RSHR-BENCH: BENCHMARKING MLLM UNDERSTANDING AND VISUAL-TOKEN PRUNING IN ULTRA-HIGH-RESOLUTION REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have made substantial progress in visual perception and reasoning. However, reliable evaluation of MLLMs on ultra-high-resolution remote sensing imagery remains limited because existing high-resolution benchmarks do not consistently include systematic human verification of visual grounding. To address this gap, we introduce the Remote Sensing High-Resolution Benchmark (RSHR-Bench), which contains 5,329 satellite, aerial, and UAV images with an average resolution of $8{,}700 \times 8{,}065$ pixels. RSHR-Bench comprises 1,932 perception and reasoning VQA items, spanning multiple-choice and open-ended formats as well as single-image, multi-image, and multi-turn settings. All questions and reference answers are independently verified by human annotators for correctness, visual grounding, and unambiguous wording. We evaluate thirty-four vision–language models on RSHR-Bench, ranging from general-purpose MLLMs to models developed specifically for remote sensing. The results show that the highest overall accuracy is only 54.1% and that most models specialized for remote sensing fail to outperform general-purpose MLLMs. We further compare twenty-two visual-token pruning methods on RSHR-Bench, XLRS-Bench-lite, and MME-RS at multiple nominal token-retention budgets. Under aggressive compression, many methods suffer substantial accuracy losses, highlighting the trade-off between token reduction and the preservation of visual understanding. We hope RSHR-Bench will serve as a challenging benchmark for advancing MLLMs on ultra-high-resolution remote sensing imagery.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) have shown steadily improving capabilities in visual perception and reasoning (Wang et al., 2025c; Bai et al., 2025b). Visual question answering (VQA) provides a direct way to assess whether these models can use images to answer specific questions (Liu et al., 2024b; Li et al., 2023). Recent general-purpose multimodal benchmarks evaluate perception, knowledge, and reasoning largely using comparatively low-resolution natural images (Lin et al., 2026; Cheng et al., 2025; Ying et al., 2024; Wang et al., 2025a). In high-resolution images, task-relevant regions may be small and widely separated across the visual field (Lee et al., 2026; Ma et al., 2025). Models therefore need to preserve local details and global context while managing longer visual-token sequences. Consequently, strong performance on benchmarks based on images of lower resolution may not reliably indicate that a model can understand much larger images. High-resolution benchmarks have begun to address this gap: HRBench (Wang et al., 2025d) targets 4K and 8K inputs, while MME-RealWorld (Zhang et al., 2025d) covers challenging high-resolution scenarios across multiple real-world domains. Satellite and UAV applications further highlight this need, as their images often cover large areas while preserving small but important visual details at multiple spatial scales.

Remote sensing VQA benchmarks have similarly trended toward larger image inputs. RSVQA (Lobry et al., 2020), RSIVQA (Zheng et al., 2021), and VRSBench (Li et al., 2024c) primarily use 512×512 images, while HRVQA (Li et al., 2024b) increases the resolution to 1024×1024 for aerial imagery. XLRS-Bench-lite (Wang et al., 2025b) extends evaluation to extremely large remote sensing images across VQA, captioning, and localization, whereas LRS-VQA (Luo et al., 2025) fo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

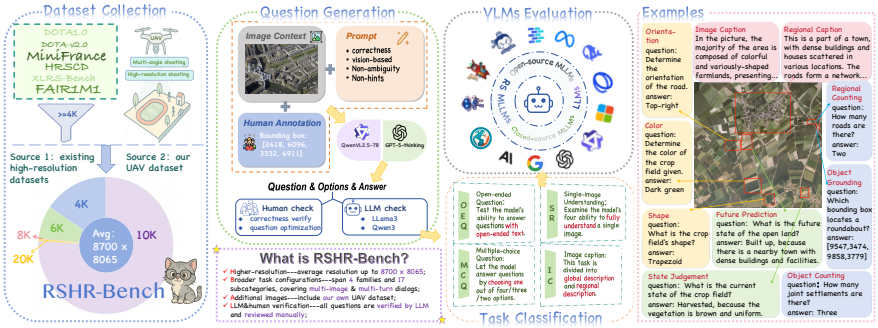


Figure 1: Overview of RSHR-Bench. We collect high-resolution satellite, aerial, and UAV imagery, generate questions, and conduct independent human verification. The benchmark includes four evaluation settings and representative examples spanning diverse remote sensing scenarios.

cuses on large-image VQA and introduces coarse-to-fine, text-guided visual token pruning to make such inputs more tractable. However, larger images do not ensure that questions and answers are visually grounded, while retaining fine details creates long visual-token sequences, motivating rigorous human verification and efficient token selection. Most existing visual-token pruning methods are evaluated primarily on comparatively low-resolution, general-domain imagery, reducing tokens before the language model (Yang et al., 2025b; Shang et al., 2025; Alvar et al., 2025), within the decoder layers of the language model (Ye et al., 2025), or across multiple stages (Liu et al., 2024a; Singh et al., 2026). Recent remote sensing efforts have begun to consider larger inputs: GeoLLaVA-8K (Wang et al., 2026) scales remote sensing MLLMs to 8K resolution with visual-token selection, while UHR-BAT (Dang et al., 2026) introduces budget-aware token compression for ultra-high-resolution remote sensing imagery. Overall, two coupled gaps remain: existing benchmarks often lack systematic human verification of visual grounding, while pruning methods lack unified evaluation of compression robustness on ultra-high-resolution remote sensing VQA.

To address these gaps, we construct the Remote Sensing High-Resolution Benchmark (RSHR-Bench) for VQA in ultra-high-resolution remote sensing imagery. Figure 1 summarizes the benchmark construction process, including image collection, question generation, and independent human verification. Its source corpus contains 5,329 images from satellite, aerial, and UAV collections, with an average native resolution of $8{,}700 \times 8{,}065$ pixels. RSHR-Bench contains 1,932 VQA items covering perception and reasoning through multiple-choice and open-ended questions, single- and multi-image inputs, and multi-turn interactions. Every final question and reference answer is independently checked by human reviewers for correctness, visual grounding, and unambiguous wording. Using this human-verified benchmark, we compare MLLMs designed for remote sensing with general-purpose MLLMs across open-source and proprietary systems. The results reveal persistent limitations: models designed for remote sensing do not consistently outperform general-purpose models, and several MLLMs perform close to chance on the multiple-choice tasks. Beyond benchmarking model capability, we use RSHR-Bench to study how visual-token compression affects accuracy under ultra-high-resolution inputs. We evaluate twenty-two pruning methods spanning pre-LLM, intra-LLM, and hybrid designs and extend the comparison to XLRS-Bench-lite and MME-RS. On RSHR-Bench, even the strongest method drops from 50.72% under dense inference to 40.73% at 0.01% nominal retention. Across the three benchmarks, the best method remains 6.98–19.33 percentage points below the corresponding dense baselines at 0.1% retention, while simple random or uniform sampling remains competitive with many specialized methods. Overall, our contributions are summarized as follows:

- We introduce RSHR-Bench, a human-verified VQA benchmark for ultra-high-resolution remote sensing imagery that covers perception and reasoning tasks.
- We comprehensively benchmark thirty-four general-purpose and remote sensing MLLMs on RSHR-Bench across diverse VQA settings. The results show that advanced models still achieve limited accuracy on ultra-high-resolution remote sensing VQA.
- We conduct a unified evaluation of twenty-two visual-token pruning methods across three high-resolution remote sensing VQA benchmarks and multiple retention budgets, showing that many methods lose substantial accuracy under aggressive compression.

2 Related Work

General Multimodal Benchmarks. Recent multimodal benchmarks have advanced the quantitative assessment of VLMs, yet many focus on narrow domains or a small set of tasks (e.g., captioning (Chen et al., 2015; Plummer et al., 2015) or VQA (Hudson & Manning, 2019; Marino et al., 2019; Goyal et al., 2017; Gurari et al., 2018; Zhou et al., 2018)). To broaden coverage, MME (Fu et al., 2026) spans 14 perceptual and cognitive tasks, MMBench (Liu et al., 2024b) offers 3,000+ questions over 20 skill dimensions (e.g., localization and social reasoning), SeedBench (Li et al., 2023) further enlarges question volume (19k), and MMT-Bench (Ying et al., 2024) brings in application-oriented data from autonomous driving and embedded AI. High-resolution benchmarks provide complementary coverage: HRBench (Wang et al., 2025d) targets 4K/8K inputs, while MME-RealWorld (Zhang et al., 2025d) includes a dedicated remote sensing domain alongside other real-world scenarios. Thus, the distinction is not the absence of remote sensing evaluation in prior work, but the combination of source-image scale, task design, interaction formats, and independent human verification of every retained item.

Remote Sensing Multimodal Benchmarks. Prior work follows three main directions. (i) Modest-resolution captioning and VQA: RSIEval (Hu et al., 2025) and LHRS-Bench (Muhtar et al., 2024) use human-curated items that mainly assess perception and short-context understanding. (ii) Reliability and structured semantics: RSSA (Pang et al., 2024) targets false-information detection, while FIT-RSRC (Luo et al., 2024) and VLEO-BENCH (Zhang & Wang, 2024) emphasize relations, localization, counting, and change understanding. Other datasets broaden remote sensing VQA through high-resolution aerial imagery, relational reasoning, and multiple sensor modalities (Lobry et al., 2020; Zheng et al., 2021; Li et al., 2024b; Wang et al., 2024; Zi et al., 2025; Boussaid et al., 2025). (iii) Scale and resolution: DOTA (Xia et al., 2018) and HRSCD (Daudt et al., 2019) illustrate the large source imagery encountered in remote sensing. XLRS-Bench-lite (Wang et al., 2025b) extends multimodal evaluation to high-resolution VQA, captioning, and localization, VRSBench (Li et al., 2024c) provides large-scale QA with object references, and MDAS (Hu et al., 2022) adds hyperspectral classification and change detection. RSHR-Bench complements these efforts by combining ultra-high-resolution source imagery, diverse interaction formats, and independent human verification of questions and reference answers within a unified protocol.

Visual Token Compression. Existing methods, surveyed broadly in (Shao et al., 2025), can be grouped by the location at which visual tokens are removed. *(i) Pre-LLM pruning.* EViT (Liang et al., 2022) and QuietPrune (Gao et al., 2026) reduce tokens in the vision encoder, while VisionZip (Yang et al., 2025b), DivPrune (Alvar et al., 2025), TopV (Yang et al., 2025a), LLaVA-PruMerge (Shang et al., 2025), VisPruner (Zhang et al., 2025b), CDPruner (Zhang et al., 2025c), and HiPrune (Liu et al., 2025) prune or merge tokens before they enter the language model. *(ii) Intra-LLM pruning.* FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024), TransPrune (Li et al., 2026a), FEATHER (Endo et al., 2025), FiT-Prune (Ye et al., 2025), VTW (Lin et al., 2025), PyramidDrop (Xing et al., 2024), DART (Wen et al., 2025), and PriorTR (Chen et al., 2026) remove visual tokens within decoder layers using attention, text relevance, token transitions, spatial coverage, or layer-adaptive schedules. *(iii) Hybrid pruning.* IVTP (Huang et al., 2024), MustDrop (Liu et al., 2024a), DUET-VLM (Singh et al., 2026), and VScan (Zhang et al., 2025a) combine token reduction across multiple stages. Recent remote sensing systems also incorporate token selection or compression for large-image inputs (Luo et al., 2025; Wang et al., 2026; Dang et al., 2026). In this work, we evaluate a range of visual-token pruning methods under a unified protocol for ultra-high-resolution remote sensing VQA across multiple token budgets, with dense inference and random and uniform sampling as controls for controlled method comparisons.

3 Method

This section describes the construction and evaluation of RSHR-Bench. Figure 1 outlines three stages: dataset collection, question generation, and independent human verification. Section 3.1 details the image sources and construction process, while Section 3.2 introduces the capability taxonomy and four settings: multiple-choice VQA, open-ended VQA, image captioning, and single-image evaluation. Together, they define the benchmark data, tasks, and evaluation protocol.

3.1 RSHR-BENCH CONSTRUCTION

Dataset Collection. We assemble a high-resolution corpus by combining six public remote sensing benchmarks with a high-altitude UAV set of $\sim 100\,\mathrm{MP}$ single-shot frames. For every source, we keep native resolution and filter images so that the long side is at least 4K pixels. From DOTA v1.0 (Xia et al., 2018) and DOTA v2.0 (Ding et al., 2021), we select large-format satellite scenes that contain dense human-made structures and long-range context. From XLRS-Bench-lite (Wang et al., 2025b), we sample high-resolution scenes with detailed spatial context. From HRSCD (Daudt et al., 2019), we retain paired views of the same areas to support temporal consistency analysis. From MiniFrance (Castillo-Navarro et al., 2022), we choose very high-resolution metropolitan tiles with rich textures and many small objects. From FAIR1M 1.0 (Sun et al., 2022), we gather wide-area airport, port, and industrial scenes with dense fine-scale targets and varied viewing angles. Our UAV collection repeatedly surveys the same geographic area from diverse viewpoints and flight attitudes, with each frame captured in a single exposure at about $100\,\mathrm{MP}$, which preserves native geometry. Our UAV scenes emphasize small objects under scale and perspective variation. Overall, our image collection is constructed to prioritize high resolution, comprehensive scene coverage, object diversity, task relevance, and varied scene structures. The native-resolution distribution of the images used in the VQA evaluation is reported in Appendix C.

Question Generation. Following recent remote sensing benchmarks (Li et al., 2024c; Wang et al., 2025b), we define 13 task families covering perception and reasoning. The resulting data support four evaluation settings: *closed-set VQA* (multiple choice), *open-ended VQA* (free-form), *image captioning*, and *single-image evaluation*. Perception tasks cover nine common query families, including color, orientation, regional grounding, and regional counting. Reasoning questions are image-specific, require direct inspection of image content, and avoid phrasing that hints at the correct option, including in future-prediction and object-state-judgement tasks. All tasks follow three principles—*correctness*, *unambiguous phrasing*, and *visual answerability*. ❶ **VQA generation.** For closed-set and open-ended VQA tasks, we prioritize small but distinctive targets. This approach emphasizes high-resolution understanding while ensuring that each target remains uniquely identifiable in context during annotation. We draft questions and answer choices through a model-assisted pipeline: perception-oriented items with provided boxes and labels are drafted with Qwen2.5-VL-7B (Bai et al., 2025b), whereas harder compositional or multi-image items are drafted with GPT-5 Thinking (Singh et al., 2025). As an additional language-only check, LLaMA 3 and Qwen3 attempt to answer each question without access to the image; questions that can be answered reliably from wording alone are revised or removed. The resulting drafts undergo human revision and independent review. ❷ **Caption generation.** For full-image captioning, GPT-5 Thinking generates a global scene summary together with descriptions of the top, bottom, left, and right regions. Caption references are then manually reviewed, with a random set of 100 references used for an additional audit (Appendix C). ❸ **Single-image evaluation.** We design ten questions per image covering both perception and reasoning, including eight open-ended questions, one local image caption, and one global image caption. We select 50 images and author 500 questions for the single-image evaluation. This separate set is written and checked by humans. The model-assisted generation and validation process consumed about 300 GPU hours. Construction and evaluation prompts are included in the appendix.

Human Verification. To ensure question quality, question authoring and verification are assigned to different annotators. Annotator **A** drafts each question or revises a model-assisted draft, checking for answer correctness, dependence on visible evidence, unambiguous wording, and the absence of unintended linguistic cues. Reviewer **B** independently inspects the source image, question, answer options, and reference answer. Items with ambiguous targets, unsupported answers, multiple plausible options, or cross-turn inconsistencies are revised and rechecked or discarded. Consequently, every retained item passes independent human review. All six trained annotators and reviewers held at least a bachelor's degree and spent approximately 300 person-hours on question authoring and verification. In a separate answerability study, human raters correctly answered 237 of 255 sampled items (92.94%, Appendix B). After editing and review, RSHR-Bench contains 1,932 question–answer items, including multi-image questions and individual turns of multi-turn dialogues.

Comparison with Prior Benchmarks. As shown in Table 1, we compare our benchmark with existing remote sensing datasets. Specifically, we consider two categories of remote sensing datasets: non-VQA and VQA. Non-VQA datasets (e.g., DOTA (Xia et al., 2018; Ding et al., 2021), DIOR (Li et al., 2020)) focus on recognition or captioning and generally lack multi-turn dialogue

and multi-image reasoning. Many earlier VQA datasets use relatively low-resolution images (typically 512×512) and are limited to single-turn, single-image interactions (e.g., RSVQA (Lobry et al., 2020), HRVQA (Li et al., 2024b), RSIVQA (Zheng et al., 2021)). The comparison reveals a scale–resolution trade-off: VQA benchmarks with tens or hundreds of thousands of images generally use inputs of only 512–1,024 pixels, whereas existing large-image VQA benchmarks contain fewer than 2,000 images and do not support multi-turn interaction. Among the VQA datasets in Table 1, RSHR-Bench is the only one that combines satellite, aerial, and UAV imagery with multi-turn questions and independent human verification of every retained item. In contrast, our benchmark emphasizes: ❶ **Higher-resolution imagery.** RSHR-Bench includes native images up to $29{,}200 \times 27{,}620$ pixels (approximately 800 million pixels), requiring models to handle small targets within large scenes. The full image corpus combines satellite, aerial, and UAV imagery spanning *tens to hundreds of megapixels*, with an average resolution of $8{,}700 \times 8{,}065$ pixels. ❷ **Independent human verification.** Every retained question, option set, and reference answer is checked against the corresponding image by a reviewer who is separate from the question author. The verified annotations include 1,578 boxed objects across 1,366 images, while RSHR-Bench contains 1,932 question–answer items. ❸ **Global image perception and broader task configurations.** We vary interaction formats (e.g., single-image multi-turn and multi-image single-turn) to probe multi-image fusion, memory, and decision-making, and broaden task coverage from fine-grained perception (color/shape cues) to cross-image reasoning (multi-region contrast, object-state judgement, future prediction) across ultra-high-resolution scenes and interaction settings.

Table 1: Comparison of existing remote sensing benchmarks with our benchmark. ✗and ✓indicate absence and presence, respectively; ◯, ✓, and ✓ indicate model-generated, human-written, and semi-automatic annotations (*i.e.*, model generation followed by human verification), respectively.

Dataset	Source	Images	Avg Res.	Max Res.	Multi-turn	Annotation Method	LLM Check
Non-VQA							
GeoPixel (Shabbir et al., 2025)	Aerial	65,463	2,560×2,560	4,960×4,960	✗	✓	✗
GeoGround (Zhou et al., 2024)	Aerial, UAV	80,000	640×640	1,024×1,024	✗	✓	✗
LuoJiaHOG (Zhao et al., 2025)	Google Earth	94,856	640×640	1,280×1,280	✗	✓	✗
DOTA v1.0 (Xia et al., 2018)	Aerial, Satellite	2,806	4,000×4,000	20,000×20,000	✗	✓	✗
DOTA v2.0 (Ding et al., 2021)	Aerial, Satellite	11,268	4,500×4,500	20,000×20,000	✗	✓	✗
FAIR1M 1.0 (Sun et al., 2022)	Satellite	15,000	5,300×5,300	10,000×10,000	✗	✓	✗
RSICD (Lu et al., 2017)	Aerial, Satellite	10,921	224×224	224×224	✗	✓	✗
RSVG (Zhan et al., 2023)	Satellite	17,402	800×800	800×800	✗	✓	✗
RRSIS-D (Yuan et al., 2024)	Satellite	17,402	800×800	800×800	✗	✓	✗
DIOR (Li et al., 2020)	Satellite	23,463	800×800	800×800	✗	✓	✗
MillionAID (Long et al., 2021)	Aerial, Satellite	1,000,848	-	31,672×31,672	✗	✓	✗
RSICap (Hu et al., 2025)	Aerial, Satellite	2,585	512×512	512×512	✗	✓	✗
VQA							
RSIEval (Hu et al., 2025)	Aerial, Satellite	100	512×512	512×512	✗	✓	✗
DVL-Bench (Xuan et al., 2026)	Aerial	3,469	1,024×1,024	1,024×1,024	✗	✓	✗
XLRS-Bench-lite (Wang et al., 2025b)	Satellite, Aerial	1,400	8,500×8,500	10,000×10,000	✗	✓	✗
LRS-VQA (Luo et al., 2025)	Satellite, Aerial	1,657	5,403×4,935	26,176×24,832	✗	✓	✗
RSIVQA (Zheng et al., 2021)	Aerial, Satellite	37,264	512×512	512×512	✗	✓	✗
VRSBench (Li et al., 2024c)	Aerial, Satellite	29,614	512×512	512×512	✓	✓	✗
RSVQA (Lobry et al., 2020)	Aerial, Satellite	11,431	499×499	512×512	✗	◯	✗
HRVQA (Li et al., 2024b)	Aerial	53,512	1,024×1,024	1,024×1,024	✗	◯	✗
TAMMI (Boussaid et al., 2025)	Aerial, Satellite	282,852	1,000×1,000	1,000×1,000	✗	◯	✗
RSVLM-QA (Zi et al., 2025)	Aerial, Satellite	6,000	512×512	512×512	✗	✓	✗
EarthVQA (Wang et al., 2024)	Satellite	6,000	512×512	512×512	✗	✓	✗
Segmentation VQA (Tosato et al., 2024)	Aerial	16,274	2,560×2,560	4,096×4,096	✗	◯	✗
RSHR-Bench	**Satellite, Aerial, UAV**	**5,329**	**8,700 × 8,065**	**29,200 × 27,620**	✓	✓	✓

3.2 Tasks and Evaluation Settings

VQA Task Suite. We evaluate nine perception categories and four reasoning families. Regional comparison includes separate multi-image and single-image variants, yielding fourteen single-turn task categories in total. Three multi-turn categories provide complementary dialogue-based evaluation. Following recent benchmarks (Luo et al., 2025; Wang et al., 2025b), we use multiple-choice VQA with the task-specific options supplied for each question, including three- and four-option formats. For the open-ended setting, GPT-4o (Hurst et al., 2024) removes the answer options and revises the wording to request an answer supported by the image. Human reviewers check that the revised question remains clear and consistent with the reference. Open-ended and multiple-choice scores are treated as complementary measurements. Option removal does not guarantee visual dependence or greater difficulty. Detailed task definitions, examples, and evaluation configurations are provided in Appendix C. These tasks separate local evidence from broader contextual reasoning. Multi-image and multi-turn variants also test consistency across views and related questions.

