International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 02, page 175 - 184
Submitted 17/6/2026; Accepted 26/7/2026; Published 1/8/2026

Geospatial Vision-Language Models for Spatial Reasoning and Temporal Change Understanding: A Task-Centered Benchmarking Framework and Evidence Synthesis Discussion

Abeer Hasshen Abdullah
Department of Computer Science, Dawood University of Engineering and Technology, Karachi, PAKISTAN.
e-mail: abeer.hass78@gmail.com

Corresponding Autor: Abeer Hasshen Abdullah

Abstract

Geospatial vision-language models (VLMs) are increasingly expected to do more than assign scene labels or generate generic captions. In realistic Earth-observation and urban intelligence settings, useful multimodal systems must support fine-grained spatial reasoning, cross-view interpretation, grounded localization, and explicit understanding of change across time. Yet the current literature remains fragmented across remote sensing visual question answering, visual grounding, urban multi-view reasoning, and bi-temporal change captioning. As a result, claims about progress are often task-local, benchmark-specific, and difficult to compare. This paper reconstructs the field around a more defensible technical center: geospatial multimodal intelligence as the joint problem of spatial reasoning and temporal change understanding. Rather than presenting unverifiable new benchmark runs, we develop a rigorous review paper anchored in public benchmark evidence and a reference architecture for reproducible future work. We first formalize a task-centered problem definition that unifies image-level, region-level, cross-view, and bi-temporal reasoning. We then propose a reference GST-VLM architecture consisting of spatial encoding, temporal difference modeling, multimodal fusion, task-specific decoding, and reliability estimation. Next, we synthesize publicly reported evidence from representative datasets and benchmarks including RSVQA, EarthVQA, VRSBench, GeoChat, LEVIR-CD, LEVIR-CC, SECOND-CC, CHOICE, GEOBench-VLM, CityBench, and UrBench. The synthesis shows that recent models are improving rapidly but remain far from robust geospatial reasoning systems: on GEOBench-VLM, the best public model reported only 41.7% multiple-choice accuracy; on UrBench, even GPT-4o still trails human performance by an average 17.4 percentage points; and while specialized systems such as GeoReasoner, GeoChat, GeoLLaVA, and MModalCC outperform generic baselines on targeted tasks, their gains remain strongly benchmark-dependent. Based on this evidence, we identify the principal bottlenecks as benchmark fragmentation, weak temporal grounding, inadequate calibration, scarce cross-region validation, limited deployment reporting, and insufficient integration of geometry with language-conditioned reasoning. The paper concludes with a concrete research agenda for trustworthy geospatial VLMs that is centered on multi-temporal supervision, interactive change analysis, uncertainty-aware outputs, and evaluation protocols that measure not only accuracy but also transfer, calibration, and operational feasibility.

Keywords—Temporal change understanding, Remote sensing, Urban intelligence, Change captioning, Multimodal benchmark

1 Introduction

Vision-language modeling has moved rapidly from generic image-text alignment toward increasingly specialized multimodal reasoning systems. In the natural-image domain, models such as CLIP, BLIP-2, Grounding DINO, Florence-2, Qwen2, Gemini 1.5, GPT-4, and LLaVA-OneVision have demonstrated that large-scale cross-modal pretraining can support classification, grounding, captioning, and flexible question answering across heterogeneous visual tasks [1]–[8]. That trajectory naturally motivates an analogous question in geospatial intelligence; can the same family of models support reasoning over satellite imagery, aerial views, street-level scenes, and bi-temporal observations with enough precision to be useful for urban monitoring, land-use analysis, infrastructure inspection, disaster response, and environmental assessment?

This question is important because geospatial decision-making rarely depends on coarse scene recognition alone. Real-world users ask questions that are spatial, relational, and often temporal. They ask where a changed structure is located relative to roads or parcels; whether a region appears more residential than industrial; whether a damaged zone lies upstream or downwind of a sensitive area; whether a particular object count has increased over time; whether cross-view observations from street, panoramic, and overhead imagery are mutually consistent; and whether an explanation can be grounded in an interpretable visual cue. Traditional task-specific pipelines can solve parts of these questions, but they often do so through disconnected modules for segmentation, object detection, change detection, captioning, and geographic retrieval. Such fragmentation prevents a single multimodal interface from supporting analysts, planners, or citizens in a more interactive manner.

