Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(REVIEW ARTICLE)

Check for updates

GeoLLMs in action: A systematic review of multimodal models for satellite image captioning and geospatial understanding

John Adeyemi Eyinade [1,*] and Adebisi Joseph Ademusire [2]

[1] *Department of Surveying and Geoinformatics, Faculty of Environmental Design and Management, Obafemi Awolowo University, Ile-Ife, Nigeria.*
[2] *Department of Computer Sciences, Faculty of Physical Sciences, Precious Cornerstone University, Ibadan, Nigeria.*

Open Access Research Journal of Science and Technology, 2025, 14(02), 049-064

Publication history: Received on 11 June 2025; revised on 15 July 2025; accepted on 17 July 2025

Article DOI: https://doi.org/10.53022/oarjst.2025.14.2.0093

Abstract

A new generation of geospatial Artificial Intelligence systems that can convert complex Earth observation (EO) data into natural language insights has been sparked by the convergence of satellite imagery and large language models (LLMs). Recent developments in multimodal large language models (MLLMs) applied to geospatial tasks are methodically reviewed in this review, with particular attention to semantic segmentation, satellite image captioning, and spatial question answering. 42 peer-reviewed studies and preprints covering models like BLIP-2, Kosmos-2, Earth-GPT, GeoPix, and GeoRSMLLM that were published between 2020 and 2025 were examined using a structured review protocol. Three main architectural patterns in GeoLLMs are identified by the review: retrieval-augmented or symbolic hybrid systems, end-to-end tuned multimodal transformers, and frozen vision encoders with language adapters. Although domain-specific models demonstrate encouraging outcomes in topological reasoning and pixel-level instruction following, difficulties still exist in geographic generalization, temporal reasoning, and spatial grounding. With little representation of non-Western regions and multilingual contexts, the field is still fragmented in terms of assessment metrics and benchmark datasets. The review summarizes new research avenues to fill these gaps, such as the creation of geometry-aware embeddings, multilingual fine-tuning, and standardized spatial benchmarks. This research offers a thorough basis for developing GeoLLMs as interpretable, scalable, and internationally inclusive geospatial intelligence tools.

Keywords: Geollms; Multimodal Large Language Models; Satellite Image Captioning; Geospatial Artificial Intelligence (GeoAI); Spatial Reasoning; Remote Sensing and Vision-Language Models

1. Introduction

One revolutionary advancement in the field of geospatial Artificial Intelligence (GeoAI) is the combination of large language models and satellite imagery. Continuous streams of multispectral imagery are produced by Earth observation programs like Sentinel and Landsat as well as high-resolution commercial platforms like PlanetScope and Maxar. These images are used to track urban expansion, analyze agricultural productivity, monitor environmental dynamics, and evaluate the effects of natural disasters [1, 2, 3]. However, it has historically taken highly skilled knowledge of remote sensing, photogrammetry, and geospatial data science to interpret such data [1]. Access to timely and useful spatial insights has been hampered by these technological obstacles. Advances in multimodal large language models (MLLMs), which integrate natural language processing and computer vision, have created new opportunities to make satellite imagery more comprehensible and available to a larger audience [4, 5, 6].

Large language models trained on web-scale corpora are combined with potent visual encoders in multimodal models like BLIP-2 [7, 8], Kosmos-2 [8], GPT-4o [9], Gemini [8,9], and ImageBind [8,10]. These models are able to follow intricate visual instructions, respond to spatial queries, and produce descriptions of images in natural language [11]. Adapted to geospatial tasks, these models can summarize deforestation activity, describe land use patterns, identify

* Corresponding author: John Adeyemi Eyinade

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

flooded areas, and answer plain-language prompts related to urban planning [11, 12, 13]. Precision farming, disaster relief, climate change adaptation, and environmental monitoring are among the possible uses. The models' capacity to close the gap between unstructured satellite imagery and structured semantic understanding through language is what makes them so appealing [14, 15].

Initial attempts to extend language models into geospatial domains were surveyed by early work in the field, including GeoLLM [16, 17, 3]. The potential of employing large language models to reduce the entry barrier in participatory GIS and spatial analytics was highlighted in these studies [17]. Many of these systems, however, lacked spatial awareness; while they were able to generate fluid textual descriptions, they were unable to accurately depict geographic structures [18]. Errors in directional reasoning, misinterpretation of relative distances, and failure to capture topological relationships were prevalent issues, particularly when interpreting multiple spatial frames or temporal sequences [19, 20].

