ROARS: A Cloud-Based Resolution-Native Framework for Large-Scale Remote Sensing Multimodal Analysis

Jingrui Zhang, Feng Liang* *Member, IEEE*, Yong Zhang, Zixuan Shangguan, Zhida Li, Xiaoyi Fan *Member, IEEE*, Victor C. M. Leung *Life Fellow, IEEE*, Guanbin Li* *Member, IEEE*, and Xiping Hu* *Senior Member, IEEE*

***Abstract*—The paradigm shift of remote sensing (RS) analysis toward cloud-based large-scale computing demands intelligent methods capable of interpreting petabyte-scale, high-resolution heterogeneous imagery. Multimodal Large Language Models (MLLMs) have emerged as a transformative solution for such data-intensive tasks, offering unprecedented capabilities in open-set perception and high-level semantic reasoning. However, directly applying general MLLM architectures to the RS domain faces significant challenges: the constraint of fixed-resolution vision encoders causes severe information loss for small objects, while the reliance on single-layer feature fusion neglects critical spatial cues preserved in intermediate layers. To address these challenges, we propose ROARS: a Cloud-Based Resolution-Native Framework for Large-Scale Remote Sensing Multimodal Analysis, tailored for high-fidelity interpretation. First, we introduce a Native-Resolution Vision Encoder tailored to process images of arbitrary aspect ratios and scales without downsampling, thereby strictly preserving pixel-level details. Second, we design a Shortcut-Based Fusion Mechanism that bridges the semantic gap by injecting multi-level visual features into the language model. Furthermore, to enhance global contextual understanding, we incorporate a Multi-Granularity Token Pooling strategy that provides fine- and coarse-grained visual representations. Extensive experiments on the VRSBench demonstrate that our framework significantly outperforms state-of-the-art approaches, particularly in fine-grained object perception and complex spatial reasoning. These results validate the proposed method as a robust and efficient core component for next-generation cloud-based RS interpretation systems. Our source code is publicly accessible.**[1]

*Feng Liang, Guanbin Li and Xiping Hu are corresponding authors.

Jingrui Zhang is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with the School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: jingrui1119@bit.edu.cn).

Yong Zhang, Zixuan Shangguan and Xiping Hu are with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: zhangyong2023, gzxshang, huxp@bit.edu.cn).

Feng Liang and Xiaoyi Fan are with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: fliang, xiaoyi.fan@smbu.edu.cn).

Zhida Li and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, and is also with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China (e-mail: lizhd6@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn).

Victor C. M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail:vleung@ece.ubc.ca).

[1]https://github.com/DistriAI/roars

***Index Terms*—Remote Sensing, Native Resolution, Multimodal Large Language Model, Vision-Language Information Fusion**

I. Introduction

REMOTE sensing (RS) image interpretation aims to extract high-value geospatial information from Earth observation data and serves as a fundamental means for understanding the physical world [1] [2]. Recently, with the exponential growth of satellite and the accumulation of petabyte-scale data, RS analysis is undergoing a profound paradigm shift from traditional local processing to cloud-based large-scale analysis, as illustrated in Fig. 1. Cloud platforms now host vast archives of high-resolution, continuously updated imagery, offering unprecedented opportunities for planetary-scale monitoring. However, effectively utilizing these cloud-hosted resources remains a significant challenge. Compared with natural images, remote sensing scenes exhibit distinct imaging mechanisms and extreme scene complexity, which pose severe obstacles for automated interpretation in cloud environments.

The challenges arise primarily from two fundamental aspects: extreme spatial resolution variation and high-level semantic complexity. First, extreme scale variation and high spatial resolution are among the most salient characteristics of remote sensing imagery. A single RS image often spans vast geographic extents, encompassing macroscopic landforms such as mountain ranges and rivers, as well as microscopic targets such as densely distributed vehicles and ships. This multi-order-of-magnitude variation in scale requires models to simultaneously possess global scene summarization capabilities and fine-grained perceptual sensitivity. Second, the semantic complexity of remote sensing scenes necessitates deep scene-level understanding. Geospatial objects in RS imagery do not exist in isolation; instead, they are embedded within intricate spatial and topological relationships, such as the co-occurrence of ports and ships or the connectivity between road networks and buildings. Traditional RS analysis models are typically constrained to closed-set perception tasks, including object detection and semantic segmentation, and lack the ability to perform contextual reasoning. Consequently, they struggle to handle open-world scenarios [3] and complex applications that

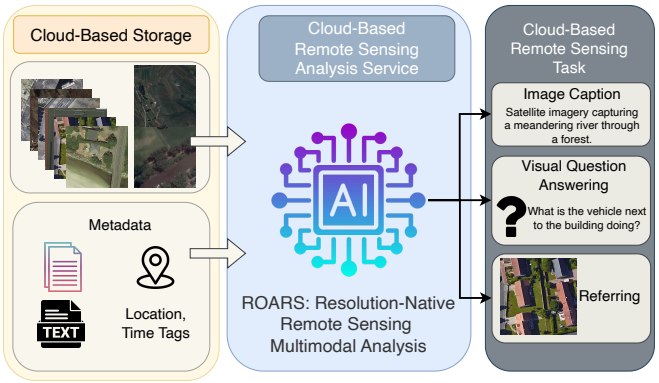

Fig. 1. Overview of the proposed ROARS framework. The system ingests remote sensing imagery alongside auxiliary metadata (e.g., location, time tags). These inputs are processed by a Multimodal Large Language Model (MLLM) deployed on a cloud computing platform to perform intelligent interpretation tasks, instantiated here by Image Captioning, Referring and Visual Question Answering.

require interpretability and decision-level reasoning, such as disaster assessment and urban planning [4] [5] [6].

In recent years, the emergence of Multimodal Large Language Models (MLLMs) has introduced a new paradigm for alleviating the aforementioned bottlenecks in remote sensing interpretation. By tightly aligning powerful large language models (LLMs) with vision encoders, MLLMs exhibit two key advantages that make them particularly promising for addressing complex RS problems. First, MLLMs demonstrate strong generalization and open-set recognition capabilities. Benefiting from large-scale pretraining, they acquire extensive world knowledge, enabling them to recognize and reason about concepts that were not explicitly observed during training. This property effectively compensates for the limited generalization ability of traditional RS models when encountering rare objects or cross-domain scenarios. Second, MLLMs enable high-level semantic reasoning and interactive analysis. Rather than functioning solely as visual feature extractors, they act as reasoning engines equipped with logical inference capabilities. Through natural language interaction, MLLMs can understand complex user intents and perform vision-grounded causal reasoning (e.g., "analyze the likelihood of flooding in this region"). This capability allows RS systems to evolve from single-purpose perception tools into intelligent assistants, better suited to the rich semantic understanding required by real-world remote sensing applications [7] [8].

Despite this promising vision, directly transferring existing general-purpose MLLM architectures into the remote sensing domain results in limited practical performance. This limitation also applies to early RS-oriented MLLMs, such as RSGPT [9] and GeoChat [7]. These limitations arise from fundamental mismatches between current architectural designs and the intrinsic properties of RS data and are primarily manifested in two aspects. **1) Resolution bottleneck and information degradation:** Most existing MLLMs adopt visual encoders originally designed for natural images (e.g., CLIP-ViT), which restrict inputs to fixed low resolutions such as 224×224 or 336×336. For large-format, detail-dependent RS imagery, this constraint necessitates aggressive downsampling or local cropping. Downsampling inevitably suppresses or eliminates small objects and fine-grained textures that are critical in RS scenes, while cropping breaks long-range spatial dependencies and contextual relationships among geospatial entities. Such preprocessing fundamentally undermines the perceptual fidelity of RS imagery. **2) Semantic discontinuity in cross-modal alignment:** At the feature fusion stage, most approaches employ token concatenation strategy, in which only the final-layer features of the vision encoder are provided to the LLM. However, prior studies indicate that deep Vision Transformer(ViT) [10] layers primarily encode global semantic information, whereas shallow and intermediate layers preserve fine-grained spatial cues [11], including object boundaries, orientations, and textures. For remote sensing tasks, such structural details (e.g., road connectivity or vehicle alignment) are essential for reliable reasoning. By discarding these lower-level cues, existing fusion schemes deprive LLMs of critical visual evidence, increasing the risk of hallucinations or erroneous predictions [12] [13].