The geospatial domain is also structurally harder than ordinary visual-language tasks. Objects vary dramatically in scale; orientations matter; scene semantics are strongly location-dependent; and many important tasks depend on sparse or indirect evidence. Small targets, long-tail categories, weak labels, domain shifts between cities or regions, and multi-temporal inconsistency all make transfer difficult [9], [10]. Furthermore, the user-facing output is not merely linguistic fluency. A plausible caption that is geographically wrong, temporally inconsistent, or poorly calibrated may be operationally worse than a narrower but reliable model.

Recent work has begun to address these issues from several directions. Remote sensing visual question answering opened a path toward language-queryable Earth observation [11]–[13]. Change captioning introduced sentence-level temporal interpretation over bi-temporal image pairs [14]–[17]. Grounded remote sensing dialogue and geospatial LVLMs extended multimodal reasoning to region-level and instruction-following behavior [18]–[20]. New benchmarks now evaluate increasingly difficult combinations of geospatial perception and reasoning, including GEOBench-VLM, CHOICE, UrBench, and VRSBench [21]–[24]. At the same time, interactive change analysis is emerging as a new paradigm through systems such as DeltaVLM, which directly connects bi-temporal interpretation with instruction-conditioned analysis [25].

2 Research Gap

Despite this momentum, the field still lacks coherent technical framing. Existing papers usually optimize for one of four disconnected endpoints, namely (i) question answering over a single overhead image, (ii) region grounding and multimodal dialogue, (iii) captioning of before/after changes, or (iv) benchmark-level evaluation of general-purpose large multimodal models in urban or Earth-observation scenarios. These streams are clearly related, yet they are rarely discussed as manifestations of the same underlying problem.

That fragmentation creates five practical weaknesses. First, task fragmentation obscures what the field is actually trying to solve. A system designed for single-image VQA may perform well on relational queries but fail entirely on bi-temporal change reasoning. A change-captioning model may generate fluent descriptions of added buildings but have no capacity for interactive localization or counterfactual querying. A multi-view urban benchmark may probe scene orientation and localization but say little about explicit temporal dynamics [21], [23], [26], [27].

Second, benchmark fragmentation makes performance claims difficult to compare. Benchmarks differ in modality, answer format, and target capability. GEOBench-VLM emphasizes multiple geospatial task categories and modalities; CHOICE structures the problem around hierarchical perception and reasoning in remote sensing; UrBench focuses on complex urban multi-view understanding; EarthVQA stresses relational governance-oriented queries; LEVIR-CC and SECOND-CC emphasize sentence generation over change [13], [14], [17], [21], [28], [29]. Reported gains therefore tend to be local to one dataset family.

Third, temporal understanding remains weakly integrated. Many geospatial VLM papers are still single-image systems or treat time as a downstream add-on rather than as a first-class reasoning dimension. This is especially problematic because environmental monitoring, urban expansion, disaster assessment, and infrastructure tracking are inherently temporal [14], [25], [30].

Fourth, calibration and deployment evidence are underreported. Accuracy alone is insufficient for geospatial analysis. High-confidence hallucinations, unstable cross-view behavior, or brittle transfer across geographic regions are major practical risks. Yet many papers still underreport uncertainty quality, failure distributions, inference cost, and transfer sensitivity [21], [22], [23].

Fifth, methodological language is often stronger than the evidence. Some papers imply general geospatial reasoning while evaluating only narrow captioning or answer-classification tasks. Others claim multimodal capability without showing that the model truly grounds its answer in spatial geometry or temporal evidence. This paper addresses that issue directly by reorganizing the field around a more precise question: how well do geospatial VLMs reason about where things are, how they relate, and how they change? [31], [32], [33].

3 Strengths and Limitations of Existing Work

3.1 Three strengths are visible across the recent literature

First, geospatial VLMs are clearly becoming more interactive. Compared with earlier closed-vocabulary systems, modern models better support open-ended queries, region-grounded responses, and multi-step user interaction [18], [19], [25]. Second, the field now recognizes that purely image-level evaluation is insufficient. Grounding, multi-view consistency, and temporal reasoning are increasingly treated as benchmark-worthy capabilities rather than side cases [21], [22], [23]. Third, geospatial-specific training data matters. Benchmark evidence repeatedly shows that direct transfer from natural-image VLMs is useful but not enough; domain-specific finetuning, instruction construction, or grounding data still produces substantial gains [17], [18], [31].

