI selection, an attention heatmap is generated in the low-resolution branch

$$P_{\text{ROI}} = \text{Topk}(\text{GaussianBlur}(A_{\text{coarse}}), k) \quad (6)$$

where P_{ROI} denotes the selected ROI set; $A_{\text{coarse}} \in \mathbb{R}^{\frac{H}{n} \times \frac{W}{n}}$ denotes the coarse attention heatmap generated from the low-resolution representation; H and W denote the height and width of the original image; n denotes the downsampling factor; $\text{GaussianBlur}(\cdot)$ denotes the smoothing operation used

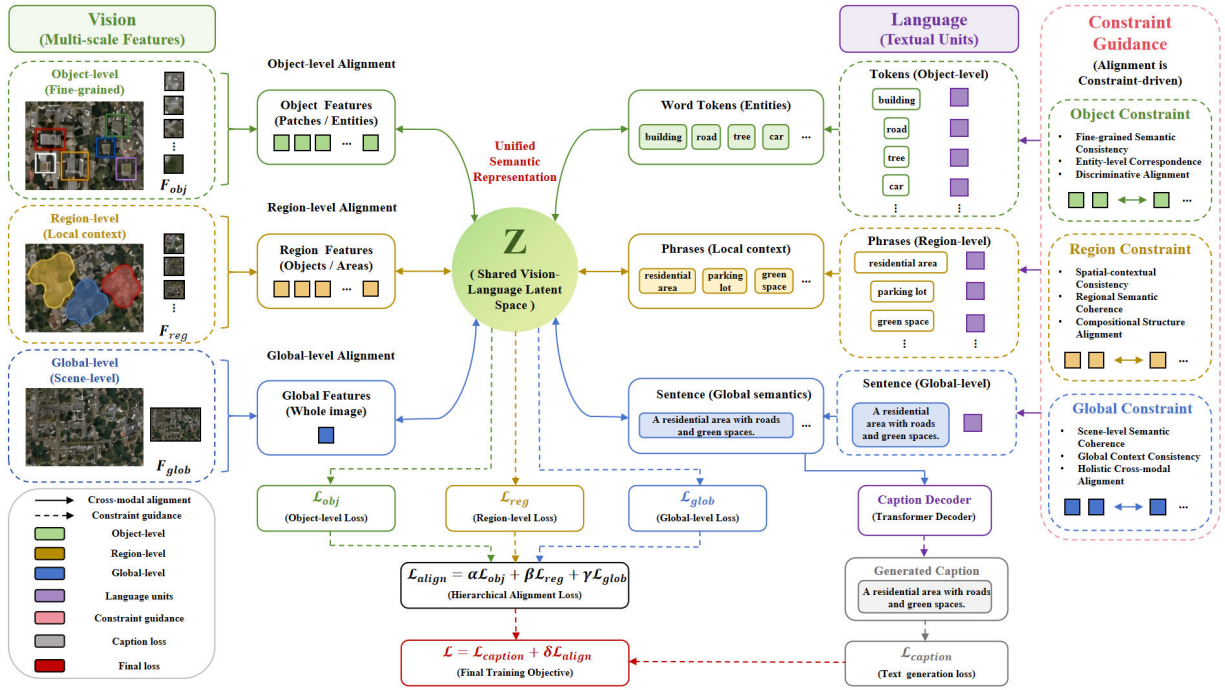


Fig. 4. Overview of the Proposed MS-VLAM. MS-VLAM performs hierarchical cross-modal alignment between multiscale visual representations and multigranularity textual semantics. Object-, region-, and global-level representations are aligned within a shared vision–language latent space under hierarchical semantic constraints. The resulting alignment objective jointly optimizes semantic consistency and caption generation.

to suppress noisy responses and preserve stable salient regions; and $\text{Topk}(\cdot, k)$ selects the top- k regions with the highest response values as candidate ROIs. The Gaussian smoothing parameter σ controls the trade-off between noise suppression and boundary preservation.

Through this design, DRIS reduces redundant computation and improves inference efficiency while preserving fine-grained visual details. More importantly, the resolution-adaptive features generated by DRIS provide structured visual inputs for the subsequent MS-VLAM alignment mechanism. Therefore, DRIS is not merely an efficiency-oriented preprocessing strategy, but also a representation mechanism that supports hierarchical vision–language alignment. More importantly, the enhanced object-aware and region-aware representations generated by DRIS directly provide reliable multiscale visual cues for the subsequent MS-VLAM, facilitating more accurate hierarchical semantic alignment across different spatial granularities.

D. Multiscale Vision–Language Alignment Mechanism

Building upon the resolution-adaptive visual representations generated by DRIS, MS-VLAM further organizes these adaptive features into object-level, region-level, and global-level semantic representations. Since the informative regions have already been selectively enhanced by DRIS, hierarchical alignment can focus on semantically important structures rather than redundant background information, resulting in more effective cross-modal correspondence. Existing vision–language models mainly rely on global or single-scale alignment strategies, which often match an entire image with a sentence-level representation. Such a design

is insufficient for RS imagery, where semantic information is inherently distributed across multiple spatial granularities, ranging from individual objects and local regions to holistic scene semantics. Consequently, semantic mismatch may occur when object-level details, regional context, and global scene understanding need to be jointly modeled.

To address this limitation, we propose an MS-VLAM, which explicitly performs hierarchical cross-modal alignment across multiple semantic levels. As illustrated in Fig. 4, MS-VLAM decomposes vision–language interaction into object-level, region-level, and global-level semantic spaces, enabling structured semantic correspondence between resolution-adaptive visual representations and textual semantics at different granularities.

The object-level alignment focuses on fine-grained entities such as buildings, aircraft, vehicles, ships, and roads. The region-level alignment models local spatial contexts and interactions among objects, such as roads adjacent to water bodies or buildings surrounded by vegetation. The global-level alignment captures holistic scene semantics, such as airports, ports, residential areas, or agricultural regions. By jointly optimizing these three levels, MS-VLAM improves fine-grained semantic correspondence while preserving global contextual coherence.

Concretely, the framework first extracts salient object representations from region proposals generated by the visual encoder and ROI pooling, then aligns them with entity-level textual features. Meanwhile, semantic masks or segmented regions are used to generate region-level visual features, which are matched with phrase-level textual embeddings. Finally, global visual representations are derived using spatial pyramid pooling (SPP) and aligned with sentence-level textual

representations. The three alignment losses are then combined to support robust caption generation, visual grounding, and multimodal reasoning.

1) *Object-Level Alignment Loss*: At the object level, visual features are extracted from the p th candidate object region identified from the visual feature map through region proposal generation and RoI pooling. The visual feature vector of the p th object is defined as

$$\mathbf{v}_{\text{obj}}^{(p)} = f_v(\text{RoI}(V, \mathbf{B}_p)) \quad (7)$$

where V denotes the overall visual feature map of the input image; $\mathbf{B}_p$ denotes the bounding box of the p th candidate object; $\text{RoI}(\cdot)$ denotes RoI pooling or RoIAlign, which extracts local visual features from V according to $\mathbf{B}_p$; and $f_v(\cdot)$ denotes a visual projection function, such as a fully connected layer or linear transformation, that maps RoI features into the shared alignment space.

The corresponding text entity feature is defined as

$$\mathbf{o}_{\text{obj}}^{(p)} = f_t(\mathbf{e}_p) \quad (8)$$

where $\mathbf{e}_p$ denotes the textual representation of the corresponding entity, such as an object category label or descriptive phrase, and $f_t(\cdot)$ denotes the text projection function that maps entity-level text features into the same alignment space as visual object features.