In this work, we propose **ROARS**, a cloud-based resolution-native framework for large-scale remote sensing multimodal analysis that structurally addresses both resolution limitations and insufficient semantic alignment. First, to address the resolution problem, we replace the vanilla ViT encoder in prevailing MLLM structure, to support arbitary resolution inputs. Specifially, we replace the conventional CLIP-ViT backbone with remote sensing native resolution encoder, enabling the direct ingestion of native-resolution RS imagery. This design allows the model to encode elongated river corridors and large-scale urban scenes without sacrificing pixel-level fidelity, thereby preserving critical information about small objects and fine-grained structures. Second, to leverage visual information at different granularities and prevent the loss of low-level spatial cues during LLM decoding, we introduce a shortcut-based vision language information fusion mechanism to overcome the limitations of single-stage fusion. Rather than relying solely on the final ViT output, shortcut connections dynamically extract intermediate-layer visual features and progressively inject these multi-level representations into the LLM. This strategy achieves comprehensive alignment across low-level details, mid-level structures, and high-level semantics. Furthermore, to enhance global scene understanding over ultra-large spatial extents, we incorporate a multi-granularity token pooling module at the top of the ViT. This module generates hierarchical global visual tokens, providing the LLM with contextual cues at multiple granularities and improving reasoning robustness in complex RS scenes.

We conduct extensive experiments on the VRSBench. The results demonstrate that the proposed framework consistently outperforms existing state-of-the-art methods across tasks including fine-grained object perception, complex scene description, and spatial relationship reasoning. In particular, our model exhibits superior performance when handling scenes characterized by dense small objects and large-format imagery, validating the effectiveness of native-resolution input and multi-level feature fusion.

The main contributions of this work are summarized as

follows:

- To our best knowledge, we are the first to introduce a native-resolution visual encoding framework into RS-MLLMs. It effectively mitigates information loss caused by fixed-size inputs and significantly improves the preservation of small objects and fine-grained structures.
- We design a shortcut-based vision language information fusion mechanism. It transcends traditional single-layer interaction, enabling deep fusion between visual details and linguistic semantics.
- We design a hierarchical vision token pooling module that supplies LLMs with multi-granularity global context. This substantially strengthenes understanding depth and reasoning robustness in open-world remote sensing scenarios.
- Extensive experiments on VRS benchmark demonstrate that our method achieves state-of-the-art performance. Specifically, our model significantly outperforms existing baselines in fine-grained perception and complex reasoning tasks, validating the effectiveness of the proposed components.

II. Related Work

A. Vision Transformer

The transition [14] [15] from CNNs to Transformers marks a shift in visual representation learning. The ViT [10] treats images as patch sequences, utilizing self-attention to capture global dependencies. However, standard ViT requires fixed square inputs (e.g., 224×224), necessitating resizing that distorts aspect ratios and discards fine-grained details critical for remote sensing. To address resolution constraints, cropping or tiling strategies [16] [17] have been explored. While these preserve local details, independent processing of subregions disrupts global spatial continuity and fragments large-scale structures (e.g., roads or rivers). Recently, NaViT [18] introduced the "Patch n' Pack" paradigm, enabling the processing of variable-length sequences from images of arbitrary resolutions. By avoiding resizing or tiling, NaViT preserves both geometric structure and spatial semantics. This design is particularly well-suited for the multi-granularity, multi-aspect-ratio nature of remote sensing imagery, offering a more principled solution than conventional fixed-resolution architectures. Motivated by this, we adopt NaViT as the vision encoder in our framework to natively handle RS images of arbitrary resolutions without information loss.

B. Multimodal Large Language Models

In recent years, extending Large Language Models (LLMs) to the multimodal domain has emerged as a prominent research direction [19]. Based on the architectural mechanisms used for visual feature integration, existing approaches can be broadly categorized into two groups: connector-based methods and direct projection-based methods.

Connector-based alignment methods introduce specialized intermediate modules to bridge the gap between a frozen vision encoder and an LLM. Flamingo [20] incorporates gated cross-attention layers into the language model, enabling effective few-shot learning on large-scale multimodal data while keeping the LLM parameters frozen. BLIP-2 [21] proposes the Q-Former, a lightweight Transformer that employs learnable query embeddings to extract language-relevant information from visual embeddings and maps it into the LLM's semantic space, thereby achieving efficient alignment.

Conversely, direct projection-based methods adopt a more streamlined architectural design by directly projecting visual features into the token embedding space of the LLM via linear layers, facilitating joint training on large-scale multimodal instruction data. LLaVA [22] pioneered this paradigm, demonstrating strong instruction-following capabilities and multimodal reasoning performance. Building upon this framework, MiniGPT-V2 [23] introduces task identifiers to support multiple vision–language tasks, including image captioning, visual question answering, and object localization, within a unified model, further enhancing its generalizability.

Our method follows the second paradigm which has been widely validated as an effective design for MLLMs. In contrast, connector-based approaches often suffer from convergence difficulties when scaled to large-scale instruction tuning.

C. Remote Sensing MLLMs

Recent advancements in Remote Sensing Multimodal Large Language Models (RS-MLLMs) have been primarily driven by two complementary efforts: large-scale instruction tuning and architectural adaptation for unified task modeling.

A significant body of work focuses on data-centric scaling. SkyEyeGPT [12] employs a staged fine-tuning strategy on the SkyEye-968k dataset, enhancing performance in multi-turn dialogue and object detection. Similarly, VHM [24] mitigates the "blind affirmation" hallucination issue by leveraging the VersaD and HnstD datasets, demonstrating effectiveness in multi-task settings. SkySenseGPT [25] further pushes this boundary by constructing a million-scale instruction dataset to benchmark relational understanding. However, these approaches rely heavily on data expansion to drive performance, often neglecting explicit structural modeling or architectural optimizations tailored to the unique spatial characteristics of remote sensing imagery.

Concurrently, other research seeks to enhance model capabilities through architectural adaptation. Early attempts like RS-GPT [9] explored LLMs for captioning and VQA but suffered from task-specific fragmentation, requiring separate training for different tasks. RS-LLaVA [26] addressed this by fine-tuning LLaVA for joint multi-task modeling; however, its reliance on a standard CLIP-ViT backbone limits its ability to capture the fine-grained spatial details inherent in geospatial scenes. To tackle high-resolution demands, GeoChat [7] adapted CLIP-ViT to support region-level interactions, yet its training data remains predominantly optical, lacking multi-sensor diversity (e.g., SAR, infrared). While EarthGPT [8] attempts to unify multi-sensor interpretation by integrating multi-granularity features, it is still constrained by conventional ViT architectures and shallow alignment modules, which limits the stability and reliability of regional spatial reasoning.

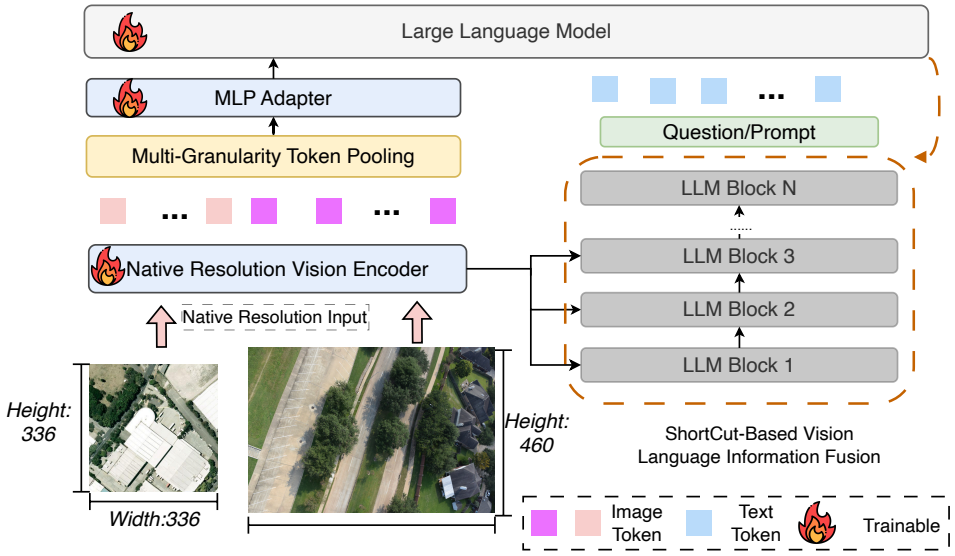

Fig. 2. The overall architecture of the proposed Vision-Language Model for processing remote sensing imagery. It is composed of three main components: (1) A Remote Sensing Native Resolution Encoder (based on NaViT or similar architecture) processes satellite images of varying sizes (e.g., 336×336 and 460×680) without aggressive resizing. (2) The visual features are passed through an MLP Adapter for dimensionality matching and space projection into the language domain. (3) The aligned visual features and the text prompt (processed by the Tokenizer) are fed into a Large Language Model (LLM), which is responsible for generating the final detailed description (Caption) of the remote sensing inputs.

