ULTRAVR: A DIAGNOSTIC ULTRA-RESOLUTION IMAGE–VQA BENCHMARK FOR EVIDENCE-GROUNDED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have advanced rapidly on visual question answering and multimodal reasoning benchmarks. Yet it remains unclear whether they can reason over ultra-resolution images, where answer-critical evidence may be tiny, subtle, spatially distant, or distributed across regions. Existing evaluations largely report final-answer accuracy, offering limited insight into whether models acquire and integrate the necessary visual evidence. We introduce **UltraVR**, a diagnostic benchmark for evidence-grounded visual reasoning over ultra-resolution images. UltraVR spans four high-value scenarios: Closed-Circuit Television (CCTV) surveillance, remote sensing (RS), whole-slide image (WSI) pathology, and industrial anomaly detection (AD). These domains pose complementary challenges, including fine-grained object grounding in crowded CCTV scenes, long-range spatial comparison in RS, multi-scale evidence navigation in WSI, and subtle irregularity detection in repetitive industrial layouts. Beyond image-question-answer triples, each UltraVR instance includes a structured ground-truth chain of thought with step-level questions, intermediate answers, and reasoning operation labels. These labels decompose reasoning into evidence grounding, local perception, quantification, evidence integration, and decision inference, enabling process-level diagnosis rather than black-box answer scoring. Using UltraVR, we evaluate frontier proprietary and open-weight VLMs and show that current models remain far from reliable on ultra-resolution reasoning. The structured annotations further localize persistent difficulties in grounding and local perception, even with correct preceding context, while downstream inference often recovers when intermediate visual facts are supplied. These findings establish UltraVR as a diagnostic testbed for measuring not only task success, but where models struggle to establish and use visual evidence.

1 INTRODUCTION

Vision-language models (VLMs) have rapidly progressed from image-level perception toward complex visual reasoning. However, many high-value applications require reasoning over ultra-resolution imagery, including public safety and urban monitoring (Feng et al., 2025a), geospatial and environmental intelligence (Liu et al., 2024), biomedical decision support (Lu et al., 2024), and industrial inspection (Jeong et al., 2023). These inputs are not merely larger versions of ordinary VQA images: their answer-critical evidence may be tiny, subtle, spatially distant, or distributed across multiple regions, making coarse image-level understanding insufficient. Solving such tasks requires models to acquire task-relevant evidence from a large image, preserve fine local details, integrate visual facts across regions or scales, and infer the answer from the accumulated evidence. This poses fundamental challenges for current VLM pipelines, which compress ultra-resolution inputs through resizing, tiling, cropping, multi-scale encoding, or token compression (Qwen Team, 2026; Wang et al., 2025c; Google, 2026; GLM-V Team et al., 2025; Wu et al., 2024; Yu et al., 2025; Aneja et al., 2026; Kimi Team et al., 2025). As a result, model errors can arise from multiple stages of the visual-to-decision pipeline, including missed evidence, lost local detail, weak cross-region integration, or incorrect final inference. Final-answer accuracy alone cannot distinguish these sources of failure.

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.

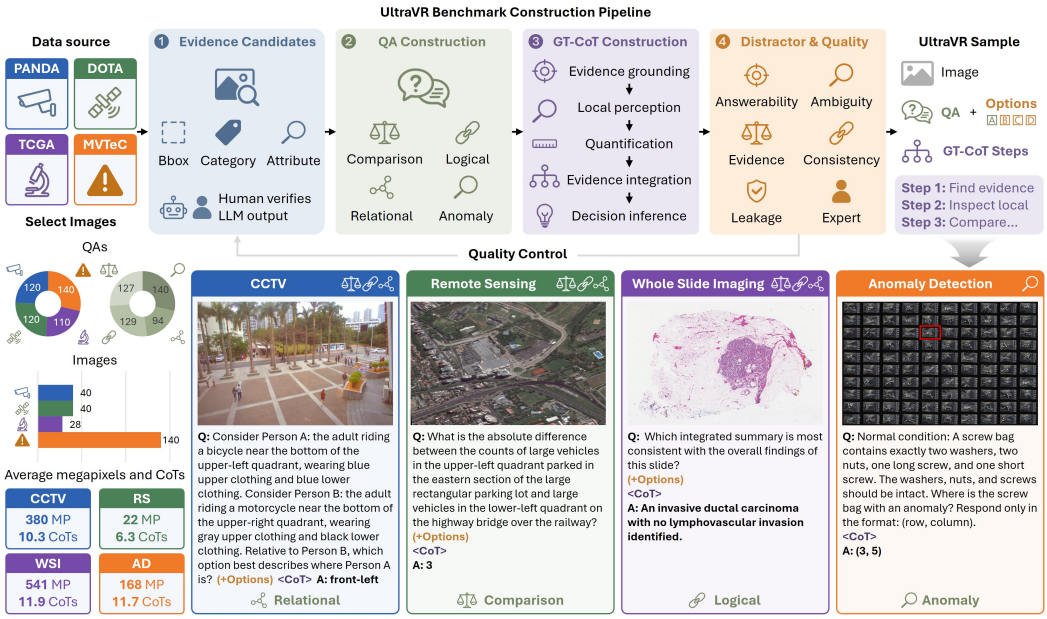


Figure 1: **Overview and construction pipeline of UltraVR.** UltraVR pairs cross-domain ultra-resolution QA with intermediate questions, verified step answers, and operation labels. The construction pipeline combines domain-specific evidence extraction and reasoning templates with checks of visual support, answer uniqueness, and GT-CoT consistency, plus expert review.