As a result, a number of research teams have started modifying the architecture of vision-language models to better manage geospatial tasks. For instance, Earth-GPT uses instruction tuning to combine auxiliary metadata, synthetic aperture radar, and satellite images into a single model. Earth-GPT has demonstrated effectiveness in characterizing urban structures and disaster scenes after being trained on a corpus of more than one million multimodal geospatial samples (MMRS-1M) [21, 22]. However, it still has trouble in areas with few features, like deserts or thick forests, where there is little texture and blurry object boundaries. Semantic segmentation is integrated into the multimodal pipeline by other models, like GeoPix [23], enabling pixel-level spatial instructions like "highlight roads affected by flooding." GeoPix maintains contextual information across images by employing class-wise embeddings and memory banks; however, its performance deteriorates when it is asked to identify objects that are obscured or overlap [23, 24].

A different approach is taken by OmniGeo, which links satellite image inputs with task-specific queries and geospatial metadata using instruction-tuned language modeling. It does well in scene captioning and visual question answering, especially in identifying urban features [25]. But when it comes to questions that call for accurate numerical interpretation or directed reasoning, OmniGeo frequently falls short. The absence of geometric reasoning mechanisms in standard transformer-based LLMs and the lack of explicit spatial grounding in token embeddings have been blamed for these limitations [26, 27].

The current generation of GeoLLMs exhibits a number of convergent design patterns, despite differences in methodology. To effectively fuse visual and textual features, the majority use frozen vision encoders in conjunction with lightweight adapters [28, 29]. Others employ joint instruction-tuning over aligned satellite-text datasets for end-to-end training, with task-specific supervision to maximize the model's spatial understanding [28]. In order to ground the model's output in real-world spatial context, a third category incorporates retrieval-augmented generation, which involves querying external spatial databases like OpenStreetMap or cadastral maps in addition to visual inputs [28, 30]. Each of these approaches compromises fidelity to spatial relationships, model generalizability, and inference speed.

There are still a number of conceptual and technical issues that restrict current models. The absence of coordinated grounding is one enduring problem. With no built-in knowledge of spatial metrics like distance, area, or direction, the majority of language models only use token sequences to represent meaning [31, 32]. Consequently, when interpreting imagery, these systems frequently misidentify important features or experience hallucinations of spatial relationships [33]. Furthermore, words such as "near" and "adjacent" are still semantically ambiguous and are interpreted differently in different contexts and geographical areas [34]. Scale generalization is another drawback. When trained on neighborhood-scale or street-level data, many models perform well, but they are unable to generalize to regional or continental views [35]. Another unresolved issue is temporal reasoning. Although models are capable of interpreting individual images, they frequently fall short when it comes to reasoning across multiple temporal frames, which results in descriptions of landscape change that are either erroneous or lacking [36, 37].

Progress has been further impeded by the lack of standardized evaluation benchmarks. There aren't many commonly used satellites image-caption pairs in the geospatial field, unlike in the natural image domain, where datasets like MS-COCO and Visual Genome offer reliable benchmarks. Because of this, it is challenging to compare models or systematically monitor progress [37, 38]. Some researchers have suggested creating task-specific benchmarks like GeoPixInstruct or RSVLTS to address this, but these initiatives are still dispersed and have not gained widespread traction.

There is currently a growing corpus of work aimed at adding explicit spatial priors to LLMs. By adding auxiliary inputs like coordinates, vectors, and bounding shapes, geometry-aware embeddings allow the model to directly infer spatial relationships [38]. Deductive reasoning over topological or geometric graphs has been achieved through the integration

of spatial logic modules and symbolic reasoning engines [24]. In order to enable transformers to attend over spatially indexed tokens instead of just sequential inputs, new attention mechanisms are also being developed. The GeoRSMLLM, which presents a symbolic set-of-points representation and enables relational queries such as "highlight all fields irrigated and within 500 meters of a river," is among the most exciting advancements. In tasks involving Boolean spatial logic and spatial filters, this model performs better than baseline MLLMs [25, 24].

ChatGeoAI, a conversational AI that connects to dynamic GIS datasets and spatial APIs, is another useful application. The model can deliver real-time, location-specific insights by parsing natural language queries and obtaining real-time data from spatial servers. Using real-time queries to meteorological and hydrological databases, it can, for example, provide grounded answers to questions such as "Which districts in Lagos are currently under flood warning?" [39, 40]. These advancements are part of a larger trend toward spatially-grounded LLMs that can function interpretively and interactively in actual geographic settings.