In summary, although current approaches have achieved significant progress through data-centric scaling and architectural adaptation, they often neglect fundamental modeling innovations. Most studies rely on standard visual backbones constrained by fixed input resolutions and rigid patch divisions. These limitations render such architectures suboptimal for remote sensing data, which is characterized by diverse scales and variable aspect ratios. To address this, we introduce a methodology designed to natively support high-resolution and structurally complex scenes.

III. Method

A. Overview

Fig. 2 illustrates the overall architecture of our model, which follows the mainstream MLLM design paradigm and consists of three main components: a vision encoder, an adaptor, and an LLM module. The workflow operates as follows to address the specific challenges highlighted in the introduction: First, the **Native Resolution Vision Encoder**(III-B) accepts remote sensing images of arbitrary aspect ratios and scales. This design directly resolves the geometric distortion issue caused by traditional resizing, ensuring that the structural integrity of large geospatial scenes (e.g., rivers, road networks) is preserved in the resulting feature sequence. Subsequently, to prevent the loss of fine-grained cues necessary for small object recognition, we design a **ShortCut-Based Fusion Mechanism** (III-C) that extracts features from different layers. Then, to counterbalance the risk of global semantic dilution caused by the lengthy token sequences from remote sensing native resolution encoder, we implement the **Multi-Granularity Token Pooling**(III-D) strategy. This module augments the visual representation by generating multi-scale summaries (e.g., 16×16 and 4×4 grids) and concatenating them with the original tokens. Finally, this composite token sequence which containing both high-resolution details and global semantic anchors is fed into the language model.

B. Remote Sensing Native Resolution Encoder

Distinct from natural images, remote sensing imagery is characterized by significant scale variations and extreme aspect ratios (e.g., elongated river corridors or strip-like scanning swaths). Standard Vision Transformers, which enforce rigid resizing to fixed resolutions (e.g., 336×336), inevitably disrupt the geometric integrity of these geospatial targets and discard fine-grained details essential for interpretation. To address these domain-specific challenges, we construct our vision encoder based on the NaViT architecture [27]. By leveraging its resolution-agnostic modeling capabilities, our proposed encoder is tailored to process satellite data in its native form, thereby ensuring robustness across the diverse spatial variances inherent in Earth observation tasks.

As illustrated in Fig. 3, to preserve the original semantic information of remote sensing inputs, we bypass the conventional resizing pipeline. For an input image of arbitrary size $H \times W$ and patch size P, our approach applies minimal padding solely to align dimensions with the patch grid (satisfying $H' \mod P = 0$ and $W' \mod P = 0$). This allows the image to be losslessly partitioned into a sequence of $L = (H'/P) \times (W'/P)$ patches. By processing images at their native resolution, the encoder effectively retains the authentic geometric structure, which is critical for recognizing small or densely packed objects in large-scale remote sensing scenes.

Furthermore, the diversity of sensors and altitudes in remote sensing requires the model to generalize to resolutions outside the training distribution. Traditional one-dimensional absolute positional embeddings often fail to extrapolate to such unseen scales. To mitigate this, we employ a Continuous Factorized Positional Embedding strategy. Instead of binding positional

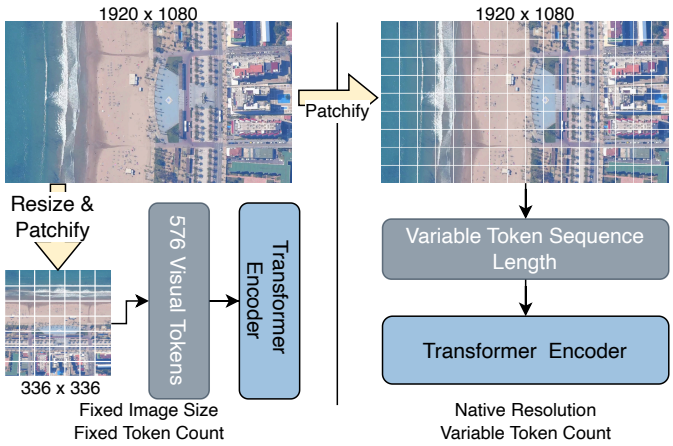

Fig. 3. Illustration of the resolution handling mechanisms in ViT v.s. ours native resolution vision encoder. Conventional ViT mandates resizing inputs to a fixed resolution (e.g., 336×336), which inevitably leads to geometric distortion and loss of fine-grained details. We overcomes this by processing images at their native resolution (1920×1080), producing a variable number of tokens. This approach preserves the original structural integrity of high-resolution inputs.

semantics to absolute pixel indices, we decouple them by normalizing spatial positions into continuous relative coordinates ($r_x, r_y \in [0, 1]$). Specifically, for a patch grid with a height of h patches and a width of w patches, we normalize the coordinates of the patch at index (i, j) to continuous relative coordinates: $r_x = \frac{i}{w}, r_y = \frac{j}{h}$. Two independent learnable functions, $\phi_x(\cdot)$ and $\phi_y(\cdot)$, encode these relative coordinates, and the final embedding is derived via additive composition:

$$\mathrm{PE}(r_x, r_y) = \phi_x(r_x) + \phi_y(r_y). \tag{1}$$

By decomposing positional encoding along height and width dimensions separately, this parameterization preserves explicit 2D spatial structure and generalizes across varying resolutions and aspect ratios, making it well-suited for remote sensing images with diverse spatial layouts. Consequently, our encoder significantly enhances extrapolation capabilities, allowing the MLLM to adapt seamlessly to the arbitrary shapes and scales typical of multi-source remote sensing data.

C. ShortCut-Based Vision Language Information Fusion

Conventional multimodal large language model (MLLM) architectures typically adopt a simplified paradigm: visual representations are extracted solely from the final layer of the vision encoder, projected through an adaptor, and subsequently injected into the large language model (LLM) [22] [28]. While this design is intuitive and computationally efficient, it relies on the implicit assumption that high-level output features sufficiently encapsulate all visual information necessary for downstream tasks. However, recent studies [29] [11] have highlighted the limitations of this paradigm, revealing that high-level features from standard encoders often lack the fine-grained locality information required for precise visual grounding. While some approaches optimize the projection layer to mitigate this [29], explicitly incorporating multi-scale features from shallow layers remains a more direct solution for the extreme scale variations in remote sensing. Unlike object-centric natural images, remote sensing data necessitates the simultaneous recognition of multi-scale targets, ranging from vast terrains to minute vehicles, where the loss of high-frequency spatial details in deep layers becomes a critical bottleneck.

Specifically, features extracted from varying depths of a vision encoder exhibit pronounced hierarchical heterogeneity in representational capacity. Shallow-layer features primarily encode local spatial structures, texture details, and precise geometric information. Intermediate-layer features progressively incorporate discriminative structural patterns. In contrast, visual encoders pretrained with classification-oriented objectives tend to produce final-layer features that emphasize abstract global semantics, such as scene- or category-level concepts. For RS tasks requiring fine-grained reasoning (e.g., distinguishing similar land-cover textures or delineating precise boundaries of road networks), relying exclusively on these high-level features often compromises the necessary spatial cues. Relying exclusively on these high-level features often compromises the spatial and textural cues critical for tasks requiring precise localization and fine-grained recognition, thereby limiting the model's fidelity in representing multi-scale visual content.

Notably, recent empirical analyses [30] [31] provide systematic evidence supporting this observation: *The best visual embeddings are not at the output of the network*. Empirical results further demonstrate that features residing in the intermediate layers of the encoder can even outperform representations learned by encoders explicitly trained for specific downstream tasks. This finding underscores the importance of leveraging multi-level intermediate features to enhance the overall performance of multi-task and multimodal models.

Driven by these insights, we propose a **shortcut-based vision language information fusion** mechanism that explicitly incorporates multi-level visual features to augment the MLLM's holistic perception of multi-scale information. Formally, we extract features from multiple stages of the visual encoder, comprising: (1) low-level spatial features rich in texture and precise positional cues; (2) mid-level features capturing local structural patterns and discriminative representations; and (3) high-level features encoding global contextual semantics. In practice, we select Layers 5, 11, and 17 from the 24-layer vision encoder, corresponding to shallow, middle, and deep stages, respectively, to provide complementary visual cues ranging from fine-grained spatial details to high-level semantic representations. This uniformly spaced selection reduces redundancy among adjacent transformer layers while preserving multi-level feature diversity for shortcut-based fusion.