The semantic similarity between visual and textual features is measured using cosine similarity

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \quad (9)$$

where $\mathbf{u}^\top \mathbf{v}$ denotes the inner product between vectors $\mathbf{u}$ and $\mathbf{v}$, and $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$ denote their Euclidean norms.

Considering that object localization quality affects alignment reliability, an IoU-based dynamic weighting strategy is introduced. The object-level alignment loss is formulated as

$$\mathcal{L}_{\text{obj}} = 1 - \frac{1}{P} \sum_{p=1}^P w_p \cdot \cos(\mathbf{v}_{\text{obj}}^{(p)}, \mathbf{o}_{\text{obj}}^{(p)}) \quad (10)$$

where P denotes the number of object instances; $\mathbf{v}_{\text{obj}}^{(p)}$ and $\mathbf{o}_{\text{obj}}^{(p)}$ denote the visual and textual features of the p th object, respectively; and w_p denotes the localization-aware weight assigned to this object.

The weight w_p is computed as

$$w_p = \frac{\text{IoU}(\hat{\mathbf{B}}_p, \mathbf{B}_p)}{\sum_{q=1}^P \text{IoU}(\hat{\mathbf{B}}_q, \mathbf{B}_q)} \quad (11)$$

where $\hat{\mathbf{B}}_p$ denotes the predicted bounding box of the p th object, $\mathbf{B}_p$ denotes the corresponding reference bounding box, and $\text{IoU}(\cdot)$ denotes the intersection-over-union metric. This weighting strategy strengthens the contribution of accurately localized objects and improves the stability of fine-grained semantic alignment.

2) *Region-Level Alignment Loss*: At the region level, semantic masks or segmented regions generated by segment anything model (SAM) or comparable segmentation models are used as candidate regions $\{R_k\}_{k=1}^K$, where K denotes the number of regions. Each region undergoes masked pooling or masked RoI operation to obtain the region-level visual feature $\mathbf{v}_k$. On the textual side, a phrase extractor generates phrase candidates $\{t_j\}_{j=1}^M$, where M denotes the number of textual phrases. Each phrase is mapped into a phrase-level embedding

$$\mathbf{p}_j = \text{PhraseEmbed}(t_j) \quad (12)$$

where $\mathbf{p}_j$ denotes the embedding of the j th phrase, and $\text{PhraseEmbed}(\cdot)$ denotes the phrase embedding function.

Since the correspondence between regions and phrases is often one-to-many or many-to-one, direct one-to-one hard matching may not be robust. Therefore, we combine hard region-phrase matching and contrastive soft matching. The hard matching loss is defined as

$$\mathcal{L}_{\text{reg}}^{\text{hard}} = 1 - \frac{1}{K} \sum_{k=1}^K \cos(\mathbf{v}_k, \mathbf{p}_{\pi(k)}) \quad (13)$$

where $\pi(k)$ denotes the phrase index assigned to the k th visual region.

For contrastive region-phrase alignment, the similarity score between the k th region and the j th phrase is computed as

$$s_{kj} = \frac{\mathbf{v}_k^\top \mathbf{p}_j}{\tau_c} \quad (14)$$

where s_{kj} denotes the similarity score, and $\tau_c > 0$ denotes the temperature parameter for contrastive learning. Here, τ_c is used to distinguish it from the DRIS threshold τ in (1).

The InfoNCE-based region-level loss is defined as

$$\mathcal{L}_{\text{reg}}^{\text{NCE}} = -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(s_{k,y(k)})}{\sum_{j=1}^M \exp(s_{kj})} \quad (15)$$

where $y(k)$ denotes the positive phrase index corresponding to the k th region.

The final region-level alignment loss is

$$\mathcal{L}_{\text{reg}} = \mu \mathcal{L}_{\text{reg}}^{\text{hard}} + (1 - \mu) \mathcal{L}_{\text{reg}}^{\text{NCE}}, \quad \mu \in [0, 1] \quad (16)$$

where μ controls the balance between hard matching and contrastive soft matching. This formulation enables the model to capture region-phrase relationships and multiobject interactions more robustly.

3) *Global-Level Alignment Loss*: At the global level, the overall scene representation is obtained by applying SPP to the visual feature map V

$$\mathbf{g} = \text{SPP}(V) \quad (17)$$

where $\mathbf{g}$ denotes the global visual vector, and $\text{SPP}(\cdot)$ denotes the SPP operation. In practice, SPP aggregates features using multiple pooling scales, such as 1×1 , 2×2 , and 4×4 , to capture scene-level context at different spatial granularities.

The corresponding textual representation is the sentence-level vector $t_{[\text{CLS}]}$, which is obtained from the classification

token (CLS) or an equivalent global textual embedding. The global-level alignment loss is defined as

$$\mathcal{L}_{\text{glob}} = 1 - \cos(\mathbf{g}, \mathbf{t}_{[\text{CLS}]}) \quad (18)$$

where $\mathcal{L}_{\text{glob}}$ measures the semantic discrepancy between the global visual context and sentence-level textual semantics. Layer normalization and L2 normalization are applied before computing similarity to improve numerical stability.

The overall hierarchical alignment objective is formulated as

$$\mathcal{L}_{\text{align}} = \alpha\mathcal{L}_{\text{obj}} + \beta\mathcal{L}_{\text{reg}} + \gamma\mathcal{L}_{\text{glob}} \quad (19)$$

where α , β , and γ are nonnegative weights controlling the contributions of object-level, region-level, and global-level alignment losses, respectively. In all experiments, α , β , and γ are treated as learnable parameters and optimized jointly with the model. To ensure stable optimization and balanced contributions from different semantic levels, the weights are normalized through a softmax operation, satisfying

$$\alpha + \beta + \gamma = 1.$$

This design improves training stability while enabling adaptive weighting across object-level, region-level, and global-level alignment objectives.

E. Caption Generation Loss and Optimization

To jointly optimize semantic alignment and language generation, the hierarchical alignment loss is combined with the caption generation loss. Given a ground-truth caption sequence $(x_1, x_2, \dots, x_T)$, the caption generation loss is defined as

$$\mathcal{L}_{\text{caption}} = - \sum_{t=1}^T \log P_L(x_t | V, x_{<t}) \quad (20)$$

where T denotes the length of the caption sequence; x_t denotes the t th token; $x_{<t}$ denotes all previously generated tokens before time step t ; V denotes the visual representation used by the decoder; and $P_L(\cdot)$ denotes the probability predicted by the language model.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{caption}} + \delta\mathcal{L}_{\text{align}} \quad (21)$$

where $\delta > 0$ is a balancing coefficient that controls the contribution of hierarchical alignment to the overall optimization objective.

This joint objective enables the model to jointly optimize language generation and multilevel vision–language alignment, leading to semantically consistent multimodal RS understanding. As a result, the learned representation is both resolution-adaptive and semantically consistent across object, region, and global levels.

IV. DATASET DESCRIPTION

RS vision–language benchmarks provide the basis for training and evaluating multimodal RS understanding models. In this work, the proposed framework is trained on the large-scale

multitask benchmark *RS-GPT4V* [53], which provides instruction following supervision across diverse RS vision–language tasks.

RS-GPT4V is a large-scale instruction-based RS vision–language benchmark containing 91 937 training images with 991 206 instruction–answer pairs and 15 999 test images with 258 419 instruction–answer pairs. The benchmark supports a variety of RS understanding tasks, including image captioning, visual question answering, detailed scene description, local-region description, and multiturn conversation.

