Semantic-Geometric Dual Compression: Training-Free Visual Token Reduction for Ultra-High-Resolution Remote Sensing Understanding

Yueying Li[1*], Fengxiang Wang[1*], Yan Li[1], Mingshuo Chen[1], Mengying Zhao[1], and Long Lan[1✉]

College of Computer Science and Technology, National University of Defense Technology,
Changsha 410073, China
long.lan@nudt.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) have demonstrated immense potential in Earth observation. However, the massive visual tokens generated when processing Ultra-High-Resolution (UHR) imagery introduce prohibitive computational overhead, severely bottlenecking their inference efficiency. Existing visual token compression methods predominantly adopt static and uniform compression strategies, neglecting the inherent *Semantic–Geometric Duality* in remote sensing interpretation tasks. Specifically, object semantic tasks focus on the abstract semantics of objects and benefit from aggressive background pruning, whereas scene geometric tasks critically rely on the integrity of spatial topology. To address this challenge, we propose DualComp, a task-adaptive dual-stream token compression framework. Dynamically guided by a lightweight pre-trained router, DualComp decouples feature processing into two dedicated pathways. In the object semantic stream, the Spatially-Contiguous Semantic Aggregator (SCSA) utilizes size-adaptive clustering to aggregates redundant background while protecting small object. In the scene geometric stream, the Instruction-Guided Structure Recoverer (IGSR) introduces a greedy path-tracing topology completion mechanism to reconstruct spatial skeletons. Experiments on the UHR remote sensing benchmark XLRS-Bench demonstrate that DualComp accomplishes high-fidelity remote sensing interpretation at an exceptionally low computational cost, achieving simultaneous improvements in both efficiency and accuracy.

Keywords: UHR Remote Sensing · Visual Token Compression · Semantic–Geometric Duality

* Equal contribution. ✉Corresponding authors.

1 Introduction

Recent advances in multimodal large language models (MLLMs) [1, 4, 35, 36, 61] have substantially improved visual understanding and reasoning, and have also accelerated progress in Earth science applications involving remote sensing data [15, 16, 26]. At the same time, ultra-high-resolution (UHR) remote sensing imagery has greatly expanded the observable detail available for Earth observation: it captures not only large geographic regions, but also fine-grained cues about human activities and natural environments [9, 45]. However, UHR imagery also produces extremely long visual token sequences in MLLMs, which significantly increase inference cost, memory usage, and latency [34]. As a result, efficient processing of UHR remote sensing imagery has become a major bottleneck for practical remote sensing MLLMs.

To mitigate the burden of long visual sequences in MLLMs [11, 20], existing MLLM approaches mainly follow three directions. In the general domain, recent methods typically reduce visual tokens through token merging on the vision side [30, 47, 55], token pruning during LLM decoding [10, 46, 49, 50], or instruction-aware compression guided by cross-modal relevance [2, 13, 48, 53]. These strategies have also inspired early remote sensing adaptations [24, 39], where representative systems such as GeoLLaVA-8K [39] introduce visual token compression, ZoomEarth [22] tackles UHR remote sensing through active perception. Despite their differences, these approaches largely treat efficiency as a problem of reducing or reallocating visual inputs, yet still lack an explicit task-adaptive principle for deciding what should be compressed and what should be preserved in remote sensing interpretation. This limitation becomes particularly critical in UHR remote sensing, where the value of visual evidence is highly task-dependent.

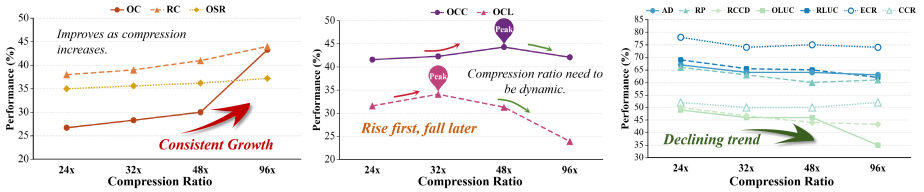


Fig. 1: Pilot study reveals semantic–geometric duality under token compression. We evaluate a UHR remote-sensing MLLM across multiple compression ratios and report performance changes for three representative groups of subtasks: (a) semantic-dominant improves as compression increases; (b) balanced peaks at a moderate ratio; (c) geometric-dominant drops under strong compression. This highlights that background tokens can be either redundant noise or indispensable geometric evidence, depending on task intent.

A key observation of this work is that UHR remote-sensing interpretation exhibits a pronounced *semantic–geometric duality*. We validate this via a pilot study on a representative UHR remote-sensing MLLM by sweeping compression

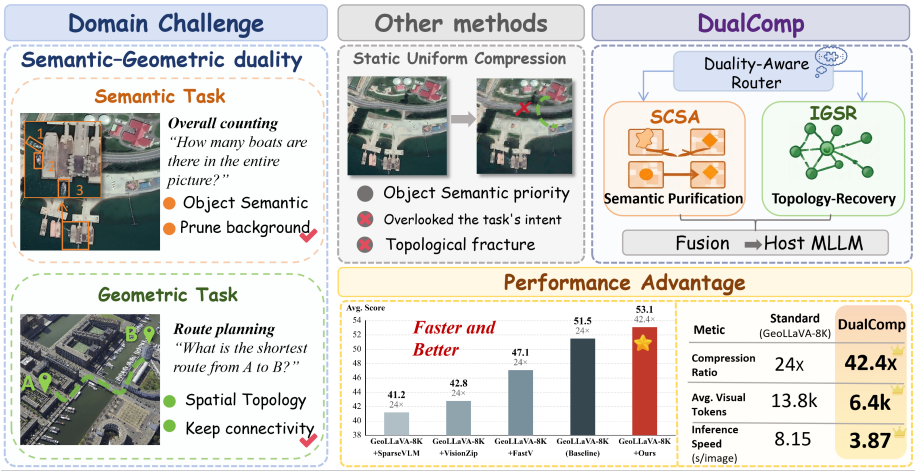


Fig. 2: DualComp achieves performance improvements under high compression ratios through semantic–geometric dual-stream adaptive compression. We reveal a pervasive semantic–geometric duality in UHR remote sensing MLLMs.

strengths and tracking subtask performance as background-redundant tokens are progressively removed. As shown in Fig. 1, the curves consistently fall into three regimes: **(i)** semantic-dominant tasks improve with stronger compression (*semantic purification*); **(ii)** balanced tasks peak at moderate compression; and **(iii)** geometric-dominant tasks degrade under strong compression, indicating the necessity of preserving context and topology. Overall, remote-sensing interpretation exhibits a pronounced *semantic–geometric duality*: semantic understanding relies more on object-related foreground evidence, whereas geometric reasoning requires preserving sufficient background context and topology—therefore, the utility of the same visual token can dynamically switch between the two with task intent, making token importance inherently task-dependent.