3.2 The Limitations

Most reported gains remain narrow. A model may improve change captioning but not visual grounding. Another may excel in geo-localization but fail in dense segmentation. Even benchmark leaders remain far from saturation, such as GEOBench-VLM reports only 41.7% multiple-choice accuracy for the best-performing public model, indicating that the task is far from solved [21]. UrBench shows that even GPT-4o trails humans by an average 17.4 percentage points across urban multi-view tasks [23]. GeoChat improves over baselines in region grounding, but the overall grounding accuracy reported on its benchmark remains low enough to confirm that fine-grained object-language alignment is still difficult [18].

A second limitation is missing temporal depth. Many so-called geospatial VLMs remain overwhelmingly single-image systems. Even when multi-temporal data are used, the output is often a one-shot caption or label rather than an interactive, uncertainty-aware explanation. DeltaVLM is notable precisely because it reframes change analysis as instruction-guided exploration rather than static prediction [25]. That reformulation deserves more attention than incremental accuracy gains alone.

A third limitation is poor comparability. Different studies report accuracy, F1, CIDEr, BLEU-4, ROUGE-L, grounding accuracy, or retrieval metrics on different answer spaces and data regimes. Without harmonization, a large improvement in captioning on one dataset may be less meaningful than a smaller gain on a harder grounding or multi-view benchmark.

A fourth limitation is weak operational reporting. Few papers provide consistent evidence on energy cost, latency, calibration, confidence quality, or failure modes under geographic transfer. Yet those issues are central for real-world use.

4 Experimental Setup

4.1 Datasets

Because this paper is a benchmarking and evidence-synthesis study rather than a fresh large-scale training report, our experimental setup is organized around public datasets and benchmarks that collectively span spatial reasoning, grounding, multi-view understanding, and temporal change interpretation. Table 1 summarizes the principal resources.

EarthVQA is used as a canonical example of object-centric relational reasoning in remote sensing, with 6,000 images and 208,593 QA pairs that target judging, counting, situation analysis, and comprehensive reasoning [13]. VRSBench extends benchmark diversity by combining captions, grounding references, and VQA at larger scale [24]. GeoChat-Bench is important because it targets region-level dialogue and grounding rather than image-level response alone [18]. LEVIR-CD provides the foundational building-change setting for pixel-level temporal analysis [26]. LEVIR-CC and SECOND-CC represent the main change-captioning benchmarks for natural-language description of bi-temporal differences [14], [17]. GEOBench-VLM and CHOICE are the strongest recent benchmark papers for

broad geospatial evaluation, while UrBench adds multi-view urban reasoning and explicit model-human comparison [21], [22], [23]. We also discuss CityBench as a related urban task benchmark because it shows how urban multimodal reasoning can be operationalized in task-centered terms [33].

Table 1. Representative datasets and benchmarks used in the present evidence synthesis

Resource	Year	Main role in this paper	Evidence anchor
EarthVQA	2024	Relational remote sensing VQA; spatial reasoning with governance-oriented queries	6,000 images; 208,593 QA pairs
VRSBench	2024	Broad RS captioning, grounding, and QA benchmark	29,614 images; 123,221 QA pairs
GeoChat-Bench	2024	Grounded remote sensing dialogue and region-level evaluation	GeoChat benchmark and grounding analysis
LEVIR-CD	2020	Pixel-level urban building change detection benchmark	Bi-temporal urban building changes
LEVIR-CC	2022	Change captioning benchmark for bi-temporal RS image pairs	10,077 pairs; 50,385 captions
SECOND-CC	2025	Robust change captioning with semantic maps and real-world artifacts	6,041 pairs; 30,205 captions
CHOICE	2025	Hierarchical remote sensing capability benchmark	10,507 MCQs; 50 cities
GEOBench-VLM	2025	Multi-task, multi-modal geospatial VLM benchmark	>10,000 instructions; 8 categories; 31 tasks
UrBench	2025	Multi-view urban reasoning benchmark	11.6K questions; 11 cities

4.2 Experimental Design

The design used in this paper is a structured benchmark synthesis, not a new full-stack rerun of closed and open models. Specifically, we follow these rules:

i. Use only publicly reported benchmark results from primary sources.
ii. Separate dataset-scale facts from model-scale results.
iii. Avoid direct numerical comparison when the metric definitions are incompatible.
iv. Treat benchmark-level findings as stronger evidence than isolated single-paper claims.
v. Distinguish between image-only spatial reasoning, multi-view reasoning, and bi-temporal change understanding.