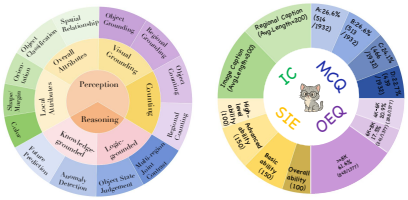


Figure 2: RSHR-Bench capability taxonomy and evaluation settings.

Capability Taxonomy. Figure 2 summarizes the complementary perception and reasoning capabilities evaluated by RSHR-Bench. Perception evaluates object attributes and spatial structure, covering color, shape, orientation, classification, spatial relationships, grounding, and counting. Reasoning uses image-derived information for anomaly interpretation, future prediction, regional comparison, and object state judgement. These tasks span single- and multi-image inputs with single-turn and multi-turn interactions, assessing whether models can connect local observations with contextual and temporal inferences across spatial scales while remaining consistent across related views.

Evaluation Settings. The benchmark comprises four evaluation settings that differ in answer format and evaluation scope while drawing on the capability taxonomy above. (i) *Multiple-choice VQA* requires selecting an answer from a fixed candidate set. (ii) *Open-ended VQA* requires a free-form answer without candidate options. This setting removes explicit option cues but depends on reference quality and the scoring rubric. (iii) *Image captioning* requires concise, accurate descriptions of regions and whole images. (iv) *Single-image evaluation* applies a fixed ten-item suite to each selected image, jointly assessing perception, reasoning, and captioning across images with different native resolutions. These settings assess recognition, localization, compositional reasoning, and holistic understanding in ultra-high-resolution remote sensing imagery.

4 Experiments

This section evaluates model capabilities and visual-token pruning on ultra-high-resolution remote sensing imagery. Section 4.1 benchmarks models for remote sensing and general applications across multiple-choice VQA, open-ended VQA, captioning, and multi-turn questions. Section 4.2 compares pruning methods across retention budgets and remote sensing benchmarks.

4.1 Evaluation on RSHR-Bench

Experimental setup. We evaluate 34 vision–language models on RSHR-Bench, including fifteen models specialized for remote sensing, twelve general-purpose open-source models, and seven closed-source models. Each model answers the same set of multiple-choice questions. A prediction is counted as correct only when the selected option matches the human-verified reference answer. We use the same rules to compute accuracy for every task and for the combined results. The models retain their original image-processing and decoding configurations, so the comparison reflects practical task performance rather than a matched-compute setting. We additionally evaluate open-ended VQA and image captioning to examine free-form generation. Every evaluated question and reference answer has been independently checked by human reviewers for correctness, visual grounding, and unambiguous wording under a shared evaluation protocol.

Evaluation Metrics. All entries in Table 2 are percentages based on whether the selected option matches the reference answer. P-Avg., R-Avg., and Overall Avg. (shown as Avg. in the table) combine predictions from perception, single-turn reasoning, and all single-turn tasks, respectively. Each question receives equal weight, so these values differ from simple means of the individual task scores. Multi-turn questions are excluded from these averages. MAD, MTFP, and MOSJ report accuracy for individual turns. MTEM@1 counts a group as correct only when all three turns are answered correctly, computes this rate separately for MAD, MTFP, and MOSJ, and macro-averages the three task-level rates. Across Tables 2, 3, and 4, the five highest distinct scores in each comparable column are marked by 1st, 2nd, 3rd, 4th, and 5th, with ties sharing the same color.

Comparison on RSHR-Bench (Multiple-Choice VQA). To assess how current vision–language models handle the full range of RSHR-Bench capabilities, Table 2 compares 34 models on nine perception tasks, five single-turn reasoning tasks, and three multi-turn tasks. GPT-5.5 achieves the highest perception and overall accuracies, at 54.2% and 54.1%, respectively. GPT-5 achieves the highest MTEM@1 at 52.6%. Among models specialized for remote sensing, RSCoVLM performs best in perception, reasoning, overall accuracy, and MTEM@1, reaching 33.3%, 52.7%, 36.0%, and 36.8%, respectively. EarthDial reaches 73.5% in multi-turn object-state judgement. The strongest

remote sensing model remains well below the best general-purpose model in overall accuracy. The best overall score is only 54.1%. Only slightly more than half of the multi-turn groups are answered entirely correctly, highlighting the difficulty of fine-grained and compositional understanding in ultra-high-resolution imagery across diverse remote sensing scenes.

Table 2: Comparison on RSHR-Bench for multiple-choice VQA. P-Avg., R-Avg., and Avg. give equal weight to each question in perception, single-turn reasoning, and all single-turn tasks, respectively. MTEM@1 macro-averages the all-turn success rates across the three multi-turn tasks.

Model	Perception									Reasoning					Weighted Average			Multi-turn			MTEM @1
	COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN	AD	FP	MRJC	MRJCS	OSJ	P-Avg.	R-Avg.	Avg.	MAD	MTFP	MOSJ	
Remote Sensing VLMs																					
EarthDial (Soni et al., 2025)	41.0	22.0	21.0	30.0	32.5	30.5	27.1	18.0	31.0	42.0	30.0	29.5	32.0	52.0	29.0	38.2	30.2	56.7	60.0	73.5	30.7
GeoLLaVA-8K (Wang et al., 2026)	25.0	24.0	25.0	25.0	25.0	25.0	21.4	25.0	25.0	24.0	0.0	0.0	34.0	22.0	24.7	18.2	23.8	25.0	24.7	47.7	7.0
VHM (Pang et al., 2025)	25.5	25.0	26.0	26.5	55.0	25.0	22.9	25.0	25.0	26.0	24.0	26.5	34.0	28.0	29.6	27.9	29.4	45.0	53.3	46.2	16.7
EarthMind-4B (Shu et al., 2025)	50.5	23.0	19.0	28.0	34.0	31.5	31.4	24.5	29.0	52.0	36.0	50.0	44.0	46.0	31.4	45.0	33.3	65.0	61.3	76.5	35.1
GeoChat-7B (Kuckreja et al., 2024)	47.5	14.0	35.0	33.5	24.5	25.0	28.6	16.5	33.0	50.0	38.0	40.0	38.0	40.0	28.9	41.4	30.6	56.7	62.0	56.1	20.2
LHRS-Bot-Nova (Muhtar et al., 2024)	40.0	30.0	27.0	37.0	34.5	27.5	17.1	26.5	26.0	46.0	48.0	30.0	30.0	56.0	31.1	43.6	32.8	68.3	58.0	65.9	24.6
MF-RSVLM-7B-SFT (Dang et al., 2025)	36.5	32.0	17.0	38.0	24.0	25.0	22.9	25.0	29.0	30.0	32.0	50.0	30.0	38.0	28.5	34.1	29.3	45.0	38.0	74.2	21.1
RS-LLaVA-7B (Bazi et al., 2024)	33.0	16.0	29.0	31.5	32.5	29.0	24.3	20.5	40.0	56.0	50.0	30.0	38.0	56.0	28.8	48.2	31.5	60.0	59.3	57.6	20.2
RSCoVLM-7B-2512 (Li et al., 2026b)	46.5	22.0	16.0	40.0	46.0	29.0	47.1	16.0	30.0	74.0	46.0	55.0	30.0	60.0	33.3	52.7	36.0	66.7	64.7	77.3	36.8
RSThinker (Liu et al., 2026)	39.0	31.0	32.0	35.0	33.0	30.0	31.4	23.0	24.0	52.0	46.0	25.0	24.0	64.0	31.3	44.5	33.1	61.7	62.7	77.3	35.1
RSUniVLM (Liu & Lian, 2024)	33.5	22.0	26.0	23.0	22.5	24.0	20.0	21.5	31.0	40.0	46.0	50.0	26.0	28.0	25.0	36.4	26.5	41.7	44.7	57.6	12.3
SkySenseGPT-7B (Luo et al., 2024)	49.5	17.0	28.0	36.0	26.0	23.0	25.7	21.0	35.0	52.0	34.0	35.0	36.0	40.0	29.9	40.0	31.3	66.7	59.3	65.2	25.4
TEOChat (Irvin et al., 2025)	40.0	26.0	17.0	27.5	24.5	24.0	25.7	12.5	27.0	58.0	34.0	50.0	38.0	58.0	25.2	47.3	28.2	60.0	58.0	67.4	24.6
GeoPixel-7B (Shabbir et al., 2025)	33.0	20.0	27.0	28.0	27.5	26.0	24.3	27.0	26.0	34.0	24.0	50.0	32.0	40.0	27.2	34.1	28.2	38.3	37.3	75.0	15.8
FALCON-8B (Yao et al., 2025)	39.0	23.0	32.0	27.0	30.0	28.0	21.4	20.0	32.0	52.0	26.0	35.0	38.0	50.0	28.5	40.9	30.2	60.0	60.7	68.9	25.4
Open-source VLMs																					
InternVL2.5-8B (Chen et al., 2023)	25.5	22.0	26.0	26.0	22.5	24.5	30.0	22.5	20.0	26.0	20.0	22.5	34.0	20.0	24.2	25.0	24.2	25.0	28.7	35.6	1.5
InternVL3.5-8B (Wang et al., 2025c)	21.5	28.0	18.0	21.5	29.0	28.5	30.0	26.5	25.0	20.0	16.0	29.0	34.0	26.0	25.3	24.5	25.1	30.0	22.7	40.2	6.1
MiniCPM-V-2.6 (Yao et al., 2024)	21.5	28.0	30.0	24.0	19.5	29.5	34.3	22.0	29.0	26.0	30.0	35.0	32.0	30.0	25.1	30.0	25.8	26.7	23.3	31.1	2.4
Phi-3.5-Vision (Abdin et al., 2024)	25.0	24.0	25.0	25.0	23.5	25.0	22.9	25.0	25.0	24.0	22.0	23.5	30.0	22.0	24.6	24.4	24.6	28.3	24.7	47.0	7.0
Qwen2.5-VL-7B (Bai et al., 2025b)	29.5	25.0	22.0	28.0	25.0	24.5	24.3	26.5	22.0	26.0	28.0	25.0	10.0	20.0	25.8	21.4	25.2	21.7	24.0	10.6	0.0
DeepSeek-VL (Lu et al., 2024)	22.5	22.0	21.0	25.0	20.5	26.0	28.6	20.5	22.0	22.0	28.0	50.0	32.0	20.0	22.9	27.7	23.6	20.0	23.3	33.3	9.1
VILA-HD (Shi et al., 2025)	40.0	22.0	22.0	37.0	35.5	26.0	21.4	24.5	24.0	58.0	30.0	55.0	32.0	58.0	29.9	45.5	32.0	65.0	57.3	57.6	24.8
LLaVA-OneVision (Li et al., 2024a)	46.5	36.0	23.0	36.5	49.5	25.0	27.1	13.5	38.0	62.0	20.0	50.0	38.0	62.0	33.4	45.9	35.2	68.3	58.7	72.7	29.8
Qwen3-VL-32B-Instruct (Bai et al., 2025a)	50.5	27.0	20.0	20.5	55.0	28.5	27.1	19.0	31.0	76.0	46.0	45.0	34.0	64.0	32.4	54.1	35.4	70.0	71.3	83.3	48.2
Qwen3-VL-4B-Instruct (Bai et al., 2025a)	50.0	36.0	16.0	30.5	45.5	32.0	45.7	19.0	44.0	82.0	46.0	50.0	48.0	68.0	35.2	60.0	38.6	70.0	64.0	80.3	43.9
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	24.5	25.0	25.0	25.0	23.5	26.5	25.7	25.5	27.0	24.0	26.0	45.0	36.0	20.0	25.2	28.2	25.6	28.3	25.3	10.6	0.0
Qwen3.5-9B (Qwen Team, 2026)	25.5	23.0	25.0	24.0	24.0	25.0	22.9	26.0	21.0	24.0	24.0	50.0	34.0	24.0	24.4	28.6	25.0	23.3	28.0	22.0	0.8
Closed-source VLMs																					
Claude-3.5-Sonnet	40.0	16.0	22.0	30.0	35.5	25.0	22.9	22.0	25.0	70.0	10.0	30.0	34.0	64.0	28.0	43.2	30.1	75.0	66.0	77.3	43.9
GPT-5 (Singh et al., 2025)	29.0	10.0	23.0	23.0	37.0	24.5	31.4	20.0	23.0	74.0	58.0	35.0	34.0	66.0	25.2	55.9	29.4	78.3	73.3	86.4	52.6
GPT-4o (Hurst et al., 2024)	49.5	23.0	15.0	35.5	30.5	28.0	27.1	22.5	41.0	68.0	56.0	30.5	32.0	64.0	31.4	53.2	34.3	70.0	72.0	84.1	47.4
GPT-4o-mini (Hurst et al., 2024)	41.5	16.0	29.0	31.5	31.5	32.0	28.6	19.5	32.0	54.0	54.0	31.5	48.0	54.0	29.9	50.6	32.7	78.3	68.0	75.0	41.2
Gemini-2.5-Pro (Comanici et al., 2025)	55.0	18.0	31.0	40.0	41.5	32.5	45.7	25.0	25.0	66.0	32.0	41.5	38.0	50.0	36.1	46.0	37.4	56.7	60.0	57.6	32.6
GPT-5.5 (OpenAI, 2026)	71.5	49.0	41.0	69.0	62.0	55.0	64.3	25.5	42.0	64.0	44.0	65.0	34.0	66.0	54.2	53.2	54.1	70.0	71.3	84.8	43.9
Claude-3.7-Sonnet (Anthropic, 2025)	48.0	29.0	24.0	31.0	46.0	30.0	37.1	23.5	27.0	68.0	54.0	60.0	34.0	64.0	33.8	55.5	36.8	71.7	66.0	78.8	38.6

Comparison on RSHR-Bench (Open-Ended VQA and Captioning). Open-ended VQA evaluates free-form responses without candidate options. Qwen3-8B scores the responses on a 1–100 scale for the category-level analysis. We test perception, reasoning, and multi-turn questions across the three model families. VILA-HD records the highest mean perception score at 44.0, while GPT-4o-mini records the highest mean reasoning score at 63.1 and the highest MTEM@80 at 42.1. Captioning is evaluated on a shared 1,000-image subset. InternVL3.5-8B is the strongest open-source model on BLEU-2–4, METEOR, and ROUGE-L. Its BLEU-4 score is only 2.5, indicating limited exact overlap with longer reference phrases. Phi-3.5-Vision obtains the highest open-source BLEU-1 score at 36.5. Metrics based on overlap with reference captions do not directly measure factual correctness. The appendix separately reports open-source results on the full 3,913-image caption set; because the image populations differ, scores from the subset and full set are not directly comparable. Complete results and evaluation protocols are provided in Appendix A.

Visual-Token Count versus Task Accuracy. To determine whether models that process more visual tokens perform better, Figure 3 compares task-macro accuracy across the fourteen single-turn categories with the mean visual-token count for the 22 configurations with available profiles. These comprise twelve remote-sensing models and ten general-purpose open-source models; closed-source models are omitted because comparable token counts are unavailable. Under the evaluated input settings, the Qwen3-VL and Qwen3.5 variants process approximately 17K visual tokens per question, yet their task-macro accuracies range from 26.6% to 43.8%. Qwen3-VL-4B and Qwen3-VL-8B reach 43.8% and 42.8%, respectively. RSCoVLM reaches 39.8% with about 1.4K tokens, exceeding the 37.7% achieved by LLaVA-OneVision with about 24.6K tokens, while LHRS-Bot-Nova reaches 34.0% with only 272 tokens. These results show that token count alone does not determine performance, motivating a direct evaluation of how much visual understanding remains at a given token budget.

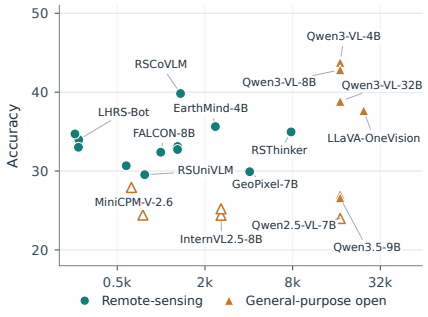


Figure 3: Task accuracy versus mean visual-token count.

4.2 VISUAL-TOKEN PRUNING UNDER ULTRA-HIGH-RESOLUTION INPUTS

Experimental setup. We reproduce the visual-token pruning algorithms from their original papers and official codebases in a common Qwen3-VL-8B interface that fixes the model, input pipeline, numerical precision, retention controller, and option scoring. All pruning methods use the same Qwen3-VL-8B model without fine-tuning or additional training. We group methods as Pre-LLM, Intra-LLM, or Hybrid according to their pruning location and use dense inference as the reference. The evaluation covers RSHR-Bench, the XLRS-Bench-lite single-answer subset, and MME-RS. The nine-budget RSHR-Bench main sweep uses 70%, 50%, 30%, 20%, 10%, 2%, 0.2%, 0.05%, and 0.01% retention. The four-budget cross-benchmark comparison uses 5%, 1%, 0.5%, and 0.1%. Realized retention and the number of tokens processed across layers vary by method, so the budgets do not represent matched computation across pruning strategies.

Comparison on RSHR-Bench across Retention Budgets. Table 3 compares 22 reproduced methods at nine nominal retention budgets on RSHR-Bench. Dense Qwen3-VL-8B reaches 50.72% accuracy without pruning. FiT-Prune achieves the highest accuracy at 70% retention (51.40%), and VScan leads at 50% retention (51.60%). PyramidDrop leads at the remaining seven compressed budgets. PyramidDrop reaches 51.45% at 30% retention, 49.22% at 2% retention, 44.41% at 0.2% retention, and 40.73% at 0.01%, compared with 50.72% for dense inference. At 0.01%, FiT-Prune reaches 39.96%. At the nominal 0.01% setting, PyramidDrop retains about 0.080% of the final visual tokens, but the cumulative number processed across decoder layers remains about 26.24% of dense inference. The nominal budget describes final retention, and PyramidDrop keeps more intermediate tokens than methods that prune before the language model. At moderate retention, PyramidDrop exceeds dense inference by at most 0.73 percentage points. Overall, PyramidDrop maintains accuracy most consistently as retention decreases, although it processes more intermediate tokens than methods that prune before the language model.

Table 3: RSHR-Bench accuracy (%) for 22 methods on the nine-budget main RSHR-Bench grid and 1,932 questions. Colors indicate the five highest distinct scores per column.