Motivated by this observation, we argue that remote sensing MLLMs should move from passive information reduction to *task-adaptive information scheduling*. To this end, we propose **DualComp** (Duality-Aware Token Compression), a plug-and-play dynamic visual token compression framework designed for UHR remote sensing. DualComp decomposes compression into two complementary pathways: a *semantic stream* that preserves object-level semantic fidelity, and a *geometric stream* that preserves scene-level structural fidelity. Specifically, we design a lightweight parasitic router that infers task intent from the user instruction and dynamically allocates token budgets between semantic and geometric evidence. The semantic stream is implemented by a Spatially-Contiguous Semantic Aggregator (SCSA), which compresses redundant background while preserving critical object-level information, and the geometric stream is implemented by an Instruction-Guided Structure Recoverer (IGSR), which preserves and reconstructs task-relevant geometric structures. The compression modules are training-free,

while the lightweight router is pretrained offline and frozen at inference, enabling plug-and-play token reduction with no updates to the host MLLM weights and improved efficiency and performance under high compression.

In summary, our main contributions are as follows:

1. We identify and quantitatively validate a pervasive *semantic–geometric duality* in UHR remote sensing MLLMs, and show that existing unified and static visual token compression strategies are fundamentally mismatched to the heterogeneous demands of remote sensing tasks.
2. We propose **DualComp**, a task-intent-aware dynamic visual token compression framework that explicitly models task demands through a lightweight router and adaptively preserves semantic and geometric evidence via the dual-stream design of SCSA and IGSR.
3. We develop a plug-and-play compression pipeline that significantly reduces the inference cost of UHR remote sensing imagery while consistently improving overall task performance, demonstrating the effectiveness of task-intent-aware compression for remote sensing MLLMs.

2 Related Work

2.1 Visual Token Compression for MLLMs

As MLLMs become increasingly capable of processing high-resolution images, the visual token sequence grows rapidly, while the quadratic complexity of Transformer self-attention [37] makes inference cost a critical bottleneck. To address this issue, one line of work converts high-resolution patch grids into more compact representations through explicit downsampling [17], spatial transformations [6, 7, 44, 62], or lightweight projections [2, 13, 48, 61] including LLaVA-OneVision [17], Qwen2.5-VL [7], InternVL2 [44], Blip-2 [18], and Honeybee [25], but these approaches usually require architectural modifications and additional training overhead. Another line of research explores training-free token reduction [10, 47, 55, 56, 59], primarily on the visual encoder side and the LLM decoding side, such as ToMe [8] and VisionZip [47]. During decoding, methods such as FastV [10], SparseVLM [59], PyramidDrop [46], and ATP-LLaVA [51] accelerate inference by pruning tokens at specific layers or in staged manners based on attention signals. In addition, approaches such as SparseVLM [59] and AdaFV [14] explicitly leverage text queries for cross-modal matching and selection, while VisionTrim [52] also uses text guidance to perform context-aware token merging within a more complete MLLM pipeline. Overall, most of these studies treat token compression as a single-dimensional problem of importance estimation, yet they often overlook the semantic–geometric duality inherent in ultra-high-resolution remote sensing scenes.

2.2 MLLMs in UHR Remote Sensing Understanding

General-purpose multimodal large language models (MLLMs), such as LLaVA [21] and Intern-S1 [5], have demonstrated strong visual understanding and instruction-following capabilities, driving the rapid development of remote sensing MLLMs.

Early efforts mainly focused on aligning and adapting remote sensing visual encoders to general-purpose LLMs, such as RSGPT [15], SkyEyeGPT [54], GeoChat [16], and EarthGPT [58]. Subsequent studies further improved data construction, alignment strategies, and instruction-following ability, including VHM [28], RS-CapRet [32], EarthMarker [57], LHRS-Bot-Nova [19], RSUni-VLM [23], EarthMind [31], EarthDial [33], RingMoGPT [42], and EarthVL [41].However, in ultra-high-resolution (UHR) remote sensing scenarios, these models often struggle to accurately localize task-relevant fine-grained regions within vast pixel spaces. To address this challenge, some studies have introduced visual token compression and selection strategies within MLLMs, such as GeoLLaVA-8K [39], ImageRAG [60], and RFM [24]. Another line of work shifts toward a tool-use paradigm, as exemplified by ZoomEarth [22] and VICoT-Agent [38], which acquire local details through chain-of-thought-driven multi-step zooming or retrieval. Yet, these approaches often require additional interaction rounds and incur substantial token overhead, making them still constrained by the trade-off between efficiency and scalability in UHR settings.

3 Method

3.1 Preliminary Analysis

Visual Token Compression in UHR MLLMs. When MLLMs process UHR remote sensing images, AnyRes-style multi-cropping [20] and dense patch-based representations often produce a massive number of visual tokens, leading to substantial memory and latency overhead. Existing static compression baselines typically shorten the visual sequence by uniformly clustering and merging local visual features without modifying the backbone model [10,39,46,47,59]. However, as illustrated in Fig. 1, our empirical analysis under extreme compression ratios (24×–96×) reveals a systematic limitation of this *static and uniform* paradigm in UHR remote sensing scenarios: it fails to account for the intrinsic semantic–geometric duality of remote sensing tasks, resulting in clear mismatches across different task intents.