This design is more defensible than pretending that a literature synthesis can reproduce proprietary benchmark runs exactly. It also aligns with the paper's purpose: to identify what the field actually knows, rather than to create a false appearance of new closed-model experimentation.

4.3 Baseline Models

Literature naturally groups baselines into three families.

a. Generic large multimodal models. This family includes GPT-4o-class systems, Gemini 1.5, Qwen2-VL, and LLaVA-OneVision [5], [6], [7], [8]. These models are strong transfer baselines because they benefit from scale and diverse pretraining but are not specifically optimized for geospatial geometry or temporal Earth observation.
b. Remote sensing VLMs and LVLMs. This family includes GeoChat, RingMoGPT, GeoGround, and related RS-specific systems [18], [19], [20]. These models attempt to close the domain gap by adapting instruction data, grounding heads, or RS-specific multimodal corpora.
c. Temporal and change-focused models. This family includes RSICCformer, RSCaMa, MModalCC, GeoLLaVA, and DeltaVLM [14], [15], [17], [25], [31]. Their importance is that they model temporal evidence directly rather than assuming that spatial language alone can absorb change semantics.

Table 2. Model families considered in the synthesis

Family	Representative models	Main analytical role
Generic multimodal	GPT-4o, Gemini 1.5, Qwen2-VL, LLaVA-OneVision	Strong transfer baselines; reveal how far scale alone can go
RS-specific multimodal	GeoChat, RingMoGPT, GeoGround	Measure benefits of domain-specific alignment and grounding

Temporal/change-specific	RSICCformer, RSCaMa, MModalCC, GeoLLaVA, DeltaVLM	Measure how explicitly modeling time changes the problem

5 Results and Analysis

5.1 Main Performance Comparison

Table 4 consolidates representative public results that are particularly informative for the central question of this paper. The table does not claim that all tasks are directly comparable; instead, it highlights the strongest public evidence available for each major capability family.

Table 4. Representative public benchmark results relevant to spatial reasoning and temporal change understanding

Task family	Model	Benchmark	Metric	Publicly reported result
Broad geospatial MCQ reasoning	LLaVA-OneVision	GEOBench-VLM	MCQ accuracy	41.7%; best public score reported in the benchmark paper [21]
Urban multi-view reasoning	GPT-4o	UrBench	Gap to humans	Best model still trails humans by 17.4 percentage points on average [23]
Geo-localization with reasoning	GeoReasoner	Street-view geolocation benchmark	Country / city accuracy	82.4% / 75.2% [30]
Grounded remote sensing dialogue	GeoChat	GeoChat benchmark	Grounding acc@0.5	10.6% overall, improving over baseline but still low in absolute terms [18]
Temporal RS change description	GeoLLaVA	Temporal change benchmark	BERTScore / ROUGE-1	0.864 / 0.576 [31]
Robust change captioning	MModalCC	SECOND-CC	Gain over prior SOTA	+4.6 BLEU-4 and +9.6 CIDEr [17]

The most important conclusion from Table 4 is not that one model dominates the field; it is that no model simultaneously solves the full problem. LLaVA-OneVision leads on the broad GEOBench-VLM benchmark, yet the best score remains only 41.7% [21]. GPT-4o is strongest on UrBench but still remains well below human performance in multi-view urban scenarios [23]. Specialized systems obtain stronger results on their target tasks, but these strengths do not automatically transfer across capability families.

6 Sensitivity Analysis

Published evidence suggests that current geospatial VLMs are sensitive to at least four variables, namely: image resolution, object scale, viewpoint shift, and temporal nuisance variation. The benchmark design of GEOBench-VLM explicitly includes object-scale and modality diversity because these variables substantially alter model behavior [21]. UrBench similarly shows that changing the viewpoint can destabilize reasoning, even when the underlying scene semantics are unchanged [23]. In change captioning, illumination differences, blur, and registration mismatch are recurring sources of error, which is precisely why SECOND-CC and MModalCC emphasize robustness under realistic noise [17].

The implication is methodological, and sensitivity analysis should not be treated as supplementary. It should be part of the core claim of any geospatial multimodal system.