Compared with traditional single-task datasets, RS-GPT4V provides.

- 1) *Large-Scale Data Distribution*, which supports robust learning of resolution-adaptive visual perception in DRIS.
- 2) *Multilevel Semantic Supervision*, which aligns well with the hierarchical design of MS-VLAM.
- 3) *Multitask Learning Capability*, which improves multimodal understanding and generalization performance.

To ensure fair and direct comparison with existing methods, image captioning performance is evaluated on the official test splits of four widely used benchmark datasets, including *remote sensing image caption dataset (RSICD)*, *Northwestern Polytechnical University Captions (NWPU-Captions)*, *University of California, Merced Captions (UCM-Captions)*, and *Sydney-Captions*.

For visual grounding evaluation, we adopt the official test split of *DIOR-based remote sensing visual grounding dataset (DIOR-RSVG)*, which provides fine-grained region–text correspondence annotations for cross-modal localization assessment.

To further investigate the cross-modality generalization capability of the proposed framework, we additionally employ the recently released *HyperCap* benchmark [58]. HyperCap is a hyperspectral image captioning benchmark constructed from four representative hyperspectral RS datasets, namely *Indian Pines*, *Botswana*, *Houston13*, and *Kennedy Space Center (KSC)*. These datasets cover a wide range of land-cover categories, spectral characteristics, and scene complexities, comprising 200, 145, 48, and 176 spectral bands with 16, 14, 7, and 13 land-cover classes, respectively. In total, the benchmark contains 21 237 image–text pairs, providing a representative testbed for evaluating vision–language models on hyperspectral RS imagery.

V. EXPERIMENTS

A. Experimental Setup

1) *Datasets*: To comprehensively evaluate the proposed framework, experiments are conducted on multiple widely used RS vision–language benchmarks. Specifically, RS-GPT4V is adopted for model training, while image captioning performance is evaluated on four standard benchmark datasets, including RSICD, NWPU-Captions, UCM-Captions, and Sydney-Captions. For visual grounding, the DIOR-RSVG benchmark is employed to assess fine-grained cross-modal localization capability. In addition, the recently released

HyperCap benchmark is utilized for hyperspectral image captioning evaluation to further investigate the cross-modality generalization capability of the proposed framework. Furthermore, qualitative visualization analyses are conducted on both optical and hyperspectral RS imagery to investigate the effectiveness and interpretability of the proposed DRIS and MS-VLAM mechanisms. These datasets cover diverse scene types, sensing characteristics, spatial resolutions, and semantic complexities, enabling comprehensive evaluation of image captioning, visual grounding, multimodal reasoning, and cross-modality generalization capability.

2) *Implementation Details*: The proposed framework is implemented in PyTorch with compute unified device architecture (CUDA) acceleration and mixed-precision (bfloat16) training. All experiments are conducted on an NVIDIA A800 GPU with 80GB memory. The AdamW optimizer is adopted with an initial learning rate of 3×10^{-4} . During training, a warm-up schedule of 100 steps is applied, followed by a linear learning-rate decay strategy.

The framework is built upon the GeoChat architecture. CLIP ViT-L/14 is employed as the vision encoder, while vicuna-7b-v1.5 serves as the language decoder. To stabilize the initial cross-modal semantic space, a small subset (1%) of the COCO dataset is used for image-text prealignment. This stage is not task-specific and can be replaced by other image-text datasets without affecting the overall framework design.

For reproducibility, we summarize the implementation details of the proposed framework in Algorithm 1. The DRIS module follows a token-level coarse-to-fine visual encoding paradigm. Given an input image, a low-resolution version is first generated at 224×224 resolution. This low-resolution image is then upsampled via bilinear interpolation to 504×504 to construct a coarse branch aligned with the high-resolution scale. Both the original high-resolution image and the interpolated coarse image are processed by a shared CLIP vision encoder, producing two sets of visual tokens.

The saliency score of each visual token is computed as the L2 norm of its corresponding high-resolution feature. The top- K_d tokens with the highest saliency scores are selected as informative regions. A binary token mask is constructed to fuse coarse and high-resolution tokens, where selected tokens retain high-resolution features while the remaining tokens are replaced by coarse features. This design eliminates the need for external object detectors or bounding box annotations.

For MS-VLAM, hierarchical multiscale fusion is performed over the DRIS-enhanced tokens. Specifically, local token-level features are first projected into the LLM embedding space. Region-level features are obtained by selecting the top- K_m tokens based on feature norms and aggregating them via mean pooling, while global-level features are computed using global mean pooling over all tokens. The final visual representation is obtained via fixed-weight fusion of object-level, region-level, and global-level features, controlled by α , β , and γ .

During training, the fused multiscale visual representation is inserted into the LLM input sequence. The model is optimized using only the caption generation loss. Only LoRA parameters and lightweight non-LoRA trainable modules are updated, while the vision encoder and language model remain frozen.

Algorithm 1 DRIS-MS-VLAM LoRA Training Pipeline

Require: Training dataset $D = \{(I, T)\}$; vision encoder E_v ; LLM M_l ; DRIS parameters (s_l, s_h, K_d); fixed MS-VLAM fusion weights α, β, γ

Ensure: Learned LoRA parameters W_{loa} and non-LoRA trainable weights $W_{\text{non-loa}}$

- 1: Initialize M_l with LoRA adapters
 - 2: Initialize vision encoder E_v
 - 3: Integrate DRIS and MS-VLAM modules into the framework
 - 4: **for** each image-text pair $(I, T) \in D$ **do**
 - 5: **DRIS: Coarse-to-fine visual encoding**
 - 6: Generate low-resolution image I_{low} with size $s_l \times s_l$
 - 7: Resize I_{low} to high-resolution size $s_h \times s_h$ to obtain I_{coarse}
 - 8: Extract coarse visual tokens $F_{\text{low}} \leftarrow E_v(I_{\text{coarse}})$
 - 9: Extract high-resolution visual tokens $F_{\text{high}} \leftarrow E_v(I)$
 - 10: Compute saliency scores $s_j \leftarrow \|F_{\text{high},j}\|_2$
 - 11: Select Top- K_d salient tokens Ω
 - 12: Construct token-level mask m_j
 - 13: Fuse DRIS visual tokens $F_{\text{dris},j} \leftarrow (1 - m_j)F_{\text{low},j} + m_jF_{\text{high},j}$
 - 14: **MS-VLAM: Hierarchical multiscale fusion**
 - 15: Compute local token features $H_{\text{local}} \leftarrow \text{LN}(W_{\text{local}}F_{\text{dris}})$
 - 16: Select Top- K_m region tokens
 - 17: Compute region feature H_{reg} by mean pooling
 - 18: Compute global feature H_{global} by mean pooling
 - 19: Fuse multilevel features:
 - 20: $H_{\text{img}} \leftarrow \alpha H_{\text{local}} + \beta H_{\text{reg}} + \gamma H_{\text{global}}$
 - 21: Insert H_{img} into LLM input sequence
 - 22: Compute caption loss $\mathcal{L}_{\text{caption}}$
 - 23: Update W_{loa} and $W_{\text{non-loa}}$
 - 24: **end for**
 - 25: **return** $W_{\text{loa}}, W_{\text{non-loa}}$
-

3) *Comparison Protocol*: Since existing RS vision-language models are developed with different visual encoders, language decoders, and parameter tuning strategies, achieving completely identical experimental settings across all compared methods is generally infeasible. Therefore, all quantitative comparisons in this work follow the official benchmark protocols reported in the corresponding publications. Specifically, all methods are evaluated on the official benchmark datasets using the standard test splits and widely adopted evaluation metrics without modifying the original evaluation procedures.