Method	Type	70%	50%	30%	20%	10%	2%	0.2%	0.05%	0.01%
EViT (Liang et al., 2022)	Pre-LLM	46.38	45.19	43.48	43.12	40.79	37.06	35.20	35.46	35.46
QuietPrune (Gao et al., 2026)	Pre-LLM	50.36	50.67	48.19	46.38	43.94	39.34	35.97	35.56	35.56
VisionZip (Yang et al., 2025b)	Pre-LLM	50.16	50.36	48.91	47.00	44.10	39.70	35.35	35.04	35.04
DivPrune (Alvar et al., 2025)	Pre-LLM	50.67	49.43	48.65	47.46	46.27	42.49	35.51	35.25	35.09
TopV (Yang et al., 2025a)	Pre-LLM	50.98	50.00	49.59	49.02	46.64	40.79	35.71	35.56	35.56
LLaVA-PruMerge (Shang et al., 2025)	Pre-LLM	49.02	46.84	44.20	42.55	40.53	37.32	35.87	36.13	36.13
VisPruner (Zhang et al., 2025b)	Pre-LLM	49.84	48.40	46.74	45.13	44.98	39.54	36.34	36.08	36.02
CDPruner (Zhang et al., 2025c)	Pre-LLM	50.78	50.05	48.55	45.86	42.91	38.35	34.94	35.46	35.46
HiPrune (Liu et al., 2025)	Pre-LLM	51.19	50.78	50.16	49.33	46.89	38.72	35.56	35.56	35.56
FastV (Chen et al., 2024)	Intra-LLM	49.43	48.76	47.62	45.29	42.60	37.22	35.30	35.46	35.35
SparseVLM (Zhang et al., 2024)	Intra-LLM	50.16	50.36	48.91	46.64	46.07	40.22	35.40	35.20	35.09
TransPrune (Li et al., 2026a)	Intra-LLM	50.47	50.36	48.45	47.36	44.72	40.58	36.39	36.28	36.23
FEATHER (Endo et al., 2025)	Intra-LLM	50.47	50.36	48.45	47.20	44.57	40.37	36.44	37.01	37.01
FiT-Prune (Ye et al., 2025)	Intra-LLM	51.40	50.98	51.35	50.67	47.83	43.37	40.48	39.96	39.96
VTW (Lin et al., 2025)	Intra-LLM	50.41	46.74	36.85	34.58	35.56	35.71	35.71	35.71	35.71
PyramidDrop (Xing et al., 2024)	Intra-LLM	50.72	51.35	51.45	50.78	50.78	49.22	44.41	42.49	40.73
DART (Wen et al., 2025)	Intra-LLM	50.57	50.10	48.14	48.08	46.17	41.36	36.75	36.75	36.75
PriorTR (Chen et al., 2026)	Intra-LLM	50.26	48.96	48.45	47.93	45.96	41.77	37.37	36.85	36.85
IVTP (Huang et al., 2024)	Hybrid	50.78	50.31	47.20	46.22	44.72	38.35	35.40	35.20	35.46
MustDrop (Liu et al., 2024a)	Hybrid	51.09	51.29	50.10	47.88	46.48	43.53	36.08	35.77	35.71
DUET-VLM (Singh et al., 2026)	Hybrid	49.07	47.15	44.82	43.69	42.60	39.75	36.28	35.92	35.56
VScan (Zhang et al., 2025a)	Hybrid	51.19	51.60	51.14	50.62	50.26	45.19	37.63	36.34	35.51

Comparison across Remote-Sensing Benchmarks. To test whether the pruning trends observed on RSHR-Bench generalize to other remote sensing evaluations, Table 4 compares 22 reproduced methods at nominal retention budgets of 5%, 1%, 0.5%, and 0.1% on the XLRS-Bench-lite single-answer subset, RSHR-Bench, and MME-RS. Across all 22 methods, PyramidDrop has the highest accuracy at every compressed setting, reaching 45.47%, 42.70%, and 38.67% at 0.1% retention on the three benchmarks, respectively. These values are 6.98, 8.02, and 19.33 percentage points below their dense references. The remaining methods vary substantially across benchmarks and budgets. Notably, random external sampling reaches 37.68% on RSHR-Bench at 0.5% retention, outperforming 17 of the 22 specialized pruning methods at the same budget. Full random and uniform sampling results across all evaluated datasets are reported in Appendix D.

Table 4: Cross-benchmark pruning accuracy (%) for twenty-two methods. Colors indicate the five highest distinct scores among all twenty-two methods in each compressed-budget column.

		XLRS-Bench-lite					RSHR-Bench					MME-RS				
Method	**Type**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**
EViT (Liang et al., 2022)	Pre-LLM	52.45	42.35	39.36	37.18	34.70	50.72	39.44	37.27	35.82	35.14	58.00	31.33	33.33	29.33	30.67
QuietPrune (Gao et al., 2026)	Pre-LLM	52.45	45.03	41.11	39.09	35.40	50.72	42.18	38.41	37.11	36.54	58.00	42.67	31.33	30.67	28.00
VisionZip (Yang et al., 2025b)	Pre-LLM	52.45	43.59	37.32	35.91	33.79	50.72	41.72	37.68	35.46	34.99	58.00	39.33	36.00	33.33	28.67
DivPrune (Alvar et al., 2025)	Pre-LLM	52.45	44.93	39.56	36.61	33.99	50.72	43.79	39.75	37.78	34.94	58.00	44.00	35.33	32.67	30.00
TopV (Yang et al., 2025a)	Pre-LLM	52.45	45.13	39.80	38.29	35.70	50.72	43.27	37.32	36.90	36.28	58.00	42.67	34.00	36.00	33.33
LLaVA-PruMerge (Shang et al., 2025)	Pre-LLM	52.45	44.63	40.30	39.40	37.01	50.72	39.39	36.75	35.77	36.54	58.00	37.33	34.67	30.00	30.67
VisPruner (Zhang et al., 2025b)	Pre-LLM	52.45	44.77	40.44	39.26	35.77	50.72	42.08	38.15	37.01	35.40	58.00	42.67	36.67	30.67	30.67
CDPruner (Zhang et al., 2025c)	Pre-LLM	52.45	42.62	35.54	34.46	33.62	50.72	40.53	37.47	34.99	34.99	58.00	34.67	26.00	29.33	25.33
HiPrune (Liu et al., 2025)	Pre-LLM	52.45	47.38	39.19	36.38	33.62	50.72	43.32	38.61	36.02	35.66	58.00	42.67	32.67	30.67	26.00
FastV (Chen et al., 2024)	Intra-LLM	52.45	42.15	34.23	33.86	33.59	50.72	40.94	36.34	35.40	35.35	58.00	34.67	31.33	26.67	24.67
SparseVLM (Zhang et al., 2024)	Intra-LLM	52.45	44.46	38.72	36.07	33.86	50.72	43.74	37.68	35.77	35.20	58.00	45.33	37.33	35.33	29.33
TransPrune (Li et al., 2026a)	Intra-LLM	52.45	43.52	39.60	38.83	34.77	50.72	43.53	39.18	37.47	36.90	58.00	41.33	37.33	37.33	30.00
FEATHER (Endo et al., 2025)	Intra-LLM	52.45	42.82	39.93	38.46	34.70	50.72	42.24	38.82	37.16	37.01	58.00	43.33	37.33	37.33	30.00
FiT-Prune (Ye et al., 2025)	Intra-LLM	52.45	45.64	42.21	41.61	41.04	50.72	44.15	41.51	41.05	40.22	58.00	36.67	37.33	38.00	37.33
VTW (Lin et al., 2025)	Intra-LLM	52.45	33.99	33.96	33.96	33.96	50.72	35.66	35.71	35.71	35.71	58.00	28.00	28.00	28.00	28.00
PyramidDrop (Xing et al., 2024)	Intra-LLM	52.45	50.64	48.09	47.38	45.47	50.72	50.00	48.86	46.01	42.70	58.00	49.33	50.00	46.00	38.67
DART (Wen et al., 2025)	Intra-LLM	52.45	47.11	41.17	35.81	35.81	50.72	44.00	40.48	36.75	36.75	58.00	44.67	35.33	26.67	26.67
PriorTR (Chen et al., 2026)	Intra-LLM	52.45	46.21	39.09	36.51	33.89	50.72	44.57	39.80	38.72	36.85	58.00	45.33	32.00	34.00	30.67
IVTP (Huang et al., 2024)	Hybrid	52.45	44.19	38.79	36.21	33.72	50.72	43.84	37.32	37.27	35.35	58.00	41.33	30.00	28.00	28.00
MustDrop (Liu et al., 2024a)	Hybrid	52.45	45.50	39.93	38.89	34.70	50.72	44.62	40.01	36.59	36.02	58.00	43.33	34.67	30.00	28.67
DUET-VLM (Singh et al., 2026)	Hybrid	52.45	42.72	40.44	39.66	36.24	50.72	41.56	38.56	37.06	36.13	58.00	32.67	32.00	29.33	26.67
VScan (Zhang et al., 2025a)	Hybrid	52.45	49.30	43.42	39.90	35.57	50.72	49.48	41.41	38.98	36.02	58.00	48.67	35.33	32.00	30.67

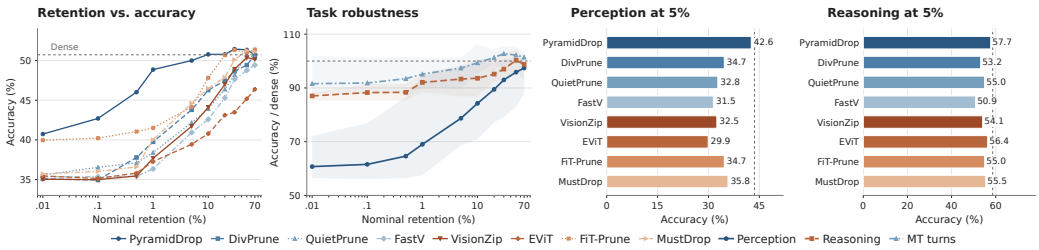


Figure 4: Visual-token pruning for eight representative methods on RSHR-Bench. From left to right: overall accuracy across retention budgets, task-group accuracy normalized by dense inference, and perception and reasoning accuracy at 5% retention. Dashed lines denote dense performance.

Analysis of Pruning Robustness across Retention Budgets and Task Groups. Figure 4 separates overall retention trends from task-specific degradation for eight representative methods. The first panel shows that PyramidDrop remains closest to dense accuracy as retention decreases, whereas most selected methods that prune before the language model degrade more sharply under aggressive compression. The second panel normalizes each task group's accuracy by its dense reference and summarizes the mean and variation across methods. Perception degrades earlier and more severely than reasoning and multi-turn performance, suggesting that pruning first removes fine-grained evidence needed for attribute and localization questions. The two right panels report raw accuracy at 5% retention. PyramidDrop performs best, reaching 42.6% on perception and 57.7% on reasoning, although both results remain below their dense references. Overall, robust pruning should preserve local visual details together with broader scene cues across task groups and retention budgets.

5 Conclusion

We construct RSHR-Bench to evaluate vision–language understanding in ultra-high-resolution remote sensing imagery across perception, reasoning, multi-image, and multi-turn tasks. We evaluate 34 models specialized for remote sensing or general applications and find that current systems remain limited in fine-grained perception and consistent reasoning. We also test 22 visual-token pruning methods across RSHR-Bench, XLRS-Bench-lite, and MME-RS, showing substantial accuracy degradation under aggressive compression and limited advantages over simple sampling. RSHR-Bench provides a common testbed for diagnosing these limitations and developing efficient MLLMs that preserve fine-grained visual evidence at extreme image resolutions. In future work, we will expand the benchmark to more remote sensing scenes, interaction formats, and task types, together with broader evaluations of efficient multimodal processing.

AI USE STATEMENT

We used generative AI tools in two limited ways. During benchmark construction, Qwen2.5-VL-7B, GPT-5, and GPT-4o helped prepare initial drafts of questions, captions, and open-ended reformulations, while LLaMA 3 and Qwen3 supported language-only answerability checks. Human annotators reviewed every retained question, option set, caption, and reference answer. During paper preparation, AI-assisted tools were also used for language editing and LaTeX debugging. The authors checked the experimental results, technical claims, citations, and final wording and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

We have tried to make the experiments easy to reproduce. The main paper and appendix describe the image sources, filtering rules, annotation process, task definitions, prompts, scoring rules, model input settings, and pruning configurations. All reported comparisons use fixed evaluation data and consistent scoring rules. We also distinguish nominal retention from the visual tokens actually processed by each pruning method. We plan to release the benchmark annotations, evaluation scripts, prompts, prediction files used for scoring, and pruning configurations, subject to the licenses of the source imagery and model providers.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

Anthropic. Claude 3.7 sonnet system card, 2025. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang

Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9), 2024. doi: 10.3390/rs16091477.

Hichem Boussaid, Lucrezia Tosato, Flora Weissgerber, Camille Kurtz, Laurent Wendling, and Sylvain Lobry. Visual question answering on multiple remote sensing image modalities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2310–2319. IEEE, 2025.

Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: the minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, 111(9):3125–3160, 2022. doi: 10.1007/s10994-020-05943-y.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zengjie Chen, Yuxiang Cai, Jingcai Guo, Taotao Cai, Jianwei Yin, and Zhi Chen. Accelerating multimodal large language models with prior-corrected token reduction. *arXiv preprint arXiv:2606.24156*, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646. IEEE, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Yunkai Dang, Donghao Wang, Jiacheng Yang, Yifan Jiang, Meiyi Zhu, Yuekun Yang, Cong Wang, Qi Fan, Wenbin Li, and Yang Gao. Fuse-rsvlm: Feature fusion vision-language model for remote sensing. *arXiv preprint arXiv:2512.24022*, 2025.

Yunkai Dang, Minxin Dai, Yuekun Yang, Zhangnan Li, Wenbin Li, Feng Miao, and Yang Gao. Uhrbat: Budget-aware token compression vision-language model for ultra-high-resolution remote sensing. In *Forty-third International Conference on Machine Learning*, 2026.

Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187: 102783, 2019. doi: 10.1016/j.cviu.2019.07.003.

Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7778–7796, 2021.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835. IEEE, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Tianxiao Gao, Shanwei Zhao, Shuo Fang, Shiai Zhu, and Chenguang Ma. Quietprune: Query-guided early token pruning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3553–3562, 2026.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jingliang Hu, Rong Liu, Danfeng Hong, Andrés Camero, Jing Yao, Mathias Schneider, Franz Kurz, Karl Segl, and Xiao Xiang Zhu. Mdas: A new multimodal benchmark dataset for remote sensing. *Earth System Science Data Discussions*, 2022:1–26, 2022.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European conference on computer vision*, pp. 214–230. Springer, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jeremy Irvin, Emily Liu, Joyce Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. In *International Conference on Learning Representations*, volume 2025, pp. 68883–68911, 2025.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840. IEEE, 2024.

Jewon Lee, Wooksu Shin, Seungmin Yang, Ki-Ung Song, DongUk Lim, Jaeyeon Kim, Tae-Ho Kim, and Bo-Kyeong Kim. Ergo: efficient high-resolution visual understanding for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 8750–8769, 2026.

Ao Li, Yuxiang Duan, Jinghui Zhang, Congbo Ma, Yutong Xie, Gustavo Carneiro, Mohammad Yaqub, and Hu Wang. Transprune: Token transition pruning for efficient large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39529–39538, 2026a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

Kun Li, George Vosselman, and Michael Ying Yang. Hrvqa: A visual question answering benchmark for high-resolution aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 214:65–81, 2024b.

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

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024c.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5334–5342, 2025.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 6535–6563, 2026.

Jizhihui Liu, Feiyi Du, Guangdao Zhu, Niu Lian, Jun Li, and Bin Chen. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models. *arXiv preprint arXiv:2508.00553*, 2025.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024a.

Xu Liu and Zhouhui Lian. Rsunivlm: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566, 2020.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):2183–2195, 2017. doi: 10.1109/TGRS.2017.2776321.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9206–9217. IEEE, 2025.

Xianzhi Ma, Jianhui Li, Changhua Pei, and Hao Liu. Geomag: A vision-language model for pixel-level fine-grained remote sensing image parsing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5441–5450, 2025.

Kenneth Marino, Mohammad Rastegar, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2024.

OpenAI. GPT-5.5 system card, April 2026. URL `https://openai.com/index/gpt-5-5-system-card/`.

Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*, 2(3):5, 2024.

Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6381–6388, 2025.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *arXiv preprint arXiv:2501.13925*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *arXiv preprint arXiv:2507.20198*, 2025.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. *arXiv preprint arXiv:2503.19903*, 2025.

Yan Shu, Bin Ren, Zhitong Xiong, Danda Pani Paudel, Luc Van Gool, Begüm Demir, Nicu Sebe, and Paolo Rota. Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. *arXiv preprint arXiv:2506.01667*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Aditya Kumar Singh, Hitesh Kandala, Pratik Prabhanjan Brahma, Zicheng Liu, and Emad Barsoum. Duet-vlm: Dual stage unified efficient token reduction for vlm training and inference. *arXiv preprint arXiv:2602.18846*, 2026.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, Martin Weinmann, Stefan Hinz, Cheng Wang, and Kun Fu. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. doi: 10.1016/j.isprsjprs.2021.12.004.

Lucrezia Tosato, Hichem Boussaid, Flora Weissgerber, Camille Kurtz, Laurent Wendling, and Sylvain Lobry. Segmentation-guided attention for visual question answering from remote sensing images. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2750–2754. IEEE, 2024.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336. IEEE, 2025b.

Fengxiang Wang, Mingshuo Chen, Yueying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Shan Boqi, Long Lan, Yulin Wang, et al. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *Advances in Neural Information Processing Systems*, 38:159185–159218, 2026.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. *Proceedings of the AAAI conference on artificial intelligence*, 38(6):5481–5489, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025d.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. Dynamicvl: Benchmarking multimodal large language models for dynamic city understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19803–19813. IEEE, 2025a.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025b.

Kelu Yao, Nuo Xu, Rong Yang, Yingying Xu, Zhuoyan Gao, Titinunt Kitrungrotsakul, Yi Ren, Pu Zhang, Jin Wang, Ning Wei, et al. Falcon: A remote sensing vision-language foundation model. *arXiv preprint arXiv:2503.11070*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Yaqi Xie, Katia Sycara, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *arXiv preprint arXiv:2505.22654*, 2025a.

Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7839–7849. IEEE, 2024.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025b.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025c.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025d.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Yuanxin Zhao, Mi Zhang, Bingnan Yang, Zhan Zhang, Jujia Kang, and Jianya Gong. Luojiahog: A hierarchy oriented geo-aware image caption dataset for remote sensing image–text retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 222:130–151, 2025.

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7590–7598, 2018.

Yue Zhou, Mengcheng Lan, Xiang Li, Litong Feng, Yiping Ke, Xue Jiang, Qingyun Li, Xue Yang, and Wayne Zhang. Geoground: A unified large vision-language model for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024.

Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. Rsvlm-qa: A benchmark dataset for remote sensing vision language model-based question answering. *arXiv preprint arXiv:2508.07918*, 2025.

APPENDIX CONTENTS

A Additional Evaluation Results and Input Settings 19

A.1 Open-Ended Visual Question Answering 19

A.2 Captioning Results Across Model Families 19

A.3 Single-Image Evaluation Results 20

B Human Annotation and Evaluation 21

B.1 Annotation and Independent Review 21

B.2 Human Answerability Check 21

B.3 Qwen3-8B Judge Calibration 21

B.4 Human Verification 21

B.5 Annotation Challenges and Quality Controls 22

C Details of Our Benchmark 23

C.1 Model Input Resolution Policies 23

C.2 Native Image-Resolution Distribution 23

C.3 Perception and Reasoning Tasks 24

C.4 Task Taxonomy and Answer Formats 24

C.5 Interpreting Multi-Turn Exact Match 25

C.6 Caption Reference Construction and Verification 25

C.7 Single-Image Evaluation Protocol 26

C.8 Captioning Results on the Full Image Set 26

C.9 UAV Data Collection 27

D Visual-Token Pruning Details 27

D.1 Visual-Token Profiling Protocol 27

D.2 Pruning Protocol and Comparison Scope 27

D.3 Measured Inference Efficiency 30

E Supplementary Cross-Dataset Evaluation 30

E.1 Extended Retention Sweeps on External Benchmarks 30

E.2 Extended Retention Sweeps for the Remaining Methods 31

F Case Studies 32

G Data Construction Prompts 37

H Evaluation and Captioning Prompts 49

A Additional Evaluation Results and Input Settings

RSHR-Bench contains 5,329 satellite, aerial, and UAV images with an average native resolution of $8{,}700 \times 8{,}065$ pixels, and its VQA evaluation comprises 1,932 independently human-verified items. The main evaluation compares 34 vision–language models, with GPT-5.5 reaching the highest overall accuracy of only 54.1%; models specialized for remote sensing do not consistently outperform general-purpose MLLMs. The pruning study evaluates 22 methods on RSHR-Bench, XLRS-Bench-lite, and MME-RS across multiple retention budgets. The appendix provides task-specific breakdowns and extended analyses under these same benchmark totals and evaluation settings.