Recent benchmarks have made important progress in evaluating either multimodal reasoning or large-image understanding, but they still leave an evaluation gap for traceable reasoning over ultra-resolution images. Visual reasoning benchmarks test spatial, physical, knowledge-intensive, and puzzle-like reasoning (Feng et al., 2025b; Jiang et al., 2026; Li et al., 2025; Zhou et al., 2025), but they are often built on standard images, diagrams, videos, or controlled task settings rather than tasks that require recovering sparse evidence from very large images. High-resolution benchmarks stress image scale and detail preservation (Wu & Xie, 2024; Wang et al., 2024; Ma et al., 2024; Zhang et al., 2025), but many focus on perception, search, grounding, or regional utilization rather than evidence-based inference. Recent large-image reasoning benchmarks move closer to this setting (Wang et al., 2025b; Dang et al., 2025; Li et al., 2026), but their supervision is mainly task-level QA: they show whether a model reaches the correct answer, while providing limited support for tracing which intermediate visual operation fails. A benchmark for ultra-resolution visual reasoning should therefore measure not only whether a model answers correctly, but also whether it can acquire, perceive, integrate, and use the necessary visual evidence.

To address this gap, we introduce **UltraVR** (see Fig. 1), a multi-domain diagnostic benchmark for evidence-grounded visual reasoning over ultra-resolution images. UltraVR spans four domains: CCTV surveillance, remote sensing (RS), whole-slide imaging (WSI), and anomaly detection (AD), which instantiate complementary forms of large-image evidence acquisition and reasoning. CCTV, RS, and WSI are organized around comparison, logical verification, and relational inference, while AD focuses on logical and structural anomaly reasoning.

A central feature of UltraVR is its structured ground-truth chain of thought (GT-CoT). Each question is paired with a sequence of verifiable intermediate steps, where each step contains a step question, a ground-truth step answer, and an operation label. These operation labels abstract the visual-to-decision process into five shared operations: evidence grounding (GND), local perception (PER), quantification (QUA), evidence integration (INT), and decision inference (INF). This annotation design turns each QA item into a traceable diagnostic sample, enabling process-level evaluation beyond black-box final-answer scoring.

Using UltraVR, we benchmark frontier proprietary and open-weight VLMs under three complementary evaluation axes: end-to-end QA, visual-evidence access evaluation, and GT-CoT process-level evaluation with operation-level diagnostics. End-to-end performance remains limited. The visual-

evidence access evaluation shows that local evidence improves accuracy for most models, but substantial errors remain. The process-level evaluation shows that explicit reasoning guidance alone brings limited gains, whereas correcting preceding answers substantially improves task performance. Operation-level analysis clarifies this contrast: grounding and local perception remain difficult even with correct preceding context, while later operations are generally answered more accurately given verified premises. Together, these findings reveal a gap between establishing visual evidence and reasoning from it, highlighting reliable visual-fact acquisition as a persistent challenge within the evaluated decompositions.

Our contributions are: (1) We introduce UltraVR, a multi-domain diagnostic benchmark for evidence-grounded ultra-resolution visual reasoning across CCTV, RS, WSI, and AD. (2) We design a unified reasoning taxonomy covering comparison, logical verification, relational inference, and anomaly reasoning, with domain-specific templates that require visually grounded evidence acquisition rather than text-only priors. (3) We provide structured GT-CoT annotations with verified intermediate answers and operation labels, enabling controlled step-level evaluation across five operation categories. (4) We uncover a recurring gap between visual evidence acquisition and evidence-conditioned reasoning, highlighting grounding and local perception as persistent challenges under the evaluated decompositions.

2 Related Work

2.1 Visual Reasoning Benchmarks

General multimodal benchmarks have substantially advanced the evaluation of VLM reasoning. MME evaluates perception and cognition abilities with manually designed instruction-answer pairs (Fu et al., 2023). MMBench provides a multiple-choice suite for broad multimodal capability assessment (Liu et al., 2023). MMMU evaluates college-level multimodal problems that require domain knowledge and deliberate reasoning (Yue et al., 2023). Recent benchmarks further target more specific forms of visual reasoning. REASONMAP evaluates fine-grained spatial reasoning over transit maps (Feng et al., 2025b). Pix2Fact studies fine-grained visual grounding, multi-hop reasoning, and external knowledge in high-resolution daily-life images (Jiang et al., 2026). QuantiPhy evaluates whether VLMs can infer quantitative physical properties from videos (Li et al., 2025). MIRA studies visual reasoning problems where intermediate visual representations, such as sketches or diagrams, are important for solving the task (Zhou et al., 2025). These works reveal important reasoning limitations of VLMs, but they do not primarily target evidence acquisition, fine-detail preservation, and cross-region integration under ultra-resolution visual inputs. UltraVR complements them by constructing evidence-grounded reasoning tasks over heterogeneous ultra-resolution domains.

2.2 Ultra-Resolution VLM Evaluation

A growing line of work evaluates VLMs under high-resolution inputs. V* introduces guided visual search for locating important details in high-resolution and crowded images (Wu & Xie, 2024). HR-Bench evaluates MLLM perception on 4K and 8K images and shows that downsampling can remove answer-critical information (Wang et al., 2024). GigaGrounding extends visual grounding to gigapixel-level large-scale scenes with large spatial context, scale variation, and multi-hop referring expressions (Ma et al., 2024). HRScene covers diverse high-resolution scenes and studies regional utilization under large-image inputs (Zhang et al., 2025). These benchmarks establish the importance of large-image evaluation, but their main focus is perception, search, grounding, or regional utilization rather than process-level reasoning diagnosis. Recent work moves closer to ultra-resolution reasoning, but differs from UltraVR in how the benchmark instances are constructed. XLRS-Bench constructs a large ultra-high-resolution remote-sensing benchmark by organizing RS images into VQA, captioning, and grounding annotations across perception and reasoning dimensions (Wang et al., 2025b). RSHR-Bench further revisits high-resolution RS reasoning by designing visually dependent QA tasks and applying text-only adversarial filtering with human verification to reduce language-prior shortcuts (Dang et al., 2025). UR-Bench constructs ultra-high-resolution QA over humanistic and natural scenes by composing Micro, Regional, and Global questions with increasing spatial integration requirements, and studies an agent framework for tool-based image exploration (Li et al., 2026). These benchmarks are closest to our setting in image scale and reasoning motivation. UltraVR builds each instance around domain-specific evidence candidates and annotates its visual-

reasoning chain with GT-CoT steps and operation labels, supporting failure localization beyond task-level accuracy.