For the proposed RAH-VLA, the implementation is built upon the GeoChat framework while preserving the original benchmark evaluation protocol. The adaptive resolution strategy introduced by DRIS is applied only during visual feature extraction and does not alter the benchmark definition, inference procedure, or evaluation metrics. Consequently, the reported performance improvements originate from the proposed DRIS and MS-VLAM modules rather than differences in the evaluation protocol.

Whenever available, baseline results are reproduced from the official implementations; otherwise, the officially reported

TABLE II
IMAGE CAPTIONING RESULTS ON NWPU-CAPTIONS, RSICD, UCM, AND SYDNEY DATASETS

Evaluation Dataset	Method	Visual Encoder	Text Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
NWPU-Captions	MLCA-NET [59]	VGG16	LSTM	0.745	0.624	0.541	0.478	0.337	0.601	1.164	0.285
	RS-CapRet [60]	CLIP-Cap-4	LlamaV2	0.871	0.786	0.713	0.650	0.439	0.775	1.919	0.320
	RS-CapRet _{finetuned} [60]	CLIP-Cap-4	LlamaV2	0.871	0.787	0.717	0.656	0.436	0.776	1.929	0.311
	RAH-VLA(Ours)	CLIP ViT-L/14	vicuna-7b-v1.5	0.889	0.812	0.748	0.689	0.458	0.791	2.061	0.334
RSICD	MLCA-NET [59]	VGG16	LSTM	0.757	0.634	0.539	0.461	0.351	0.646	2.356	0.444
	RS-GPT [13]	EVA-G	Vicuna	0.703	0.542	0.440	0.368	0.301	0.533	1.029	NA
	SkyEyeGPT [61]	EVA-G	LlamaV2-Chat	0.867	0.767	0.673	0.600	0.354	0.626	0.837	NA
	RS-CapRet [60]	CLIP-Cap-4	LlamaV2	0.741	0.622	0.529	0.455	0.376	0.649	2.605	0.484
	RS-CapRet _{finetuned} [60]	CLIP-Cap-4	LlamaV2	0.720	0.599	0.506	0.433	0.370	0.633	2.502	0.474
	RAH-VLA(Ours)	CLIP ViT-L/14	vicuna-7b-v1.5	0.772	0.662	0.578	0.505	0.392	0.672	2.781	0.501
UCM	MLCA-NET [59]	VGG16	LSTM	0.826	0.770	0.717	0.668	0.435	0.772	3.240	0.473
	RS-GPT [13]	EVA-G	Vicuna	0.861	0.791	0.723	0.657	0.422	0.783	3.332	NA
	SkyEyeGPT [61]	EVA-G	LlamaV2-Chat	0.907	0.857	0.816	0.784	0.462	0.795	2.368	NA
	RS-CapRet [60]	CLIP-Cap-4	LlamaV2	0.833	0.760	0.699	0.645	0.447	0.786	3.429	0.525
	RS-CapRet _{finetuned} [60]	CLIP-Cap-4	LlamaV2	0.843	0.779	0.722	0.670	0.472	0.817	3.548	0.525
	RAH-VLA(Ours)	CLIP ViT-L/14	vicuna-7b-v1.5	0.918	0.868	0.829	0.798	0.489	0.832	3.701	0.547
Sydney	MLCA-NET [59]	VGG16	LSTM	0.831	0.742	0.659	0.580	0.390	0.711	2.324	0.409
	RS-GPT [13]	EVA-G	Vicuna	0.823	0.753	0.686	0.622	0.414	0.748	2.731	NA
	SkyEyeGPT [61]	EVA-G	LlamaV2-Chat	0.919	0.856	0.809	0.774	0.466	0.777	1.811	NA
	RS-CapRet [60]	CLIP-Cap-4	LlamaV2	0.782	0.688	0.611	0.545	0.383	0.704	2.390	0.423
	RS-CapRet _{finetuned} [60]	CLIP-Cap-4	LlamaV2	0.787	0.700	0.628	0.564	0.388	0.707	2.392	0.434
	RAH-VLA(Ours)	CLIP ViT-L/14	vicuna-7b-v1.5	0.927	0.872	0.826	0.793	0.482	0.789	2.864	0.458

This table reports image captioning performance on four benchmark remote sensing datasets, including NWPU-Captions, RSICD, UCM-Captions, and Sydney-Captions. Evaluation metrics include BLEU-1/2/3/4, METEOR, ROUGE-L, CIDEr, and SPICE, where higher values indicate better performance. The proposed method consistently achieves the best or highly competitive results across different datasets and evaluation metrics, demonstrating its effectiveness for remote sensing image captioning.

results from the corresponding publications are adopted under the same benchmark settings. Although different baseline methods are originally pretrained or fine-tuned using their own training configurations, all quantitative comparisons are performed on the same public benchmark datasets and official evaluation splits. This protocol ensures that all comparisons are conducted under comparable evaluation conditions and provides a fair assessment of the proposed framework.

4) *Evaluation Metrics*: We adopt standard evaluation metrics for image captioning, including BLEU-1/2/3/4, metric for evaluation of translation with explicit ordering (METEOR), ROUGE-L, CIDEr, and semantic propositional image caption evaluation (SPICE). For visual grounding tasks, Accuracy@0.5 is employed to evaluate localization performance. For RS-GPT4V-Instruct evaluation, GPT-4V-based scores on complex reasoning, detailed description, and overall performance are reported. These metrics provide a comprehensive assessment of caption quality, grounding accuracy, and high-level cross-modal understanding capability.

B. Main Results

To comprehensively evaluate the effectiveness of the proposed framework, we conduct experiments on three representative RS vision-language tasks, including image captioning, visual grounding, and multimodal reasoning. Quantitative comparisons are performed on multiple public benchmarks, while GPT-4V-based evaluation is adopted to assess high-level reasoning and descriptive capabilities. The results consistently demonstrate the superiority of the proposed framework across different tasks and evaluation settings.

1) *Image Captioning Performance*: As shown in Table II, the proposed framework achieves the best overall performance across all four benchmark datasets, including NWPU-Captions, RSICD, UCM-Captions, and Sydney-Captions.

Specifically, on the NWPU-Captions benchmark, the proposed method achieves a BLEU-4 score of 0.689 and a CIDEr score of 2.061, outperforming the strongest baseline RS-CapRet_{finetuned} by 5.0% and 6.8%, respectively. Similar improvements are consistently observed on the remaining datasets.

On the UCM-Captions benchmark, our method further achieves a BLEU-4 score of 0.798 and a CIDEr score of 3.701, achieving the best performance among all compared methods. Moreover, the proposed framework consistently obtains the highest METEOR, ROUGE-L, and SPICE scores across most evaluation settings, demonstrating superior capability in generating accurate, semantically coherent, and context-aware descriptions for RS imagery.

The consistent improvements across different datasets indicate that the proposed framework effectively captures both fine-grained visual details and global scene semantics. This performance gain can be attributed to the complementary effects of DRIS and MS-VLAM. Specifically, DRIS enhances visual representation quality through resolution-adaptive processing, enabling efficient perception of both large-scale scene context and local details. Meanwhile, MS-VLAM strengthens hierarchical semantic correspondence across object-level, region-level, and global-level representations, thereby improving cross-modal semantic consistency. Consequently, the proposed framework achieves more effective vision-language interaction and delivers substantial improvements in RS image captioning performance.