A.1 Open-Ended Visual Question Answering

Open-Ended VQA Evaluation on RSHR-Bench. To evaluate whether MLLMs can answer questions without relying on candidate options, we conduct an open-ended VQA evaluation using free-form responses. We use Qwen3-8B at decoding temperature $T = 0$ as the automatic scorer to compare each response with its human-verified reference on a 1–100 scale. Table 5 reports the mean Qwen3-8B score within each of nine perception, five reasoning, and three multi-turn categories; the threshold of 80 is not applied to these category columns. P-Mean and R-Mean are arithmetic means over the corresponding category-level scores. MTEM@80 counts a group as successful only when all three turns have valid stored scores of at least 80, computes this rate separately for MAD, MTFP, and MOSJ, and equally averages the three task-level rates. VILA-HD achieves the highest P-Mean at 44.0, while GPT-4o-mini achieves the highest R-Mean at 63.1 and the highest MTEM@80 at 42.1. The Qwen3-8B scoring prompt and its temperature-sensitivity analysis are reported in Section B.3.

Table 5: Open-ended results on RSHR-Bench scored by Qwen3-8B at $T = 0$. All perception, reasoning, and multi-turn task columns report mean scores on a 1–100 scale, without thresholding. P-Mean and R-Mean are arithmetic means over the nine perception and five reasoning category scores, respectively. MTEM@80 counts groups with valid numeric scores for all three turns, computes the all-turn success rate at threshold 80 within each of MAD, MTFP, and MOSJ, and then macro-averages the three rates. Missing, nonnumeric, or out-of-range stored scores are invalid and are not imputed.

Model	Perception									P-Mean	Reasoning					R-Mean	Multi-turn			MTEM@80
	COL	SHP	DET	OC	REL	OGD	RG	OCN	RCN		AD	FP	MRJC	MRJCS	OSJ		MAD	MTFP	MOSJ	
Remote sensing VLMs																				
EarthDial (Soni et al., 2025)	34.3	35.0	42.4	50.8	52.7	28.6	25.5	24.0	45.3	37.6	55.4	62.6	20.0	58.7	57.9	50.9	65.2	65.7	65.9	11.0
GeoChat (Kuckreja et al., 2024)	35.7	40.5	41.1	47.0	54.9	24.6	22.6	35.5	44.2	38.5	49.8	39.9	24.5	24.0	37.6	35.2	53.9	58.3	55.3	2.4
VHM (Pang et al., 2025)	49.3	42.9	39.6	63.0	48.2	24.4	23.5	29.1	54.3	41.6	60.6	62.7	38.8	35.2	51.4	49.7	60.9	63.1	71.3	15.1
Open source VLMs																				
InternVL2.5-8B (Chen et al., 2023)	53.2	37.6	31.6	63.6	59.1	26.9	30.1	46.1	41.8	43.3	74.3	74.7	35.0	48.5	61.8	58.9	67.6	71.3	72.4	20.1
InternVL3.5-8B (Wang et al., 2025c)	37.1	44.6	36.8	68.1	54.2	26.9	31.7	41.1	34.3	41.6	66.0	67.6	68.0	43.0	58.6	60.6	70.0	71.1	66.7	17.9
MiniCPM-V-2.6 (Yao et al., 2024)	49.3	38.6	37.8	60.8	58.9	23.2	29.0	35.0	33.0	40.6	71.3	70.9	51.8	43.1	60.2	59.5	63.5	75.1	74.1	25.7
Phi-3.5-Vision (Abdin et al., 2024)	40.4	32.2	39.8	61.1	56.5	28.9	26.6	36.3	32.3	39.3	67.5	67.1	43.6	40.1	55.2	54.7	64.5	70.8	75.1	21.2
Qwen2.5-VL-7B (Bai et al., 2025b)	35.1	34.6	30.0	55.5	57.9	26.0	30.2	33.2	28.3	36.8	58.8	37.4	67.1	33.5	51.4	49.6	55.9	49.6	54.2	1.6
DeepSeek-VL (Lu et al., 2024)	54.1	35.2	34.2	63.5	60.2	25.4	24.5	38.5	33.9	41.1	70.4	60.1	46.8	35.2	55.1	53.5	67.7	68.4	60.2	17.9
VILA-HD (Shi et al., 2025)	58.4	40.7	33.4	56.8	67.8	27.7	29.6	43.9	37.9	44.0	62.3	60.1	62.4	32.5	56.3	54.7	71.1	74.1	78.6	28.3
Closed source VLMs																				
GPT-5 (Singh et al., 2025)	25.6	26.6	37.6	49.0	52.4	20.9	29.9	34.8	27.4	33.8	70.7	62.5	48.2	44.6	60.1	57.2	75.0	76.8	59.8	40.1
GPT-4o (Hurst et al., 2024)	55.6	35.0	33.8	56.5	46.6	29.9	35.4	35.1	40.8	41.0	74.4	41.0	28.6	55.8	53.8	50.7	54.5	74.9	76.0	30.4
GPT-4o-mini (Hurst et al., 2024)	29.3	37.6	31.8	62.2	45.9	28.4	23.0	30.1	24.3	34.7	70.5	72.2	65.0	43.9	63.9	63.1	78.0	74.9	82.3	42.1
Gemini-2.5-Pro (Comanici et al., 2025)	34.4	25.5	23.3	38.0	51.2	18.4	17.3	25.4	16.6	27.8	42.7	30.8	60.4	29.0	41.7	40.9	35.0	20.3	41.9	0.8

A.2 Captioning Results Across Model Families

Cross-Family Image-Captioning Comparison. Following recent remote-sensing evaluations (Wang et al., 2025b; Li et al., 2024c), we compare image captions against the references using BLEU-1–4, METEOR, and ROUGE-L in Table 6. This table uses the shared 1,000-image subset evaluated across open- and closed-source model families. These metrics measure unigram overlap, longer phrase overlap, and sequence matching, capturing different aspects of reference agreement. GPT-4o-mini obtains 49.7 BLEU-1, 4.8 BLEU-4, and 33.7 METEOR. Among the open-source models, Phi-3.5-Vision leads BLEU-1 at 36.5, whereas InternVL3.5-8B leads BLEU-2–4, METEOR, and ROUGE-L. Even the strongest BLEU-4 score is low relative to BLEU-1, reflecting limited exact agreement on longer phrases. These results measure similarity to the generated references and do not independently establish factual correctness. The full-set results in Table 12 use all 3,913 captioning images; because the two tables evaluate different image sets, their absolute scores are not directly comparable.

Table 6: Captioning results on the shared 1,000-image subset using text-generation metrics.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
InternVL2.5-8B (Chen et al., 2023)	11.0	4.7	1.7	0.6	17.0	17.3
InternVL3.5-8B (Wang et al., 2025c)	33.7	14.4	6.1	2.5	26.4	23.5
MiniCPM-V-2.6 (Yao et al., 2024)	2.9	1.2	0.5	0.2	11.8	13.6
Phi-3.5-Vision (Abdin et al., 2024)	36.5	14.0	5.6	2.1	25.1	22.3
Qwen2.5-VL-7B (Bai et al., 2025b)	0.1	0.0	0.0	0.0	3.8	6.1
DeepSeek-VL (Lu et al., 2024)	2.7	1.2	0.5	0.2	11.8	14.6
VILA-HD (Shi et al., 2025)	28.0	10.0	4.2	1.7	21.6	21.8
GPT-5 (Singh et al., 2025)	43.3	15.3	5.0	1.6	30.1	20.6
GPT-4o-mini (Hurst et al., 2024)	49.7	21.8	10.0	4.8	33.7	25.8
GPT-4o (Hurst et al., 2024)	43.8	18.1	7.9	3.6	29.2	23.1
Gemini-2.5-Pro (Comanici et al., 2025)	9.1	3.0	0.9	0.3	7.5	7.1

A.3 Single-Image Evaluation Results

Following the single-image protocol described in Section C.7, we evaluate fifteen VLMs spanning remote-sensing, general-purpose open-source, and closed-source model families on 50 images divided into five native image-scale groups. All models receive the same ten-item suite for each image, covering perception, reasoning, and local and global captioning. Table 7 reports the aggregate score within each image-scale group, and the final column gives the arithmetic mean across the 4K, 6K, 8K, 100M, and 200M groups with equal weight. GLM-4V-9B achieves the highest score in every group and the highest five-group mean of 37.4. MiniCPM-V-2.6 ranks second overall at 22.0 and reaches 33.0 on the 8K group, but its score falls to 3.0 on the 200M group. InternVL2.5-8B and InternVL3.5-8B obtain similar means of 21.0 and 20.8, respectively. Among the evaluated closed-source models, GPT-5 performs best with a mean of 17.4. The remote-sensing models remain below the strongest general-purpose models, with VHM, EarthDial, and GeoChat reaching means of 10.4, 9.5, and 3.8. More broadly, performance at 4K–8K does not consistently transfer to the 100M and 200M groups, indicating that preserving useful fine-grained evidence remains difficult as total image area increases.

Table 7: Single-image success rates (%) across five native image-scale groups. Each of the eight question-answer items and two captioning items per image has equal weight and is successful when its Qwen3-8B score is at least 80. The denominator is the number of valid item scores for that model and group. The 4K, 6K, and 8K groups are defined by long-side resolution, whereas 100M and 200M are defined by total pixel count. Mean gives equal weight to the five group scores.

Model	4K	6K	8K	100M	200M	Mean
Remote sensing VLMs						
EarthDial (Soni et al., 2025)	8.9	11.1	11.1	7.8	8.8	9.5
GeoChat (Kuckreja et al., 2024)	5.6	1.1	3.3	5.6	3.3	3.8
VHM (Pang et al., 2025)	8.9	12.2	10.0	15.6	5.5	10.4
General-purpose open-source VLMs						
InternVL2.5-8B (Chen et al., 2023)	16.0	24.0	24.0	19.0	22.0	21.0
InternVL3.5-8B (Wang et al., 2025c)	19.0	23.0	22.0	23.0	17.0	20.8
MiniCPM-V-2.6 (Yao et al., 2024)	25.0	27.0	33.0	22.0	3.0	22.0
Phi-3.5-Vision (Abdin et al., 2024)	14.0	17.0	14.0	12.0	14.0	14.2
Qwen2.5-VL-7B (Bai et al., 2025b)	17.8	13.3	16.7	11.1	10.0	13.8
DeepSeek-VL (Lu et al., 2024)	24.0	18.0	25.0	17.0	17.0	20.2
VILA-HD (Shi et al., 2025)	16.0	14.0	17.0	14.0	14.0	15.0
GLM-4V-9B (Glm et al., 2024)	35.0	29.0	44.0	39.0	40.0	37.4
Closed-source VLMs						
GPT-5 (Singh et al., 2025)	13.0	17.0	21.0	20.0	16.0	17.4
GPT-4o-mini (Hurst et al., 2024)	10.0	14.0	20.0	11.0	8.0	12.6
GPT-4o (Hurst et al., 2024)	8.0	12.0	13.0	8.0	8.0	9.8
Gemini-2.5-Pro (Comanici et al., 2025)	7.0	9.0	11.0	4.0	4.0	7.0

B Human Annotation and Evaluation

B.1 Annotation and Independent Review

Annotation and Independent Review. To separate question authoring from verification, we assign annotation and review responsibilities to different groups. For the single-image annotation subset described here, three annotators, all of whom held at least a bachelor's degree, spent approximately 50 person-hours drawing target boxes and preparing question–answer annotations. The single-image setting includes five image-scale groups: 4K, 6K, and 8K denote long-side scales, whereas 100M and 200M denote image pixel counts. Task-dependent sampling provides more detailed imagery for questions involving shape, boundaries, or small targets. Within an image, annotators select identifiable objects with different relative scales, enabling evaluation of local visual details at varying levels of difficulty. Reviewers check the source images, target boxes, questions, options, and reference answers, revising or discarding ambiguous or unsupported items. Every retained item is independently verified for correctness, visual grounding, and unambiguous wording.

B.2 Human Answerability Check

Human Answerability across 17 Task Types. To assess answerability, we sample 15 image–query instances from each of 17 task types, giving 255 instances across nine perception categories, five reasoning categories, and three multi-turn task categories. Raters receive the task instructions, images, and questions without access to reference answers or model predictions. They are distinct from the original annotators and reviewers. Accuracy is computed as the fraction of correctly answered instances within each task type, with 237 correct answers out of 255 instances overall. Color, shape, orientation, spatial relations, regional grounding, object counting, both multi-region joint contrast tasks, and multi-turn object-state judgement each reach 15/15. Anomaly interpretation reaches 10/15, and future prediction reaches 12/15, indicating that these categories remain challenging even for human raters. Because each category contains only 15 assessed instances, these observations should not be treated as precise estimates of human performance over the entire benchmark.

B.3 Qwen3-8B Judge Calibration

Qwen3-8B Judge Calibration at Two Temperatures. To examine scoring stability, we compare Qwen3-8B with three human raters on 100 open-ended responses in Table 8, using the same scoring prompt at both temperatures. At $T = 0$, the Qwen3-8B mean is 53.00, within the range of the three human means (51.5–58.0). Increasing T to 0.6 lowers the Qwen3-8B mean by 2.25 points to 50.75. This difference indicates modest sensitivity to the decoding temperature. All reported open-ended benchmark results use Qwen3-8B at $T = 0$; the $T = 0.6$ setting is included only as a sensitivity check. The table reports aggregate means, so it does not establish per-response agreement, inter-rater reliability, or agreement on the correctness threshold of 80. Those quantities require paired response-level scores, which are not summarized here.

Table 8: Mean scores for 100 open-ended responses under a fixed Qwen3-8B scoring prompt. T is the Qwen3-8B decoding temperature. Human means are repeated for reference and do not depend on T.

Setting	Qwen3-8B	Human 1	Human 2	Human 3
$T = 0$	53.00	51.5	58.0	53.5
$T = 0.6$	50.75	51.5	58.0	53.5

B.4 Human Verification

Independent Human Verification. Question authors and reviewers are assigned to different groups. Reviewers inspect the source image, target region, question wording, answer options, and reference answer for factual correctness, visible support, clarity, and a unique best answer. Items with ambiguous targets, unsupported answers, multiple plausible options, or cross-turn inconsistencies are

revised and independently rechecked or discarded. All final benchmark questions and reference answers are retained only after this verification process.

B.5 Annotation Challenges and Quality Controls

Controls for Common Annotation Issues. To address these four sources of difficulty, we combine corpus screening with independent human review of the images, questions, options, and reference answers. First, clusters of similar objects can make targets hard to distinguish, so screening emphasizes varied scenes and identifiable instances while retaining their surrounding context. Second, category names or descriptive words may unintentionally suggest an answer. Human reviewers revise or remove questions whose wording reveals the correct option without requiring the relevant visual evidence. Third, questions about causes or plausible future states can admit multiple interpretations from a static image. We therefore retain only items for which reviewers identify one option as best supported by the visible evidence. Performance on these tasks measures agreement with the verified reference interpretation and does not validate causal inference or forecasting accuracy. Reviewers select examples with discriminative visual cues and remove questions whose alternatives cannot be separated reliably. Finally, small boxed targets and crowded counting regions can produce incorrect candidate answers, requiring direct inspection and correction. These checks reduce specific annotation errors while retaining uncertainty where the available image does not support a unique interpretation.

C Details of Our Benchmark

C.1 Model Input Resolution Policies

Comparison of Model Input-Resolution Policies. Table 9 summarizes model families and evaluation configurations across four groups. It distinguishes supported input envelopes, encoder views, processing canvases, crop or tile budgets, configurable pixel budgets, and API acceptance limits. Square-equivalent dimensions are explanatory conversions of pixel budgets, not independent per-edge limits. Crop and tile sizes describe local processing units, not full source-image bounds.

Table 9: Image input and processing specifications by model family. Processing sizes are not universal native-image resolution limits.

Model / configuration	Concrete input	Processing policy / scope
Remote-sensing models		
EarthDial	448×448 per tile, up to 40 tiles	Dynamic tiles plus a global thumbnail.
GeoChat family	504×504 encoder view	Fixed visual-encoder input.
GeoLLaVA-8K	Up to $8K \times 8K$ supported input	Tiling with visual-token selection.
VHM	336×336 encoder view	Fixed CLIP-L/14 input.
EarthMind-4B	448×448 per tile, up to 12 tiles	Dynamic tiles plus a global thumbnail.
LHRS-Bot-Nova	336×336 encoder view	Fixed SigLIP-L/14 input.
MF-RSVLM-7B-SFT	672×672 processing canvas	Global and local 336×336 views.
RS-LLaVA-7B	336×336 encoder view	Fixed CLIP-336 input.
RSCoVLM-7B-2512	1.02-megapixel processor budget	Dynamic resizing with Zoom-in Chain.
RSThinker	11.76-megapixel processor budget	Dynamic resizing to the pixel budget.
RSUniVLM	384×384 encoder view	Per-view SigLIP input with multi-image support.
SkySenseGPT-7B	504×504 encoder view	Resize before visual encoding.
TEOChat	224×224 center crop	Short-side resize followed by center crop.
GeoPixel-7B	Up to 4K HD supported input	Variable-aspect-ratio processing.
FALCON-8B	384×384 per crop, up to 16 crops	Multi-crop encoding with register tokens.
General-purpose open models		
InternVL2.5/3/3.5 family	448×448 per tile, up to 12 tiles	Dynamic tiles plus a global thumbnail.
MiniCPM-V-2.6	Up to 1.8 MP supported input	Variable-aspect-ratio processing.
Phi-3.5-Vision	336×336 per crop, 16 crops	Multi-crop input plus a global thumbnail.
Qwen2.5-VL-7B	12.85-megapixel processor budget	Dynamic resizing to the pixel budget.
DeepSeek-VL	1024×1024 reference input	Fixed reference resolution.
LLaVA-OneVision	Up to 2304×2304 grid canvas	Any-resolution crop grid.
Qwen3-VL family	16.78-megapixel processor budget	Dynamic resizing to the pixel budget.
Qwen3.5-9B	16.78-megapixel processor budget	Dynamic resizing to the pixel budget.
VILA-HD	1512×1512 or 3780×3780	Separate 1.5K and 4K variants.
Closed-source models		
Claude-3.5-Sonnet / Claude-3.7-Sonnet	8000×8000 API acceptance	Standard processing uses smaller bounds.
GPT-5	Long edge $\leq$ 6000	Patch-based image processing.
GPT-5.5	API file-size limits	Aspect-ratio-preserving tiled processing.
GPT-4o / GPT-4o-mini	API file-size limits	Aspect-ratio-preserving tiled processing.
Gemini-2.5-Pro	API file-size limits	File-size and image-count limits.
Additional model families discussed in related work		
LHRS-Bot	224×224 encoder view	Fixed CLIP ViT-L/14 input.
EarthGPT	Multiscale visual features	Combined transformer and convolutional features.
LLaVA-NeXT	Multiple aspect-ratio canvases	Aspect-ratio-dependent processing.
Monkey	Up to 1344×896 supported input	Patch-based high-resolution encoding.
LLaVA-UHD	672×1008 paper maximum	Adaptive image slicing.
Mini-Gemini	Standard and HD branches	Dual-resolution processing.
Cambrian	Multiple square encoder views	Encoder-specific input resolutions.
SliME	2016×2016 grid canvas	Tiled grid processing.

C.2 Native Image-Resolution Distribution

Native Image-Resolution Distribution. The complete RSHR-Bench source corpus contains 5,329 images and reaches a maximum native resolution of 806.5 megapixels. The VQA evaluation uses 1,377 distinct image files from this corpus, each counted once regardless of how many questions refer to it. Table 10 summarizes this VQA image pool using two complementary partitions: native long-side length and total image area. In total, 835 images (60.6%) have a long side of at least 10,000 pixels, while 835 images (60.6%) contain at least 100 megapixels. The two panels describe the same images and therefore should not be added together.

Table 10: Native resolutions of 1,377 distinct RSHR-Bench VQA images, each counted once. Left: long-side length $L = \max(W, H)$. Right: image area $A = WH/10^6$ in megapixels (MP).