3 ULTRAVR: A DIAGNOSTIC BENCHMARK

3.1 DESIGN PRINCIPLES AND TASK TAXONOMY

UltraVR is designed as a cross-domain diagnostic benchmark for evidence-grounded visual reasoning over ultra-resolution images, with dataset statistics summarized in Fig. 1 and detailed in Table D.3 (Appendix D.7). The benchmark follows three design principles. First, each question should require visually grounded evidence rather than being answerable from language priors or coarse scene context alone. Second, answer-critical evidence should reflect the core challenges of ultra-resolution imagery: it can be tiny, subtle, spatially distant, or distributed across regions and scales. Third, each instance should support process-level diagnosis through verifiable intermediate visual facts, allowing evaluation beyond final-answer accuracy.

UltraVR defines visual reasoning as locating task-relevant evidence, perceiving local details, integrating facts across regions or scales, and inferring an answer from the resulting evidence.

UltraVR spans four high-value ultra-resolution domains: CCTV, RS, WSI, and AD. These domains instantiate complementary evidence challenges. CCTV requires grounding fine-grained people or vehicles in crowded wide-field scenes. RS requires comparing small objects and spatial structures across long-range aerial layouts. WSI requires navigating multi-scale pathology evidence and integrating diagnostic findings. AD requires detecting subtle violations in repetitive industrial grids. For CCTV, RS, and WSI, questions are organized into three reasoning categories: comparison, logical verification, and relational inference. Comparison requires comparing quantities, attributes, or interpretations across objects or regions; logical verification checks whether a visually grounded condition or rule is satisfied; and relational inference reasons about spatial, structural, or semantic relationships between visual entities. For AD, we use anomaly reasoning, where the model must identify a local logical or structural pattern that violates the global regularity of the grid.

3.2 STRUCTURED GT-COT AND OPERATION LABELS

A central annotation component of UltraVR is the structured ground-truth chain of thought (GT-CoT), which is one verified evidence-to-answer decomposition whose intermediate visual facts jointly support the final answer. This structure makes the evidence supporting an answer explicit and verifiable, enabling systematic evaluation of how models acquire, integrate, and reason over visual evidence. Each GT-CoT trace is represented as a sequence of reasoning steps, and each step contains a *step question*, a *ground-truth step answer*, and an *operation label*. The step question specifies the evidence-seeking or reasoning action, the step answer records the verified intermediate visual fact, and the operation label identifies the role of the step in the visual-to-decision chain. Unlike free-form rationales, this structured format enables process-level evaluation: models can be assessed not only by whether they reach the correct final answer, but also by which intermediate operation fails.

We define five operation types shared across domains. **Evidence grounding (GND)** locates task-relevant image evidence, such as objects, regions, grid cells, ROIs, or landmarks. **Local perception (PER)** recognizes local properties of the acquired evidence, such as attributes, orientations, morphology, defect patterns, or visual states. **Quantification (QUA)** converts visual observations into numerical or ordinal information, such as counts, size comparisons, differences, or ratios. **Evidence integration (INT)** combines evidence across objects, regions, scales, or previous steps. **Decision inference (INF)** derives the task-level conclusion by applying task-specific criteria, domain knowledge, logical conditions, pathology references, anomaly definitions, or answer-option rules to the accumulated evidence. This shared taxonomy abstracts heterogeneous CCTV, RS, WSI, and AD tasks into a common visual-to-decision schema, enabling operation-level evaluation across domains.

3.3 BENCHMARK CONSTRUCTION AND QUALITY CONTROL

UltraVR follows a shared construction protocol across domains, as shown in Fig. 1. For each domain, we first select ultra-resolution images whose relevant evidence is spatially distributed or locally subtle.

We then extract domain-specific evidence candidates, instantiate reasoning-oriented QA templates, annotate GT-CoT traces with operation labels, and construct plausible distractors or constrained answer spaces. The detailed construction pipeline is provided in Appendix A. Detailed construction templates, filtering rules, and domain-specific annotation procedures are provided in Appendix D.

CCTV surveillance. The CCTV split is built from PANDA (Wang et al., 2020), where wide-field surveillance frames contain dense pedestrians, small distant people, occlusions, and long-range spatial layouts. This domain evaluates whether VLMs can ground people or vehicles in crowded scenes, perceive local appearance or orientation cues, and reason about counts or relations across distant regions. We use person and vehicle annotations to form evidence candidates such as semantic object bundles, quadrant-level groups, and pairs of people with visually grounded descriptions. Questions cover comparison, logical verification, and relational inference: models compare counts across bundles, verify count-and-region claims, or infer the relative position of one person with respect to another. The corresponding GT-CoT traces ground the relevant boxes, perceive local cues when needed, quantify regional evidence, and integrate the evidence into the final decision.

Remote sensing. The RS split is constructed from DOTA 1.5 (Xia et al., 2018), which contains aerial imagery with small objects distributed over large geographic layouts. This domain evaluates long-range spatial comparison and fine-grained object grounding: relevant evidence may consist of tiny vehicles, ships, bridges, storage tanks, or sports fields located in different regions of a large image. We parse object annotations into region-level evidence candidates and instantiate questions that compare object counts, verify orientation- or presence-based conditions, or infer spatial relations between selected objects across regions. Unlike ordinary object recognition, these tasks require models to jointly preserve small targets and global layout, making them sensitive to downsampling and patch fragmentation.