2) *Visual Grounding Performance*: As shown in Table III, the proposed framework achieves the best visual grounding performance on the DIOR-RSVG dataset, obtaining an Accuracy@0.5 score of 40.27. Compared with the strongest baseline MoE-LoRA, which achieves an Accuracy@0.5 score of 37.86, the proposed method improves localization accuracy by 2.41 percentage points. More importantly, our framework

TABLE III
PERFORMANCE OF MODELS ON DIOR-RSVG DATASET

Method	Accuracy@0.5
Qwen-vl-Chat	25.05
LLaVA-1.5	9.52
Full-FT	36.31
LoRA	33.15
MoE-LoRA	37.86
RAH-VLA(Ours)	40.27

This table presents the accuracy (Pr@0.5) of different models on the DIOR-RSVG dataset. The models are trained and tested using the standard DIOR dataset split, with higher values indicating better performance. Our proposed method achieves the best performance, demonstrating its effectiveness for remote sensing visual grounding.

TABLE IV
GPT-4V-BASED PERFORMANCE EVALUATION

Method	Complex Reasoning & Conversation	Detailed Description	Overall Score
LLaVA-1.5	5.210	5.088	5.194
Qwen-vl-Chat	2.648	2.282	2.599
InternLM-XC2	5.312	4.392	5.189
Full-FT	6.270	6.530	6.304
LoRA	6.061	6.374	6.103
MoE-LoRA	6.108	6.468	6.156
RAH-VLA(Ours)	6.512	6.781	6.574

This table reports the GPT-4V evaluation scores of various models on the RS-GPT4V-Instruct dataset, focusing on two core tasks: complex reasoning & conversation, and detailed description. Scores range from 1 to 10, with higher values representing more accurate and high-quality responses. The overall score is computed as a comprehensive evaluation metric. Our method consistently achieves the best performance across all evaluation aspects.

substantially outperforms general multimodal models such as Qwen-vl-Chat and LLaVA-1.5, demonstrating stronger adaptation capability to RS scenarios.

These results indicate that the proposed hierarchical vision-language alignment strategy effectively improves fine-grained spatial correspondence between textual expressions and visual regions. In particular, the object-level and region-level alignment mechanisms provide more accurate localization cues, enabling the model to better identify and ground RS objects under complex spatial distributions and large-scale variations. The consistent improvement further validates the effectiveness of combining DRIS and MS-VLAM for RS visual grounding.

3) Cross-Modal Reasoning and Description Performance: The GPT-4V-based evaluation results on the RS-GPT4V-Instruct benchmark are summarized in Table IV. The proposed framework achieves the best performance across all evaluation dimensions, obtaining scores of 6.512 for complex reasoning and conversation, 6.781 for detailed description, and 6.574 overall. Compared with the strongest baseline Full-FT, the proposed method improves the overall score from 6.304 to 6.574, demonstrating consistent gains in both reasoning and descriptive capabilities.

These results indicate that the proposed framework is not limited to conventional image captioning or visual grounding, but also effectively enhances high-level cross-modal semantic understanding and reasoning ability. The improvement can be attributed to the complementary effects of DRIS and MS-VLAM. DRIS improves the perception of informative visual content through resolution-adaptive processing, while

MS-VLAM strengthens semantic interaction between visual and textual modalities through hierarchical alignment at the object, region, and global levels.

As a result, the proposed framework generates more accurate, comprehensive, and semantically coherent descriptions while demonstrating stronger reasoning capability in complex RS scenarios. Overall, the results further validate the effectiveness of RAH-VLA across captioning, grounding, and advanced multimodal understanding tasks.

4) Generalization to Hyperspectral RS: To further evaluate the transferability of the proposed framework beyond conventional optical RS imagery, we conduct a quantitative cross-domain generalization study on the HyperCap benchmark.

Table V summarizes the captioning results. Compared with the original GeoChat backbone, the proposed framework consistently improves all evaluation metrics across the four hyperspectral datasets. Specifically, the overall BLEU-1, METEOR, ROUGE-L, and CIDEr scores increase by 24.9%, 43.9%, 33.2%, and 149.8%, respectively. Particularly, the CIDEr score is improved by more than 100% on every individual dataset, with the largest gain reaching 190.9% on the KSC dataset.

These consistent improvements demonstrate that the proposed DRIS and MS-VLAM are not limited to optical RS imagery, but can also generalize effectively to hyperspectral image captioning. The results provide quantitative evidence supporting the strong cross-domain generalization capability of the proposed framework across heterogeneous sensing modalities.

TABLE V
GENERALIZATION PERFORMANCE ON THE HYPERCAP BENCHMARK

Dataset	Method	BLEU-1 $\uparrow$	METEOR $\uparrow$	ROUGE-L $\uparrow$	CIDEr $\uparrow$
Botswana	GeoChat-7B	0.0245	0.0563	0.0332	0.0128
	RAH-VLA (Ours)	0.0304 ($\uparrow$24.3%)	0.0815 ($\uparrow$44.7%)	0.0433 ($\uparrow$30.5%)	0.0370 ($\uparrow$188.1%)
Houston13	GeoChat-7B	0.0239	0.0572	0.0324	0.0146
	RAH-VLA (Ours)	0.0339 ($\uparrow$41.7%)	0.1028 ($\uparrow$79.6%)	0.0515 ($\uparrow$59.1%)	0.0383 ($\uparrow$162.7%)
Indian Pines	GeoChat-7B	0.0263	0.0581	0.0341	0.0155
	RAH-VLA (Ours)	0.0319 ($\uparrow$21.3%)	0.0758 ($\uparrow$30.4%)	0.0433 ($\uparrow$27.0%)	0.0329 ($\uparrow$111.8%)
KSC	GeoChat-7B	0.0087	0.0206	0.0123	0.0082
	RAH-VLA (Ours)	0.0109 ($\uparrow$25.0%)	0.0349 ($\uparrow$69.3%)	0.0169 ($\uparrow$37.8%)	0.0239 ($\uparrow$190.9%)
Overall	GeoChat-7B	0.0214	0.0485	0.0284	0.0124
	RAH-VLA (Ours)	0.0267 ($\uparrow$24.9%)	0.0698 ($\uparrow$43.9%)	0.0378 ($\uparrow$33.2%)	0.0310 ($\uparrow$149.8%)

The values in parentheses denote the relative improvement over the GeoChat baseline. The proposed RAH-VLA consistently improves all evaluation metrics across four representative hyperspectral datasets, demonstrating strong cross-modality generalization capability.