These multi-scale features are first aligned in dimensionality and representation space via a shared adaptor. The procedural details of this fusion strategy are formalized in Algorithm 1. They are then injected into several shallow layers of the LLM through shortcut connections. As illustrated in Fig. 4, the fusion operation is implemented by directly adding the aligned feature vectors to the corresponding visual token positions. This design enables the effective integration of multi-level

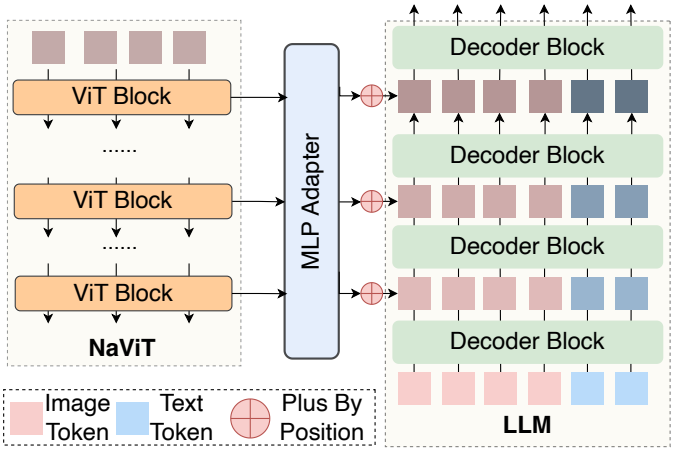

Fig. 4. Illustration of the proposed **Shortcut-Based Vision-Language Information Fusion** mechanism. Unlike traditional architectures that only utilize the final output of the vision encoder, our method extracts intermediate features from selected ViT layers within the Remote Sensing Native Resolution Encoder to capture hierarchical visual semantics. These features are projected via MLP Adapters [22] and injected into the corresponding shallow Decoder Blocks of the LLM [32]. The fusion is implemented as an element-wise addition ($\oplus$) applied specifically to the image tokens based on their positions, thereby enriching the LLM's initial representations with fine-grained visual details.

Algorithm 1 Pseudocode of ShortCut-Based Vision Language Information Fusion.

Require: $\mathbf{H}$: Hidden States In LLM Decoder Layer.
Require: v_{pos}: the position of visual tokens.
Require: $\mathbf{V}$, $\mathbf{V}^{\mathrm{S}}$: Original visual tokens, List of Hidden embeddings from the intermediate layers of the vision encoder.

function FORWARD($\mathbf{H}, \mathbf{V}, \mathbf{V}^{\mathrm{S}}, v_{pos}$)
  for (idx, DecoderLayer) in enumerate(layers) **do**
    if idx ≥ 1 and idx $\leq$ len($\mathbf{V}^{\mathrm{S}}$) **then**
      $\mathbf{H}[v_{pos}]$ += $\mathbf{V}^{\mathrm{S}}[idx]$
    end if
    $\mathbf{H}$ = DecoderLayer($\mathbf{H}$)
  end for
end function

visual information while introducing negligible computational overhead.

Through this shortcut-based fusion mechanism, the LLM gains access to both fine-grained local details and high-level global context during the early stages of reasoning. This effectively alleviates the information bottleneck induced by relying solely on high-level visual features, facilitating a more comprehensive, robust, and scale-aware multimodal understanding.

D. Multi-Granularity Token Pooling

To enable the Large Language Model (LLM) to perceive visual information with both high-resolution precision and global semantic context, we propose a Multi-Granularity Feature Fusion Strategy, as shown in Fig. 5. Instead of relying on a single scale, we process the output from the vision encoder through three parallel branches, constructing a hierarchical visual representation. Given the raw token sequence $\mathbf{F} \in \mathbb{R}^{L \times D}$ extracted by the vision encoder, we generate three distinct sets of visual tokens: **1) Original Token Branch**: We preserve the original output tokens from the vision encoder without any downsampling. This ensures that the LLM retains access to the finest visual details, such as small objects and texture information. **2) Fine-grained Pooling Branch** (16×16): We apply an adaptive pooling operation to compress the features into a 16×16 grid. This branch serves as an intermediate representation, reducing sequence length while maintaining local spatial structures. **3) Coarse-grained Pooling Branch** (4×4): We further downsample the features into a 4×4 grid. This branch acts as a global semantic summarizer, allowing the model to quickly grasp the dominant content and layout of the image.

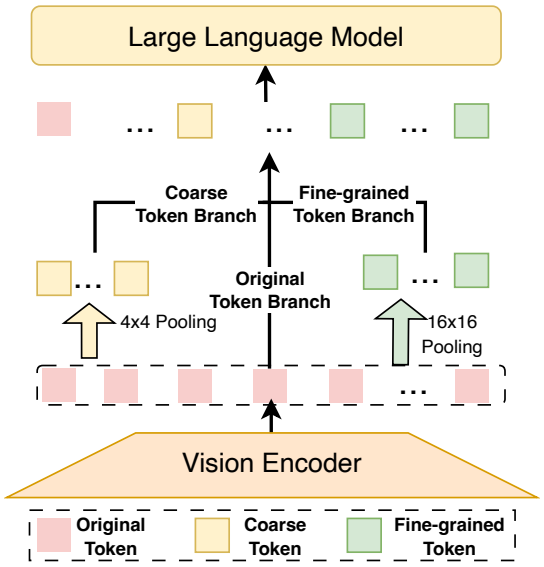

Fig. 5. Illustration of the Multi-Granularity Token Pooling module. The visual features are organized into three levels of granularity: original tokens, 16×16 pooled tokens, and 4×4 pooled tokens. These sequences are concatenated to form a comprehensive visual representation that provides the LLM with both fine-grained details and global context.

Formally, let $\mathbf{F} \in \mathbb{R}^{L \times D}$ denote the output feature sequence from the vision encoder, where L is the sequence length and D is the embedding dimension. To apply spatial pooling, we first restore the spatial structure by reshaping the sequence into a 2D feature map, denoted as $\mathbf{F}_{2D} = \mathrm{Reshape}(\mathbf{F})$, where $\mathbf{F}_{2D} \in \mathbb{R}^{H' \times W' \times D}$ and $L = H' \times W'$. Let $\mathcal{P}_{k \times k}(\cdot)$ represent the adaptive pooling operation with a target output size of $k \times k$, and $\mathrm{Flatten}(\cdot)$ denote the operation that collapses the spatial dimensions back into a sequence. We denote the output feature sequences from the original, fine-grained, and coarse-grained branches as $\mathbf{Z}_{orig}$, $\mathbf{Z}_{fine}$, and $\mathbf{Z}_{coarse}$, respectively. These features are formulated as:

$$\mathbf{Z}_{orig} = \mathbf{F}, \tag{2}$$

$$\mathbf{Z}_{fine} = \mathrm{Flatten}\left(\mathcal{P}_{16 \times 16}(\mathbf{F}_{2D})\right), \tag{3}$$

$$\mathbf{Z}_{coarse} = \mathrm{Flatten}\left(\mathcal{P}_{4 \times 4}(\mathbf{F}_{2D})\right). \tag{4}$$

Finally, we concatenate these sequences along the temporal dimension to obtain the multi-granularity visual representation

$\mathbf{Z}_{input}$:

$$\mathbf{Z}_{input} = \text{Concat}\left([\mathbf{Z}_{orig}, \mathbf{Z}_{fine}, \mathbf{Z}_{coarse}], \text{dim} = 1\right). \quad (5)$$

The resulting $\mathbf{Z}_{input}$ serves as the primary visual prompt for the MLLM. After being projected via the MLP adaptor, it is fed into the LLM together with the textual tokens. Crucially, this module functions synergistically with the Shortcut-Based Fusion Mechanism described in III-C. While the Multi-Granularity Pooling operates at the input level to construct a structured global-to-local context from the encoder's final layer, the Shortcut Mechanism operates at the feature level by injecting intermediate features directly into the LLM's decoder layers. This dual-pathway design ensures that the LLM is not only grounded by a comprehensive initial visual context ($\mathbf{Z}_{input}$) but also continuously supplemented with fine-grained spatial cues during the generation process.