Whole-slide imaging. The WSI split uses breast tumor whole-slide images from the TCGA-BRCA cohort (Weinstein et al., 2013), where diagnostic evidence is distributed over gigapixel tissue slides and must be interpreted through pathology morphology. Unlike CCTV and RS, WSI does not provide dense object boxes; instead, evidence is organized through slide-level diagnostic attributes and expert-reviewed regions of interest. Questions ask models to compare clinically confusable interpretations, verify integrated pathology summaries, or reason about relationships between pathological components such as invasive and in-situ regions. The GT-CoT follows a structured diagnostic workflow: low-power screening, ROI selection, local morphology inspection, criterion selection, and evidence-to-conclusion integration. This design evaluates whether VLMs can navigate multi-scale pathology evidence rather than only match a diagnostic label.

Anomaly detection. The AD split simulates ultra-resolution industrial inspection by composing large repetitive grids from normal and anomalous samples extracted from MVTec LOCO AD (Bergmann et al., 2022). Each image contains many visually similar normal cells and a rare cell that violates a logical or structural normality rule. We decompose each category-level normality description into visual conditions and annotate which condition is violated by the anomalous cell. The GT-CoT verifies local or structural conditions and identifies the cell that violates the global regularity.

Quality control. We apply quality-control checks at both the QA and GT-CoT levels. At the QA level, we verify visual answerability, single-answer validity, evidence visibility, distractor plausibility, and text-only leakage. At the GT-CoT level, we check that each intermediate answer is well defined, supported by the specified evidence, and consistent with the final answer. Domain-specific validation is also applied: WSI samples are reviewed by pathology experts, AD samples are checked for unique anomaly placement and normality-rule consistency, and CCTV/RS samples are verified against their evidence annotations. Samples are revised or discarded when the answer is ambiguous, the evidence is not visually supported, or the question can be solved without image inspection. Appendix B.3 details the domain-specific human evaluation of visual answerability and answer uniqueness.

4 Evaluation Protocols

UltraVR defines three complementary evaluation axes. End-to-end QA tests task solving under default inference. Visual-evidence access conditions compare visual inputs to test evidence accessibility

and preservation. GT-CoT process evaluation uses intermediate steps to localize errors across the visual-to-decision chain.

4.1 Models and Metrics

We evaluate both proprietary frontier VLMs and open-weight VLMs. The proprietary models include GPT-5.5, GPT-5.4, Gemini-3.1-Pro, and Seed2.0 (OpenAI, 2026b;a; Google DeepMind, 2026; ByteDance Seed, 2026). The open-weight models include Kimi-K2.6, Qwen3.5, InternVL3.5, Gemma-4, GLM-4.5V, DeepSeek-VL2, MiniCPM-V-4.5, Phi-4-Reasoning-Vision, and Kimi-VL-A3B-Thinking (Moonshot AI, 2026; Qwen Team, 2026; Wang et al., 2025c; Google, 2026; GLM-V Team et al., 2025; Wu et al., 2024; Yu et al., 2025; Aneja et al., 2026; Kimi Team et al., 2025). This model pool covers different parameter scales, visual-token budgets, and high-resolution image processing strategies. Implementation and visual-budget details are provided in Appendices D.1 and D.5.

CCTV, RS, and WSI use four-choice questions; AD requires the anomalous cell's row and column. For end-to-end QA, we report category-level and domain-level accuracy, together with a domain-balanced macro average over CCTV, RS, WSI, and AD. For the end-to-end evaluation, we additionally report text-only accuracy, abstention rate, and the visual gap between the full-image Direct QA setting and the text-only control. For process-level evaluation, we report final-answer accuracy under different GT-CoT conditions, operation-wise accuracy/error, and first-error distributions. Answer extraction is described in Appendix D.3, and all metric definitions are given in Appendix D.4.

Tool-enabled inference. Each model processes images under its visual budget and uses a shared crop-and-zoom tool to select regions from the retained native-resolution image or a previous crop; returned regions pass through the same image-processing pipeline. No GT crop, GT-CoT step, or evidence coordinate is provided. All models share a budget of $B = 6$ tool-use rounds, stopping on a final answer or receiving one final answer request with tools disabled when the budget is exhausted (Appendix D.2).

4.2 Visual-Evidence Access Conditions

We define five visual-input conditions to evaluate how access to visual evidence affects UltraVR performance. **V0: Text-only.** The model receives only the question and answer options. This condition measures language-prior leakage and answer-option bias. **V1: Thumbnail.** The model receives an externally downsampled global view of the full image. This condition tests whether coarse global context is sufficient for the task. **V2: Full image.** The model receives the original full image under its default preprocessing pipeline. This is the standard visual-input condition for end-to-end evaluation. **V3: Local evidence.** The model receives domain-specific local visual evidence without the original full image. WSI provides cropped ROI regions; AD provides a 5×5 cell sheet cropped from the original 10×10 grid, with the anomalous cell randomly placed within the crop; CCTV and RS provide quadrant-level crops rather than fine-grained answer-localizing crops. Crops are extracted from native-resolution source images before upload and model-specific preprocessing. **V4: Full + local evidence.** The model receives both the original full image and the local evidence package. This condition tests whether models can use localized evidence together with global context.

4.3 GT-CoT Process Evaluation

We define six evaluation settings (S0–S5) to examine the difficulty of establishing visual facts and reasoning from them. These settings vary reasoning guidance and execution format, and compare model-predicted with verified preceding answers in step-wise evaluation. **S0: Direct QA.** The model receives the full image, final question, and answer options, and directly predicts the answer. **S1: Generic CoT.** The model receives the same input as S0 plus a generic step-by-step reasoning instruction. **S2: Schema-CoT.** The model receives the full image, final question, answer options, and the task-aware GT-CoT step schema, and is asked to produce all intermediate step answers and the final answer in a single response. **S3: Few-shot GT-CoT.** The model receives in-context examples with complete ground-truth reasoning trajectories before answering the current test sample; no ground-truth intermediate answer for the test sample is revealed. **S4: Pred-Step.** The model answers structured sub-questions sequentially, conditioning each step on its own previous predictions.