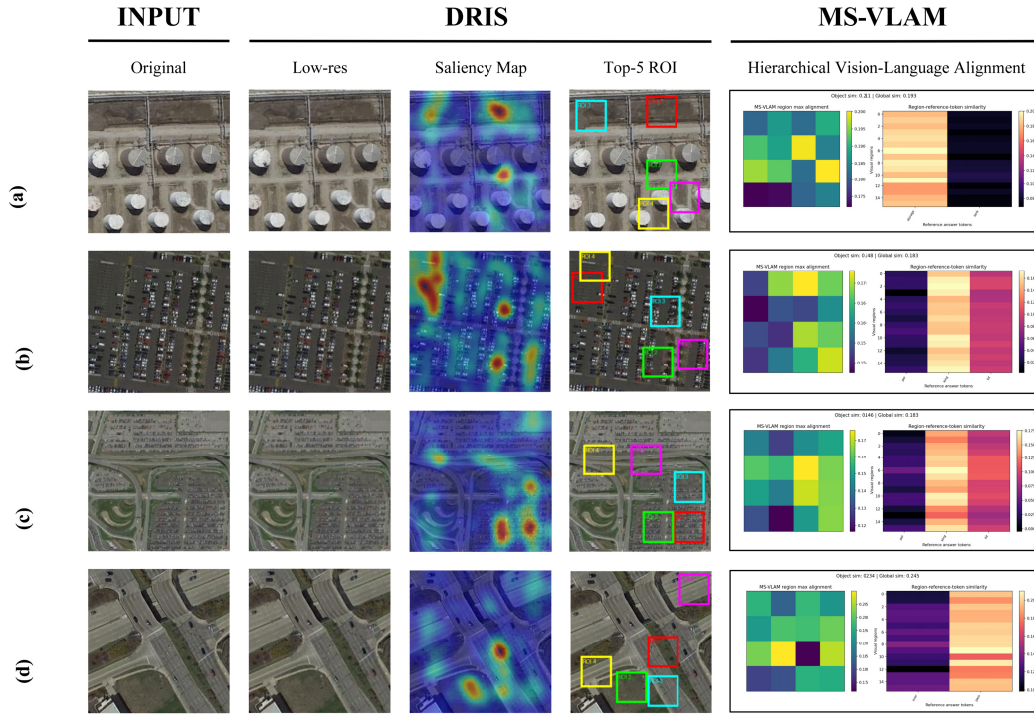


Fig. 5. *Qualitative Visualization of the Proposed DRIS and MS-VLAM Mechanisms.* For each representative RS scene, the figure shows the original image, low-resolution global perception, saliency map, selected top-5 ROIs, and the corresponding hierarchical vision–language alignment results. Warmer colors in the saliency maps indicate regions with higher semantic importance, while brighter colors in the alignment maps indicate stronger semantic correspondence between visual regions and textual tokens. (a) Storage tank. (b) Parking lot-I. (c) Parking lot-II. (d) Urban overpass.

C. Visualization Analysis

To further analyze the behavior and interpretability of the proposed framework, we provide qualitative visualization results on both optical RS imagery and hyperspectral RS imagery.

1) *DRIS Visualization:* As shown in Fig. 5, the low-resolution branch first captures global scene context and generates saliency maps for region importance estimation. The selected top-5 ROIs consistently concentrate on semantically informative structures, such as storage tanks, parking areas, and road intersections, while large homogeneous background regions receive relatively low attention. This observation

demonstrates that DRIS can effectively allocate computational resources according to scene complexity and preserve fine-grained semantic details through adaptive high-resolution refinement. The visualization results further verify that the proposed resolution-adaptive strategy successfully balances global contextual perception and local detail preservation.

2) *MS-VLAM Visualization:* The alignment visualizations in Fig. 5 further illustrate the effectiveness of MS-VLAM in establishing hierarchical vision–language correspondence. The region–token similarity maps show that semantically relevant textual tokens consistently exhibit stronger responses to informative visual regions, indicating successful cross-modal

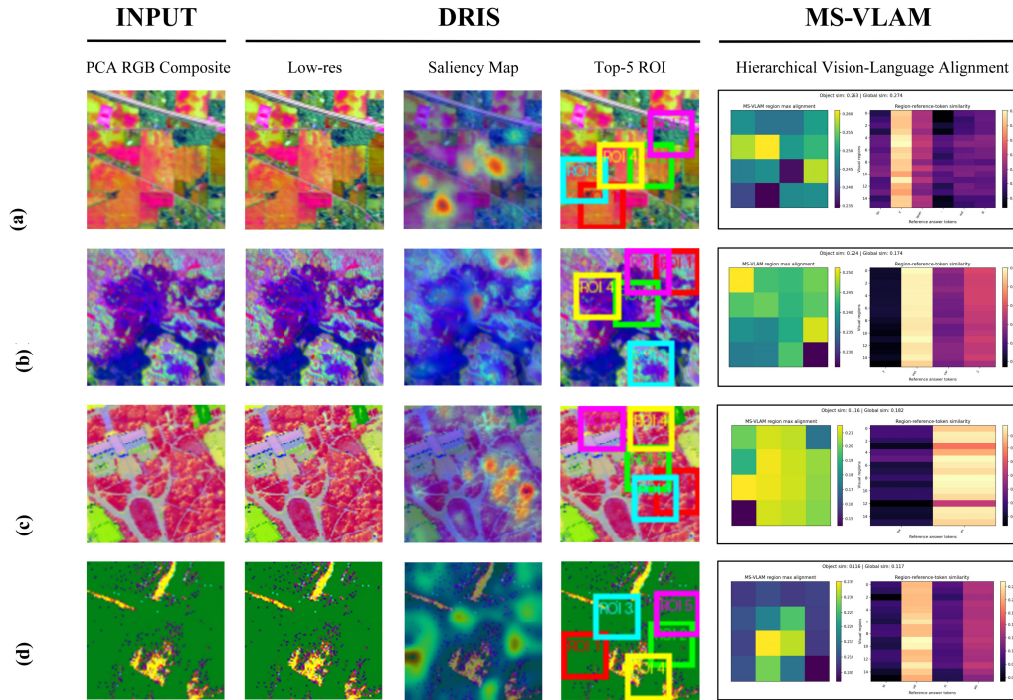


Fig. 6. *Hyperspectral Visualization of the Proposed DRIS and MS-VLAM Mechanisms.* The figure presents qualitative results on four representative hyperspectral datasets, including Indian Pines, Botswana, Houston13, and KSC. For each example, the figure shows the PCA-based red, green, and blue (RGB) composite visualization, low-resolution global perception, saliency map, selected top-5 ROIs, and the corresponding hierarchical vision–language alignment results. Warmer colors in the saliency maps indicate regions with higher semantic importance, while brighter colors in the alignment maps indicate stronger semantic correspondence between visual regions and textual tokens. (a) Indian Pines soybean-notill. (b) Botswana firescar2. (c) Houston13 trees. (d) KSC mud flats.

semantic matching. Moreover, consistent alignment patterns can be observed across different scene categories, suggesting that the learned vision–language correspondence is not dependent on a specific scene type. These observations demonstrate that the proposed hierarchical alignment mechanism effectively captures semantic relationships between visual content and textual descriptions across multiple semantic granularities, thereby improving cross-modal semantic consistency and interpretability.

3) Hyperspectral Generalization Analysis: To further evaluate the cross-modality generalization capability of the proposed framework, we additionally conduct qualitative visualization analysis on the *HyperCap* benchmark, a hyperspectral image captioning dataset constructed from four representative hyperspectral datasets, namely *Indian Pines*, *Botswana*, *Houston13*, and *KSC*. These datasets were not involved during the original optical RS training process, as shown in Fig. 6.

The visualization results demonstrate that the proposed DRIS remains effective under hyperspectral imaging conditions. Specifically, the saliency maps consistently highlight semantically informative regions, while the selected top-5 ROIs are concentrated on discriminative land-cover structures rather than homogeneous background areas. This observation indicates that the resolution-adaptive region selection strategy is not restricted to optical RS imagery and can generalize to hyperspectral representations with rich spectral characteristics.