IV. Experiment

A. Experiment Setup

We conduct all experiments on VRSBench [33], a comprehensive vision-language benchmark specifically designed for remote sensing image understanding. We select VRSBench as our primary evaluation benchmark for two key reasons: (1) Multi-Task Diversity: It unifies three core tasks, including Visual Question Answering (VQA), Visual Grounding and Image Captioning, to provide a holistic assessment of MLLM capabilities. (2) Fine-Grained Spatial Evaluation: Unlike general-domain benchmarks, VRSBench-VQA includes specific reasoning categories for Quantity, Size, and Position. This makes it an ideal platform to evaluate the efficacy of our Native Resolution Encoder in preserving geometric fidelity and handling the extreme scale variations inherent in satellite imagery.

To ensure reproducibility and fair comparison, we strictly adhere to the official split protocols of VRSBench. We use the designated training set for model fine-tuning and evaluate our method on the official test set without any test-time augmentation. No additional data cleaning, filtering, or resampling strategies are applied to the training data. To optimize training efficiency, we reorganize the data samples by merging multiple QA pairs associated with the same image into a single multi-turn dialogue format.

To comprehensively evaluate the effectiveness of our proposed framework, we compare it against a series of representative Multimodal Large Language Models reported in the VRSBench benchmark [33]. These baselines encompass mainstream architectures including LLaVA [22], MiniGPT-v2 [23], GeoChat [7], GPT-4V [34], and Mini-Gemini [28], serving as strong references for general-domain adaptation capabilities. To ensure a strictly fair comparison, we cite the results directly from the official VRSBench publication. As detailed in their benchmarking protocol, all baseline models(with the exception of GPT-4V, which was evaluated in zero-shot mode due to its closed-source nature) were fine-tuned on the VRSBench training set for 5 epochs to ensure sufficient domain alignment.

B. Training Recipe

Experiments are conducted using four NVIDIA A800 GPUs under CUDA 13.0 version. Each GPU processes a batch size of 8, resulting in a global batch size of 64 through gradient accumulation. Our model is initialized from Qwen3-VL-4B [35] and trained for 5 epochs on the VRSBench training split. We use AdamW as the optimizer with the following hyperparameters: 1) learning rate: 1×10^{-5}; 2) weight decay: 0.01; 3) betas: $(0.9, 0.95)$. Unless otherwise stated, we adopt mixed-precision training and standard PyTorch optimization settings.

We extract intermediate visual features from layers 5, 11, and 17 of the vision encoder. These features are aligned using our adaptor module and injected into the lower layers of the LLM through the short-based vision language information fusion mechanism. This fusion facilitates the propagation of multi-level spatial and semantic cues from the vision encoder into the language backbone.

C. Main Result

To comprehensively evaluate multimodal understanding in the context of remote sensing imagery, we follow the standard benchmark protocol defined by VRSBench [33] and adopt its three core downstream tasks: **visual question answering**, **referring** and **captioning**. Each task targets a distinct dimension of visual–linguistic reasoning within complex Earth observation scenarios.

1) VRSBench-VQA (Visual Question Answering)

This task evaluates the model's ability to answer open-ended or multiple-choice questions grounded in remote sensing imagery. The benchmark spans ten diverse categories, ranging from basic perception (e.g., Object Category, Presence, Color) to fine-grained spatial understanding (Quantity, Size, Position) and high-level cognition (Scene, Reasoning). Success in this task requires not only identifying visual elements but also performing compositional reasoning across multimodal inputs.

In Table I, we report the quantitative results on the VRSBench-VQA task. Overall, our method achieves the highest performance across the benchmark, securing the best accuracy in 8 out of the 10 reasoning dimensions. Specifically, our model attains an overall accuracy of 76.8%, demonstrating a solid advantage over the strongest baseline, Mini-Gemini (72.8%). This consistent improvement highlights the robustness of our architecture in handling diverse visual–linguistic reasoning scenarios.

A key advantage of our model lies in categories demanding precise object-level localization and scale estimation. Our method achieves 74.1%, 63.2%, and 73.1% in Quantity, Size, and Position, respectively, consistently surpassing the baselines. Notably, the gains in the Size ($+5.2\%$) and Position ($+5.1\%$) categories are particularly significant. We primarily attribute these gains to the *Native Resolution Vision Encoder*. By processing images at their original scales and aspect ratios, our encoder prevents the geometric distortion caused by the rigid resizing operations in traditional ViT-based models, thereby preserving the precise spatial cues required for these tasks.

TABLE I
QUANTITATIVE RESULTS ON THE VRSBENCH VISUAL QUESTION-ANSWERING TASK. WE REPORT CATEGORY-WISE ACCURACY ACROSS ELEVEN REASONING DIMENSIONS, INCLUDING CATEGORY, PRESENCE, QUANTITY, COLOR, SHAPE, SIZE, POSITION, DIRECTION, SCENE, AND REASONING, AS WELL AS THE OVERALL ACCURACY ("ALL"). THE PERFORMANCE OF ALL COMPARISON METHODS IS DIRECTLY REFERENCED FROM [33].

Method	Category	Presence	Quantity	Color	Shape	Size	Position	Direction	Scene	Reasoning	All
#VQAs	5435	7789	6374	3550	1422	1011	5829	477	4620	902	
GeoChat w/o sft	48.5	85.9	19.2	17.0	18.3	32.0	43.4	42.1	44.2	57.4	40.8
GPT-4V	67.0	87.6	45.6	71.0	70.8	54.3	67.2	50.7	69.8	72.4	65.6
MiniGPT-v2	61.3	26.0	46.1	51.0	41.8	11.2	17.1	12.4	49.3	21.9	33.8
LlaVA-1.5	86.9	91.8	58.2	69.9	72.2	61.5	69.5	56.7	**83.9**	73.4	72.4
GeoChat	86.5	92.1	56.3	70.1	73.8	60.4	69.3	53.5	83.7	73.5	72.0
Mini-Gemini	**87.8**	92.1	58.8	74.0	75.3	58.0	68.0	56.7	83.2	74.4	72.8
ROARS(Ours)	84.6	**95.1**	**74.1**	**77.0**	**81.0**	**63.2**	**73.1**	**60.8**	79.5	**79.9**	**76.8**

Our approach also demonstrates clear superiority in attribute reasoning, achieving 77.0% in Color and 81.0% in Shape. These improvements validate the effectiveness of our *shortcut-based vision language information fusion* mechanism. Unlike traditional architectures that rely solely on high-level semantics, our shortcut connections explicitly inject shallow-layer features that are rich in color and edge information into the LLM. When combined with the high-fidelity inputs from the native resolution encoder, this design ensures that fine-grained spectral and textural patterns are not lost during the encoding process.

In the Reasoning category, which requires multi-step inference, our model achieves 79.9%, outperforming Mini-Gemini (74.4%) by a margin of 5.5%. This performance underscores the synergy of our *Multi-Granularity Token Pooling* strategy. By simultaneously providing the LLM with original-resolution tokens for detail and pooled tokens for global context, the model can effectively integrate semantic cues across different scales. This holistic design enables complex decision-making by leveraging comprehensive visual evidence.

Regarding general perception, we achieve the highest accuracy in object Presence detection (95.1%), reflecting robust existence verification capabilities. While the baseline performs slightly better in the Scene category (83.2% vs. 79.5%), likely due to the strong global semantic priors of the fixed-resolution CLIP backbone, our method maintains a more balanced performance profile across fine-grained and global tasks.

Overall, the consistent improvements across diverse categories demonstrate the collective efficacy of our core innovations. The native resolution encoder ensures geometric fidelity, the shortcut fusion retains fine-grained attributes, and the multi-granularity pooling facilitates complex reasoning. Together, they validate the architecture's suitability for structurally complex remote sensing imagery.

2) VRSBench-Ref (Referring Expression Comprehension)

This task requires models to localize a specific target object described by a natural language query. It rigorously evaluates the ability to map linguistic attributes to precise spatial coordinates. In remote sensing, this is uniquely challenging due to dense object clusters, extreme scale variations, and the need to disambiguate visual targets based on subtle spatial or attribute cues.

In Table II, we present the quantitative results on the VRSBench Referring Expression task. Our method achieves state-of-the-art performance across all metrics, significantly outperforming existing multimodal baselines. Specifically, we achieve an overall accuracy of 68.9% at Acc@0.5 and 47.8% at Acc@0.7. Compared to the strongest baseline, GeoChat [7], our model delivers substantial improvements of $+19.1\%$ and $+27.9\%$, respectively.