Table 1: **End-to-end UltraVR evaluation.** Text-only Acc./Abs. rate gives the accuracy and abstention rate (%) when images are withheld, pooled across domains. Bold marks the best direct score within each model group.

Model	Max Visual Budget	Max Res. (px)	CCTV				RS				WSI				AD			Summary		
			Comp.	Logic	Rel.	Overall	Comp.	Logic	Rel.	Overall	Comp.	Logic	Rel.	Overall	L-Anom.	S-Anom.	Overall	Macro Acc.	Text-only Acc./Abs. rate	Visual Gap
Closed-source frontier VLMs																				
GPT-5.5 (OpenAI, 2026b)	–	3200^2	38.9	**34.2**	**80.0**	**51.8**	27.5	40.0	**65.0**	44.2	36.2	**71.4**	42.9	52.7	**34.8**	25.9	**30.8**	**44.9**	3.9/95.7	+41.0
GPT-5.4 (OpenAI, 2026a)	–	3200^2	41.7	28.9	42.5	37.7	**47.5**	12.5	47.5	35.8	14.9	55.1	**50.0**	37.3	28.8	11.1	20.8	32.9	3.5/94.2	+29.4
Gemini-3.1-Pro (Google DeepMind, 2026)	2240	2260^2	**50.0**	28.9	60.0	46.5	22.5	30.0	51.3	34.5	**68.1**	53.1	14.3	54.5	27.5	**32.4**	30.0	41.4	3.1/94.0	+38.3
w/ tool	2240	2260^2	40.0	30.0	67.5	45.8	35.0	55.0	60.0	50.0	42.6	55.1	21.4	45.5	21.7	31.0	26.4	41.9	3.1/94.0	+38.8
Seed2.0 (ByteDance Seed, 2026)	–	6000^2	25.0	20.0	55.0	33.3	37.5	**55.0**	55.0	**49.2**	55.3	63.3	35.7	**56.4**	15.9	23.9	20.0	39.7	4.3/91.4	+35.4
w/ tool	–	6000^2	40.0	15.0	40.0	31.7	47.5	57.5	47.5	50.8	46.8	59.2	42.9	51.8	26.1	28.2	27.1	40.4	4.3/91.4	+36.1
Open-weight large VLMs																				
Qwen3.5-27B (Qwen Team, 2026)	16384	4096^2	33.3	26.3	35.0	31.6	2.5	**45.0**	**50.0**	**32.5**	38.3	**73.5**	21.4	51.8	**22.7**	**18.5**	**20.8**	34.2	5.7/82.2	+28.5
w/ tool	16384	4096^2	43.3	18.2	76.0	43.2	30.3	62.9	53.6	49.0	30.8	51.4	36.4	40.2	28.1	36.1	32.4	41.2	5.7/82.2	+35.5
Kimi-K2.6 (Moonshot AI, 2026)	4216	2974^2	**47.5**	**35.0**	50.0	44.2	20.0	42.5	35.0	**32.5**	**55.3**	69.4	**64.3**	**62.7**	18.8	11.3	15.0	**38.6**	6.3/84.9	+32.3
w/ tool	4216	2974^2	41.7	12.5	72.5	42.2	40.7	52.5	47.4	47.6	45.5	53.8	57.1	51.2	45.3	28.6	37.9	44.7	6.3/84.9	+38.4
InternVL3.5-38B (Wang et al., 2025c)	32768	5068^2	25.0	21.1	55.0	34.2	15.0	27.5	40.0	27.5	46.8	63.3	21.4	50.9	4.5	1.9	3.3	29.0	13.9/42.4	+15.1
Gemma-4-26B (Google, 2026)	1120	1606^2	41.7	28.9	**72.5**	**48.2**	2.5	20.0	42.5	21.7	34.0	57.1	28.6	43.6	12.1	5.6	9.2	30.7	2.9/95.7	+27.8
GLM-4.5V (GLM-V Team et al., 2025)	12288	3103^2	45.0	25.0	37.5	35.8	**25.0**	32.5	32.5	30.0	38.2	50.3	35.4	41.3	5.8	5.6	5.7	28.2	3.5/76.9	+24.7
Open-weight small VLMs																				
DeepSeek-VL2 (Wu et al., 2024)	2101	1152^2	13.9	26.3	35.0	25.4	**35.0**	**42.5**	55.0	**44.2**	**44.7**	46.9	14.3	41.8	0.0	0.0	0.0	27.8	16.3/36.7	+11.5
MiniCPM-V 4.5 (Yu et al., 2025)	640	1344^2	19.4	**31.6**	57.5	36.8	12.5	32.5	45.0	30.0	38.3	55.1	57.1	48.2	1.4	0.0	0.7	28.9	18.6/14.5	+10.3
Phi-4-RV-15B (Aneja et al., 2026)	3600	960^2	**38.9**	26.3	52.5	39.5	30.0	32.5	**62.5**	41.7	19.1	**69.4**	21.4	41.8	**2.9**	0.0	**1.4**	**31.1**	3.7/92.2	+27.4
Kimi-VL-A3B-Thinking (Kimi Team et al., 2025)	4096	1792^2	22.2	26.3	**72.5**	**41.2**	0.0	20.0	45.0	21.7	34.0	**69.4**	**64.3**	**53.6**	0.0	**1.4**	0.7	29.3	10.0/62.9	+19.3
Mean (w/o tool)			34.0	27.6	54.2	38.9	21.3	33.3	48.2	34.3	40.2	61.3	36.2	49.0	13.5	10.6	12.2	33.6	7.4/74.1	+26.2

S5: GT-Prefix. At step t, all previous intermediate answers are replaced by ground-truth intermediate answers, while the model still has to answer the current step and final question.