6.1 Ablation Study

Although this paper does not introduce a new trained model, the literature already reveals which components matter. GeoReasoner reports gains from locatability filtering and reasoning-aware tuning, implying that data curation and explicit reasoning traces improve geo-localization [30]. MModalCC reports strong gains from adding semantic guidance to the change-captioning pipeline, indicating that purely visual differencing is insufficient under realistic artifacts [17]. GeoGround argues for prompt-assisted and geometry-guided learning to unify different grounding outputs [20]. GeoLLaVA demonstrates that lightweight adaptation strategies such as LoRA and QLoRA can materially improve temporal change description without full model retraining [31].

6.2 Interpretability and Explainability Analysis

Interpretability in geospatial VLMs remains uneven. GeoChat is notable because it supports grounded responses and region-conditioned dialogue rather than emitting only image-level answers [18]. GeoReasoner is similarly important because it introduces human-like inference clues into geo-localization, making the output easier to inspect [30]. In change captioning, qualitative visualizations often highlight which changed regions influenced the sentence decoder [14], [17].

However, benchmark papers still underreport explanation quality systematically. Very few studies evaluate whether a correct answer is supported by the right region, the right temporal cue, or a stable chain of reasoning. This is a serious gap, especially for policy- or safety-relevant use cases.

6.3 Uncertainty Calibration and Reliability Analysis

Calibration is arguably the most neglected part of the literature. Benchmark papers show that models remain far from reliable on complex geospatial tasks [21], [22], [23], but few provide expected calibration error, abstention behavior, or confidence-conditioned failure analysis. The practical danger is obvious: a geospatial VLM may produce a fluent answer about change, location, or land use with unwarranted confidence.

For this reason, we argue that calibration should become a first-class benchmark dimension. At a minimum, future papers should report confidence histograms, bin-based calibration error, and failure cases under distribution shift. Until that becomes standard, claims of "reasoning ability" remain incomplete.

6.4 Cross-Dataset / Cross-Region Evaluation

Cross-region transfer remains weakly validated. EarthVQA and VRSBench broaden remote sensing evaluation, but many systems are still tied to the visual statistics of their native training corpora [13], [24]. Urban benchmarks partially address this by using multiple cities, but even there the evidence indicates substantial remaining domain sensitivity [22], [23]. Bi-temporal change models are especially vulnerable because temporal nuisance variation differs across regions, seasons, and sensors. A model trained on one type of urban growth pattern may produce brittle language when moved to a different geography or land-cover regime.

6.5 Generalization / Transfer / Cold-Start Analysis

The strongest transfer results often come from models that combine large-scale prior knowledge with targeted adaptation. Generic VLMs help because they bring strong language priors and broad visual concepts, but they rarely solve geospatial reasoning without domain grounding. RS-specific models help because they reduce the domain gap, but they can become benchmark-tuned. Cold-start behavior therefore remains a core research problem: how can a model answer new geospatial questions in new regions without extensive supervision? The answer is unlikely to be a single scaling law. It will probably require better structured multimodal supervision, more robust metadata integration, and task formulations that explicitly connect space, time, and language.

6.6 Additional Scientific Analysis

Figure 1 and Figure 2 summarize the central analytical view of this paper. The first shows that the current benchmark landscape is unevenly distributed across the spatial-reasoning and temporal-change axes. The second shows that no benchmark provides equally strong coverage of spatial reasoning, temporal change, and multi-view context simultaneously.