More specifically, UHR remote sensing tasks lie on a semantic–geometric demands. For clarity, we describe three representative regimes:

- **Semantic-dominant tasks.** These tasks rely more heavily on object attributes and instance-level statistics, such as counting, presence detection, and relative relationships between targets. In essence, they focus more on *what is present*. In such cases, large homogeneous background regions often function mainly as noise, and the model relies more on recognizing discrete objects than on preserving precise topology or long-range connectivity.
- **Semantically–geometrically balanced tasks.** These tasks require both object-level semantic cues and fine-grained local structure. Typical examples include object classification and object color recognition, where moderate background compression can reduce irrelevant noise, but excessive compression may also remove discriminative details or context needed for reliable

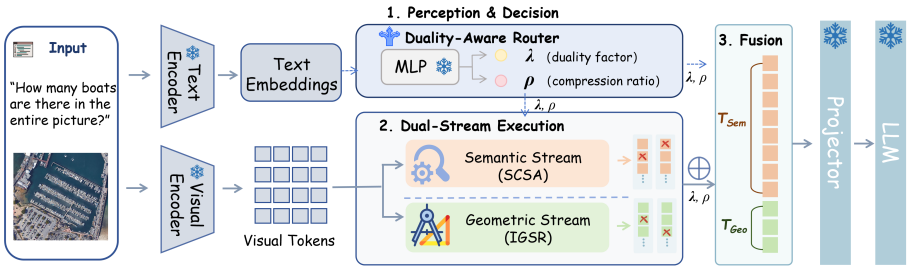


Fig. 3: Overview of DualComp. Text-conditioned routing outputs (λ, ρ), steering SCSA (semantic) and IGSR (geometric) token reduction and fusion.

prediction. As a result, these tasks typically favor a balanced preservation of semantic and geometric evidence rather than an extreme compression policy.

– **Geometric-dominant tasks.** These tasks rely more heavily on spatial organization and structural reasoning, including path planning, land-use or functional zone classification, and boundary understanding. In essence, they focus more on *where objects are* and *how spatial structures are organized*. Their success depends heavily on spatial integrity, continuous paths, precise boundaries, and contextual relations across regions. Static compression guided mainly by semantic saliency may therefore discard or over-merge tokens carrying critical structural evidence, causing severe performance degradation.

These observations indicate that, in UHR remote sensing interpretation, the set of high-value tokens changes dynamically with task intent, and a single compression ratio is insufficient for all tasks.

3.2 Our Approach

To address the compression mismatch caused by the semantic–geometric duality of ultra-high-resolution (UHR) remote sensing tasks, we propose DualComp, a task-adaptive dual-stream visual token compression framework. DualComp formulates visual token compression as an instruction-conditioned adaptive allocation problem and implements it through a dynamic *Perception–Decision–Execution* pipeline.

As illustrated in Fig. 3, the overall workflow of DualComp consists of two stages:

1. **Perception and Decision: Duality-Aware Router.** A lightweight *duality-aware router* parses the textual instruction and predicts a task-specific compression policy, including the overall retention strength and the relative preference between semantic and geometric evidence. In this way, the router determines both how aggressively the visual tokens should be compressed and how the retained budget should be scheduled across the two streams.
2. **Dual-Stream Execution and Fusion: SCSA and IGSR.** In the execution stage, DualComp decomposes visual token compression into two complementary streams. The semantic stream uses the Spatially-Contiguous Semantic

Aggregator (SCSA) to compress redundant background regions while preserving object-level semantics. The geometric stream uses the Instruction-Guided Structure Recoverer (IGSR) to preserve connectivity, boundaries, and other geometry-critical structures. The two streams are then fused by simple feature-level operations and directly fed into the MLLM without additional learnable projection layers.

Duality-Aware Router. The Router serves as the control center of DualComp. Given a textual instruction, it predicts two task-specific control variables: a duality factor $\lambda \in [0, 1]$ and an overall compression ratio $\rho \in [\rho_{\min}, 1]$. Here, a smaller λ indicates a stronger semantic preference, while a larger λ indicates a stronger geometric preference, and ρ controls the overall retention strength. Given the initial upper bound of visual tokens $N_{\max}$, the framework determines the total retention budget as $n_{\text{keep}} = N_{\max}\rho$, and then allocates it to the semantic and geometric streams as $n_{\text{sem}} = n_{\text{keep}}(1 - \lambda)$ and $n_{\text{geo}} = n_{\text{keep}}\lambda$.

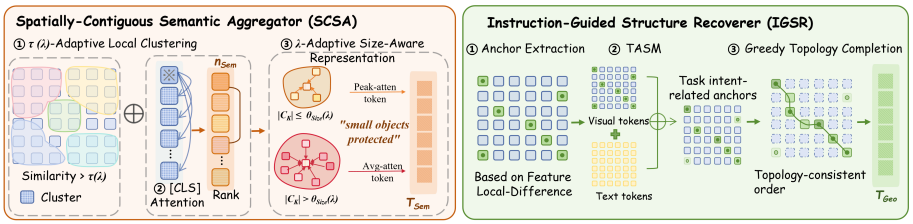
To minimize overhead, we adopt a parasitic design that attaches the Router directly to the text embedding output of the host MLLM. Text features are compressed into a compact instruction representation and fed into a shared multilayer perceptron (MLP) with two independent Sigmoid heads, which predict λ and ρ, respectively. The Router contains only about 1M parameters, remains frozen during inference, and preserves the plug-and-play nature of DualComp.

Since task duality preference lacks manually annotated labels, we construct an offline label generation pipeline with a dual-perspective annotation scheme for Router training. Specifically, we obtain λ_{llm} through Likert-scale probing with the host LLM and derive λ_{rule} from expert-designed linguistic rules. The final supervision signal is defined as $\lambda_{\text{gt}} = \alpha \cdot \lambda_{\text{llm}} + (1 - \alpha) \cdot \lambda_{\text{rule}}$.

Spatially-Contiguous Semantic Aggregator (SCSA) To reduce background redundancy when the task emphasizes object semantics, we introduce SCSA as the semantic stream of DualComp. Given the Router-allocated budget n_{sem}, SCSA performs training-free compression by aggregating homogeneous background regions while preserving instance-level information for small objects. As illustrated in Fig. 4(a), it consists of three stages.