A defining characteristic of our results is the robustness at stricter evaluation thresholds. Standard MLLMs often struggle to generate precise bounding boxes for small, densely packed remote sensing objects, leading to sharp performance drops when the IoU threshold increases from 0.5 to 0.7 (e.g., LLaVA-1.5 and Mini-Gemini drop significantly). In contrast, our method maintains exceptional accuracy at the Acc@0.7 threshold (47.8%), more than doubling the performance of GeoChat (19.9%). We attribute this high-precision capability to the Native Resolution Vision Encoder. By processing images at their original scales without aggressive resizing, our model avoids the blurring of object boundaries and geometric distortion, thereby generating tight, accurate bounding boxes that align closely with ground-truth regions.

The *Non-Unique* subset poses a critical challenge, requiring the model to distinguish a target object from visually similar distractors based on complex linguistic attributes (e.g., "the white car next to the tree"). While existing methods degrade severely in these scenarios (e.g., Mini-Gemini drops from 41.1% in Unique to 22.3% in Non-Unique at Acc@0.5), our model exhibits superior resilience, achieving 64.9% (Acc@0.5) and 44.4% (Acc@0.7) in the Non-Unique setting. This success verifies the efficacy of our shortcut-based vision language information fusion mechanism. By injecting fine-grained textural and spatial features from shallow layers into the LLM, our model can effectively resolve subtle visual differences and perform unambiguous grounding even in cluttered environments. To provide a more intuitive understanding of the grounding performance, we further include several visualization examples in Fig. 6. These examples show that our method can accurately localize the referred objects under different remote sensing scenarios, including small objects, dense object distributions, and visually similar distractors. The predicted bounding boxes are well aligned with the target

TABLE II
QUANTITATIVE RESULTS ON THE VRSBENCH VISUAL GROUNDING TASK. WE REPORT THE ACCURACY AT IOU THRESHOLDS OF 0.5 AND 0.7 ACROSS UNIQUE, NON UNIQUE, AND ALL SETTINGS. THE RESULTS DEMONSTRATE THAT OUR METHOD SIGNIFICANTLY OUTPERFORMS EXISTING BASELINES ON ALL METRICS, EXHIBITING A PARTICULARLY SUBSTANTIAL PERFORMANCE ADVANTAGE AT THE HIGH-PRECISION THRESHOLD (ACC@0.7). THE PERFORMANCE OF ALL COMPARISON METHODS IS DIRECTLY REFERENCED FROM [33].

Method	Unique		Non Unique		ALL	
	Acc@0.5	Acc@0.7	Acc@0.5	Acc@0.7	Acc@0.5	Acc@0.7
GeoChat w/o sft	20.7	5.4	7.3	1.7	12.9	3.2
GPT-4V	8.6	2.2	2.5	0.4	5.1	1.1
MiniGPT-v2	40.7	18.9	32.4	15.2	35.8	16.8
LlaVA-1.5	51.1	16.4	34.8	11.5	41.6	13.6
GeoChat	57.4	22.6	44.5	18.0	49.8	19.9
Mini-Gemini	41.1	9.6	22.3	4.9	30.1	6.8
ROARS(Ours)	**74.5**	**52.6**	**64.9**	**44.4**	**68.9**	**47.8**

Fig. 6. Bounding box visualization results. Green solid boxes denote ground truth (GT), and red dashed boxes denote predicted boxes (Pred).

regions, demonstrating the effectiveness of our method in producing precise and unambiguous grounding results.

3) VRSBench-Cap (Captioning)

This task requires generating a detailed and contextually coherent natural-language description for a given remote sensing image. Compared with conventional natural-image captioning, VRSBench-Cap emphasizes fine-grained object attributes, spatial relationships, man-made structural patterns, and environmental context, presenting a significantly more challenging scenario for multimodal models. To comprehensively assess the generation quality, we employ a diverse set of evaluation metrics. specifically:

- BLEU (1-4) [36]: Measures n-gram precision, focusing on the exact lexical overlap between the generated caption and ground-truth references. High scores typically indicate strict adherence to the reference phrasing.
- Meteor [37]: Goes beyond exact matching by considering synonyms, stemming, and paraphrasing. It correlates better with human judgment regarding semantic correctness.
- ROUGE_L [38]: Evaluates the longest common subsequence, capturing sentence-level structural similarity.
- CIDEr [39]: Designed specifically for image captioning, this metric computes the consensus between the candidate and reference sentences using TF-IDF weighting for n-grams. Unlike BLEU, it prioritizes the accuracy of rare, informative concepts (e.g., specific objects or attributes) over common function words, making it highly sensitive to the semantic content of the image.
- CLAIR [40]: A recently proposed metric that uses LLMs to evaluate the semantic matching and factual accuracy of the caption, providing a more robust assessment of "understanding" than traditional n-gram metrics.

In Table III, we report the quantitative results on the VRS-Bench Captioning task. Our method demonstrates a distinct advantage in semantic understanding and flexible expression, as evidenced by the following observations: **1)** Superior Se-

TABLE III
QUANTITATIVE RESULTS ON THE VRSBENCH IMAGE CAPTIONING TASK. WE EVALUATE PERFORMANCE USING STANDARD METRICS (BLEU, METEOR, ROUGE_L, CIDER) AND THE SEMANTIC METRIC CLAIR. OUR METHOD DEMONSTRATES SUPERIOR SEMANTIC ALIGNMENT, ACHIEVING A STATE-OF-THE-ART METEOR SCORE OF 30.5, SIGNIFICANTLY OUTPERFORMING EXISTING BASELINES. THE PERFORMANCE OF ALL COMPARISON METHODS IS DIRECTLY REFERENCED FROM [33].

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	Rouge_L	CIDEr	CLAIR	Avg_L
GeoChat w/o sft	13.9	6.6	3.0	1.4	7.8	13.2	0.4	0.42	36
GPT-4V	37.2	22.5	13.7	8.6	20.9	30.1	19.1	**0.83**	**67**
MiniGPT-v2	36.8	22.4	13.9	8.7	17.1	30.8	21.4	0.73	37
LlaVA-1.5	**48.1**	31.5	21.2	**14.7**	21.9	**36.9**	**33.9**	0.78	49
GeoChat	46.7	30.2	20.1	13.8	21.1	35.2	28.2	0.77	52
Mini-Gemini	47.6	31.1	20.9	14.3	21.5	36.8	33.5	0.77	47
ROARS(Ours)	46.4	**32.0**	**21.3**	9.7	**30.5**	35.8	33.0	0.79	56

TABLE IV
ABLATION STUDY ON THE IMPACT OF INPUT RESOLUTION AND SHORTCUT-BASED VISION-LANGUAGE INFORMATION FUSION. WE COMPARE OUR NATIVE RESOLUTION APPROACH AGAINST THE MINI-GEMINI BASELINE AND FIXED-RESOLUTION VARIANTS (224×224 AND 336×336). ROARS UTILIZES THE NATIVE RESOLUTION MECHANISM TO PRESERVE ORIGINAL ASPECT RATIOS. THE BEST RESULTS(ACC.) ARE HIGHLIGHTED IN BOLD.

Resolution	Category	Presence	Quantity	Color	Shape	Size	Position	Direction	Scene	Reasoning	All
#VQAs	5435	7789	6374	3550	1422	1011	5829	477	4620	902	
Mini-Gemini(Baseline)	**87.8**	92.1	58.8	74.0	75.3	58.0	68.0	56.7	**83.2**	74.4	72.8
224×224	84.0	92.7	72.0	75.5	74.9	60.4	69.8	55.1	78.9	73.5	73.7
336×336	84.3	94.0	72.9	75.7	76.7	60.7	71.6	57.7	79.0	73.5	74.4
ROARS(w/o ShortCut)	84.9	93.2	72.8	76.1	79.3	61.7	70.5	60.3	79.0	77.4	75.5
ROARS(Ours)	84.6	**95.1**	**74.1**	**77.0**	**81.0**	**63.2**	**73.1**	**60.8**	79.5	**79.9**	**76.8**

mantic Alignment (Meteor): The most striking result is our performance on the Meteor metric, where our method achieves a score of 30.5. This represents a remarkable improvement of $+8.6$ percentage over the strongest baseline, LLaVA-1.5 (21.9), and significantly outperforms GPT-4V (20.9). Unlike BLEU, which penalizes non-exact word matches, Meteor rewards the use of synonyms and semantic variants. This high score indicates that our model captures the core semantic content of remote sensing scenes (e.g., identifying land use types and object interactions) more accurately than competitors, even when using different vocabulary. **2)** Diversity vs. Rigid Matching (BLEU vs. Meteor): We observe an interesting trade-off between BLEU-4 and Meteor. While our BLEU-4 score (9.7) is lower than LLaVA-1.5, our BLEU-1 and BLEU-2 scores remain highly competitive. This discrepancy suggests that our model does not simply "memorize" or reproduce the rigid sentence templates found in the training data. Instead, it generates more diverse and natural descriptions. The combination of a lower BLEU-4 and high Meteor score is a common characteristic of models that prioritize semantic richness over verbatim copying. **3)** High-Level Comprehension (CLAIR): Furthermore, in terms of the CLAIR metric, which reflects AI-based semantic evaluation, our method achieves 0.79. This surpasses widely-used MLLMs like LLaVA-1.5 [22] and GeoChat [7], and is comparable to the proprietary GPT-4V (0.83). This confirms that, from a high-level semantic perspective, the captions generated by our model are factually accurate and contextually relevant, effectively handling the complexities of remote sensing imagery.