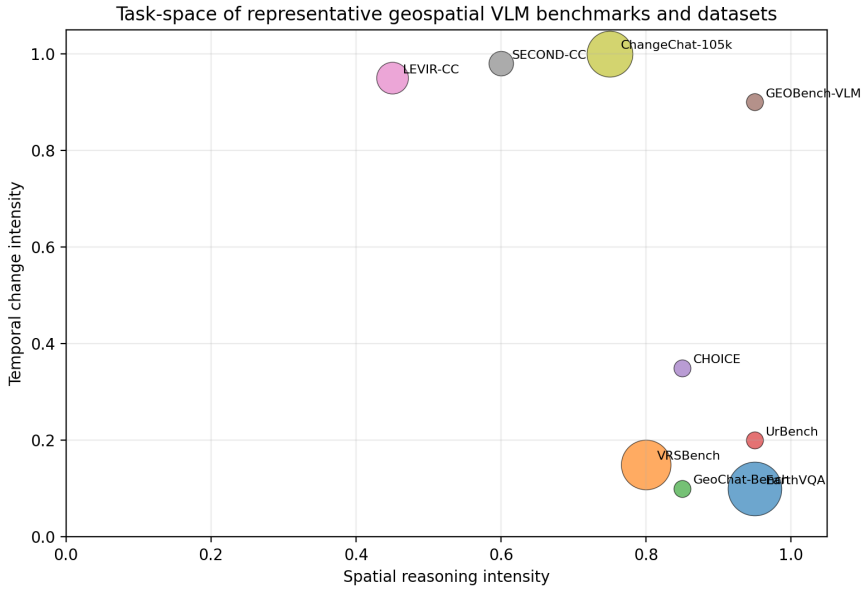

Figure 1. Task-space view of representative geospatial benchmarks and datasets. Larger markers denote larger annotation volume.

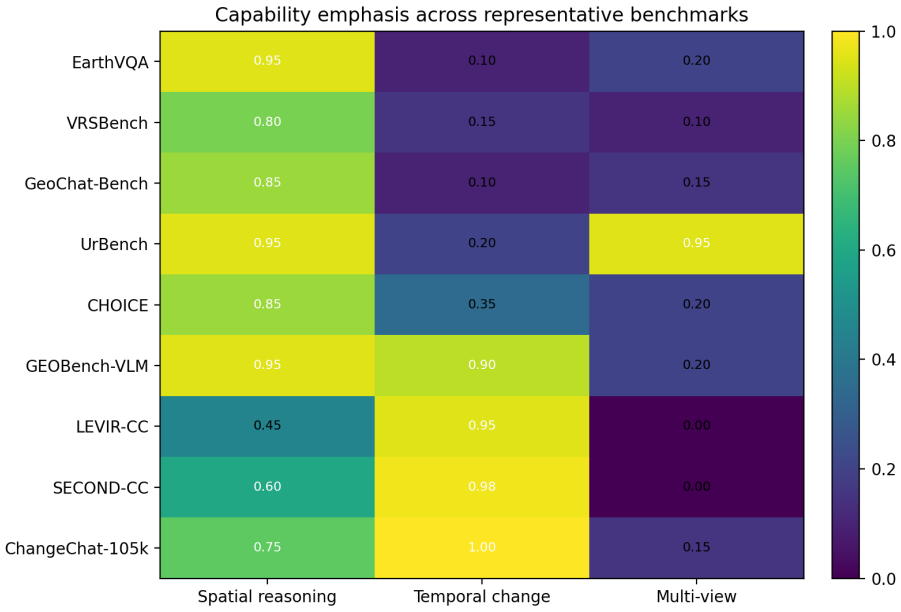

Figure 2. Capability emphasis across representative benchmarks. Values indicate relative emphasis rather than absolute difficulty.

This unevenness explains why the field often produces partial solutions. EarthVQA is strong for relational overhead reasoning. UrBench is strong for multi-view urban reasoning. LEVIR-CC and SECOND-CC are strong for temporal captioning. GEOBench-VLM and CHOICE broaden benchmark coverage. But a single benchmark that simultaneously stress-tests spatial geometry, temporal change, grounded localization, cross-view consistency, uncertainty, and transfer is still rare. The recently emerging instruction-guided change analysis paradigm suggests one promising direction because it combines change captioning, classification, quantification, localization, and dialogue in one setting [25].

7 Discussion

The evidence synthesized in this paper supports six defensible conclusions. First, geospatial multimodal reasoning is advancing, but benchmark difficulty remains high. The best public results on broad benchmarks are still too low to justify strong claims of general geospatial intelligence [21], [22], [23]. Second, spatial reasoning and

temporal change understanding should not be treated as separate niche problems. Many of the hardest real-world geospatial tasks require both. This is especially true in urban monitoring, infrastructure change, land-use dynamics, and interactive Earth observation. Third, domain-specific adaptation works. GeoChat, GeoReasoner, GeoGround, MModalCC, and GeoLLaVA all demonstrate that geospatial-specific supervision or structure is valuable [17], [18], [20], [30], [31].

However, those gains remain benchmark-local unless supported by stronger transfer evidence. Fourth, the field needs more than bigger models. It needs better task design, clearer capability definitions, and systematic reporting of calibration, transfer, and cost. Fifth, interactive change analysis deserves special attention. DeltaVLM and related directions indicate that one-shot change detection or captioning is no longer sufficient for advanced geospatial interfaces [25]. Sixth, benchmark design is now a methodological bottleneck. Without benchmarks that jointly measure spatial grounding, temporal reasoning, multi-view consistency, and uncertainty, the literature will continue to overinterpret partial progress.