1.* $\tau(\lambda)$*-Adaptive Local Clustering. SCSA first groups locally similar tokens into spatially contiguous clusters to compress redundant background regions. We compute cosine similarity using frozen CLIP [29] visual features and introduce a dynamic threshold $\tau(\lambda)$ controlled by the duality factor λ. For a token i, merging is allowed only when the maximum similarity within its local neighborhood exceeds the threshold:

$$\text{parent}(i) = \begin{cases} \arg\max_{j \in \mathcal{N}(i),\, j<i} \cos(f_i, f_j), & \text{if } \max\cos > \tau(\lambda), \\ i, & \text{otherwise.} \end{cases} \tag{1}$$



(a) Structure of SCSA module (b) Structure of IGSR module

Fig. 4: Two compression modules in DualComp. (a) SCSA aggregates redundant background while preserving semantic cues. (b) IGSR recovers topological cues via instruction-guided completion.

Here, $\tau(\lambda)$ is a monotonically increasing mapping, encouraging more aggressive aggregation when λ is small and preserving finer local granularity as λ increases. Detailed clustering implementation is provided in the appendix.

2. CLS Attention-Based Cluster Scoring. Given the resulting clusters $\{c_1, c_2, \ldots, c_K\}$, SCSA selects the top n_{sem} clusters according to semantic importance. We use the `[CLS]`-to-patch attention in a pretrained ViT as a measure of global semantic relevance [56]:

$$a_{[\text{CLS}]} = \text{Softmax}\left(\frac{q_{[\text{CLS}]} K_v^T}{\sqrt{d}}\right) \tag{2}$$

Cluster importance is then computed by cumulative attention:

$$\text{Importance}(c_k) = \sum_{i \in c_k} a_{[\text{CLS}],i} \tag{3}$$

This yields a compact set of semantically important clusters for subsequent representation.

3. λ-Adaptive Size-Aware Representation. We further apply a size-aware representation strategy controlled by a dynamic threshold $\theta_{\text{size}}(\lambda)$. For small clusters ($|c_k| \leq \theta_{\text{size}}$), we retain the original token with the highest `[CLS]` attention. For large clusters ($|c_k| > \theta_{\text{size}}$), we use an attention-weighted average as a summary token. Together, these steps enable SCSA to balance background compression and small-object preservation under different task intents, under the unified control of the Router's λ.

Instruction-Guided Structure Recoverer (IGSR) To preserve structural evidence for scene geometric tasks, DualComp introduces IGSR to utilize the Router-allocated budget n_{geo}. IGSR performs topological reconstruction on the compressed feature grid to recover connectivity, boundaries, and other geometry-critical structures. As illustrated in Fig. 4(b), it is a parameter-free module consisting of three steps.

1. Feature Local-Difference Anchor Extraction. IGSR first estimates geometric saliency from local feature variation:

$$S_{\text{geo}}(i) = \left\| F(i) - \text{AvgPool}_{3\times 3}(F)(i) \right\|_2^2 \tag{4}$$

To ensure spatial coverage, IGSR retains the highest-scoring token in each subregion as a structural anchor. Detailed anchor selection is provided in the appendix.

2. Text-Aware Structural Modulation. To make reconstruction instruction-aware, IGSR further introduces a text-aware structural modulation signal (TASM). Specifically, we compute a text relevance score S_{text} using CLIP [29] text–vision similarity and use it to modulate the structural cost field:

$$S_{\text{struct}}(i) = \tilde{S}_{\text{geo}}(i) \cdot \Big(1 + \beta \cdot \tilde{S}_{\text{text}}(i)\Big) \tag{5}$$

where $\tilde{S}_{\text{geo}}$ and $\tilde{S}_{\text{text}}$ are normalized scores. This biases reconstruction toward instruction-relevant structures, while naturally reducing to purely geometric reconstruction when the text signal is weak.

3. Parallel Greedy Topology Completion. Given structural anchor pairs (A_i, A_{i+1}), IGSR recovers their connecting path on the discrete feature grid through a fully parallel *Greedy Path Tracing* strategy. At each step, the next node is selected from the n-neighborhood $\mathcal{N}_n(v_t)$ as the candidate with the highest structural cost, under a Chebyshev-distance-decreasing constraint:

$$v_{t+1} = \arg \max_{u \in \mathcal{N}_n(v_t)} S_{\text{struct}}(u), \quad \text{s.t. } D_{\text{Chebyshev}}(u, A_{i+1}) < D_{\text{Chebyshev}}(v_t, A_{i+1}) \tag{6}$$

Implemented with parallel tensor operations, this step restores structural connectivity with minimal additional overhead.

Together, these steps restore critical structural evidence, allowing IGSR to provide the geometric stream of DualComp with efficient structural fidelity.

Dual-Stream Fusion and Sequence Unrolling After SCSA produces the condensed semantic features T_{sem} and IGSR reconstructs the geometrically continuous features T_{geo}, DualComp fuses the two streams and feeds them into the host MLLM. To preserve plug-and-play compatibility, no additional projection or normalization layers are introduced. Instead, we adopt a strategy based on λ-guided fusion and topological sequence unrolling.

1. λ-Encoded Dual-Stream Fusion. We perform fusion through λ-based weighted concatenation, $T_{\text{fused}} = \text{Concat}\big((1-\lambda)\,T_{\text{sem}},\, \lambda\,T_{\text{geo}}\big)$, which injects task intent at both the token-allocation and feature-magnitude levels: $(1-\lambda) : \lambda$ controls the semantic–geometric token ratio, while the weights softly suppress the non-dominant stream under strong task preference.

2. Topological Sequence Unrolling. Under extreme compression, token dropping and clustering may disrupt the original 2D spatial structure. DualComp alleviates this issue through *topological sequence unrolling*, where IGSR outputs a 1D geometric token sequence T_{geo} ordered by spatial connectivity. When fed into the host LLM, its native relative positional encoding (e.g., RoPE) can capture local continuity along this sequence. Without modifying positional encoding parameters, the host LLM can therefore leverage its autoregressive context modeling to reason over underlying 2D connectivity.