5 Benchmark Results and Diagnostic Analysis

5.1 End-to-End UltraVR Evaluation

The end-to-end evaluation uses S0/V2 to assess whether models can acquire visual evidence and produce the correct answer under their default inference pipeline, without local evidence packages or reasoning scaffolds.

Table 1 reports domain-level accuracy and domain-balanced macro accuracy. For Gemini-3.1-Pro, Qwen3.5-27B, Kimi-K2.6, and Seed2.0, secondary rows also report tool-enabled inference configurations with iterative image crops, detailed in Appendix D.2. We also include the text-only control V0, where the image is removed and only the question and answer options are provided. The abstention rate measures the fraction of V0 responses in which the model states that the answer cannot be determined without the image. The visual gap is each row's macro accuracy minus its model's pooled text-only accuracy.

Observations. UltraVR remains far from saturated by current VLMs. Under the standard S0/V2 setting, the strongest model, GPT-5.5, reaches only 44.9% macro accuracy, and the average across the 13 direct baselines is 33.6%, indicating substantial headroom for ultra-resolution visual reasoning. Appendix D.7 reports dataset statistics, image-clustered bootstrap confidence intervals, and ranking stability under subsampling. The text-only control further supports the visual dependence of UltraVR: removing the image reduces the mean accuracy to 7.4%, while stronger instruction-following models often abstain when visual evidence is unavailable. Performance also varies substantially across domains and reasoning categories, and larger visual budgets or higher maximum resolutions do not consistently predict better accuracy. These results establish UltraVR as a challenging and visually grounded benchmark, and motivate the following diagnostics on visual-evidence access and process-level reasoning behavior.

5.2 Visual-Evidence Access Evaluation

The visual-evidence access evaluation tests whether errors arise from limited access to answer-critical evidence or unreliable use of evidence once surfaced. Table 2 compares V1–V4, defined in Section 4.2.

Observations. First, V2 improves over V1 on average, but gains vary across models; Seed2.0 performs better with the thumbnail. Second, V3 is the strongest visual-input condition for most models and reaches 37.2% mean accuracy, suggesting that default full-image inference is limited by

Table 2: **Visual-evidence access evaluation.** We report accuracy (%) under the visual-input conditions V1–V4 defined in Section 4.2. The diagnostic gaps are $\Delta_{\text{res}} = \text{V2} - \text{V1}$, $\Delta_{\text{access}} = \text{V3} - \text{V2}$, and $\Delta_{\text{aug}} = \text{V4} - \text{V2}$.

Type	Model	Accuracy				Diagnostic gap		
		Thumbnail (V1)	Full image (V2)	Local evidence (V3)	Full + local evidence (V4)	Δ_{res}	Δ_{access}	Δ_{aug}
Closed	GPT-5.5 (OpenAI, 2026b)	36.0	44.9	**50.1**	43.0	+8.9	+5.2	-1.9
	GPT-5.4 (OpenAI, 2026a)	29.2	32.8	**38.4**	33.7	+3.6	+5.6	+0.9
	Gemini-3.1-Pro (Google DeepMind, 2026)	36.7	41.4	**48.8**	39.4	+4.7	+7.4	-2.0
	Seed2.0 (ByteDance Seed, 2026)	41.4	39.7	**51.2**	45.8	-1.7	+11.5	+6.1
Open-Large	Qwen3.5-27B (Qwen Team, 2026)	24.5	33.8	**37.8**	24.7	+9.3	+4.0	-9.1
	Kimi-K2.6 (Moonshot AI, 2026)	33.2	38.6	**39.5**	38.8	+5.4	+0.9	+0.2
	InternVL3.5-38B (Wang et al., 2025c)	28.2	28.4	27.6	**29.0**	+0.2	-0.8	+0.6
	Gemma-4-26B (Google, 2026)	29.4	30.2	**38.0**	29.8	+0.8	+7.8	-0.4
	GLM-4.5V (GLM-V Team et al., 2025)	22.7	27.3	**30.8**	24.5	+4.6	+3.5	-2.8
Open-Small	DeepSeek-VL2 (Wu et al., 2024)	27.8	26.4	**29.4**	26.1	-1.4	+3.0	-0.3
	MiniCPM-V 4.5 (Yu et al., 2025)	21.8	27.3	**31.2**	30.4	+5.5	+3.9	+3.1
	Phi-4-RV-15B (Aneja et al., 2026)	27.1	29.5	**30.2**	30.0	+2.4	+0.7	+0.5
	Kimi-VL-A3B-Thinking (Kimi Team et al., 2025)	26.3	27.5	**30.4**	25.5	+1.2	+2.9	-2.0
Mean		29.6	32.9	37.2	32.4	+3.3	+4.3	-0.5

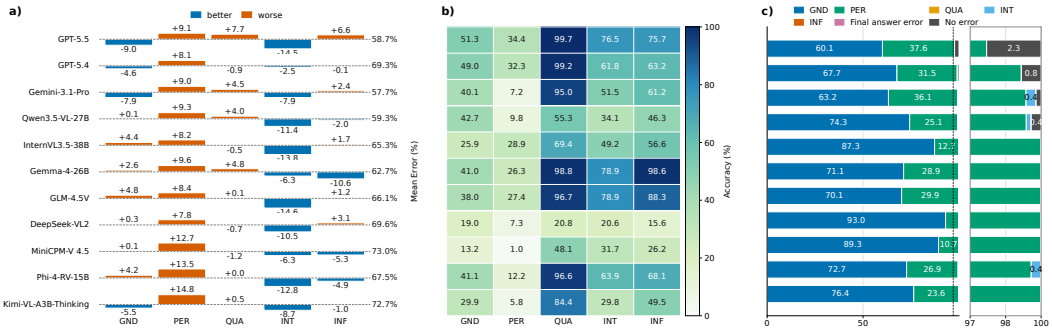


Figure 2: **Operation-level evaluation with GT-CoT.** (a) Under single-response schema completion (S2), each bar shows an operation's error deviation from the model's mean step error. (b) S5 reports operation accuracy with verified preceding answers. (c) S4 uses model-predicted prefixes to locate the first error in each sequential trajectory. The inset magnifies rare final-answer-only and no-error cases.