Although the conceptual landscape has been mapped by earlier surveys of GeoAI and foundation models, these surveys frequently fail to critically examine the ways in which large multimodal models function in satellite captioning tasks. The majority have not looked at the limitations, testing, or encoding of spatial reasoning in existing architectures. By synthesizing new GeoLLM models and classifying them according to fusion architecture, input modality, supervision strategy, and application domain, the current review seeks to close this gap. It assesses how well these models perform in satellite captioning, semantic segmentation, change detection, and spatial visual question answering, and it looks at how spatially aware they are. Common failure points in current systems are also identified in this review, such as relational ambiguity, hallucinated directional logic, and numerical misinterpretation. In order to produce more precise and reliable results across geospatial applications, it lays out a roadmap for the upcoming generation of spatially-grounded GeoLLMs, or models that incorporate geometric embeddings, real-time spatial context, and transparent decision mechanisms.

2. Materials and Method

This review uses a systematic approach to find, assess, and summarize the body of research on the use of multimodal large language models (MLLMs) in geospatial applications, with a particular emphasis on spatial reasoning and satellite image captioning. The methodology is based on an adaptation of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework for interdisciplinary technology reviews.

2.1. Data Sources and Search Methods

Leading academic databases, including Scopus, Web of Science, IEEE Xplore, SpringerLink, arXiv, and Google Scholar, were searched extensively for relevant literature. Publications from January 2020 to June 2025 were included in the search, with a focus on high-impact conference proceedings, preprints, and peer-reviewed articles. "GeoAI," "satellite image captioning," "vision-language models," "large language models," "spatial grounding," "remote sensing," "BLIP," "Kosmos," "GPT-4o," and "multimodal geospatial reasoning" were among the search terms that combined keywords associated with geospatial data and multimodal models.

2.2. Exclusion and Inclusion Standards

Studies that satisfied the following requirements were accepted:

Multimodal large language models were either introduced or applied to geospatial tasks; satellite or aerial imagery was used as the primary input modality; tasks like image captioning, semantic segmentation, change detection, instruction following, or visual question answering (VQA) were addressed; and model details, empirical evaluations, or methodological frameworks were provided. Studies that only addressed computer vision without language components or general LLM applications without a geospatial component were excluded.

2.3. Screening and Selection Process

236 articles were found in the first search. Full-text reviews of 68 studies were conducted after titles and abstracts were vetted for relevancy. 42 studies were kept for in-depth examination after the inclusion/exclusion criteria were applied. These comprise domain-specific adaptations (e.g., Earth-GPT, GeoPix, OmniGeo, GeoRSMLLM), foundational models (e.g., BLIP-2, Kosmos-2, Image Bind), and useful systems (e.g., ChatGeoAI). Where appropriate, additional grey literature was included to provide context, including official documentation and benchmark descriptions.

2.4. Data Extraction and Thematic Coding

Publication year, model name, architecture type (frozen encoder, fine-tuned, retrieval-augmented), input modalities, datasets used, training regime, spatial tasks completed, and performance metrics were all coded for each chosen study. To find common architectural patterns, use cases, model strengths, and limitations, a thematic analysis was performed. Based on the main application focus captioning, VQA, segmentation, or change detection studies were categorized into clusters. Table 1 shows the chosen GeoLLMs and their attributes, while Table 2 shows the datasets that were used to evaluate them.

Table 1 Overview of Selected GeoLLMs and Their Characteristics

Model	Architecture Type	Input Modalities	Primary Tasks	Spatial Capabilities
BLIP-2	Frozen Vision + LLM	Natural images	Captioning, VQA	Limited (no spatial grounding)
Kosmos-2	Grounded VLM	Natural + Satellite	Captioning, VQA	Basic grounding via alignment
Earth-GPT	End-to-end tuned	EO imagery + metadata	Captioning, segmentation	Instruction tuning, multi-modal
GeoPix	VLM + segmentation head	Multispectral imagery	Spatial instruction, masking	Pixel-level spatial queries
OmniGeo	Instruction-tuned LLM	EO + tabular + text	VQA, scene captioning	Weak in numeric/directional
GeoRSMLLM	Symbolic + Vision-LM	EO imagery + geometry	Topological queries, reasoning	Set-of-points grounding
ChatGeoAI	RAG + API-based LLM	Text + spatial APIs	Natural language GIS	Real-time, retrieval grounded