4 Experiments

4.1 Experimental Setup

Implementation Details and Benchmark. We deploy DualComp on a UHR remote-sensing-specific MLLM and further validate its generality on Qwen2.5-VL [7]. For fair comparison, both experimental tracks follow the default settings of their respective original papers under the same evaluation protocol. We adopt XLRS-Bench [40], currently the largest and highest-resolution multimodal benchmark for remote sensing, as our primary evaluation platform. It contains ultra-high-resolution remote sensing images and covers 13 fine-grained VQA subtasks, spanning the full spectrum from object-semantic-oriented tasks (e.g., counting and object classification) to scene-geometric-oriented tasks (e.g., route planning and anomaly detection).

4.2 Main Results

Overall Performance Comparison. Tab. 1 shows that generic token reduction methods (VisionZip, SparseVLM, and FastV) consistently underperform on UHR remote sensing tasks, indicating that task-agnostic importance heuristics often discard remote-sensing-specific evidence. In contrast, DualComp achieves the best overall accuracy (53.10%), outperforming both generic baselines and the RS-tailored static compression baseline GeoLLaVA-8K (51.50%, +1.6 points). This suggests that task-intent-aware scheduling is more effective than fixed compression policies for balancing semantic and geometric evidence.

The largest gains appear on geometry-sensitive tasks, where performance depends on connectivity and structural integrity. With the geometric stream IGSR, DualComp improves *Route Planning* from 66% to 72% and *Overall Land Use Classification* from 49% to 53%, with similar gains on other geometry-heavy reasoning subtasks. This confirms that preserving topology-critical evidence is essential under aggressive compression.

DualComp also remains competitive on semantic-dominant tasks. The semantic stream SCSA improves *Regional Counting* (38% $\rightarrow$ 45%) and *Object Color* (31.6% $\rightarrow$ 34.0%), while maintaining comparable performance on most object-centric subtasks. The OMS accuracy remains virtually unchanged across methods, which we attribute to a limitation of the current evaluation set; the same

Table 1: Experimental results on the perception and reasoning dimensions on XLRS-Bench. The sub-tasks include **Perception**: Overall/Regional Counting (OC/RC), Overall/Regional Land Use Classification (OLUC/RLUC), Object Classification (OCC), Object Color (OCL), Object Motion State (OMS), and Object Spatial Relationship (OSR); and **Reasoning**: Anomaly Detection and Interpretation (AD), Environmental Condition Reasoning (ECR), Route Planning (RP), Regional Counting with Change Detection (RCCD), and Counting with Complex Reasoning (CCR). 'Avg.' represents the macro average accuracy across sub-tasks. Rankings are calculated column-wise: green (1st, bold), light green (2nd), and yellow (3rd). Tie scores are treated identically.

Method	Compression	Perception								Reasoning					Avg.
Sub-tasks (L-3 Capability)		**OC**	**RC**	**OSR**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	
Closed-source MLLMs															
Claude 3.7 Sonnet [3]	-	27.6	22.7	27.6	17.4	68.4	30.5	29.9	63.6	64.8	78.4	34.5	27.8	32.6	40.5
Gemini 2.5 Pro [12]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
GPT-5.2 [27]	-	30.0	37.0	34.0	17.0	70.5	43.0	41.4	**68.3**	74.0	76.0	52.0	36.7	38.0	47.5
Open-source MLLMs															
LLaVA-Next [20]	-	26.7	40.0	30.0	5.0	67.0	28.8	32.8	66.7	69.0	78.0	27.0	35.0	36.0	41.7
LLaVA-Next+RFM+DIP [24]	4×	36.7	41.0	25.6	2.0	54.5	33.9	34.0	53.3	70.0	76.0	24.0	**53.3**	44.0	42.2
InternVL3-8B [62]	-	**40.0**	39.0	25.2	10.0	71.5	44.5	30.8	65.0	**77.0**	82.0	36.0	21.7	50.0	45.6
Qwen2-VL-7B [43]	-	26.7	40.0	31.8	11.0	73.0	35.9	34.6	61.7	70.0	81.0	35.0	46.7	48.0	45.8
InternVL2.5-8B [44]	-	38.3	37.0	21.6	10.0	77.0	33.4	35.5	65.0	73.0	**83.0**	34.0	50.0	43.0	46.2
Qwen2.5-VL-7B [7]	-	33.3	40.0	36.2	31.0	77.0	40.6	40.5	66.7	68.0	72.0	27.0	38.3	45.0	47.4
Qwen3-VL-8B [6]	-	21.7	**50.0**	30.4	26.0	**81.5**	**46.6**	**43.1**	66.7	74.0	79.0	37.0	43.3	51.0	50.0
Qwen2.5-VL-72B [7]	-	33.3	47.0	34.0	39.0	80.0	45.3	42.1	65.0	71.0	74.0	37.0	43.3	42.0	50.2
Intern-S1-mini [5]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51.6
Remote Sensing MLLMs															
GeoChat [16]	-	16.7	29.0	24.2	2.0	23.0	21.1	16.8	35.0	33.0	43.0	10.0	-	21.0	22.9
ZoomEarth [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.2
GeoLLaVA-8K [39]	-	26.7	38.0	35.0	49.0	69.0	41.6	31.6	65.0	67.0	78.0	66.0	50.0	**52.0**	51.5
GeoLLaVA-8K+VisionZip [47]	24×	23.3	39.0	38.6	37.0	49.0	44.1	30.0	65.0	36.0	38.0	62.0	46.7	47.0	42.8
GeoLLaVA-8K+FastV [10]	24×	23.3	41.0	37.4	46.0	50.5	43.8	31.6	65.0	53.0	60.0	65.0	46.7	49.0	47.1
GeoLLaVA-8K+SparseVLM [59]	24×	21.7	39.0	**38.8**	38.0	32.5	33.8	26.5	65.0	43.0	43.0	62.0	45.0	47.0	41.2
GeoLLaVA-8K+Ours	**42.4×**	26.7	45.0	37.4	**53.0**	69.5	43.6	34.0	65.0	69.0	79.0	**72.0**	48.3	49.0	**53.1**

invariance is observed on Qwen-based backbones. A few hybrid tasks still show limited gains or slight drops, suggesting that tasks jointly requiring fine-grained instance cues and broader context remain more sensitive to budget partitioning. Nevertheless, DualComp achieves the best overall performance across the semantic–geometric remote sensing tasks.