evidence access, compression, or retrieval. Third, V4 improves over V2 for Seed2.0 but does not consistently outperform V2 or V3 across models, indicating that simply adding local evidence to the full image is not a robust solution. Overall, the visual-evidence access evaluation shows that localized evidence reduces, but does not eliminate, the ultra-resolution reasoning bottleneck. Appendix D.6 further discusses why adding the full image in V4 can reduce accuracy relative to local evidence alone (V3) and whether few-shot GT-CoT demonstrations help.

5.3 Process-Level Evaluation with GT-CoT

Reasoning from visual facts and establishing those facts from an image pose different demands. Within the reference decompositions, we examine whether explicit reasoning guidance improves task solving, how predictions change when earlier answers are corrected, and which operations remain difficult under correct preceding context. Table 3 presents the protocol-level comparisons, while Figure 2 provides the operation-level diagnostics.

Protocol-level evaluation. Table 3 compares accuracy across S0–S5. S1 is slightly below S0 on average, suggesting that a generic "think step by step" instruction does not reliably improve ultra-resolution reasoning. S2 averages 31.8% accuracy, while S3 remains close to S0 (32.8% versus 32.9%), indicating that explicit GT-CoT schemas and few-shot GT trajectories do not consistently improve over direct prediction. We therefore use S2 as the canonical schema-based setting for operation-level comparison. S4 is only slightly above S2, showing that sequential execution provides limited benefit when later steps depend on previous model-predicted states. In contrast, S5 reaches 82.1% average accuracy, 49.7 percentage points above S4. Unlike the prompting and execution

Table 3: **GT-CoT process-level evaluation.** We report accuracy (%) under the evaluation settings S0–S5 defined in Section 4.3. The diagnostic gaps are $\Delta_{\text{text}} = \text{S1} - \text{S0}$, $\Delta_{\text{schema}} = \text{S2} - \text{S0}$, $\Delta_{\text{iter}} = \text{S4} - \text{S2}$, and $\Delta_{\text{GT}} = \text{S5} - \text{S4}$.

Type	Model	Direct QA (S0)	Generic CoT (S1)	Schema-CoT (S2)	Few-shot GT-CoT (S3)	Pred-Step (S4)	GT-Prefix (S5)	Δ_{text}	Δ_{schema}	Δ_{iter}	Δ_{GT}
Closed	GPT-5.5 (OpenAI, 2026b)	44.9	45.3	42.7	47.2	38.8	89.2	+0.4	-2.2	-3.9	+50.4
	GPT-5.4 (OpenAI, 2026a)	32.8	34.3	26.5	30.0	28.4	86.0	+1.5	-6.3	+1.9	+57.6
	Gemini-3.1-Pro (Google DeepMind, 2026)	41.4	39.1	45.5	34.1	41.5	89.6	-2.3	+4.1	-4.0	+48.1
	Seed2.0 (ByteDance Seed, 2026)	39.7	39.4	38.2	44.1	38.9	87.7	-0.3	-1.5	+0.7	+48.8
Open-Large	Qwen3.5-27B (Qwen Team, 2026)	33.8	39.2	40.5	44.2	41.4	53.6	+5.4	+6.7	+0.9	+12.2
	Kimi-K2.6 (Moonshot AI, 2026)	38.6	36.6	32.3	34.2	35.0	87.8	-2.0	-6.3	+2.7	+52.9
	InternVL3.5-38B (Wang et al., 2025c)	28.4	31.7	31.7	34.5	27.2	80.8	+3.3	+3.3	-4.5	+53.6
	Gemma-4-26B (Google, 2026)	30.2	25.6	31.2	32.8	27.2	88.1	-4.6	+1.0	-4.0	+60.9
	GLM-4.5V (GLM-V Team et al., 2025)	27.3	27.5	28.8	30.1	24.3	85.4	+0.2	+1.5	-4.5	+61.1
Open-Small	DeepSeek-VL2 (Wu et al., 2024)	26.4	26.2	23.0	18.2	25.0	72.3	-0.2	-3.4	+2.0	+47.3
	MiniCPM-V 4.5 (Yu et al., 2025)	27.3	23.8	20.9	25.9	27.5	78.7	-3.5	-6.4	+6.6	+51.2
	Phi-4-RV-15B (Aneja et al., 2026)	29.5	26.0	27.3	27.5	29.3	87.4	-3.5	-2.2	+2.0	+58.1
	Kimi-VL-A3B-Thinking (Kimi Team et al., 2025)	27.5	25.8	25.1	23.6	36.3	80.2	-1.7	-2.4	+11.2	+43.9
Mean		32.9	32.3	31.8	32.8	32.4	82.1	-0.6	-1.1	+0.5	+49.7

changes tested in S1–S4, S5 replaces model-predicted preceding answers with verified ones while retaining the same reference decomposition as S4. The key distinction is therefore between specifying a reasoning procedure and supplying reliable intermediate information: the former yields limited average gains in the tested settings, whereas the latter substantially changes task performance. We next examine which operations remain difficult even when preceding answers are correct.