Furthermore, the MS-VLAM alignment visualizations exhibit stable region–token correspondence across different

hyperspectral scenes. Despite the substantial distribution differences between optical and hyperspectral imagery, the proposed hierarchical alignment mechanism is still able to establish meaningful semantic relationships between visual regions and textual descriptions. The region–token similarity maps show that semantically relevant textual tokens consistently produce stronger responses to informative hyperspectral regions, indicating effective cross-modal semantic matching.

Overall, these qualitative results provide evidence that the proposed framework can maintain effective region selection and hierarchical semantic alignment across heterogeneous sensing modalities. This observation suggests that the proposed framework generalizes beyond conventional optical RS imagery and remains applicable to hyperspectral RS scenarios.

D. Ablation Study

To evaluate the contribution of each component, we conduct ablation experiments on the DIOR-RSVG dataset, as shown in Table VI. Accuracy@0.5 is used as the evaluation metric.

1) Component-Level Analysis: We first analyze the individual contributions of DRIS and MS-VLAM. The baseline model (Row A) achieves an Accuracy@0.5 of 33.15.

When DRIS is introduced (Row B), the performance improves to 35.62, indicating that resolution-adaptive processing enhances feature quality by preserving fine-grained details. When MS-VLAM is applied independently (Row C), the accuracy increases to 37.04, showing that hierarchical alignment significantly improves cross-modal correspondence.

TABLE VI
ABLATION STUDY ON DIOR-RSVG (ACCURACY@0.5)

ID	DRIS	MS-VLAM	MS-VLAM Sub-modules	Accuracy@0.5
A	✗	✗	—	33.15 (LoRA baseline)
B	✓	✗	—	35.62
C	✗	✓	full	37.04
D	✓	✓	full	40.27 (Ours)
<i>Ablating MS-VLAM components (with DRIS enabled)</i>				
E	✓	✓	w/o Object-level	38.91
F	✓	✓	w/o Local-region-level	39.10
G	✓	✓	w/o Global-level	38.76

Notes: (1) “✓” denotes the module is enabled; “✗” denotes it is disabled. (2) Row A is the LoRA baseline without any proposed module. (3) Rows E–G respectively remove one alignment level while keeping the other two. (4) All results are averaged over 3 runs with standard deviation < 0.15 .

When both DRIS and MS-VLAM are jointly applied (Row D), the model achieves the best performance of 40.27. This demonstrates that the two modules are complementary: DRIS improves visual representation, while MS-VLAM aligns these representations with textual semantics at multiple levels.

2) *Submodule Analysis of MS-VLAM:* We further analyze the contributions of each alignment level in MS-VLAM by removing one component at a time while keeping DRIS enabled.

Removing object-level alignment (Row E) reduces the accuracy to 38.91. Removing local-region-level alignment (Row F) leads to a result of 39.10, while removing global-level alignment (Row G) decreases the performance to 38.76.

These results indicate that all three alignment levels contribute to performance improvement. In particular, object-level alignment improves fine-grained correspondence, while global-level alignment provides complementary scene-level context.

3) *Discussion:* Overall, the ablation results demonstrate that both resolution-adaptive visual perception and hierarchical vision–language semantic alignment play indispensable roles in multimodal RS understanding. Specifically, DRIS enhances visual representation quality by adaptively preserving informative local details while maintaining global contextual awareness, whereas MS-VLAM strengthens semantic consistency through hierarchical alignment across object-, region-, and global-level representations. Their joint optimization consistently achieves the best performance, indicating that adaptive visual perception and hierarchical semantic alignment are highly complementary for robust vision–language understanding in RS scenarios.

Moreover, the complementary improvements observed in the ablation study further confirm that DRIS and MS-VLAM are not independent modules, but rather two tightly coupled components within a unified representation learning framework. By emphasizing semantically informative regions and suppressing redundant background information, DRIS produces higher-quality resolution-adaptive multiscale visual representations that serve as reliable inputs for hierarchical semantic alignment. Building upon these enhanced representations, MS-VLAM establishes more accurate cross-modal correspondence at multiple semantic granularities.

TABLE VII
EFFICIENCY COMPARISON WITH BASELINE METHODS

Method	FLOPs (G)	Time (ms)	Memory (GB)
Baseline	286.4	164.2	21.7
Token Pruning	238.7	139.5	19.4
RAH-VLA(Ours)	176.3	105.8	16.2

This table reports the computational efficiency of different methods in terms of FLOPs, inference latency, and GPU memory consumption. Lower values indicate higher computational efficiency.

Consequently, the two modules reinforce each other during end-to-end optimization, explaining why their joint integration consistently outperforms configurations that employ either module alone.

E. Efficiency Analysis

To evaluate the computational efficiency of the proposed DRIS, we compare the proposed framework with both the baseline model and a representative token-pruning strategy. The comparison includes three commonly used efficiency metrics, namely floating-point operations (FLOPs), inference latency, and GPU memory consumption.

As shown in Table VII, the proposed method consistently achieves the lowest computational cost among all compared approaches. Specifically, compared with the baseline model, DRIS reduces FLOPs from 286.4G to 176.3G, inference latency from 164.2 to 105.8 ms, and memory consumption from 21.7 to 16.2 GB. These correspond to relative reductions of 38.4%, 35.6%, and 25.3%, respectively.

Compared with the token-pruning strategy, the proposed framework further reduces FLOPs, inference time, and memory usage while maintaining effective multimodal understanding performance. This improvement can be attributed to the resolution-adaptive processing mechanism, which allocates high-resolution computation only to semantically informative regions while processing homogeneous regions at lower resolution. Consequently, redundant computation is significantly reduced without sacrificing important visual details.

These results demonstrate that DRIS effectively balances computational efficiency and semantic fidelity, validating the

effectiveness of the proposed resolution-adaptive paradigm for large-scale multimodal RS understanding.

VI. CONCLUSION

This article presents *RAH-VLA*, a resolution-adaptive hierarchical vision-language alignment framework for multimodal RS image interpretation. To address the limitations of fixed-resolution processing and single-scale semantic alignment, we propose a DRIS and an MS-VLAM, which jointly enable resolution-adaptive visual perception and hierarchical cross-modal semantic alignment. Extensive experiments demonstrate that the proposed framework consistently improves image captioning, visual grounding, and cross-modal reasoning performance while effectively reducing computational redundancy. In addition, qualitative visualization analyses further verify the strong cross-modality generalization capability of the proposed framework. Although the proposed framework has demonstrated promising performance across optical and hyperspectral RS scenarios, further investigation is required to extend its applicability to more diverse sensing modalities and larger-scale vision-language foundation models. Future work will focus on lightweight model design, efficient inference strategies, and more advanced multimodal reasoning mechanisms for large-scale RS applications.

ACKNOWLEDGMENT

The authors would like to thank the developers and maintainers of the public remote sensing (RS) datasets used in this study for making these valuable resources available to the research community.