Table 2: Comparison of computational efficiency with different token reduction methods based on GeoLLaVA-8K.

Metrics	GeoLLaVA-8K	w/ VisionZip	w/ FastV	w/ SparseVLMs	w/ Ours
Compression Ratio	24×	24×	24×	24×	**42.4×**
Tokens per Grid	24	24	24	24	**14.2**
Avg. Visual Token Volume	13.8k	13.8k	13.8k	13.8k	**6.4k**
TFLOPs of LLM	198.1	198.7	198.1	186.1	**99.8**
Inference Speed (s/image)	8.15	8.41	19.56	7.84	**3.87**
Visual Encoding + Compression	4.28	5.08	8.59	4.72	**1.52**
LLM Generation	3.87	3.33	10.97	3.13	**2.25**
Avg. Score	51.5%	42.75%	47.10%	41.17%	**53.10%**

Inference Efficiency and Acceleration As shown in Tab. 2, DualComp achieves the best overall performance while also delivering the highest efficiency. Unlike competing methods, which all use a fixed 24× compression ratio, DualComp reaches 42.4×, reducing the average visual token volume from 14.0k to 6.4k, and lowering LLM computation from 198.06 TFLOPs to 99.75 TFLOPs.

More importantly, DualComp achieves the fastest end-to-end inference without sacrificing accuracy. It runs at 3.87 s/image, significantly faster than VisionZip (8.41 s/image), FastV (19.56 s/image), and SparseVLM (7.84 s/image), while still achieving the best overall accuracy (53.10%). This speedup is achieved through a lightweight visual compression stage (1.52 s) and a shorter LLM generation stage (2.25 s).

Unlike methods such as FastV and SparseVLM, which perform token compression after visual tokens are passed to the LLM, DualComp reduces and redistributes visual evidence before the generation phase. VisionZip adopts a hybrid strategy, preserving high-value tokens while merging the remaining ones based on similarity, but it incurs significant visual-side overhead, especially with larger images. In contrast, DualComp not only compresses more aggressively but also achieves the highest performance with the lowest runtime.

4.3 Ablation Study

To evaluate the contribution of each component in DualComp, we conduct ablations on the dual-stream design, explicit topology completion, topological sequence unrolling, and text-aware structural modulation.

Dual-Stream Architecture We first compare the full model with *SCSA-only (w/o geometric stream)* and *IGSR-only (w/o semantic stream)*. The full model achieves the best overall accuracy (53.1%), outperforming SCSA-only (50.27%) and IGSR-only (51.06%), confirming the complementarity between the

Table 3: Ablation results of DualComp on XLRS-Bench. We compare the full model with variants that remove or modify key components, including *SCSA-only* (w/o geometric stream), *IGSR-only* (w/o semantic stream), *Top-K* (w/o topology), *TASM-off* (w/o text-aware modulation), and *Index-Reorder* (w/o topological unrolling). The subtask abbreviations follow Table 1.

Method	**Perception**								**Reasoning**					**Avg.**
Sub-tasks	**OC**	**RC**	**OSR**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	
SCSA-only	25.0	44.0	36.0	50.0	69.0	40.0	26.5	**65.0**	62.0	77.0	67.0	45.0	47.0	50.3
IGSR-only	25.0	42.0	34.8	50.0	68.0	43.4	28.6	**65.0**	65.0	78.0	71.0	45.0	48.0	51.1
Top-K	25.0	41.0	34.4	52.0	65.0	38.9	31.3	**65.0**	65.0	76.0	65.0	45.0	50.0	50.5
TASM-off	25.0	44.0	35.8	48.0	66.5	42.8	31.8	**65.0**	67.0	77.0	71.0	45.0	47.0	51.2
Index-Reorder	25.0	42.0	35.4	51.0	66.0	43.9	32.1	**65.0**	67.0	77.0	**72.0**	45.0	48.0	51.5
DualComp	**26.7**	**45.0**	**37.4**	**53.0**	**69.5**	**43.6**	**34.0**	**65.0**	**69.0**	**79.0**	**72.0**	48.3	49.0	**53.1**

semantic and geometric streams. As illustrated in Tab. 3, this complementarity is consistently reflected across geometric-dominant, semantically–geometrically balanced, and semantic-dominant tasks.

Removing the geometric stream mainly hurts geometric-dominant tasks. For example, RP drops from 72.00% to 67.00%, and AD decreases from 69.00% to 62.00%, indicating that semantic aggregation alone cannot adequately preserve topology-critical evidence for scene-level reasoning. In contrast, removing the semantic stream mainly affects semantic-dominant tasks. For instance, RC falls from 45.00% to 42.00%, and OC decreases from 26.70% to 25.00%, showing that geometric recovery alone is insufficient for preserving object-level semantic cues under compression.

The complementarity is even more evident on semantically–geometrically balanced tasks that depend on both fine-grained object semantics and structural context. On OCC, the full model achieves 43.63%, outperforming both SCSA-only (40.00%) and IGSR-only (43.38%). These results show that neither stream alone is sufficient to cover the full range of semantic and geometric demands in UHR remote sensing, and that the best performance is achieved only when both streams are jointly preserved.

Explicit Topology Completion To assess the role of explicit topology recovery, we construct *Top-K (w/o topology)*, which keeps only the top-scoring structural tokens without path connection. This variant yields the lowest overall accuracy among all ablations (50.5%). As shown in Tab. 3, the largest degradation appears on RP, which drops sharply from 72.00% to 60.00%; RLUC also decreases from 69.50% to 65.00%. This confirms that retaining isolated structural tokens is insufficient for geometric reasoning, and that explicit topology completion is critical for preserving connectivity under aggressive compression.