Long-side interval (pixels)	Images
$4{,}000 \leq L < 6{,}000$	288
$6{,}000 \leq L < 8{,}000$	241
$8{,}000 \leq L < 10{,}000$	13
$L \geq 10{,}000$	835
Total	**1,377**

Image-area interval (MP)	Images
$A < 25$	304
$25 \leq A < 100$	238
$100 \leq A < 200$	809
$A \geq 200$	26
Total	**1,377**

C.3 Perception and Reasoning Tasks

Perception and Reasoning Tasks. Perception covers local attributes such as color, shape or margin, and orientation, classification of land-use categories and aerial targets, spatial relationships, object and regional grounding, and object and regional counting. These tasks require identifying visual attributes, locating referents, understanding spatial relations, or counting relevant instances. Object counting covers the entire image, whereas regional counting supplies a reference bounding box and counts only qualifying instances within that box. Reasoning covers anomaly detection and interpretation, future prediction, multi-region joint contrast, and object state judgement. Anomaly interpretation relates visible patterns to plausible explanations such as flooding or construction, while state judgement asks whether context supports a static or dynamic interpretation. Future prediction includes 50 single-image multi-turn groups with three questions each and 50 temporal pairs depicting the same location across years. Regional comparison additionally tests information integration across regions or images. A high score in one task group does not establish proficiency in the other.

C.4 Task Taxonomy and Answer Formats

Task Taxonomy and Answer Formats. Table 11 summarizes these fields for all task families. For VQA, the counts denote evaluated question turns after filtering and review, not candidate annotations. The perception and reasoning rows sum to 1,932 turns, whereas the captioning count denotes 3,913 images shared by the global and regional caption settings and is not included in the VQA total. The taxonomy preserves distinctions between related tasks: object and regional grounding differ in target scope, while future prediction is evaluated through single-image multi-turn dialogs and multi-image single-turn questions. The question-generator column records the model used to assist initial drafting. All retained references are human-reviewed, but this review does not alter their drafting provenance. VQA tasks include both multiple-choice and open-ended formats, whereas captioning uses free-form text. Scores should therefore be interpreted with respect to differences in answer space, interaction structure, item count, and evaluation protocol.

Abbreviations. COL, SHP, DET, OC, REL, OGD, RG, OCN, and RCN denote Color Detection, Shape/Margin Recognition, Orientation Detection, Object Classification, Object Spatial Relationship, Object Grounding, Regional Grounding, Object Counting, and Regional Counting, respectively. AD denotes single-turn Anomaly Detection, while MAD denotes its multi-turn variant. FP is the shortened main-table label for FP2I, the two-image single-turn Future Prediction task. MTFP denotes Multi-Turn Future Prediction. MRJC denotes Multi-Region Joint Contrast across multiple images, whereas MRJCS denotes the single-image, multi-box variant. Both variants use four options (A/B/C/D), with option D reserved for "Cannot compare". OSJ and MOSJ denote single-turn and multi-turn Object State Judgement, respectively. OCAP and RCAP denote Overall Caption and Regional Caption, and MTEM denotes Multi-Turn Exact Match.

Table 11: Overview of task taxonomy, annotation sources, and QA formats in our benchmark. L1/L2/L3 denote coarse-, mid-, and fine-grained task levels, respectively, and *Attr.* lists the attribute codes for each L3 task. VQA counts refer to evaluated question turns: 1,370 perception, 220 single-turn reasoning, and 342 turns in 114 three-turn groups, totaling 1,932. Reasoning rows combine their single- and multi-turn variants. The captioning count is 3,913 images shared by the global and regional caption settings, not additional VQA questions.

L1-task	L2-task	L3-task	Attr.	Reference review	Question generator	Question type	Answer type	#Eval.
Perception	Local Attributes	Color Detection	COL	Human reviewed	Qwen2.5-VL-7B	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	200
		Shape/Margin Recognition	SHP	Human reviewed	Qwen2.5-VL-7B	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	100
		Orientation Detection	DET	Human reviewed	GPT-5	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	100
	Overall Attributes	Object Classification	OC	Human reviewed	Qwen2.5-VL-7B	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	200
		Object Spatial Relationship	REL	Human reviewed	Qwen2.5-VL-7B	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	200
	Visual Grounding	Object Grounding	OGD	Human reviewed	GPT-5	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	200
		Regional Grounding	RG	Human reviewed	GPT-5	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	70
	Counting	Object Counting	OCN	Human reviewed	GPT-5	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	200
		Regional Counting	RCN	Human reviewed	Qwen2.5-VL-7B	Single-image single-turn	Multiple Choice (A/B/C/D) Open-ended text	100
Reasoning	Logic-grounded Reasoning	Multi-region Joint Contrast	MRJC MRJCS	Human reviewed	GPT-5	Multi-image single-turn Single-image multi-box	Multiple Choice (A/B/C/D) Open-ended text	70
		Object State Judgement	OSJ MOSJ	Human reviewed	GPT-5	Single-image single-turn Single-image multi-turn	Multiple Choice (A/B/C/D) Open-ended text	182
	Knowledge-grounded Reasoning	Anomaly Detection	AD MAD	Human reviewed	GPT-5	Single-image single-turn Single-image multi-turn	Multiple Choice (A/B/C/D) Open-ended text	110
		Future Prediction	FP2I MTFP	Human reviewed	GPT-5	Single-image multi-turn Multi-image single-turn	Multiple Choice (A/B/C/D) Multiple Choice (A/B/C) Open-ended text	200
Image Caption	Overall Caption Regional Caption		OCAP RCAP	Human reviewed Human reviewed	GPT-5 GPT-5	Single-image single-turn Single-image single-turn	Open-ended text Open-ended text	3913

C.5 Interpreting Multi-Turn Exact Match

Interpretation of Multi-Turn Exact Match. Let $\mathcal{C} = \{\mathrm{MAD}, \mathrm{MTFP}, \mathrm{MOSJ}\}$ denote the three multi-turn task categories, whose full sets contain 20, 50, and 44 three-turn groups, respectively. Let $\mathcal{D}_c$ contain the evaluated groups in category c; for the stored-score open-ended results, it includes only groups with valid numeric scores in $[1, 100]$ for all three turns. Missing, nonnumeric, and out-of-range scores are invalid, and no replacement score is assigned. Let $s_{c,d,i}$ be the score for turn i in group d of category c. For multiple-choice evaluation, $s_{c,d,i} \in \{0, 1\}$ equals one when the selected option is correct. For open-ended evaluation, $s_{c,d,i} \in [1, 100]$ is the Qwen3-8B score. Given a threshold τ on the corresponding score scale, we compute the all-turn success rate within each task and macro-average the three rates:

$$\mathrm{MTEM@}\tau = \frac{100}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{D}_c|} \sum_{d \in \mathcal{D}_c} \mathbf{1}\left[\min_{1 \le i \le 3} s_{c,d,i} \ge \tau\right]. \tag{1}$$

Multiple-choice evaluation uses $\tau = 1$ and is reported as MTEM@1, while the recorded open-ended protocol uses $\tau = 80$ and is reported as MTEM@80. Equation 1 applies the threshold to the lowest turn score before averaging across groups within each task and then gives equal weight to the three task categories. It therefore distinguishes complete success on a group from a high average turn score without allowing the larger MTFP subset to dominate the aggregate. For example, open-ended scores $[95, 88, 80]$ pass MTEM@80 but fail MTEM@90. By contrast, $[95, 95, 79]$ fails MTEM@80 even though its mean score is approximately 89.7, because one turn falls below the threshold. For multiple-choice evaluation, $[1, 1, 1]$ passes MTEM@1, whereas $[1, 0, 1]$ fails. Both metrics use the same task-macro aggregation. MTEM@1 requires the correct option on every turn. MTEM@80 requires every Qwen3-8B score to reach 80 and does not use literal string matching. Each group contributes one binary outcome within its task, and each task contributes equally to the final metric, so a strong answer on one turn cannot compensate for an error on another.

C.6 Caption Reference Construction and Verification

Caption Reference Construction and Verification. To evaluate coverage at both spatial scales, we construct references for 3,913 images with long-side resolution of at least 4K, predominantly around 100 megapixels, spanning ports, airports, forests, and urban and rural scenes. Following remote-sensing caption evaluation (Lobry et al., 2020; Wang et al., 2025b; Luo et al., 2025), we develop a prompt template for individual images and apply GPT-5 (Singh et al., 2025) to generate references for the collection. Each reference follows the final captioning prompt in Section H.3: it

contains a global summary and descriptions of the upper-left, upper-right, lower-left, and lower-right quadrants, covering visible objects, attributes, spatial relations, and scene function. The instructions discourage vague quantifiers and request concrete visual details. Human reviewers inspect the generated references for grounding, coverage, fluency, redundancy, and image-grounded consistency and revise unsupported content before evaluation. We additionally draw a random set of 100 reviewed references for a separate audit of the annotations and evaluation procedure. For evaluation, the global summary and four quadrant descriptions are treated together as one structured caption for each image and compared with the corresponding complete reference using BLEU, METEOR, and ROUGE-L; the quadrants are not counted as separate evaluation samples. The additional audit complements the item-level human review and does not replace it.

C.7 Single-Image Evaluation Protocol

Single-Image Evaluation Protocol. To hold the visual source fixed across questions, we select 50 images spanning resolutions from at least 4K to approximately 2×10^8 pixels. Each image receives ten questions, yielding 500 question–answer pairs fully written and checked by annotators and reviewers, all of whom held at least a bachelor's degree. The questions cover four capability groups: image- and region-level description; basic attributes such as color, shape, orientation, and object class; relations, grounding, and counting; and higher-order state judgement, anomaly interpretation, and future prediction. Global and regional captions test whether a model can relate an overall scene description to localized visual details, while the remaining questions probe specific targets. All ten responses per image—eight question answers, one regional caption, and one overall caption—are scored against their corresponding human-verified references by Qwen3-8B at $T = 0$ on the same 1–100 scale. A valid item is counted as successful when its score is at least 80. For model m and image-scale group g, let $\mathcal{V}_{m,g}$ denote the valid item scores and let $s_{m,j}$ be the score for item j. The reported group score is

$$\mathrm{Score}_{m,g} = \frac{100}{|\mathcal{V}_{m,g}|} \sum_{j \in \mathcal{V}_{m,g}} \mathbf{1}[s_{m,j} \geq 80]. \tag{2}$$

Thus, every valid question-answer or caption item receives equal weight, while missing or invalid scores are excluded from the denominator. The reported Mean is $\frac{1}{5}\sum_g \mathrm{Score}_{m,g}$, so the five image-scale groups receive equal weight regardless of their numbers of valid items. Because these groups contain different images, the protocol measures performance across image groups and is not a controlled resizing experiment on identical inputs.

C.8 Captioning Results on the Full Image Set

Comparison on the Full Caption Set. To extend evaluation beyond the shared 1,000-image subset used for the cross-family comparison, Table 12 compares open-source models on all 3,913 images with GPT-5 references. The subset and full-set tables use the same metric definitions but different image populations, so they should not be treated as repeated evaluations on identical data. In particular, the InternVL2.5-8B values of 11.0 BLEU-1 and 0.6 BLEU-4 on the shared subset and 33.3 BLEU-1 and 2.8 BLEU-4 on the full set reflect these distinct evaluation sets rather than two measurements of the same sample collection. We compute BLEU-1–4, METEOR, and ROUGE-L, with higher values indicating greater reference agreement. The ranking changes across metrics because matching individual words and matching longer phrases capture different aspects of the generated descriptions.

Table 12: Captioning results for open-source models on the full 3,913-image set.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
InternVL2.5-8B (Chen et al., 2023)	33.3	14.4	6.2	2.8	26.2	24.1
InternVL3.5-8B (Wang et al., 2025c)	33.4	14.3	6.0	2.4	26.2	23.4
DeepSeek-VL (Lu et al., 2024)	2.7	1.2	0.5	0.2	11.8	14.6
GLM-4V-9B (Glm et al., 2024)	37.3	11.5	3.5	1.1	23.8	19.0
MiniCPM-V-2.6 (Yao et al., 2024)	2.9	1.2	0.5	0.2	11.7	13.6
Phi-3.5-Vision (Abdin et al., 2024)	36.5	14.0	5.6	2.1	25.1	22.3
VILA-HD (Shi et al., 2025)	28.1	10.0	4.2	1.7	21.6	21.8

C.9 UAV Data Collection

UAV Data Collection. To complement public remote-sensing imagery with repeated observations of the same scene, we collect data in a public park and a nearby drone operation site. Each sequence is recorded during a continuous flight at heights between 120 m and 150 m, following a smooth trajectory that observes the area over an extended period. The camera captures multiple headings and upward- and downward-looking views, producing a continuous stream with changing viewpoints. Individual frames contain approximately 10^8 pixels, retaining small objects within a broad scene context. We use these images for captioning and visual question answering, where repeated views provide contextual information for temporal and multi-view reasoning. Scene coverage is limited to the recorded park and operation-site environments, so this collection should not be interpreted as covering all UAV deployment scenarios.

D Visual-Token Pruning Details

D.1 Visual-Token Profiling Protocol

Cross-Model Visual-Token Profiling. To support Figure 3, we obtain visual-token counts from prediction records, documented adapter configurations, or processor replay on source-image dimensions without generating new answers. The profiling set contains 1,590 single-turn questions; multi-turn questions are excluded. We count visual positions entering the language model after encoder-side pooling, merging, projection, and context truncation, while excluding text tokens. Multi-image inputs follow each model's actual preprocessing route. Table 13 reports 22 independently profiled configurations and distinguishes measured runs, processor replays, deterministic adapter lengths, and archival profiles, with the last shown as hollow markers. Because input policies differ across models, these results characterize visual-token length rather than controlled latency or the causal effect of token count.

Table 13: Visual LLM positions per single-turn question and task-macro accuracy for the 22 model configurations with independent token profiles. Token statistics use 1,590 profiling questions; macro accuracy equally weights the 14 single-turn task scores. R: prediction records; P: processor/adapter replay; D: deterministic adapter configuration; A: archival processor profile without an exact matched multiple-choice run manifest (hollow markers in the figure).

Model	Mean	Median	P95	Macro acc. (%)	Source
Remote sensing					
EarthMind-4B	2,358.1	2,560.0	2,816.0	35.64	P
GeoChat-7B	1,296.0	1,296.0	1,296.0	33.11	D
LHRS-Bot-Nova	272.0	272.0	272.0	33.97	D
MF-RSVLM-7B-SFT	576.0	576.0	576.0	30.67	D
RS-LLaVA-7B	256.0	256.0	256.0	34.70	D
RSCoVLM-7B-2512	1,359.2	1,296.0	1,296.0	39.83	P
RSThinker	7,804.7	7,396.0	7,476.0	34.96	P
RSUniVLM	771.3	730.0	730.0	29.54	D
SkySenseGPT-7B	1,296.0	1,296.0	1,296.0	32.73	D
TEOChat	270.5	256.0	256.0	33.01	D
GeoPixel-7B	4,055.4	4,081.0	4,081.0	29.91	R
FALCON-8B	993.7	1,088.0	1,088.0	32.39	R
General open					
InternVL2.5-8B	2,578.4	2,560.0	3,328.0	24.39	A
MiniCPM-V-2.6	625.1	640.0	640.0	27.91	A
Phi-3.5-Vision	749.7	757.0	757.0	24.42	A
Qwen2.5-VL-7B	16,966.5	16,224.0	16,384.0	23.99	A
LLaVA-OneVision-Qwen2-7B-OV	24,603.6	27,135.0	27,135.0	37.65	P
Qwen3-VL-32B-Instruct	16,955.6	16,384.0	16,384.0	38.83	P
Qwen3-VL-4B-Instruct	16,955.6	16,384.0	16,384.0	43.76	P
Qwen3-VL-8B	16,955.6	16,384.0	16,384.0	42.83	P
Qwen3-VL-8B-Thinking	16,955.6	16,384.0	16,384.0	27.05	P
Qwen3.5-9B	16,955.6	16,384.0	16,384.0	26.60	P

D.2 Pruning Protocol and Comparison Scope

Evaluated Implementations. The pruning comparison reproduces each algorithm from its original paper and official codebase within a common evaluation interface. In total, the study evaluates 22 pruning methods using the same Pre-LLM, Intra-LLM, and Hybrid categories defined in

the main paper. All configurations use the same Qwen3-VL-8B model, BF16 model computation, and FP32 projection onto option-token rows. Compatibility wrappers standardize tensor interfaces, retention control, and answer scoring while preserving each method's token-importance rule and pruning schedule. No fine-tuning or additional method-specific training is performed. *Pre-LLM pruning* includes EViT and QuietPrune, which reduce tokens in the vision encoder, and VisionZip, DivPrune, TopV, LLaVA-PruMerge, VisPruner, CDPruner, and HiPrune, which prune or merge tokens before they enter the language model. *Intra-LLM pruning* includes FastV, SparseVLM, TransPrune, FEATHER, FiT-Prune, VTW, PyramidDrop, DART, and PriorTR, which remove visual tokens within decoder layers. *Hybrid pruning* includes IVTP, MustDrop, DUET-VLM, and VScan, which combine token reduction across multiple stages. Random and uniform controls prune within the vision encoder or before the LLM, as specified by their internal/external names. Internal pruning uses zero-based vision block 8. Let $r \in (0, 1]$ denote the target overall nominal retention ratio, with table entries such as 5% representing $r = 0.05$. The joint implementations allocate retention $\sqrt{r}$ to each of their two pruning stages, so the nominal product of the stagewise ratios equals r before integer rounding and minimum-token constraints. PyramidDrop prunes after approximately one-quarter, one-half, and three-quarters of the decoder depth. FiT-Prune rescales a fixed layerwise pruning recipe to the Qwen3-VL depth. The common interface enables consistent evaluation, while differences in pruning stage and intermediate token usage mean that nominal retention alone does not define equal computation.

Budget Accounting. Final retention, visual-token layer usage, and latency measure different resources. The main RSHR-Bench sweep uses nine nominal retention settings: 70%, 50%, 30%, 20%, 10%, 2%, 0.2%, 0.05%, and 0.01%. The cross-benchmark matrix instead uses four settings—5%, 1%, 0.5%, and 0.1%—together with dense references. Let n be the number of evaluated questions and L the number of decoder layers. For question $i \in \{1, \ldots, n\}$, let N_i be the post-merger visual-token count before pruning, let $N_{i,\ell}$ be the visual-token count entering decoder layer $\ell \in \{1, \ldots, L\}$, and let $N_{i,\text{final}}$ be the visual-token count retained after the final pruning operation. We distinguish the mean final-retention ratio $\frac{1}{n}\sum_i N_{i,\text{final}}/N_i$ from the mean layer-token ratio $\frac{1}{n}\sum_i \sum_{\ell=1}^{L} N_{i,\ell}/(LN_i)$. These values average the per-question ratios and are not ratios of aggregate token counts. Methods retaining visual tokens apply integer rounding and a minimum of one token, so nominal 0.01% retention does not imply an achieved 0.01% ratio. VTW instead maps the nominal budget to a withdrawal layer, then removes all visual positions. Its final-retention ratio is zero while its layer-token ratio remains positive. Table 14 illustrates the resulting differences using the recorded full-set RSHR-Bench results. At nominal 5%, PyramidDrop and FiT-Prune have the same final retention as QuietPrune but use substantially more visual-token layers. Consequently, accuracy differences in Tables 3 and 4 cannot by themselves establish an efficiency advantage. The layer-token measure also excludes vision-encoder, text-token, and selection overhead and is not a FLOP count.

Table 14: Budget accounting on all 1,932 RSHR-Bench questions. Percentages are means of per-question ratios. Decoder layer-token values come from the reproduced runs. QuietPrune's values equal its recorded final retention because pruning precedes all LLM layers. VTW's nominal budget controls withdrawal depth and does not directly specify final retention.

Method	Nominal (%)	Final retention (%)	LLM layer-token (%)
QuietPrune	5	5.0546	5.0546
FiT-Prune	5	5.0546	10.8401
PyramidDrop	5	5.0546	38.8936
VTW	5	0	5.5556
QuietPrune	0.01	0.0796	0.0796
FiT-Prune	0.01	0.0796	6.1699
PyramidDrop	0.01	0.0796	26.2448
VTW	0.01	0	2.7778

Evaluation Scope and Scoring. The pruning matrix uses a shared Qwen3-VL-8B backbone and a common answer-compatible subset for XLRS-Bench-lite. Its RSHR-Bench score is the accuracy across all questions, computed from option logits in one decoding step. The score includes multi-turn questions. It does not use MTEM over dialogs and is not restricted to single-turn ques-

tions. The XLRS-Bench-lite runs predict a single option for each of 3,080 questions, whereas its 100-question overall-land-use task permits multiple answers. For Table 4, we therefore exclude that entire task by dataset metadata and evaluate the same remaining 2,980 single-answer questions for every configuration. This criterion concerns the required answer format and is independent of prediction correctness. Three references in the excluded task happen to contain one option and remain excluded under the same rule. We check all 88 compressed configurations and the dense reference for unique source indices, question text, categories, image references, model configuration identifiers, FP32 option scoring, and original aggregate counts. We then recompute accuracy by matching each predicted option to the original single-answer reference. No new predictions or model inference are generated, and the original 3,080-question first-letter scores are not mixed into the published table. The resulting comparison covers single-answer XLRS tasks only. Assessing the excluded task requires a consistent protocol for generating and scoring sets of answers across all methods. MME-RS uses the 150 remote sensing questions in MME-RealWorld-Lite. The input images exclude ROI, detector, and external localization cues.