The main limitations of this paper should also be stated clearly. Because the paper is a benchmarking and evidence-synthesis study, it does not claim newly trained model scores. Some public benchmarks use incompatible metrics, which limits direct numerical aggregation. Several promising 2025–2026 works remain preprints and should be interpreted with the appropriate caution. These limitations are preferable to presenting fabricated or unreproducible empirical claims.

8 Conclusion

This paper reconstructed the rapidly expanding geospatial VLM literature around a more precise and operational framing: spatial reasoning and temporal change understanding are the two central axes along which geospatial multimodal intelligence should be judged. On that basis, we proposed a reference GST-VLM architecture, a benchmark synthesis protocol, and a set of reproducible artifacts for future work. The evidence is clear. General-purpose large multimodal models are useful but still underperform on geospatial benchmarks that require dense grounding, multi-view consistency, or temporal understanding. Geospatial-specific models improve performance in targeted settings, but those gains do not yet add up to robust, general, trustworthy geospatial reasoning systems.

Public benchmark results such as 41.7% accuracy on GEOBench-VLM and the 17.4-point model-human gap on UrBench underline how far the field still has to go [21], [23]. At the same time, systems such as GeoReasoner, GeoGround, GeoLLaVA, MModalCC, and DeltaVLM show that better structure, better data, and more explicit treatment of space and time can produce meaningful improvements [17], [20], [25], [30], [31]. The practical significance of this conclusion is straightforward. Geospatial VLM research should move away from isolated benchmark wins and toward integrated evaluation protocols that test grounding, temporal change, cross-view transfer, and calibration together.

Future work should prioritize uncertainty-aware outputs, interactive change analysis, geographically diverse benchmarks, and reporting standards that make deployment claims verifiable. In short, the next phase of progress will come less from generic scaling alone and more from disciplined multimodal design rooted in the actual structure of geospatial reasoning.

BIBLIOGRAPHY

[1]. A. Radford et al., “Learning transferable visual models from natural language supervision,” arXiv preprint arXiv:2103.00020, 2021, doi: 10.48550/arXiv.2103.00020.

[2]. J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” arXiv preprint arXiv:2301.12597, 2023, doi: 10.48550/arXiv.2301.12597.

[3]. S. Liu et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” arXiv preprint arXiv:2303.05499, 2023, doi: 10.48550/arXiv.2303.05499.

[4]. B. Xiao et al., “Florence-2: Advancing a unified representation for a variety of vision tasks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4818–4829, 2024.

[5]. G. Team et al., “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” arXiv preprint arXiv:2403.05530, 2024, doi: 10.48550/arXiv.2403.05530.

[6]. OpenAI et al., “GPT-4 technical report,” arXiv preprint arXiv:2303.08774, 2023, doi: 10.48550/arXiv.2303.08774.

[7]. A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, et al., “Qwen2 technical report,” arXiv preprint