Topological Sequence Unrolling To evaluate the effect of sequence organization, we construct *Index-Reorder (w/o topological unrolling)*, which preserves the topology paths recovered by greedy path tracing but reorders the selected tokens by their original spatial indices before feeding them into the LLM. Tab. 3 shows that its overall accuracy drops from 53.1% to 51.49%. Although it remains clearly better than *Top-K (w/o topology)* (51.49% vs. 49.96%), it still underperforms the full model, indicating that performance depends not only on recovering the right structural tokens, but also on organizing them in a topology-consistent order. This verifies the contribution of topological sequence unrolling in improving geometric continuity modeling.

Text-Aware Structural Modulation Finally, we construct *TASM-off (w/o text-aware modulation)*, which removes text–vision similarity modulation and relies only on local feature differences to build the structural cost field. As shown in the Tab. 3, the overall accuracy decreases from 53.1% to 51.22%, showing that text priors further improve the alignment between structure selection and

Table 4: Performance comparison on XLRS-Bench based on Qwen2.5-VL-7B.

Method	Compression	Perception								Reasoning					Avg.
Sub-tasks		**OC**	**RC**	**OSR**	**OLUC**	**RLUC**	**OCC**	**OCL**	**OMS**	**AD**	**ECR**	**RP**	**RCCD**	**CCR**	
Qwen2.5-VL-7B	–	33.3	40.0	36.2	31.0	77.0	40.6	40.5	66.7	68.0	72.0	27.0	38.3	45.0	47.4
+ *VisionZip*	1.43×	35.0	38.0	36.8	33.0	78.5	39.0	40.6	66.7	68.0	73.0	25.0	38.3	46.0	47.5
+ *VisionZip*	10×	36.7	42.0	34.0	33.0	78.5	38.6	40.3	66.7	60.0	72.0	24.0	38.3	46.0	46.9
+ ***Ours***	**10.24×**	**38.3**	36.0	35.4	**36.0**	72.5	**41.9**	**40.6**	**66.7**	**71.0**	**74.0**	**29.0**	**38.3**	43.0	**47.9**

task intent. The drop is more evident on scene understanding tasks, e.g., OLUC decreases from 53.00% to 48.00% and RLUC from 69.50% to 66.50%. Still, TASM-off remains stronger than several other ablated variants, suggesting that text-aware modulation is an important enhancement rather than a prerequisite for geometric recovery.

4.4 Further Analysis: Transferability to General-Purpose MLLMs

To further evaluate the generality of DualComp, we transplant it to the general-purpose model Qwen2.5-7B. As shown in Tab. 4, DualComp improves the overall accuracy from 47.40% to 47.90%, showing that the proposed framework remains effective beyond remote-sensing-specific backbones. This gain reflects more than backbone compatibility: it suggests that the semantic–geometric duality modeled by DualComp is intrinsic to UHR remote sensing tasks themselves, and therefore remains beneficial even on a general-purpose MLLM.

Under the same 10× compression ratio, DualComp also outperforms VisionZip-10× (47.90% vs. 46.92%), with especially clear gains on AD (71% vs. 60%), ECR (74% vs. 72%), and RP (29% vs. 24%). Moreover, even compared with a VisionZip variant using a smaller compression ratio, DualComp still achieves higher overall accuracy (47.90% vs. 47.53%), indicating that its advantage comes from more effective semantic–geometric scheduling rather than from a looser compression setting. Overall, these results show that DualComp can be effectively adapted to general-purpose MLLMs while maintaining stable gains on UHR remote sensing tasks.

5 Conclusion

In this paper, we target a key bottleneck in scaling MLLMs to ultra-high-resolution (UHR) remote-sensing imagery: the prohibitive inference cost from visual token explosion and the mismatch of static compression policies to task-heterogeneous interpretation. Through a pilot study, we reveal a pronounced semantic–geometric duality: semantic understanding can benefit from background denoising, whereas geometric reasoning depends on preserving background context, structural continuity, and topology as critical evidence. To address this, we propose DualComp, a task-intent-aware dual-stream token compression framework: the semantic stream uses SCSA to compress redundant background while retaining object-related

evidence, and the geometric stream employs IGSR to recover structural and connectivity cues under high compression for topology-sensitive reasoning. The compression modules are parameter-free and training-free at deployment, and the router is pretrained offline and frozen at inference, enabling plug-and-play integration without updating host MLLM weights. Extensive results on XLRS-Bench show that DualComp substantially reduces tokens and end-to-end latency while improving accuracy across both semantic- and geometry-dominant tasks, validating the effectiveness of task-aware compression for UHR remote-sensing understanding.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376282) and the Science and Technology Innovation Program of Hunan Province (No.2025RC3117).

References

1. Achiam, J., Adler, S., faces, S.A., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS **35**, 23716–23736 (2022)
3. Anthropic: Anthropic ai (2023), `https://www.anthropic.com`
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
5. Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., et al.: Intern-s1: A scientific multimodal foundation model. arXiv preprint arXiv:2508.15763 (2025)
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
8. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
9. Castillo-Navarro, J., Le Saux, B., Boulch, A., Audebert, N., Lefèvre, S.: Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. Machine Learning **111**(9), 3125–3160 (2022)
10. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: ECCV. pp. 19–35 (2024)

11. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
13. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. NeurIPS **36**, 49250–49267 (2023)
14. Han, J., Du, L., Wu, Y., Zhou, X., Du, H., Zheng, W.: Adafv: Rethinking of visual-language alignment for vlm acceleration. arXiv preprint arXiv:2501.09532 (2025)
15. Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., Li, X.: Rsgpt: A remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **224**, 272–286 (2025)
16. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: CVPR. pp. 27831–27840 (2024)
17. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
19. Li, Z., Muhtar, D., Gu, F., Zhang, X., Xiao, P., He, G., Zhu, X.: Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. arXiv preprint arXiv:2411.09301 (2024), `https://arxiv.org/abs/2411.09301`
20. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
22. Liu, R., Fu, B., Song, J., Li, K., Li, W., Xue, L., Qiao, H., Zhang, W., Meng, D., Cao, X.: Zoomearth: Active perception for ultra-high-resolution geospatial vision-language tasks. arXiv preprint arXiv:2511.12267 (2025)
23. Liu, X., Lian, Z.: Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. arXiv preprint arXiv:2412.05679 (2024), `https://arxiv.org/abs/2412.05679`
24. Luo, J., Zhang, Y., Yang, X., Wu, K., Zhu, Q., Liang, L., Chen, J., Li, Y.: When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In: ICCV. pp. 9206–9217 (2025)
25. Menzel, R.: The honeybee as a model for understanding the basis of cognition. Nature Reviews Neuroscience **13**(11), 758–768 (2012)
26. Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P.: Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In: ECCV. pp. 440–457 (2024)
27. OpenAI: Introducing gpt-5.2. Tech. rep., OpenAI (2025), `https://openai.com/index/introducing-gpt-5-2/`