D.2.1 Random and Uniform Sampling Baselines

Comparison on RSHR-Bench. Table 15 reports random and uniform sampling baselines on a separate ten-budget RSHR-Bench grid: 70%, 50%, 30%, 20%, 10%, 5%, 1%, 0.5%, 0.1%, and 0.01%. This grid differs from the nine-budget method sweep in Table 3. These simple controls remain competitive at moderate retention. Uniform internal sampling reaches 51.09% at 70% retention, exceeding eighteen of the 22 pruning methods and matching MustDrop. At 20% retention, it reaches 48.76% and exceeds seventeen pruning methods. The four controls obtain 35.14%–35.77% at 0.01% retention, where the strongest pruning methods retain a clearer advantage. The strong results at moderate budgets show that a pruning method should be compared with simple sampling under the same retention setting.

Table 15: Random and uniform sampling baselines on the separate ten-budget RSHR-Bench grid. Dense inference is included as a reference.

Method	Type	100%	70%	50%	30%	20%	10%	5%	1%	0.5%	0.1%	0.01%
Dense Qwen3-VL-8B	–	50.72	–	–	–	–	–	–	–	–	–	–
Random external	Pre-LLM	–	50.93	50.62	48.55	47.46	44.72	41.77	38.51	37.68	36.28	35.77
Random internal	Pre-LLM	–	50.98	50.26	49.74	46.69	45.34	42.60	36.34	35.87	35.04	35.14
Uniform external	Pre-LLM	–	50.72	50.05	49.02	48.03	44.41	43.27	38.30	37.53	35.71	35.51
Uniform internal	Pre-LLM	–	51.09	50.67	49.22	48.76	45.08	42.18	37.53	36.90	35.61	35.35

Comparison across Remote-Sensing Benchmarks. Table 16 compares the same controls on XLRS-Bench-lite, RSHR-Bench, and MME-RS. Uniform external sampling is particularly strong on XLRS-Bench-lite, while random external sampling reaches 38.51% on RSHR-Bench at 1% retention and uniform external sampling reaches 32.67% on MME-RS at 0.1% retention. These comparisons show that random or uniform selection can outperform many of the 22 evaluated pruning methods in several settings, even though PyramidDrop remains the strongest method across the full cross-benchmark matrix.

Table 16: Random and uniform sampling baselines across three remote-sensing benchmarks on the four-budget shared grid. Dense inference is included as a reference.

		XLRS-Bench-lite ($n = 2{,}980$)					**RSHR-Bench** ($n = 1{,}932$)					**MME-RS** ($n = 150$)				
Method	**Type**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**	**100%**	**5%**	**1%**	**0.5%**	**0.1%**
Dense Qwen3-VL-8B	–	52.45	–	–	–	–	50.72	–	–	–	–	58.00	–	–	–	–
Random external	Pre-LLM	52.45	43.96	39.46	37.55	35.47	50.72	41.77	38.51	37.68	36.28	58.00	37.33	35.33	32.67	28.67
Random internal	Pre-LLM	52.45	41.91	39.03	36.95	34.43	50.72	42.60	36.34	35.87	35.04	58.00	34.67	28.67	29.33	30.00
Uniform external	Pre-LLM	52.45	45.54	41.54	40.00	33.83	50.72	43.27	38.30	37.53	35.71	58.00	38.67	31.33	31.33	32.67
Uniform internal	Pre-LLM	52.45	43.62	38.39	36.54	33.29	50.72	42.18	37.53	36.90	35.61	58.00	33.33	24.67	24.67	26.67

D.3 Measured Inference Efficiency

Comparison of Measured Inference Efficiency. To compare practical efficiency, Table 17 reports dense inference and reproduced implementations adapted from the official QuietPrune and FastV codebases. All measurements use the fixed `dev64` subset under the same hardware and software configuration. This subset is distinct from the 1,932-question accuracy evaluation. The evaluator processes one question at a time with eager attention for deterministic comparison, BF16 model computation, and FP32 option-row scoring. Timing synchronizes CUDA at the boundaries and measures end-to-end evaluation latency, including image loading, preprocessing, forward computation, option scoring, and prediction serialization. Model loading is excluded, and no separate warm-up iterations are discarded. At the nominal 5% setting, both pruning methods achieve an average final retention of 66.08 visual tokens, compared with 1,307.80 tokens for dense inference. QuietPrune reduces tokens in the vision encoder, whereas FastV computes shallow decoder attention and prunes within the language model. QuietPrune reduces time per sample from 4.3042 to 3.4587 seconds (19.6%), while FastV reaches 3.6488 seconds (15.2%). Compared with the dense throughput of 0.233 samples/s, they achieve 0.289 and 0.274 samples/s, respectively. Peak memory decreases from 22,132 MiB to 20,958 MiB for QuietPrune and 21,045 MiB for FastV. The equal final-token counts therefore do not imply equal layerwise computation: FastV processes the dense visual sequence in its initial decoder layers before pruning. The observed gap between latency reduction and final-token reduction motivates reporting measured time and memory alongside token budgets, while these 64-question measurements do not establish speedups for the full benchmark or other hardware and attention kernels.

Table 17: Measured inference efficiency on the 64-question RSHR-Bench development subset. Time is reported in seconds per question.

Method	Time	Samples/s	Peak MiB	Final tokens
Dense	4.3042	0.233	22132	1307.80
QuietPrune (Gao et al., 2026) (5%)	3.4587	0.289	20958	66.08
FastV (Chen et al., 2024) (5%)	3.6488	0.274	21045	66.08

E Supplementary Cross-Dataset Evaluation

The cross-dataset study covers the same 22 pruning methods summarized in the abstract: five methods are reported in Table 18, and the remaining seventeen are reported in Tables 19 and 20. All experiments follow a single answer-candidate protocol in which each prediction is evaluated using the original candidate set provided for its question. XLRS-Bench-lite uses the same 2,980 single-answer questions for every method after excluding the 100 multi-select questions, while MME-RS retains all available answer options. Across the extended sweeps, accuracy generally declines as retention becomes more aggressive, consistent with the main-paper finding that substantial compression remains difficult for current pruning methods.

E.1 Extended Retention Sweeps on External Benchmarks

Cross-Dataset Rankings across Retention Budgets. To characterize compression beyond the four budgets shown in the main cross-benchmark table, we evaluate the five additional methods on a separate ten-budget grid—70%, 50%, 30%, 20%, 10%, 5%, 1%, 0.5%, 0.1%, and 0.01%—on XLRS-Bench-lite and MME-RS. Table 18 reports all 100 conditions using the same Qwen3-VL-8B model without fine-tuning. XLRS-Bench-lite includes 2,980 single-answer questions and excludes all 100 multi-select questions; its binary questions are scored over A/B. MME-RS includes 150 questions and preserves all available options, including E. On XLRS-Bench-lite, HiPrune leads at 70% and 50% with 52.45% and 52.55%, whereas DART leads at 0.01% with 35.81%. On MME-RS, CD-Pruner leads at 70% with 61.33%, while PriorTR leads at 0.01% with 30.67%. These transitions support reporting complete budget sweeps rather than selecting one operating point. Each MME-RS question changes accuracy by approximately 0.67 percentage points, so small differences on this dataset warrant cautious interpretation. Nominal retention alone does not match layerwise work or measured latency across methods.

Table 18: Complete external-dataset sweeps for the five additional pruning methods. Accuracy (%) is computed over each question's actual source options. All columns are nominal retention rates, and all methods use the same Qwen3-VL-8B model without fine-tuning.

Method	Type	70%	50%	30%	20%	10%	5%	1%	0.5%	0.1%	0.01%
XLRS-Bench-lite: single-answer subset ($n = 2{,}980$)											
CDPruner (Zhang et al., 2025c)	Pre-LLM	52.01	51.21	48.86	47.01	43.62	42.62	35.54	34.46	33.62	33.49
HiPrune (Liu et al., 2025)	Pre-LLM	52.45	52.55	51.64	51.88	49.53	47.38	39.19	36.38	33.62	33.46
DART (Wen et al., 2025)	Intra-LLM	52.05	52.21	52.05	51.51	49.30	47.11	41.17	35.81	35.81	35.81
PriorTR (Chen et al., 2026)	Intra-LLM	51.41	51.14	50.30	49.93	48.19	46.21	39.09	36.51	33.89	33.89
VScan (Zhang et al., 2025a)	Hybrid	52.32	52.01	52.15	52.25	51.48	49.30	43.42	39.90	35.57	34.97
MME-RS ($n = 150$)											
CDPruner (Zhang et al., 2025c)	Pre-LLM	61.33	59.33	58.00	52.67	43.33	34.67	26.00	29.33	25.33	27.33
HiPrune (Liu et al., 2025)	Pre-LLM	58.00	55.33	52.00	49.33	52.00	42.67	32.67	30.67	26.00	26.67
DART (Wen et al., 2025)	Intra-LLM	59.33	58.67	53.33	52.00	46.67	44.67	35.33	26.67	26.67	26.67
PriorTR (Chen et al., 2026)	Intra-LLM	55.33	56.67	50.00	45.33	44.00	45.33	32.00	34.00	30.67	30.67
VScan (Zhang et al., 2025a)	Hybrid	59.33	58.00	60.00	54.00	54.67	48.67	35.33	32.00	30.67	27.33

E.2 Extended Retention Sweeps for the Remaining Methods

Shared Answer-Candidate Protocol and Evaluation Standard. Tables 19 and 20 report the remaining seventeen methods on the same ten-budget appendix grid defined above. Together with the five methods in Table 18, these results complete the 22-method comparison reported in the abstract. For XLRS-Bench-lite, every method is evaluated on the same 2,980 single-answer questions, with the 100 multi-select questions excluded consistently. For MME-RS, every method is evaluated using the complete original candidate set for each question. PyramidDrop decreases from 50.64% at 5% retention to 41.54% at 0.01% on XLRS-Bench-lite, illustrating the substantial accuracy loss under aggressive compression. Table 4 combines all 22 methods in a single ranking under the same protocol.

Table 19: Ten-budget pruning accuracy (%) for seventeen methods on XLRS-Bench-lite ($n = 2{,}980$), grouped by the pruning locations defined in the main paper. All methods use the original answer candidates and the shared evaluation protocol. Columns are nominal retention rates.

Method	Type	70%	50%	30%	20%	10%	5%	1%	0.5%	0.1%	0.01%
EViT (Liang et al., 2022)	Pre-LLM	49.36	47.82	47.08	45.44	43.79	42.35	39.36	37.18	34.70	34.09
QuietPrune (Gao et al., 2026)	Pre-LLM	52.99	51.95	50.17	48.93	46.64	45.03	41.11	39.09	35.40	33.79
VisionZip (Yang et al., 2025b)	Pre-LLM	52.25	51.68	50.17	48.19	45.50	43.59	37.32	35.91	33.79	33.79
DivPrune (Alvar et al., 2025)	Pre-LLM	52.15	51.54	51.51	50.91	48.22	44.93	39.56	36.61	33.99	33.32
TopV (Yang et al., 2025a)	Pre-LLM	51.88	51.85	51.61	51.28	48.96	45.13	39.80	38.29	35.70	34.66
LLaVA-PruMerge (Shang et al., 2025)	Pre-LLM	48.89	47.79	47.68	46.88	46.07	44.63	40.30	39.40	37.01	35.91
VisPruner (Zhang et al., 2025b)	Pre-LLM	51.98	51.14	51.24	49.60	46.58	44.77	40.44	39.26	35.77	35.67
FastV (Chen et al., 2024)	Intra-LLM	51.71	50.91	48.29	47.08	45.54	42.15	34.23	33.86	33.59	33.56
SparseVLM (Zhang et al., 2024)	Intra-LLM	52.25	51.68	50.17	48.56	46.11	44.46	38.72	36.07	33.86	33.72
TransPrune (Li et al., 2026a)	Intra-LLM	52.18	51.34	50.23	48.83	45.64	43.52	39.60	38.83	34.77	34.63
FEATHER (Endo et al., 2025)	Intra-LLM	52.18	51.34	50.23	48.93	45.64	42.82	39.93	38.46	34.70	34.73
FiT-Prune (Ye et al., 2025)	Intra-LLM	52.65	52.58	51.28	50.07	47.21	45.64	42.21	41.61	41.04	40.67
VTW (Lin et al., 2025)	Intra-LLM	52.32	48.86	34.50	33.09	33.09	33.99	33.96	33.96	33.96	33.96
PyramidDrop (Xing et al., 2024)	Intra-LLM	52.28	52.28	51.74	51.51	51.17	50.64	48.09	47.38	45.47	41.54
IVTP (Huang et al., 2024)	Hybrid	52.11	51.44	50.00	48.02	46.64	44.19	38.79	36.21	33.72	33.46
MustDrop (Liu et al., 2024a)	Hybrid	52.55	52.08	51.11	49.87	47.55	45.50	39.93	38.89	34.70	33.96
DUET-VLM (Singh et al., 2026)	Hybrid	49.50	48.83	47.38	46.07	44.36	42.72	40.44	39.66	36.24	34.66

Table 20: Ten-budget pruning accuracy (%) for seventeen methods on MME-RS ($n = 150$), grouped by the pruning locations defined in the main paper. All methods use each question's complete original candidate set and the shared evaluation protocol. Columns are nominal retention rates.

Method	Type	70%	50%	30%	20%	10%	5%	1%	0.5%	0.1%	0.01%
EViT (Liang et al., 2022)	Pre-LLM	48.67	41.33	34.00	32.67	36.67	31.33	33.33	29.33	30.67	27.33
QuietPrune (Gao et al., 2026)	Pre-LLM	56.67	54.00	53.33	50.00	46.00	42.67	31.33	30.67	28.00	26.00
VisionZip (Yang et al., 2025b)	Pre-LLM	58.67	56.67	54.67	49.33	43.33	39.33	36.00	33.33	28.67	28.67
DivPrune (Alvar et al., 2025)	Pre-LLM	60.00	56.67	54.67	52.00	44.67	44.00	35.33	32.67	30.00	28.00
TopV (Yang et al., 2025a)	Pre-LLM	56.67	58.00	57.33	52.67	48.00	42.67	34.00	36.00	33.33	28.67
LLaVA-PruMerge (Shang et al., 2025)	Pre-LLM	54.00	52.00	47.33	43.33	40.67	37.33	34.67	30.00	30.67	32.67
VisPruner (Zhang et al., 2025b)	Pre-LLM	55.33	54.67	49.33	44.67	46.00	42.67	36.67	30.67	30.67	30.00
FastV (Chen et al., 2024)	Intra-LLM	57.33	54.67	51.33	47.33	35.33	34.67	31.33	26.67	24.67	26.67
SparseVLM (Zhang et al., 2024)	Intra-LLM	58.67	56.67	54.67	50.67	45.33	45.33	37.33	35.33	29.33	28.00
TransPrune (Li et al., 2026a)	Intra-LLM	59.33	58.00	54.00	52.67	46.67	41.33	37.33	37.33	30.00	28.00
FEATHER (Endo et al., 2025)	Intra-LLM	59.33	58.00	54.00	53.33	48.67	43.33	37.33	37.33	30.00	30.00
FiT-Prune (Ye et al., 2025)	Intra-LLM	58.67	58.67	54.00	52.67	42.67	36.67	37.33	38.00	37.33	36.67
VTW (Lin et al., 2025)	Intra-LLM	58.00	57.33	30.00	29.33	26.00	28.00	28.00	28.00	28.00	28.00
PyramidDrop (Xing et al., 2024)	Intra-LLM	58.67	60.00	56.00	55.33	54.00	49.33	50.00	46.00	38.67	34.67
IVTP (Huang et al., 2024)	Hybrid	60.00	58.67	53.33	50.67	44.67	41.33	30.00	28.00	28.00	26.00
MustDrop (Liu et al., 2024a)	Hybrid	58.67	58.00	57.33	54.67	53.33	43.33	34.67	30.00	28.67	30.00
DUET-VLM (Singh et al., 2026)	Hybrid	53.33	46.67	40.67	36.67	37.33	32.67	32.00	29.33	26.67	28.00

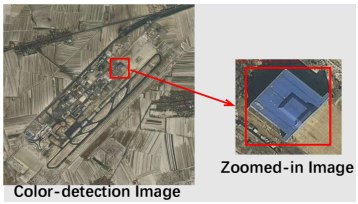
F Case Studies

Representative annotations connect local visual cues to the benchmark's perception and reasoning tasks. We present these cases to illustrate the required visual information and the interpretation of the reference answers.

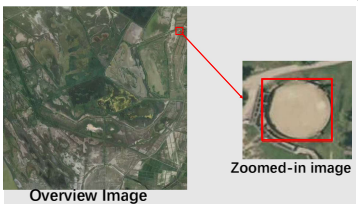
F.1 Fine-grained Visual Attributes

Fine-grained attribute recognition depends on local cues that occupy only a small part of an ultra-high-resolution scene. Figure 5 presents four examples covering color, boundary shape, orientation, and object classification, each pairing the full image with a magnified target region.

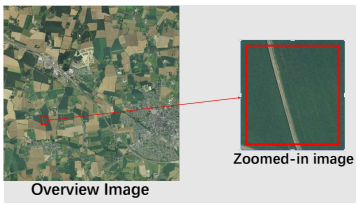
Category: Color Detection
Image Resolution: 4701 x 4946
Question: Determine the color of the airport within the given reference bounding
box in the image. Bounding box: [2814, 1628, 2975, 1790].
Options:
A: brown
B: blue
C: white
D: gray
Correct Answer: B



Category: Shape/Margin Recognition
Image Resolution: 10000 x 10000
Question: What is the most precise description of the main object's boundary
characteristic or shape within the given reference bounding box in the image?
Bounding box: [9541, 1273, 9610, 1344].
Options:
A: irregular
B: circular
C: elliptical
D: square
Correct Answer: B



Category: Orientation Detection
Image Resolution: 10000 x 10000
Question: Determine the orientation of the road network within the given
reference bounding box in the image. Bounding box: [2096, 6332, 2454, 6732].
Options:
A: right
B: top
C: top-right
D: bottom-right
Correct Answer: D



Category: Object Classification
Image Resolution: 10000 x 10000
Question: Determine the category of the main object within the given reference
bounding box in the image. Bounding box: [2382, 5403, 4489, 6853].
Options:
A: coastal wetland
B: sparse cropland area
C: urban residential zone
D: dense forest
Correct Answer: B



Figure 5: Examples of color, shape, orientation, and object-classification questions.

F.2 Spatial Relations, Grounding, and Counting

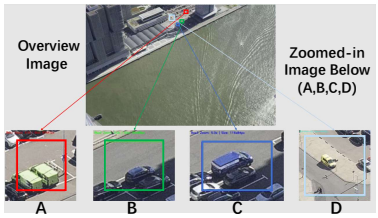
Figure 6 presents examples of spatial relations, object and regional grounding, and counting.

Category: Object Spatial Relationship
Image Resolution: 4183 x 8115
Question: In the image, where is the elliptical stadium with the track located
relative to the rectangular gray building with a white roof?
Options:
A: To the left
B: To the right
C: Below
D: Above
Correct Answer: A



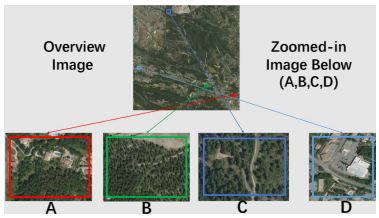
Category: Object Grounding
Image Resolution: 7360 x 4912

Question: Which bounding box best localizes a green van?
Options:
A: [3806, 270, 3893, 362]
B: [3662, 656, 3761, 742]
C: [3500, 758, 3616, 842]
D: [3237, 418, 3430, 619]
Correct Answer: A



Category: Regional Grounding
Image Resolution: 10000 x 10000

Question: Which bounding box best localizes the area that is
a commercial complex dominated by roads and parking lots,
surrounded by vegetation and residential areas?
Options:
A: [9072, 8359, 9467, 8637]
B: [6514, 7541, 7036, 7805]
C: [3102, 436, 3480, 640]
D: [336, 5876, 636, 6271]
Correct Answer: D



Category: Object Counting
Image Resolution: 7360 x 4912
Question: How many basketball courts are there in the entire image?
Options:
A: 4
B: 3
C: 5
D: 1
Correct Answer: D

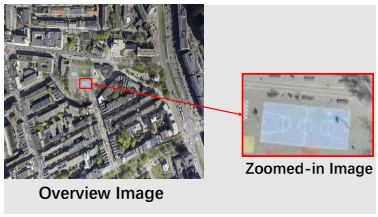


Figure 6: Examples of spatial relations, object and regional grounding, and object counting.