Operation-level evaluation. The GT-CoT annotations further allow UltraVR to localize difficulties across the operations in each reference decomposition, as shown in Figure 2. We first analyze S2, where the model completes the full GT-CoT schema in a single response. Figure 2(a) reports relative operation difficulty by comparing each model's operation error with its own mean step error. PER is consistently harder than the model average, while later operations such as INT are often easier, indicating that explicit decomposition alone does not ensure accurate local perception. We then analyze S5, where verified preceding answers allow us to assess each current operation under correct prior context. Figure 2(b) shows that GND and especially PER remain difficult under this condition, whereas downstream operations such as QUA, INT, and INF are generally more accurate given verified premises. Finally, Figure 2(c) analyzes S4, where the model answers the intermediate questions sequentially with its own predicted prefixes. Most trajectories first fail at GND or PER, with few first errors in later operations. Together, these results reveal a gap between establishing visual evidence and reasoning from supplied premises. Grounding and local perception remain challenging even with correct preceding context, while downstream operations are generally answered more accurately under verified premises. This pattern highlights reliable visual-fact acquisition as a persistent challenge within the evaluated decompositions. Appendix D.8 assesses the stability of this diagnosis by evaluating alternative reasoning paths with different visual cues, abstractions, and reasoning orders for the same tasks.

6 Discussion

UltraVR diagnoses failures in ultra-resolution visual reasoning beyond final-answer accuracy. Our structured diagnostics show that stronger performance on downstream operations can coexist with persistent weaknesses in grounding and local perception, even when preceding answers are correct. This gap makes reliable evidence acquisition a key target for improving ultra-resolution reasoning, alongside the ability to reason from established facts. UltraVR provides a framework for evaluating progress on both challenges under verified evidence-to-answer decompositions.

7 Limitations

UltraVR covers four domains using selected templates and constrained answer formats. Its process-level findings concern verified decompositions, whose choice can affect operation difficulty. The alternative-path evaluation tests this sensitivity on selected evidence routes. Broader domains, open-ended responses, and more diverse decompositions remain for future evaluation.

AI USE STATEMENT

In this work, we used AI tools for polishing writing and assisting dataset annotation, including the construction of candidate questions and structured GT-CoT annotations from source annotations and task-specific templates. Beyond this annotation assistance, we have not used AI tools to support retrieval and discovery, to support research ideation and execution, or to prove mathematical claims.

REPRODUCIBILITY STATEMENT

We provide the anonymous dataset on Kaggle and the anonymous evaluation code. Appendices A–C describe domain-specific benchmark construction, quality control and human evaluation, and the definitions and assignments of reasoning operations. Appendices D–F provide implementation and evaluation details, visual-budget configurations, supplementary diagnostic and stability analyses, representative examples, and the prompts used for GT-CoT annotation construction.

REFERENCES

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15B technical report. *arXiv preprint arXiv:2603.03975*, 2026.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 2022.

ByteDance Seed. Seed2.0. Model documentation, 2026. URL https://seed.bytedance.com/en/seed2.

Yunkai Dang, Meiyi Zhu, Donghao Wang, Yizhuo Zhang, Jiacheng Yang, Qi Fan, Yuekun Yang, Wenbin Li, Feng Miao, and Yang Gao. A benchmark for ultra-high-resolution remote sensing mllms. *arXiv preprint arXiv:2512.17319*, 2025.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. Urbanllava: A multi-modal large language model for urban intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6209–6219, 2025a.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025b.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Qi Ji, Junhui Ji, et al. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Google. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed: 2026-05-06.

Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Accessed: 2026-05-06.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19606–19616, 2023. doi: 10.1109/CVPR52729.2023.01878.

Yifan Jiang, Cong Zhang, Bofei Zhang, Yifan Yang, Bingzhang Wang, and Yew-Soon Ong. From pixels to facts: Benchmarking multi-hop reasoning for fine-grained visual fact checking. *arXiv preprint arXiv:2602.00593*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. *arXiv preprint arXiv:2512.19526*, 2025.

Siqi Li, Xinyu Cai, Jianbiao Mei, Nianchen Deng, Pinlong Cai, Licheng Wen, Yufan Shen, Xuemeng Yang, Botian Shi, and Yong Liu. Ur-bench: A benchmark for multi-hop reasoning over ultra-high-resolution images. *arXiv preprint arXiv:2601.08748*, 2026.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 2024. doi: 10.1038/s41591-024-02856-4.

Tao Ma, Bing Bai, Haozhe Lin, Heyuan Wang, Yu Wang, Lin Luo, and Lu Fang. When visual grounding meets gigapixel-level large-scale scenes: Benchmark and approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Moonshot AI. Kimi-K2.6. Model card, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a. Accessed: 2026-05-06.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026b. Accessed: 2026-05-06.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. Accessed: 2026-05-06.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10599–10609, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tian_Identifying_and_Mitigating_Position_Bias_of_Multi-image_Vision-Language_Models_CVPR_2025_paper.html.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9cf6139382f98623d08cc595622f3fb1-Abstract-Conference.html.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

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, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint arXiv:2408.15556*, 2024.

Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David J. Brady, Qionghai Dai, and Lu Fang. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3268–3278, 2020.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a264726ebd222124514a32bf0143b83d-Abstract-Conference.html.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Yusen Zhang, Wenliang Zheng, Aashrith Madasu, Peng Shi, Ryo Kamoi, Hao Zhou, Zhuoyang Zou, Shu Zhao, Sarkar Snigdha Sarathi Das, Vipul Gupta, Xiaoxin Lu, Nan Zhang, Ranran Haoran Zhang, Avitej Iyer, Renze Lou, Wenpeng Yin, and Rui Zhang. Hrscene: How far are vlms from effective high-resolution image understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Yejin Choi, James Zou, Chaorui Deng, Shen Yan, Haoqi Fan, Cihang Xie, Huaxiu Yao, and Qinghao Ye. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought. *arXiv preprint arXiv:2511.02779*, 2025.