28. Pang, C., Weng, X., Wu, J., Li, J., Liu, Y., Sun, J., Li, W., Wang, S., Feng, L., Xia, G.S., He, C.: Vhm: Versatile and honest vision language model for remote sensing image analysis. arXiv preprint arXiv:2403.20213 (2024), https://arxiv.org/abs/2403.20213
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
30. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: ICCV. pp. 22857–22867 (2025)
31. Shu, Y., Ren, B., Xiong, Z., Paudel, D.P., Gool, L.V., Demir, B., Sebe, N., Rota, P.: Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. arXiv preprint arXiv:2506.01667 (2025), https://arxiv.org/abs/2506.01667
32. Silva, J.D., Magalhães, J., Tuia, D., Martins, B.: Large language models for captioning and retrieving remote sensing images. arXiv preprint arXiv:2402.06475 (2024), https://arxiv.org/abs/2402.06475
33. Soni, S., Dudhane, A., Debary, H., Fiaz, M., Munir, M.A., Danish, M.S., Fraccaro, P., Watson, C.D., Klein, L.J., Khan, F.S., Khan, S.: Earthdial: Turning multi-sensory earth observations to interactive dialogues. arXiv preprint arXiv:2412.15190 (2025), https://arxiv.org/abs/2412.15190
34. Sun, S., Zhang, Y., Song, H., Guo, Z., Chen, C., Zhang, Y., Yao, Y., Liu, Z., Sun, M.: Llava-uhd v3: Progressive visual compression for efficient native-resolution encoding in mllms. arXiv preprint arXiv:2511.21150 (2025)
35. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
36. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
38. Wang, C., Luo, Z., Liu, R., Ran, C., Fan, S., Chen, X., He, C.: Vicot-agent: A vision-interleaved chain-of-thought framework for interpretable multimodal reasoning and scalable remote sensing analysis. arXiv preprint arXiv:2511.20085 (2025)
39. Wang, F., Chen, M., Li, Y., Wang, D., Wang, H., Guo, Z., Wang, Z., Shan, B., Lan, L., Wang, Y., et al.: Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. arXiv preprint arXiv:2505.21375 (2025)
40. Wang, F., Wang, H., Guo, Z., Wang, D., Wang, Y., Chen, M., Ma, Q., Lan, L., Yang, W., Zhang, J., et al.: Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In: CVPR. pp. 14325–14336 (2025)
41. Wang, J., Zhong, Y., Chen, Z., Zheng, Z., Ma, A., Zhang, L.: Earthvl: A progressive earth vision-language understanding and generation framework. arXiv preprint arXiv:2601.02783 (2026), https://arxiv.org/abs/2601.02783
42. Wang, P., Hu, H., Tong, B., Zhang, Z., Yao, F., Feng, Y., Zhu, Z., Chang, H., Diao, W., Ye, Q., Sun, X.: Ringmogpt: A unified remote sensing foundation model for vision, language, and grounded tasks. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–20 (2025). https://doi.org/10.1109/TGRS.2024.3510833
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

44. Wang, W., Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Zhu, J., Zhu, X., Lu, L., Qiao, Y., et al.: Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. arXiv preprint arXiv:2411.10442 (2024)
45. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
46. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024)
47. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: CVPR. pp. 19792–19802 (2025)
48. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
49. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: AAAI. vol. 39, pp. 22128–22136 (2025)
50. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: Atp-llava: Adaptive token pruning for large vision language models. In: CVPR. pp. 24972–24982 (2025)
51. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: Atp-llava: Adaptive token pruning for large vision language models. In: CVPR. pp. 24972–24982 (2025)
52. Yu, H., Li, W., Qu, X., Wang, S., Chen, J., Zhu, J.: Visiontrim: Unified vision token compression for training-free mllm acceleration. arXiv preprint arXiv:2601.22674 (2026)
53. Zeng, Q.S., Li, Y., Wang, Q., Jiang, P.T., Wu, Z., Cheng, M.M., Hou, Q.: A glimpse to compress: Dynamic visual token pruning for large vision-language models. arXiv preprint arXiv:2508.01548 (2025)
54. Zhan, Y., Xiong, Z., Yuan, Y.: Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS Journal of Photogrammetry and Remote Sensing **221**, 64–77 (2025)
55. Zhang, C., Ma, K., Fang, T., Yu, W., Zhang, H., Zhang, Z., Xie, Y., Sycara, K., Mi, H., Yu, D.: Vscan: Rethinking visual token reduction for efficient large vision-language models. arXiv preprint arXiv:2505.22654 (2025)
56. Zhang, Q., Cheng, A., Lu, M., Zhuo, Z., Wang, M., Cao, J., Guo, S., She, Q., Zhang, S.: [CLS] attention is all you need for training-free visual token pruning: Make vlm inference faster. arXiv preprint arXiv:2412.01818v1 (2024)
57. Zhang, W., Cai, M., Zhang, T., Li, J., Zhuang, Y., Mao, X.: Earthmarker: A visual prompting multi-modal large language model for remote sensing. arXiv preprint arXiv:2407.13596 (2024), `https://arxiv.org/abs/2407.13596`
58. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–20 (2024)
59. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:2410.04417 (2024)
60. Zhang, Z., Shen, H., Zhao, T., Guan, Z., Chen, B., Wang, Y., Jia, X., Cai, Y., Shang, Y., Yin, J.: Imagerag: Enhancing ultra high resolution remote sensing imagery analysis with imagerag. arXiv preprint arXiv:2411.07688 (2024)

61. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
62. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., bit, Y.D., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)