D. Ablation Study

Effectiveness of Native Resolution Input. To validate the effectiveness of our native resolution encoding strategy, we conducted ablation experiments on the VQA task. It is important to note that simply replacing remote sensing native resolution encoder with a fixed-resolution encoder (e.g., standard CLIP-ViT) would introduce confounding factors related to differences in pre-training data scales and parameter distributions. To ensure a fair comparison where the input resolution is the sole variable, we retained the remote sensing native resolution encoder architecture but imposed fixed-resolution constraints during the data preprocessing stage. Specifically, we resized all input images to the standard input dimensions of CLIP-ViT-L (i.e., 224×224 and 336×336) before feeding them into our model. The results, presented in Table IV, demonstrate that the native resolution strategy consistently outperforms the fixed-resolution counterparts. Notably, we observe significant performance gains in geometry-sensitive categories such as Shape, Size, Position and Direction. These findings empirically validate that avoiding the geometric distortion and information loss induced by mandatory resizing is critical for accurate remote sensing multimodal understanding.

Effectiveness of Shortcut-Based Vision Language Information Fusion. To validate the contribution of the shortcut-based vision language information fusion mechanism, we constructed and retrained a variant model excluding this module (w/o Shortcut). This variant relies solely on the final layer output of the vision encoder, while all other training configurations remain identical to the full model to ensure a fair comparison. Comparing the results in Table IV, incorporating the shortcut fusion mechanism significantly boosts the overall

TABLE V
ABLATION STUDY ON THE EFFECT OF MULTI-LEVEL TOKEN POOLING FOR VQA. WE ANALYZE THE CONTRIBUTION OF ADAPTIVE POOLING TOKENS WITH DIFFERENT GRID. BASE DENOTES THE ORIGINAL TOKEN SEQUENCE WITHOUT ADDITIONAL POOLED TOKENS.

Method	Token Pooling			Acc
	Base	16×16	4×4	
Base	✓			81.1
16×16 Pool	✓	✓		84.0
4×4 Pool	✓		✓	82.0
ROARS(Ours)	✓	✓	✓	**84.3**

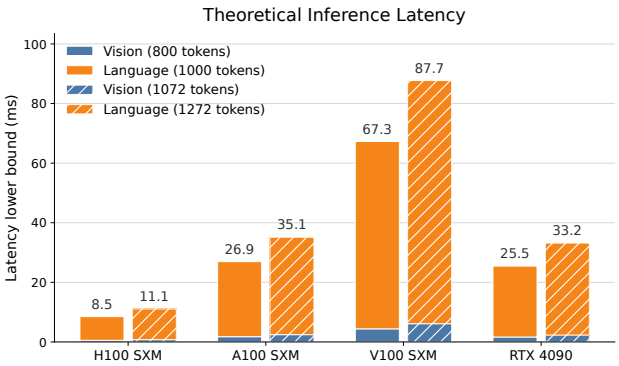

Fig. 7. Theoretical Prefill Latency Comparison with some general GPU.

accuracy from 75.5% to 76.8%.

The performance gains are particularly pronounced in tasks highly sensitive to geometric structure. A detailed analysis reveals that accuracy in Position and Size increased by +2.6% and +1.5%, respectively. This significant margin strongly supports our core hypothesis: although deep features of the vision encoder are semantically rich, they progressively lose fine-grained spatial specificity during the layer-wise abstraction process. By directly injecting shallow and intermediate features into the LLM through a shortcut mechanism, the model effectively bypasses the information bottleneck, enabling direct access to fine-grained visual cues that are critical for accurate localization and quantitative analysis.

Furthermore, the proposed fusion strategy benefits high-level cognition, with the Reasoning task improving by 2.5% (from 77.4% to 79.9%). This improvement suggests that multi-level visual integration provides the LLM with complementary visual evidence, rather than a simple aggregation of features. By jointly leveraging high-level semantic abstractions and fine-grained visual cues, the model achieves more robust reasoning in complex remote sensing scenarios. Overall, the shortcut mechanism effectively mitigates fine-grained information loss in deep features and enhances holistic perception.

Effectiveness of Multi-Granularity Token Pooling. To assess the contribution of visual representations at different granularities, we conduct an ablation study on the VQA task, with results summarized in Table V. The baseline model, which relies solely on the original token sequence (Base), achieves an accuracy of 81.1%.

Incorporating the fine-grained pooling branch (16×16) leads to a substantial performance gain of 2.9%, boosting accuracy to 84.0%. This improvement indicates that aggregating local spatial information enables the LLM to better attend to fine visual details critical for answering visually grounded questions. Similarly, introducing the coarse-grained pooling branch (4×4) yields a 0.9% increase in accuracy (82.0%), suggesting that global semantic summarization contributes positively to holistic scene understanding.

Most importantly, when both pooling strategies are jointly employed (ROARS), the model achieves the best performance of 84.3%, surpassing the baseline by a notable margin of 3.2%. These results demonstrate that the fine-grained details preserved by the 16×16 pooling and the global contextual information captured by the 4×4 pooling are complementary rather than redundant, together enabling more robust reasoning across diverse visual question types.

Besides the performance gain, Multi-Granularity Token Pooling incurs only a modest computational overhead. As shown in Fig. 7, the latency increase remains controllable across both cloud GPUs and consumer GPUs, suggesting that Multi-Granularity Token Pooling is practical for deployment.

Discussion on Ablation Robustness. We note that minor performance fluctuations may occur in individual categories when components are removed (e.g., Category accuracy slightly increases from 84.6% to 84.9% without shortcut fusion). This is attributable to feature redundancy in global tasks, where shallow-layer features may introduce minor noise for scene-level classification. Nonetheless, the full model achieves the best performance in 8 out of 10 VQA categories with a +4.0% overall gain, and consistent improvements are observed across all three tasks (VQA, Referring, and Captioning), demonstrating robust generalization.

V. DISCUSSION

A. Scope of Application

Tailored for high-resolution remote sensing, our framework leverages Native Resolution and Shortcut-Based Fusion to capture both fine-grained textures and global contexts. This ensures robustness in scale-variant scenarios, such as urban analysis, disaster assessment, and infrastructure monitoring. Unlike fixed-resolution RS-MLLMs, our method eliminates geometric distortion and information loss. Furthermore, the architecture is inherently extensible; with appropriate encoders, it can adapt to multi-source data (e.g., SAR, Infrared), serving as a versatile paradigm for all-weather remote sensing.

B. Limitations

The adoption of native resolution inevitably increases the length of visual token sequences, leading to higher computational loads compared to downsampled baselines. However, given our framework's positioning for cloud-based applications, this overhead is considered acceptable. High-performance cloud computing infrastructures possess sufficient throughput and memory capacity to sustain these demands.

VI. Conclusion

We propose the **ROARS** framework, which integrates native resolution encoding and shortcut-based fusion to address scale sensitivity and detail loss. Experiments demonstrate that our method effectively eliminates the geometric distortion inherent in fixed-resolution paradigms, significantly enhancing fine-grained perception and complex reasoning. This work provides a robust solution for high-resolution remote sensing and lays the groundwork for future research in efficiency optimization and explicit spatial modeling.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 62576213), Guangdong Province Key Projects in Artificial Intelligence (Intelligent Robots) (No. 2025ZDZX3047), the Innovation Team Project of Guangdong Province of China (2024KCXTD017), and Guangdong Hong Kong-Macau University Joint Laboratory (2023LSYS005).

References

[1] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nat.*, vol. 566, no. 7743, pp. 195–204, 2019. [Online]. Available: https://doi.org/10.1038/s41586-019-0912-1

[2] D. Wang, J. Zhang, B. Du, G. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote. Sens.*, vol. 61, pp. 1–20, 2023. [Online]. Available: https://doi.org/10.1109/TGRS.2022.3176603

[3] G. Li *et al.*, "Open-world recognition in remote sensing: Concepts, challenges, and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 2, pp. 8–31, 2024.