REFERENCES

- [1] A. Xiao et al., "Foundation models for remote sensing and Earth observation: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 4, pp. 297–324, Dec. 2025.
- [2] Q. Zhao et al., "An overview of the applications of Earth observation satellite data: Impacts and future trends," *Remote Sens.*, vol. 14, no. 8, p. 1863, Apr. 2022.
- [3] Y. Zhou, A. Ma, J. Wang, Z. Chen, and Y. Zhong, "Remote sensing meta modal representation for missing modality land cover mapping: From EarthMiss dataset to MetaRS method," *Remote Sens. Environ.*, vol. 333, Jan. 2026, Art. no. 115132.
- [4] B. Xuan, Z. Li, J. Kang, and Y. Hu, "Spatiotemporal relationship between human activities and urban heat in Chinese megacities based on multisource remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 17064–17079, 2025.
- [5] E. Dritsas and M. Trigka, "Remote sensing and geospatial analysis in the big data era: A survey," *Remote Sens.*, vol. 17, no. 3, p. 550, Feb. 2025.
- [6] P. Ghamisi et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.
- [7] X. Weng, C. Pang, and G.-S. Xia, "Vision-language modeling meets remote sensing: Models, datasets, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 3, pp. 276–323, Sep. 2025.
- [8] N. Kieu, K. Nguyen, A. Nazib, T. Fernando, C. Fookes, and S. Sridharan, "Multimodal colearning meets remote sensing: Taxonomy, state of the art, and future works," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7386–7409, 2024.
- [9] L. Xu, L. Wang, J. Zhang, D. Ha, and H. Zhang, "A review of cross-modal image-text retrieval in remote sensing," *Remote Sens.*, vol. 17, no. 24, p. 3995, Dec. 2025.
- [10] F. Samadzadegan, A. Toosi, and F. Dadrass Javan, "A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: Current status and prospects," *Int. J. Remote Sens.*, vol. 46, no. 3, pp. 1327–1402, Feb. 2025.
- [11] S. Li and H. Tang, "Multimodal alignment and fusion: A survey," *Int. J. Comput. Vis.*, vol. 134, no. 3, p. 103, Mar. 2026.
- [12] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [13] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "RSGPT: A remote sensing vision language model and benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 224, pp. 272–286, Jun. 2025.
- [14] G. Hu, Z. Wen, Y. Lv, J. Zhang, and Q. Wu, "Global-local information soft-alignment for cross-modal remote-sensing image-text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5518915.
- [15] C. Liu, J. Zhang, K. Chen, M. Wang, Z. Zou, and Z. Shi, "Remote sensing spatiotemporal vision-language models: A comprehensive survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 14, no. 1, pp. 383–423, Feb. 2026.
- [16] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 4904–4916.
- [17] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.
- [18] B. Cheng, Q. Deng, W. Gan, Y. Li, C. Wu, and Q. Wang, "A review of remote sensing image object detection: Advances and challenges in data, techniques, and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 10438–10472, 2026.
- [19] X. He et al., "Hierarchical relation learning for few-shot semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5520409.
- [20] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 9694–9705.
- [21] A. Zavras, D. Michail, B. Demir, and I. Papoutsis, "Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment," *ISPRS J. Photogramm. Remote Sens.*, vol. 228, pp. 270–287, Oct. 2025.
- [22] C. Zheng, N. Song, R. Zhang, L. Huang, Z. Wei, and J. Nie, "Scale-semantic joint decoupling network for image-text retrieval in remote sensing," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 1, pp. 1–20, Jan. 2024.
- [23] L. Bai et al., "From orbit to ground: A comprehensive review of multimodal self-supervised learning for remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 13, no. 4, pp. 346–381, Dec. 2025.
- [24] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926.
- [25] D. Hong, C. Li, X. Li, G. Camps-Valls, and J. Chanussot, "Foundation models in remote sensing: Evolving from unimodality to multimodality," *IEEE Geosci. Remote Sens. Mag.*, vol. 14, no. 2, pp. 10–35, Apr. 2026.
- [26] K. Wu et al., "A semantic-enhanced multi-modal remote sensing foundation model for Earth observation," *Nature Mach. Intell.*, vol. 7, no. 8, pp. 1235–1249, Aug. 2025.
- [27] C. Pohl and J. L. Van Genderen, "Review article multisensor image fusion in remote sensing: Concepts, methods and applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, Jan. 1998.
- [28] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.
- [29] S. Chen, H. Dong, and W. Chen, "Hyperspectral image unmixing based on scatter-constrained nonnegative matrix factorization," *J. Chongqing Univ. Posts Telecommun.*, vol. 33, no. 2, pp. 296–303, 2021.
- [30] S. Chen and Y. Zhang, "Hyperspectral image compression based on prediction and vector quantization," *J. Chongqing Univ. Posts Telecommun.*, vol. 31, no. 5, p. 688, 2019.
- [31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
- [33] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [34] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in *Proc. 3rd ERTS-1 Symp. (NASA)*, Jan. 1974, pp. 309–317.

- [35] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.
- [36] A. J. X. Guo and F. Zhu, "A CNN-based spatial feature fusion algorithm for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7170–7181, Sep. 2019.
- [37] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.
- [38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [39] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [40] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [41] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.
- [42] Y. Liu, T. Zhang, and Y. Huang, "GeoMM-SSL: Integrating geospatial object relations in multimodal self-supervised learning for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 1–13, 2025.
- [43] G. Hoxha, O. Angyal, and B. Demir, "Self-supervised cross-modal text–image time series retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5524714.
- [44] R. Shao et al., "AllSpark: A multimodal spatiotemporal general intelligence model with ten modalities via language as a reference framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5520124.
- [45] X. Jia, Z. Chen, S. Zhang, and X. Xue, "Change-LISA: Language-guided reasoning for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, 2026, Art. no. 5521307.
- [46] C. Wang, G. Fan, J. Li, M. Gan, and C. L. Philip Chen, "MGCR-Net: Multimodal graph-conditioned vision-language reconstruction network for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, 2026, Art. no. 5521140.
- [47] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519442.
- [48] A. Günther, H. Najjar, and A. Dengel, "Explainable multimodal learning in remote sensing: Challenges and future directions," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [49] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15617–15629.
- [50] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, Aug. 2022.
- [51] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 32–66, Jun. 2024.
- [52] H. Lin et al., "RS-MoE: A vision–language model with mixture of experts for remote sensing image captioning and visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5614918.
- [53] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27831–27840.
- [54] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "A unified framework with multimodal fine-tuning for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5520643.
- [55] C. Ding, J. Tang, Z. Tang, M. Deng, W. Wang, and H. Liu, "Geographical scene: The natural unit for geographical analysis and its recognition based on data with spatial and semantic features," *Ann. Amer. Assoc. Geographers*, vol. 115, no. 8, pp. 1926–1954, Sep. 2025.
- [56] J. Ge, X. Zhang, Y. Zheng, K. Guo, and J. Liang, "RSTeller: Scaling up visual language modeling in remote sensing with rich linguistic semantics from openly available data and large language models," *ISPRS J. Photogramm. Remote Sens.*, vol. 226, pp. 146–163, Aug. 2025.
- [57] C. Zheng et al., "Cross-modal progressive perspective matching network for remote sensing image-text retrieval," *IEEE Trans. Multimedia*, vol. 27, pp. 3966–3978, 2025.
- [58] A. Das et al., "HyperCap: A hyperspectral land-cover captioning dataset for vision–language models," *IEEE Geosci. Remote Sens. Mag.*, early access, Jun. 3, 2026, doi: [10.1109/MGRS.2026.3693613](https://doi.org/10.1109/MGRS.2026.3693613).
- [59] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-captions dataset and MLCA-Net for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5513212.
- [60] J. D. Silva, J. Magalhães, D. Tuia, and B. Martins, "Large language models for captioning and retrieving remote sensing images," 2024, *arXiv:2402.06475*.
- [61] Y. Zhan, Z. Xiong, and Y. Yuan, "SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *ISPRS J. Photogramm. Remote Sens.*, vol. 221, pp. 64–77, Mar. 2025.