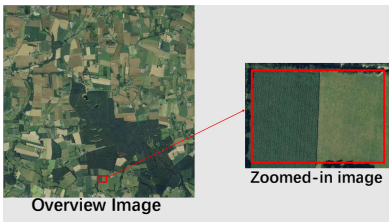
F.3 Regional Counting and Joint Comparison

Figure 7 presents regional counting and multi-region comparison examples. These tasks require preserving evidence for every referenced box or image: counting is restricted to a specified region, while comparison selects an extremum across candidates. A single salient region is insufficient; models must retain local detail and correspondence across the full candidate set to avoid discarding information needed for the final decision.

Category: Regional Counting
Image Resolution: 10000 x 10000 **Question:**
How many land blocks are there within the given reference bounding box
in the image? Bounding box: [4918, 8846, 5361, 9146].
Options:

A: 1
B: 6
C: 2
D: 4

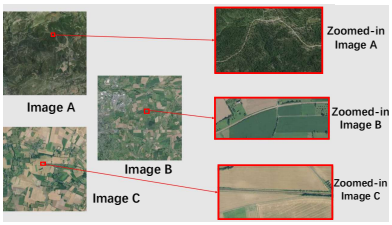
Correct Answer: C



Category: Multi-region Joint Contrast (Multi-image)
Image Resolution: 10000 x 10000
Question:
Which of the three given boxes in these three images shows the most winding road? Image A's box is [4446, 2282, 6175, 3339], Image B's box is [4404, 3882, 6046, 4453], and Image C's box is [3825, 4061, 4911, 4546]. The resolution of each image is 10000 x 10000, and Images A/B/C follow the order in which they are provided.
Options:

A: Image B
B: Image C
C: Image A
D: Cannot compare

Correct Answer: C



Category: Multi-region Joint Contrast (Single-image Multi-box)
Image Resolution: 10000 x 10000
Question:
Which bounding box best represents the cropland patch with the
highest greenness level?
Options:

A: [5054, 1461, 6196, 2453]
B: [7625, 5725, 8968, 6982]
C: [4175, 7411, 5654, 8632]
D: Cannot compare

Correct Answer: A

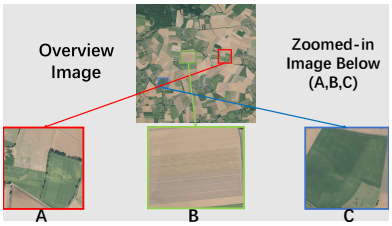


Figure 7: Examples of regional counting and multi-region comparison across images or boxes.

F.4 Future Prediction from Temporal Comparison

Figure 8 presents a two-image temporal-comparison example. The question compares imagery marked 2005 and 2012 around a road and surrounding fields, then selects the future change most consistent with the observed visual pattern. Answering the question requires direct comparison of the two images, while the predicted outcome is treated as a visually supported interpretation rather than a verified future event.

Category: Future Prediction (Two-image Single-turn)
Image Resolution: 10000 x 10000
Question:
Comparing the agricultural fields located near the central winding road in
Image A and Image B, what is the most plausible future change that might
occur in this area based on the observed visual patterns?
Options:

A: The fields are likely to revert to natural vegetation, as signs of crop reduction and overgrowth can be seen between Image A and Image B.

B: The fields are likely to be converted into residential zones, as there is a visible increase in buildings along the road from Image A to Image B.

C: The fields are likely to remain unchanged, as there is no noticeable alteration in their structure or use between Image A and Image B.

D: The agricultural fields near the central winding road are likely to
consolidate into larger plots, as several smaller fields appear to have merged between Image A and Image B.

Correct Answer: D

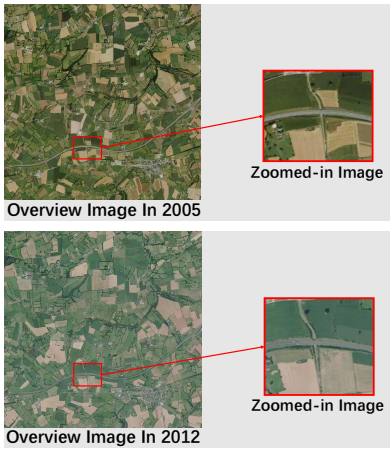


Figure 8: Example of future prediction from a two-image temporal comparison.

G Data Construction Prompts

The following templates specify the visual inputs, reference annotations, and answer formats used to construct each task.

G.1 Visual Attribute Prompts

Category: Color Detection
Model: Qwen2.5-VL-7B
Prompt: You are a dataset constructor for Remote Sensing VQA (Color Detection).
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Reference bounding box [xmin, ymin, xmax, ymax]: [{xmin}, {ymin}, {xmax}, {ymax}] - Ground-truth target color: {target_color}
Generate ONE multiple-choice QA sample that strictly follows the requirements below:
1. Inspect the given bounding box region to identify the main object. Replace "object" with a concise object category name (e.g., car, truck, ship, building). Avoid vague words like "thing" or "object".
2. Use the following question template verbatim (only substitute the "object" placeholder):
"Determine the color of the object within the given reference bounding box in the image. Image resolution: {width} x {height}. Bounding box: [{xmin}, {ymin}, {xmax}, {ymax}]."
3. Provide four answer options labeled A, B, C, D: - Use {target_color} (lowercase) as the only correct option. - Choose three distinct distractor colors that plausibly appear elsewhere in the image (other instances or surrounding materials). Keep them as simple lowercase color words (red, blue, green, yellow, white, black, gray, brown, orange, purple). - Ensure all four options are unique and the correct color appears exactly once.
4. Set the "answer" field to the key (A/B/C/D) corresponding to {target_color}.
5. Output JSON ONLY in the structure:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 9: Prompt used to construct color-detection questions from a specified bounding box.

Category: Shape/Margin Recognition
Model: Qwen2.5-VL-7B
Prompt: You are a dataset constructor for Remote Sensing VQA (Shape/Boundary Description).
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Reference bounding box [xmin, ymin, xmax, ymax]: [{xmin}, {ymin}, {xmax}, {ymax}] - Ground-truth target shape: {target_shape}
Generate ONE multiple-choice QA sample that strictly follows the requirements below:
1. Focus only on the main object inside the given bounding box. Base your choice on boundary/shape, not color or texture.
2. Use the following question template verbatim (do not change wording, only substitute numbers):
"Which is the most precise description of the main object's boundary characteristic/shape within the given reference bounding box in the image? Image resolution: {width} x {height}. Bounding box: [{xmin}, {ymin}, {xmax}, {ymax}]."
3. Provide four answer options labeled A, B, C, D: - Use {target_shape} (lowercase) as the only correct option. - Create three distinct distractors: plausible shape descriptors that could be seen elsewhere in the image (other objects) or possible but incorrect impressions inside the box (non-primary elements). - Keep options concise, lowercase words/phrases, e.g., rectangular, square, circular, oval, elongated, triangular, l-shaped, star-shaped, concave, convex, irregular, curved, linear, arc-shaped. - Ensure all four options are unique and the correct shape appears exactly once.
4. Set the "answer" field to the key (A/B/C/D) corresponding to {target_shape}.
5. Output JSON only in the structure:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 10: Prompt used to construct shape and boundary-recognition questions.

G.2 Orientation and Category Prompts

Category: Orientation Detection
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor focusing on orientation detection.
Image metadata:
- Image : {image_path} - Resolution: {width} x {height} - Reference bounding box [xmin, ymin, xmax, ymax]: [{xmin}, {ymin}, {xmax}, {ymax}] - Ground-truth orientation label (must be the correct answer): {target_orientation} - Allowed orientation words: {orientation_vocab}
Your task is to create exactly ONE multiple-choice question–answer pair that follows these strict rules:
1. Inspect the target region inside the bounding box and identify the main object (e.g., car, ship, storage tank, building). Pick a concise, specific object category name. Do not use vague words like "object" or "thing".
2. Use this exact question template (only replace the placeholder with your object phrase):
"Determine the orientation of the {{object_name}} within the given reference bounding box in the image. Image resolution: {width} x {height}. Bounding box: [{xmin}, {ymin}, {xmax}, {ymax}]." Replace {{object_name}} with your chosen object.
3. Generate four answer options labeled A, B, C, D. Every option must be a single orientation word selected from {orientation_vocab}. - Insert {target_orientation} as the only correct option.
4. Set the "answer" field to the option key (A/B/C/D) that corresponds to {target_orientation}.
5. Output JSON only in the structure:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 11: Prompt used to construct object-orientation questions from localized visual cues.

Category: Object Classification
Model: Qwen2.5-VL-7B
Prompt: You are a remote sensing VQA dataset constructor focusing on object classification.
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Reference bounding box [xmin, ymin, xmax, ymax]: [{xmin}, {ymin}, {xmax}, {ymax}] - Ground-truth target category: {target_label} - Category vocabulary (valid class names): {class_vocab}
Produce exactly ONE multiple-choice question–answer pair that obeys all rules:
1. Use this question sentence verbatim (numbers already filled in. Do not alter wording or punctuation):
"Determine the category of the main object within the given reference bounding box in the image. Image resolution: {width} x {height}. Bounding box: [{xmin}, {ymin}, {xmax}, {ymax}]."
2. Options A/B/C/D must be English only and concise (at most ≤ 6 words each): - Build the correct option by starting from the label {target_label}. You may prepend at most two adjectives to refine it (e.g., "dense {target_label} zone"). The label {target_label} must appear exactly once in the correct option. - Create three distinct distractors from exactly these two sources: (A) visually grounded objects elsewhere in the same image that are either near the box or highly salient.
(B) semantically confusable classes of the same high-level type as the target.
At least one distractor must come from (A) and at least one from (B). If (A) is not visibly available, use more from (B). - Distractors must use English descriptors that include one of the labels in {class_vocab}, but must not reuse {target_label}. Do not include any Chinese text. Do not repeat labels or duplicate phrasings across options.
3. The "answer" field must be exactly the key (A/B/C/D) pointing to the refined English description containing {target_label}.
4. Output JSON only in this schema:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 12: Prompt used to construct fine-grained object-classification questions.

G.3 Relative Spatial Relationship Prompt

Category: Relative Spatial Relationship
Model: Qwen2.5-VL-7B
Prompt: You are a remote sensing VQA dataset constructor focusing on object-level relative spatial relationships.
Your goal: From the given image **alone**, create **one** multiple-choice question of the type:
VQA-Perception → Object-Level Attributes → Relative Spatial Relationship.
What to do (strict steps):

1. Determine the image pixel resolution $(W \times H)$.
2. Select two adjacent, salient, clearly visible objects or regions in the image.
   - Objects should be distinct, identifiable, and common (e.g., brightly colored or large areas such as a stadium, forest, building, river, red plane).
   - Each object must be fully visible and not partially cropped.
   - There must be a clear distinction between the two objects and a well-defined positional relationship.
   - Both objects should be at least 10×10 pixels for clarity.
3. Formulate a question asking about the relative spatial position of Object A with respect to Object B using the exact sentence pattern:
   "In the image, where is <Object A> located relative to <Object B>? Image resolution: {w} x {h}."
   Describe the characteristics of each object in detail (e.g., "the red aircraft", "the cargo ship in motion", "the parking lot with 10 cars").

Important instructions (must follow):

- When describing the two reference objects, do not mention their global positions in the image (e.g., "top left", "bottom right").
- Instead, describe them using distinctive features such as shape, color, texture, or category (e.g., "the circular stadium with a red track", "the rectangular gray building with a white roof").
- If the two objects are in the same general area (e.g., both in the upper right), do not reveal this. The model should infer only the relative spatial relationship.
- If the objects are very large and cover most of the image, avoid repeating absolute global positions. Emphasize unique visual traits instead.
- Distinguish each object from other similar ones in the image by differences such as orientation, color, or size (e.g., multiple parking lots differentiated by the number of spaces or nearby structures).
- Questions should be grounded in the overall environmental semantics of the scene and involve relatively common objects such as ships and houses.

Example: "In the image, where is the large parking lot on the left side with 2 planes located relative to the largest airport terminal building? Image resolution: {w} x {h}."
Answer options:

- Provide exactly four options (A, B, C, D).
- Options must be directional relations: "Above", "Below", "To the left", "To the right", "Upper left corner", "Upper right corner", "Bottom left corner", "Bottom right corner".
- Randomize the order of options and ensure exactly one is correct.

Quality rules:

- The spatial relationship must be unambiguous.
- Do not select overlapping or vague objects.

Output format (JSON only, no extra text):
{
"question": "In the image, where is <Object A> located relative to <Object B>? Image resolution: {w} x {h}.",
"options": {"A": "<direction>", "B": "<direction>", "C": "<direction>", "D": "<direction>"},
"answer": "A|B|C|D"
}
Constraints:

- Output only the JSON object.
- {w} and {h} must be integers.
- The answer must be exactly one of A/B/C/D.

Figure 13: Prompt used to construct questions about relative spatial relationships.

G.4 Object and Regional Grounding Prompts

Category: Object Grounding
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor focusing on object grounding.
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Candidate bounding boxes (pixel coordinates, inclusive):
{candidate_bboxes} - Ground-truth target bounding box (must be one of the candidates): {target_bbox}
Your task is to create exactly one multiple-choice question–answer pair that follows every rule:
1. Use this question template verbatim (only substitute the placeholder with the object phrase supplied below):
"Which bounding box best localizes {object_phrase}, resolution: {width} x {height}?"
2. Provide four answer options labeled A, B, C, D:
- Each option value must match exactly one of the candidate bounding box strings listed above.
- Use every candidate bounding box exactly once (in any order).
3. Set the "answer" field to the option key (A/B/C/D) corresponding to the candidate bounding box that equals {target_bbox}.
4. Output JSON only in the structure:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 14: Prompt used to construct object-grounding questions with reference regions.

Category: Regional Grounding
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor focusing on regional grounding.
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Candidate bounding boxes (pixel coordinates, inclusive):
{candidate_bboxes} - Ground-truth target bounding box (must be one of the candidates): {target_bbox}
Your task is to create exactly one multiple-choice question–answer pair that follows every rule:
1. Use this question template verbatim (based on the region phrase supplied below and the objects within the candidate boxes, you may make appropriate refinements to substitute the placeholder):
"Which bounding box best localizes {region_phrase}, resolution: {width} x {height}?"
Regardless of what language {region_phrase} is in, convert it to full English in the final question.
2. Provide four answer options labeled A, B, C, D:
- Each option value must match exactly one of the candidate bounding box strings listed above.
- Use every candidate bounding box exactly once (in any order).
3. Set the "answer" field to the option key (A/B/C/D) corresponding to the candidate bounding box that equals {target_bbox}.
4. Output JSON only in the structure:
{
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..."},
"answer": "A|B|C|D"
}

Figure 15: Prompt used to construct regional-grounding questions from spatial descriptions.

G.5 COUNTING PROMPTS

Category: Object Counting
Model: GPT-5 thinking
Prompt: You are a dataset constructor for Remote Sensing VQA.
Your goal: From the given image alone, create one multiple-choice question of the type:
VQA-Perception → Counting & Measurement → Object Counting.
Strict steps:

1. Select one object category that is: - clearly visible and individually countable. - not too small (at least 10×10 pixels), not severely occluded, and unambiguous. - described with precise attributes (e.g., "large cargo ships", "red oval tracks", "small white cars in a parking lot"). - not vague or overly broad. Avoid labels like "ship", "vehicle", or "building". Use specific and distinctive descriptions such as "big oil tanker" or "blue-roof two-story buildings". If very small or ambiguous objects exist, ignore them and only select salient, clearly recognizable ones.
2. Count all qualifying instances across the entire image. Do not restrict the question to a selected subregion or bounding box.
3. Count the objects precisely, including every qualifying instance visible in the full image.
4. Construct the question in this format:
   "How many <detailed object description> are there in the entire image?"
5. Provide four answer options (A, B, C, D):
   - each option must be an integer. - options must be close (within $\pm$1–3 of the correct count). - randomize the position of the correct answer. - avoid trivial wide gaps such as 10/20/30/40.

Output format (JSON only, no extra text):
{
"question": "How many <objects> are there in the entire image?",
"options": {"A": <int>, "B": <int>, "C": <int>, "D": <int>},
"answer": "A|B|C|D"
}

Figure 16: Prompt used to construct object-counting questions over the full image.

Category: Regional Counting
Model: Qwen2.5-VL-7B
Prompt: You are a dataset constructor for Remote Sensing VQA.
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Reference bounding box [xmin, ymin, xmax, ymax]: [{xmin}, {ymin}, {xmax}, {ymax}]
Your goal: From the given image and specified reference bounding box, create **one** multiple-choice question of the type:
VQA-Perception → Counting & Measurement → Regional Counting.
Strict steps:

1. Inspect only the supplied reference bounding box. Select one object category within this box that is: - clearly visible and individually countable. - not too small (at least 10×10 pixels), not severely occluded, and unambiguous. - described with precise attributes (e.g., "large cargo ships", "red oval tracks", "small white cars in a parking lot"). - not vague or overly broad (avoid "ship", "vehicle", "building").
2. Count only qualifying instances located within the supplied reference bounding box. Do not count instances outside it or choose a different region.
3. Treat a boundary-crossing instance as inside only when most of the instance lies within the reference bounding box; avoid targets for which this rule remains ambiguous.
4. Construct the question using this template:
   "How many <detailed object description> are there within the given reference bounding box in the image? Image resolution: {width} x {height}. Bounding box: [{xmin}, {ymin}, {xmax}, {ymax}]."
5. Provide four answer options (A, B, C, D): - each option must be an integer. - options must be close (within $\pm$1–3 of the correct count). - randomize the position of the correct answer. - avoid wide trivial gaps such as 10/20/30/40.

Output format (JSON only, no extra text):
{
"question": "How many <objects> are there within the given reference bounding box in the image? Image resolution: ... Bounding box: [...].",
"options": {"A": <int>, "B": <int>, "C": <int>, "D": <int>},
"answer": "A|B|C|D"
}

Figure 17: Prompt used to construct regional-counting questions within a supplied reference bounding box.

G.6 MULTI-REGION COMPARISON PROMPTS

Category: Multi-region Joint Contrast (Three-Image Comparison)
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor for multi-region joint contrast.
Task: - You are given three satellite images (A, B, C), each with one reference bounding box and one contrast attribute label. - Compare only the three reference regions (A/B/C) according to the given attribute. - Produce exactly one multiple-choice question with fixed options. - The correct answer must be A, B, C, or D; use D only when the visible regions cannot be compared reliably.
Image A/B/C: - File: {image_path_A/B/C} - Resolution: {width_A/B/C} x {height_A/B/C} - Reference box [xmin, ymin, xmax, ymax]: [{xmin_A/B/C}, {ymin_A/B/C}, {xmax_A/B/C}, {ymax_A/B/C}] - Contrast attribute label: "{label_A/B/C}"
Question requirements:

1. The question should ask which image (A/B/C) shows the extremum of the attribute indicated by the labels above. For example:
   "Which of the three given boxes in these three images has the highest building density? Resolution: {width_A} x {height_A}. Image A's box: [{xmin_A}, {ymin_A}, {xmax_A}, {ymax_A}]. Image B's box: [{xmin_B}, {ymin_B}, {xmax_B}, {ymax_B}]. Image C's box: [{xmin_C}, {ymin_C}, {xmax_C}, {ymax_C}]."
   Here, "the highest building density" should be derived from the attribute implied by the labels.
   - If the label does not explicitly specify highest/lowest, choose the most natural extremum (e.g., "highest"). - Keep the question concise (no more than 25 words). - The resolution of the three images is the same, so it is sufficient to mention it once. - No matter what language the labels are in, convert them into an English question.
2. Options must be fixed strings: - A: "Image A" - B: "Image B" - C: "Image C" - D: "Cannot compare"
3. Set the "answer" to the option key (A/B/C/D) that uniquely matches the correct extremum by inspecting the three images.

Output format (JSON only, ASCII only, no commentary):
{
"question": "...",
"options": {"A": "Image A", "B": "Image B", "C": "Image C", "D": "Cannot compare"},
"answer": "A|B|C|D"
}

Figure 18: Prompt used to construct multi-image, multi-region comparison questions.