[4] S. Lobry, D. Marcos, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[5] Y. Lu, B. Wu, Z. Li, K. Li, C. Huang, H. Wang, Q. Lan, R. Chen, L. Chen, and B. Liang, "Remote sensing-oriented world model," 2025. [Online]. Available: https://arxiv.org/abs/2509.17808

[6] Z. Shangguan, J. Zhang, K. Xing, Z. Gong, Y. Zhang, Y. Xu, and F. Liang, "Training-free enhancement of satellite remote sensing vlms via geo-contrastive decoding," *Journal of Cloud Computing*, vol. 15, no. 1, p. 63, 2026. [Online]. Available: https://doi.org/10.1186/s13677-026-00885-7

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.

[8] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–20, 2024.

[9] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 224, pp. 272–286, 2025.

[10] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[12] Y. Zhan, Z. Xiong, and Y. Yuan, "Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 221, pp. 64–77, 2025.

[13] C. Pang, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, X. Weng, S. Wang, L. Feng, G. Xia, and C. He, "H2RSVLM: towards helpful and honest remote sensing large vision language model," *CoRR*, vol. abs/2403.20213, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.20213

[14] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[15] Z. Shangguan, Y. Dong, S. Guo, V. C. M. Leung, M. J. Deen, and X. Hu, "Facial expression analysis and its potentials in iot systems: A contemporary survey," *ACM Computing Surveys*, vol. 58, no. 2, pp. 1–39, 2025.

[16] M. Li, L. Shan, X. Li, Y. Bai, D. Zhou, W. Wang, K. Lv, B. Luo, and S.-B. Chen, "Global-local attention network for semantic segmentation in aerial images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5704–5711.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[18] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin *et al.*, "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2252–2274, 2023.

[19] Y. Sun, Y. Geng, P. Wei, Y. Chen, J. Yang, R. Chen, W. Zhang, and X. Shen, "Llaso: A foundational framework for reproducible research in large language and speech model," *arXiv preprint arXiv:2508.15418*, 2025.

[20] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[21] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[22] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[23] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[24] C. Pang, X. Weng, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, S. Wang, L. Feng, G.-S. Xia *et al.*, "Vhm: Versatile and honest vision language model for remote sensing image analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6381–6388.

[25] J. Luo, Z. Pang, Y. Zhang, T. Wang, L. Wang, B. Dang, J. Lao, J. Wang, J. Chen, Y. Tan *et al.*, "Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding," *arXiv preprint arXiv:2406.10100*, 2024.

[26] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, "Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.

[27] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin *et al.*, "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2252–2274, 2023.

[28] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, "Mini-gemini: Mining the potential of multi-modality vision language models," *arXiv preprint arXiv:2403.18814*, 2024.

[29] J. Cha, W. Kang, J. Mun, and B. Roh, "Honeybee: Locality-enhanced projector for multimodal llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 817–13 827.

[30] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed *et al.*, "Perception encoder: The best visual embeddings are not at the output of the network," *arXiv preprint arXiv:2504.13181*, 2025.

[31] J. Zhang, F. Liang, Y. Zhang, W. Wang, R. Zeng, and X. Hu, "Sparse shortcuts: Facilitating efficient fusion in multimodal large language models," 2026. [Online]. Available: https://arxiv.org/abs/2602.00505

[32] L. Meng, J. Yang, R. Tian, X. Dai, Z. Wu, J. Gao, and Y.-G. Jiang, "Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 23 464–23 487, 2024.

[33] X. Li, J. Ding, and M. Elhoseiny, "Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding," *Advances*

in Neural Information Processing Systems, vol. 37, pp. 3229–3242, 2024.
[34] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.
[35] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.
[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[37] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
[38] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[39] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[40] D. Chan, S. Petryk, J. Gonzalez, T. Darrell, and J. Canny, "Clair: Evaluating image captions with large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 13 638–13 646.

Jingrui Zhang received the B.S. degree in Information and Computational Science from Shandong University of Science and Technology, Shandong, China, in 2023. He is currently working toward the M.S. degree in Computer Technology with the School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China. His research interests include federated learning, multi-agent collaboration and MLLM(Multimodal Large Language Model).

Feng Liang (Member, IEEE), is an associate professor in the Artificial Intelligence Research Institute of Shenzhen MSU-BIT University. He received his Ph.D. degree in computer science from the University of Hong Kong. His research interests are in broad areas of distributed intelligence, including distributed systems, distributed machine learning, and large-scale intelligence. He has published papers in prestigious venues including IEEE TPDS, ICDE, ICCV, AAAI, HPDC, IPDPS, Information Sciences, EDBT, and ACSAC.

Yong Zhang is currently pursuing a master's degree in computer technology at Beijing Institute of Technology. He holds a bachelor's degree in computer science and technology from Beijing Wuzi University. His research interests include federated learning and large-scale multi-task learning. He has published papers at ICCV, Journal of Information and Intelligence, and International Conference on Smart Internet of Things.

Zixuan ShangGuan received the M.S. degree in Computer Science and Technology from Lanzhou University, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Medical Technology, Beijing Institute of Technology, Beijing, China. He is also a visiting student with Shenzhen MSU-BIT University, China. His research interests include affective computing, large language models, and natural language processing.

Zhida Li is currently a Ph.D. candidate at the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. He received his B.S. degree from Sun Yat-sen University in 2023. His research interests focus on computer vision and multimodal learning.

Xiaoyi Fan (Member, IEEE) is currently a Full Professor at Shenzhen MSU-BIT University, Shenzhen, China. He received the Ph.D. degree from Simon Fraser University, Burnaby, Canada, in 2018, and the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013. He was a postdoctoral research fellow (honorary) with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. He is a recipient of the NSERC Postdoctoral Fellowship in 2019. He is a recipient of the IEEE ICDCS Distinguished Paper Award (2024) and the IWQoS Best Poster Award (2024). His research interests include the Internet of Things, edge computing, and artificial intelligence.

Victor C.M. Leung (Life Fellow, IEEE) is currently with Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China. He is also an Emeritus Professor of electrical and computer engineering and the Director of the Laboratory for Wireless Networks and Mobile Systems, The University of British Columbia (UBC), Canada. His research interests include wireless networks and mobile systems. He has published widely in these areas. He is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He received the 1977 APEBC Gold Medal, the 1977–1981 NSERC Postgraduate Scholarships, the IEEE Vancouver Section Centennial Award, the 2011 UBC Killam Research Prize, the 2017 Canadian Award for Telecommunications Research, the 2018 IEEE TCGCC Distinguished Technical Achievement Recognition Award, and the 2018 ACM MSWiM Reginald Fessenden Award. He has coauthored papers that won the 2017 IEEE ComSoc Fred W. Ellersick Prize, the 2017 IEEE Systems Journal Best Paper Award, the 2018 IEEE CSIM Best Journal Paper Award, and the 2019 IEEE TCGCC Best Journal Paper Award. He has been serving on the editorial boards of the IEEE Transactions on Green Communications and Networking, IEEE Transactions on Cloud Computing, IEEE Access, IEEE Network, and several other journals. He is named in the current Clarivate Analytics list of "Highly Cited Researchers." He is a fellow of the Royal Society of Canada (Academy of Science), Canadian Academy of Engineering, and Engineering Institute of Canada, and a life fellow of IEEE.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a full Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored and coauthored more than 200 papers in top-tier academic journals and conferences. His current research interests include Cross-modal visual content understanding and generation. He was a recipient of the ICCV 2019 Best Paper Nomination Award and the CVPR 2024 Best Paper Candidate. He serves as an Area Chair for the conference of CVPR2024/CVPR2025/ICCV2025. He has been serving as a reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.

Xiping Hu (Senior Member, IEEE) received the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada. He is currently a professor with Beijing Institute of Technology, and with Shenzhen MSU-BIT University, China. He has more than 150 papers published and presented in prestigious conferences and journals, such as IEEE TPAMI/TMC/TPDS/TIP/JSAC, IEEE COMST, ACM MobiCom/MM/SIGIR/WWW, AAAI, and IJCAI. He has been serving as associate editor of IEEE TCSS, and the lead guest editors of IEEE IoT Journal and IEEE TASE etc. He has been granted several key national research projects as principal investigator. He was the Co-Founder and CTO of Bravolol Ltd., Hong Kong, a leading language learning mobile application company with over 100 million users, and listed as the top 2 language education platform globally. His research areas consist of mobile cyber-physical systems, crowd sensing and affective computing.