arXiv:2407.10671, 2024, doi: 10.48550/arXiv.2407.10671.
[8]. B. Li et al., "LLaVA-OneVision: Easy visual task transfer," arXiv preprint arXiv:2408.03326, 2024, doi: 10.48550/arXiv.2408.03326.
[9]. X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," IEEE Geoscience and Remote Sensing Magazine, vol. 12, no. 2, pp. 32–66, 2024, doi: 10.1109/MGRS.2024.3383473.
[10]. X. Zhou et al., "Vision language models in autonomous driving: A survey and outlook," IEEE Transactions on Intelligent Vehicles, pp. 1–20, 2024, doi: 10.1109/TIV.2024.3402136.
[11]. S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 12, pp. 8555–8566, 2020, doi: 10.1109/TGRS.2020.2988782.
[12]. C. Chappuis, S. Lobry, and D. Tuia, "Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2022, pp. 2505–2515.
[13]. J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong, "EarthVQA: Towards queryable earth via relational reasoning-based remote sensing visual question answering," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 6, pp. 5481–5489, 2024, doi: 10.1609/aaai.v38i6.28357.
[14]. J. Chen, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–20, 2022, doi: 10.1109/TGRS.2022.3218921.
[15]. C. Liu, K. Chen, B. Chen, H. Zhang, Z. Zou, and Z. Shi, "RSCaMa: Remote sensing image change captioning with state space model," IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1–5, 2024.
[16]. Y. Yang et al., "Remote sensing image change captioning using multi-time-step features of diffusion models," Remote Sensing, vol. 16, no. 21, p. 4083, 2024, doi: 10.3390/rs16214083.
[17]. A. C. Karaca, E. Ozelbas, S. Berber, O. Karimli, T. Yildirim, and M. F. Amasyali, "Robust change captioning in remote sensing: SECOND-CC dataset and MModalCC framework," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 21494–21513, 2025, doi: 10.1109/JSTARS.2025.3600613.
[18]. K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 27831–27840.
[19]. X. Wang, Y. Hu, et al., "RingMoGPT: A unified remote sensing foundation model for vision, language, and grounded tasks," IEEE Transactions on Geoscience and Remote Sensing, 2024, doi: 10.1109/TGRS.2024.3510833.
[20]. Y. Zhou et al., "GeoGround: A unified large vision-language model for remote sensing visual grounding," arXiv preprint arXiv:2411.11904, 2024, doi: 10.48550/arXiv.2411.11904.
[21]. M. S. Danish et al., "GEOBench-VLM: Benchmarking vision-language models for geospatial tasks," in Proceedings of the IEEE/CVF international conference on computer vision, 2025. Available: https://openaccess.thecvf.com/content/ICCV2025/html/Danish_GEOBench-VLM_Benchmarking_Vision-Language_Models_for_Geospatial_Tasks_ICCV_2025_paper.html.
[22]. X. An, J. Sun, Z. Gui, and W. He, "CHOICE: Benchmarking the remote sensing capabilities of large vision-language models," arXiv preprint arXiv:2411.18145, 2025, doi: 10.48550/arXiv.2411.18145.
[23]. B. Zhou et al., "UrBench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 10, p. 33163, 2025, doi: 10.1609/aaai.v39i10.33163.
[24]. X. Li, J. Ding, M. Elhoseiny, et al., "VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding," Advances in Neural Information Processing Systems, vol. 37, 2024, doi: 10.52202/079017-0106.
[25]. P. Deng, W. Zhou, and H. Wu, "DeltaVLM: Interactive remote sensing image change analysis via instruction-guided difference perception," arXiv preprint arXiv:2507.22346, 2025, doi: 10.48550/arXiv.2507.22346.
[26]. H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," Remote Sensing, vol. 12, no. 10, p. 1662, 2020, doi: 10.3390/rs12101662.
[27]. M. Wu, Q. Huang, S. Gao, and Z. Zhang, "Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning," International Journal of Applied Earth Observation and Geoinformation, vol. 125, p. 103591, 2023, doi:

10.1016/j.jag.2023.103591.
[28]. Y. Kang, J. Kim, J. Park, and J. Lee, "Assessment of perceived and physical walkability using street view images and deep learning technology," ISPRS International Journal of Geo-Information, vol. 12, no. 5, p. 186, 2023, doi: 10.3390/ijgi12050186.
[29]. W. Huang, J. Wang, and C. Gao, "Zero-shot urban function inference with street view images through prompting a pretrained vision-language model," International Journal of Geographical Information Science, vol. 38, no. 7, pp. 1414–1442, 2024, doi: 10.1080/13658816.2024.2347322.
[30]. L. Li, Y. Ye, B. Jiang, and W. Zeng, "GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model," arXiv preprint arXiv:2406.18572, 2024, doi: 10.48550/arXiv.2406.18572.
[31]. H. Elgendy, A. Sharshar, A. Aboeitta, Y. Ashraf, and M. Guizani, "GeoLLaVA: Efficient fine-tuned vision-language models for temporal change detection in remote sensing," arXiv preprint arXiv:2410.19552, 2024, doi: 10.48550/arXiv.2410.19552.
[32]. Y. Zhu et al., "Semantic-CC: Boosting remote sensing image change captioning with foundational knowledge and semantic guidance," arXiv preprint arXiv:2407.14032, 2024, doi: 10.48550/arXiv.2407.14032.
[33]. J. Feng et al., "CityBench: Evaluating the capabilities of large language models for urban tasks," arXiv preprint arXiv:2406.13945, 2024, doi: 10.48550/arXiv.2406.13945.