Category: Multi-region Joint Contrast (Single Image)
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor focusing on multi-region joint contrast (single image).
Image metadata:
- Image file: {image_path} - Resolution: {width} x {height} - Annotated comparison regions from the XML: {candidate_regions} Each line above has the form: [xmin, ymin, xmax, ymax] — "region description"
Your tasks:

1. Read the region descriptions carefully. They already specify the contrast attribute(s) (e.g., building density, traffic volume).
2. Craft exactly one multiple-choice question comparing these regions. The question must: - use the template verbatim (only substitute the braces with your phrase): "Which bounding box best represents the {{comparison_phrase}}" - ensure the comparison phrase stays faithful to the provided descriptions (e.g., "area with the highest building density", "cropland patch with the lowest greenness"). - if descriptions imply an extremum (highest/lowest/most/least), leverage that. Otherwise phrase the comparison naturally (e.g., "largest cluster of compact housing"). - keep the question concise (no more than 25 words).
3. Provide exactly four options: - A, B, and C use the three annotated bounding box strings exactly as provided. Do not invent new boxes. - D is the fixed string "Cannot compare".
4. Determine the correct answer by visually comparing the regions in the image according to the described attribute(s). - set "answer" to A, B, or C when one region is uniquely supported by the visible evidence. - set "answer" to D only when the regions cannot be compared reliably.
5. Output JSON only in this structure (no commentary, English only):
   {
   "question": "...",
   "options": {"A": "¡box A¿", "B": "¡box B¿", "C": "¡box C¿", "D": "Cannot compare" },
   "answer": "A|B|C|D"

Figure 19: Prompt used to compare multiple marked regions within one image.

G.7 MULTI-TURN STATE PROMPT

Category: Object State Judgement (Multi-round Dialogue)
Model: GPT-5 thinking
Prompt: You are a dataset constructor for Remote Sensing VQA.
Task: Design a three-round dialogue (user ↔ assistant) focused on **Object State Judgement** for a salient object in the image.
Strict requirements:

1. Inspect the image carefully and select one object whose state is unambiguous from visible cues.
2. Describe that object with distinctive traits so it cannot be confused with similar items nearby.
3. Craft three rounds of interaction: - Each round is a single multiple-choice question with four options labeled A/B/C/D, authored by the user. - Every option must be grounded in the image. Create three plausible distractors based on other visual cues, plus one correct answer. - The assistant responds only with the correct option letter followed by the exact option text (format: "<letter>. <option text>"). Do not output explanations, reasoning, or any extra words. - Round 1: ask about visual cues that reveal the object's condition. The four options must be obviously different from one another and must not express the same meaning. In the question, indicate the object's approximate position in the picture and describe the object in detail, but do not include any hints of the answer. - The correct choice must rely on information that requires looking at the image, not just commonsense. - Round 2: ask how those cues impact functionality or operational status. The four options must be obviously different from one another and must not express the same meaning. - Round 3: ask for the state classification. The four options can reference from the fixed list `allowed_states` below. Other options can be added based on commonsense properties of the chosen object, but they must be different from the correct answer and from each other.

Allowed states: "Moving", "Stationary", "In use / Active", "Not in use / Inactive", "Intact", "Damaged", "Under construction", "Occupied", "Empty", "Snow-covered", "Flooded", "Dry", "Vegetated", "Take-off", "Landing", "Parked", "Sailing", "Anchored", "Reversing", "Stopped", "In transit", "Under demolition", "Renovating", "Vacant", "Cruising", "Taxiing", "Departed", "Arrived", "Docked", "Moored", "In dry dock", "Drifting", "Eroded", "Burnt", "Muddy", "Irrigated", "Frozen", "Overgrown", "Drought", "Collapsed", "Unstable", "Protected", "Inaccessible", "Severe damage", "Partial damage", "Fully operational", "Non-operational"
Output format (JSON only): {
"target_object": {
"description": "<distinguishing description>",
"state_cues": ["<key cue 1>", "<key cue 2>"]
},
"conversation": [
{
"round": 1,
"user_question": "<A-D question>",
"options": {"A": "<option>", "B": "<option>", "C": "<option>", "D": "<option>"},
"assistant_answer": "<letter>. <option text>"
},
{
"round": 2,
"user_question": "<A-D question>",
"options": {"A": "<option>", "B": "<option>", "C": "<option>", "D": "<option>"},
"assistant_answer": "<letter>. <option text>"
},
{
"round": 3,
"user_question": "<A-D question using allowed states>",
"options": {"A": "<state>", "B": "<state>", "C": "<state>", "D": "<state>"},
"assistant_answer": "<letter>. <state text>"
}
]
}

Figure 20: Prompt used to construct multi-round object-state judgement dialogues.

G.8 SINGLE-TURN STATE AND ANOMALY PROMPTS

Category: Object State Judgement (Single turn)
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA dataset constructor focusing on object state judgement.
Image metadata: - Image file: {image_path} - Resolution: {width} x {height} - Allowed state vocabulary (use these strings exactly): {state_vocab}
Your task is to create exactly one multiple-choice question–answer pair following these strict requirements:
1. Inspect the image carefully and select one salient object whose state is visually unambiguous. - Describe the object's distinctive appearance (size, shape, color, context) so it cannot be confused with neighboring objects. - Identify the object's approximate location within the image (e.g., "top-left", "lower central area").
2. Compose the question using this template verbatim (only substitute the placeholders):
"What is the current state of the {{object_description}} located in the {{location_phrase}} of the image?" - Replace {object_description} with your detailed object description (ensure that no hints to the correct answer are given based on the detailed description). - Replace {location_phrase} with the location phrase (e.g., "in the upper right corner of the picture"). - Do not leak the correct answer in the question text.
3. Provide four answer options labeled A, B, C, D.
Insert the correct state exactly once. The remaining three must be plausible distractors. 4. Set the "answer" field to the option key (A/B/C/D) that corresponds to the correct state. 5. Populate the "target_object" fields: - "description": the object description from step 1. - "location": the location phrase from step 1. - "state": the correct state (identical to the answer option text). - "visual_cues": a list of two concise visual cues supporting the state judgement.
6. Output JSON only in the structure below (no extra words, explanations, or markdown):
{
"target_object": { "description": "...", "location": "...", "state": "...", "visual_cues": ["...", "..."] },
"question": "...",
"options": {"A": "...", "B": "...", "C": "...", "D": "..." },
"answer": "A|B|C|D"
}

Figure 21: Prompt used to construct single-turn object-state judgement questions.

Category: Anomaly detection and interpretation
Model: GPT-5 thinking
Prompt: You are a remote sensing VQA assistant performing Anomaly detection and interpretation for an image.
Rules:

1. Carefully observe the image and select one object whose state is unambiguous. - Describe the object with distinguishing traits (e.g., size, shape, color, position). - Provide the approximate location of the object in the image (e.g., "top-left").
2. In your question, include a detailed description of the selected object based on its visual features, including its position in the image. - Example: "The red cargo ship parked on the riverbank, with a blue warehouse visible on the right side of the image. Where is this cargo ship located in the image?" - Make sure the question and the description are clearly related to the visual elements in the image. - The question should not contain any answer hints.
3. The assistant should then ask about the most likely reason or cause for the object's current state or anomaly, based on visible cues from the image. For example: "What could be the most likely reason for the cargo ship being parked at the riverbank?"
4. Provide four options (A, B, C, D), where: - One option should be the most plausible cause, based on visual cues from the image. - The other three options are reasonable but incorrect distractors.
5. Make sure the answer to the question depends on specific visual cues in the image. The answer must not rely on commonsense knowledge or external factors not shown in the image.

Output format (JSON only, no extra text):
{
"question": "<your detailed question including object description and location>",
"options": { "A": "...", "B": "...", "C": "...", "D": "..." },
"answer": "A|B|C|D"
}

Figure 22: Prompt used to construct single-turn anomaly-detection questions.

G.9 MULTI-TURN ANOMALY PROMPT

Category: Anomaly Detection and Interpretation (Three-round Dialogue)
Model: GPT-5 thinking
Prompt: You are a dataset constructor for Remote Sensing VQA. Design a three-round dialogue (user ↔ assistant) focused on anomaly detection and interpretation for a salient object or region in the satellite image.
Requirements:

1. Inspect the image and select one object or region with a visually distinctive or unusual condition that is suitable for anomaly analysis. - The selected object or region should be clearly identifiable and in a distinct position in the image. - Try to choose relatively unique objects and avoid those with many similar counterparts in the picture. - Try to choose objects that can be observed in detail at the current resolution, whose detail cannot be detected at lower resolution.
2. First round (localization). Provide a detailed description of the object or region, including its distinguishing features and approximate location (e.g., top-left, right side). - The description must ensure that the object is unique and cannot simultaneously refer to multiple objects in the image. - The question should be: "Where is the [object/region] located in the image?" (use this wording). - The description of the [object/region] in the first-round question must be extremely detailed in order to distinguish it from other similar objects. - The description should contain the characteristics of the object and indicate its relationship with surrounding objects to make the description more detailed. - The description must be more than five words and fewer than 50 words, incorporating the inherent characteristics of the object (limited to describing color, shape, size, and orientation. It cannot describe the state) and the spatial relationships with its surrounding objects. - Avoid overly simple descriptions such as "the cargo ship". - The answer options should be the object/region's position in the image. They should only describe the position in the picture without mentioning other contents, such as the relationship with other objects in the answer. Examples: "The left side of the image." "Avoid answers such as: "The left side of the image, near the river bank."
3. Second round (anomaly identification). Identify the unusual visual condition of the selected object or region using evidence visible in the image. The question should distinguish the observed anomaly from normal conditions without asking about future changes.
   - Question example: "What unusual condition is visible in the selected [object/region] compared with its surroundings?"
   - Option examples: A. No visible anomaly is present. B. The object or region shows an unusual color, structure, or condition supported by the image. C. A different condition that is not supported by the visible evidence. D. A normal appearance inconsistent with the highlighted visual cues.
   - The correct option must describe the observed anomaly itself rather than a possible future event or its cause.
4. Third round (anomaly interpretation). Based on the anomaly identified in the second round, determine the most plausible explanation supported by the visible context.
   - Question: "What is the most plausible explanation for the observed anomaly in the selected [object/region]?" - Options: A. [Most plausible reason based on visible cues]. B. [Plausible distractor 1]. C. [Plausible distractor 2]. D. [Plausible distractor 3].
   - This round should focus on interpreting the current anomaly from visible cues rather than predicting a future state.
   - The answer must follow from the localization and anomaly evidence established in the previous rounds, not unsupported external knowledge.

The assistant responds only with the correct letter and option text ("<letter>. <option text>").
Output format (JSON only, no extra text):

```
{
"question_1": {
"user_question": "Where is the [object/region] located in the image?",
"options": { "A": "...", "B": "...", "C": "...", "D": "..." },
"correct_answer": "<letter>. <correct option text>"
},
"question_2": {
"user_question": "What unusual condition is visible in the selected object/region compared with its surroundings?",
"options": { "A": "...", "B": "...", "C": "...", "D": "..." },
"correct_answer": "<letter>. <correct option text>"
},
"question_3": {
"user_question": "What is the most plausible explanation for the observed anomaly in the selected object/region?",
"options": { "A": "...", "B": "...", "C": "...", "D": "..." },
"correct_answer": "<letter>. <correct option text>"
}
}
```

Figure 23: Prompt for multi-turn anomaly detection through localization, anomaly identification, and visually grounded interpretation.

G.10 FUTURE PREDICTION PROMPTS

Category: Future Prediction (Two-Image Temporal Comparison)
Model: GPT-5 thinking
Prompt: You are a dataset constructor for Remote Sensing VQA. Two satellite images of the same location at different times are provided (Image A = earlier year, Image B = later year). Design one multiple-choice question that requires comparing these images to predict the object's future state.
Requirements:

1. Carefully study both images and identify one salient object or region that is clearly visible in both Image A and Image B.
   - Choose an object or region whose change over time is visually noticeable and subtle. Do not rely on commonly known facts. The change must be clearly observed in the images themselves.
   - Avoid selecting objects that appear frequently or that are trivial to identify without carefully inspecting the images.
2. Compose a question that:
   - references the chosen object or region using its role or approximate location, explicitly naming "Image A" and "Image B", but without explicitly describing visual features such as colors, shapes, or textures.
   - asks about the object's most plausible future state, based on visual changes observed between the two images.
   - instructs the responder to base the answer on visual cues from both images, not on external knowledge.
3. Provide four answer options (A, B, C, D):
   - one option must describe the most plausible future outcome, grounded in the visual changes observed in the two images.
   - the other three options should be reasonable but incorrect distractors, focusing on future states or trends that are not supported by the visual cues.
   - keep the options concise, focusing on future predictions, while embedding visual observations that support the choice.
   - the correct answer should incorporate visual cues from both images and explain the future prediction.
     Example: "The observed decrease in green space near the road and the increase in settlements suggest that the settlements are likely to expand further."
4. Ensure the answer depends on contrasting Image A and Image B. The question should only be answerable by directly comparing the two images, relying on subtle visual differences to infer future predictions.
5. Output must be valid JSON only:

{
"question": "<future prediction question including visual changes from both images>",
"options": {
"A": "<future state option based on visual cues>",
"B": "<future state option>",
"C": "<future state option>",
"D": "<future state option>"
},
"answer": "A|B|C|D"
}

Figure 24: Prompt used for future prediction from two-image temporal comparisons.

H Evaluation and Captioning Prompts

H.1 Open-ended Answer Evaluation

Category: Answer Evaluation (Open-Ended VQA, No Options)
Model: Qwen3-8B
Prompt: You are an expert judge for open-ended Visual Question Answering (no options). Treat the reference answer as image-derived ground truth. Avoid rewarding unsupported specifics. However, minor non-contradictory elaborations are acceptable and should not be harshly penalized.

Assign a score from 1–100 based on:

1. Correctness/consistency with the reference (most important).
2. Usefulness/completeness: directly answers the question and includes necessary attributes/units when relevant.
3. Hallucinations/fabrication: penalize only when details are unsupported *and* material to the answer. Harmless extra context is neutral.

Lenient matching guidelines:

- Semantic equivalence counts as correct: synonyms, common paraphrases, close color groups (e.g., teal $\approx$ blue-green and maroon $\approx$ dark red), and reasonable directional mappings (top $\approx$ upper and left $\approx$ west when consistent with image coordinates).
- Numeric/count tolerance (lenient): if the reference is an integer n, allow ± 1 when $n \leq 10$. When $n > 10$, allow approximately $\pm 20\%$ rounding.
- Ranges/approximations: answers that fall within a sensible range around the reference are acceptable.
- Units: correct conversions or obvious/implicit units are acceptable. Missing a unit is a minor deduction unless it changes meaning.
- Yes/No questions: must match the reference. Hedged but correct answers (e.g., "likely yes") are acceptable.
- If the reference is "unknown/indeterminate/not inferable from the image", answers acknowledging uncertainty are preferred.
- Do not over-penalize style or grammar. Focus on factual content and alignment with the reference.

Suggested (non-binding) score bands:

- 80–100: semantically equivalent and precise, with no material hallucinations and all needed info present.
- 70–79: mostly correct, with minor omissions or small deviations and no contradictions.
- 60–69: partially correct, capturing some key information or a close numeric value with minor issues.
- 30–59: largely incorrect, contradicts reference, or contains clear/meaningful hallucinations.
- 1–29: empty, off-topic, unacceptable, or almost entirely incorrect content.

Question: {q}
Reference answer (treat as image ground truth): {ref}
Model response: {ans}

Return exactly one line with no additional text:
`score: <1-100>`

Figure 25: Qwen3-8B prompt used to score open-ended VQA responses on a 1–100 scale.

H.2 GLOBAL SCENE CAPTIONING

Category: Global Scene Description (Single Aerial/Satellite Image)
Model: GPT-5 thinking
Prompt: Given a high-resolution aerial/satellite image covering a large area with urban, rural, industrial, transport, and natural elements, produce a detailed scene description following the rules below.

1. Overall Description: Write one sentence summarizing the dominant land-use/landform mix, spatial layout, and main functions. Mention the relative positions and roles of key regions such as urban areas, farmland, industrial zones, transport corridors, ports, and green spaces (e.g., coastal layout, parallel belts along a river, clustered urban core with surrounding agriculture).

2. Subregion Description: Describe the upper-left, upper-right, lower-left, and lower-right quadrants. For each quadrant, use 1–3 sentences to state the main objects, their spatial organization, and the dominant function or usage (e.g., irrigated agriculture, port logistics, dense residential core, mixed industry/residential). Explicitly mention relationships such as "roads extending along the coastline", "residential blocks separated from fields by a highway", or "factories clustered beside a river". Use concrete phrases like "multiple factories", "a major highway", or "several residential complexes" and avoid vague terms such as "many" or "some".

3. Detailed Description Requirements: *Object classification:* Cover both natural landscapes (rivers, mountains, lakes, forests, wetlands) and man-made structures (roads, bridges, railways, airports, ports, industrial parks, residential areas, commercial zones).
Spatial relationships: Clearly describe how objects are arranged, such as roads extending along a coastline, buildings parallel to a railway, grid-like intersections of streets and blocks, houses scattered on hilly terrain, or factories clustered along a waterfront.
Function/Usage: Infer the use of each area based on visible forms, such as crop-production farmland, city centers combining residential and commercial buildings, industrial zones for production and storage, or ports used as cargo distribution hubs. Prefer concrete expressions like "multiple large factories and several container yards" or "sparsely distributed residential areas" instead of vague quantifiers.

4. Output Language: Provide the description in clear and concise English, focusing on key elements, their spatial layouts, and relationships. Avoid overly general wording and describe what is visibly supported by the scene and its inferred functions.

Example Output (for reference):
Overall Description: A coastal port city occupies the central area, with a busy deep-water terminal along the southern shoreline, a highway running parallel to the coast linking the port and inland industrial belt, and surrounding residential and commercial districts framed by farmland and natural vegetation to the north and west.
Subregion Description: In the upper-left quadrant, rectangular agricultural fields dominate, framed by straight irrigation canals running roughly north–south, with scattered farmhouses along field edges indicating intensive crop cultivation. The upper-right quadrant contains a dense residential and commercial district organized in a grid-like street pattern, with several embedded parks and a large river cutting through the area and connecting the city core to the port. The lower-left quadrant is an industrial zone with large factories laid out in parallel rows, linked by a network of roads and railways to the waterfront and flanked by warehouses along the outer edges of the complex. The lower-right quadrant shows suburban housing clusters along curving local roads, interspersed with green spaces and small parks, while a narrow river meanders through the area and joins the larger water body near the port.

Figure 26: Prompt for global and regional captioning of aerial and satellite imagery.

H.3 QUADRANT-STRUCTURED CAPTIONING

Category: Structured Captioning (Quadrant-Based Scene Description)
Model: GPT-5 thinking
Prompt: You are an expert remote-sensing analyst. For each high-resolution aerial/satellite image, produce a structured caption that strictly follows the template below (do not add extra sections):

Overall Summary:
- [One sentence describing the dominant land-use/landform mix, spatial layout, and primary functions.]

Upper-Left Quadrant:
- [Two to three sentences detailing the key natural/man-made features, their spatial organization, explicit relationships (e.g., "canals frame rectangular paddies"), and inferred usage.]

Upper-Right Quadrant:
- [Same level of detail for this quadrant.]

Lower-Left Quadrant:
- [Same level of detail for this quadrant.]

Lower-Right Quadrant:
- [Same level of detail for this quadrant.]

Object + Usage Breakdown:
- Natural: [List rivers, forests, hills, wetlands, coastlines, etc., plus their roles.]
- Man-made: [List roads, railways, bridges, ports, factories, warehouses, housing forms, etc., plus their roles.]

Guidelines:

1. Mention both natural landscapes and built infrastructure, emphasizing spatial relationships (parallel roads along rivers, clusters of warehouses near docks, farmland separated by tree belts, etc.).
2. For every quadrant, include the dominant function or usage (e.g., commercial logistics, irrigated agriculture, dense residential core, mixed industry/residential).
3. Avoid vague quantifiers such as "some" or "many". Use concrete descriptors like "five elongated warehouses", "dense grid of low-rise housing", or "narrow river meandering east–west".
4. Language must be English, concise but specific, referencing relative positions between objects and quadrants where relevant.
5. Never include placeholder text, apologies, or references to "the prompt" or "the image". Focus solely on observable visual details and inferred usage.

Figure 27: Prompt for quadrant-structured captioning of high-resolution remote-sensing imagery.