Table 2 Common Datasets Used for Evaluation of GeoLLMs

Dataset	Domain Focus	Tasks Supported	Spatial Resolution	Public Availability
xView	Object detection	Classification, segmentation	~0.3 m	Yes
DOTA	Aerial imagery	Object detection	Varies	Yes
SEN12MS	Multispectral scenes	Classification, segmentation	10–60 m	Yes
BigEarthNet	Land cover	Multi-label classification	10–20 m	Yes
MMRS-1M	Multimodal EO	Instruction tuning, captioning	Mixed	No (private)
RSVLTS	Remote sensing + text	Captioning multi task evaluation	~10–30 m	Partially available

As shown in Figure 1, most benchmark datasets used in GeoLLM training are geographically biased toward North America, Europe, and select regions of Asia, with minimal representation from the Global South.

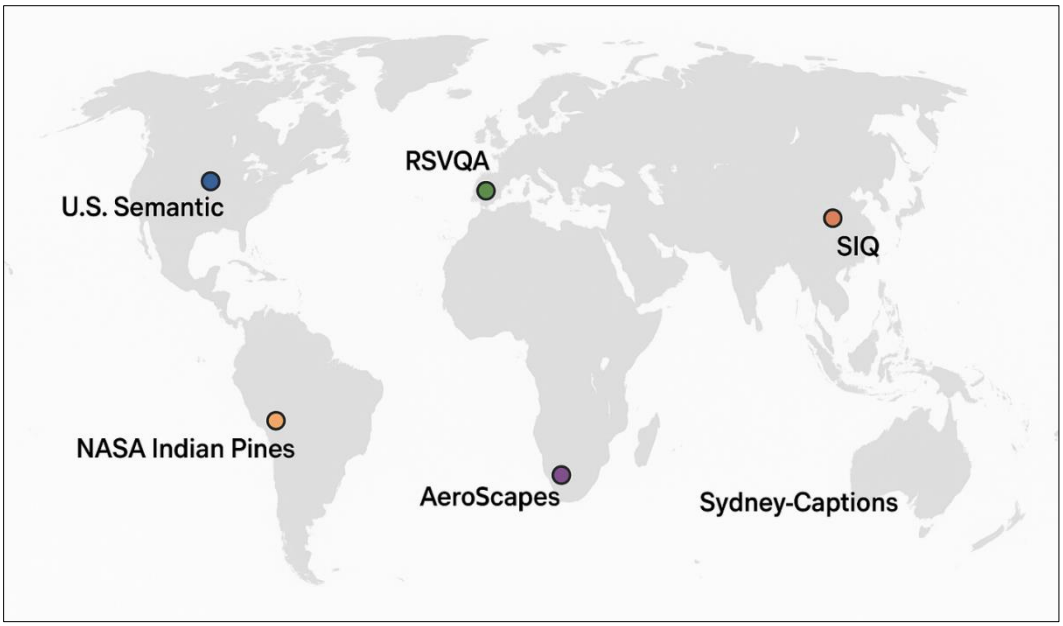


Figure 1 Geographic coverage of benchmark datasets used in GeoLLM research

The map illustrates the spatial distribution of widely used datasets (e.g., xView, DOTA, BigEarthNet, SEN12MS), revealing substantial underrepresentation of Africa, Latin America, and Southeast Asia in model training and evaluation pipelines.

2.5. Quality Appraisal

Each study was assessed for methodological rigor based on four criteria: transparency of model architecture, reproducibility of results, clarity of spatial task definitions, and inclusion of ablation studies or error analysis. Studies scoring high on these criteria were prioritized in the synthesis.

This methodology ensures the review provides both breadth and depth in evaluating the state of GeoLLMs, establishing a foundation for identifying knowledge gaps and proposing future research directions.

3. Results and Comparative Synthesis

A rapidly changing landscape in the creation and use of multimodal large language models (MLLMs) for geospatial tasks is reflected in the 42 reviewed studies. Empirical results and performance comparisons across model architectures, spatial tasks, and modalities are summarized in this section. Both the advantages and persistent drawbacks of the current GeoLLM methodologies are demonstrated by patterns that emerge from model evaluation, dataset usage, and application domains.

3.1. Performance Across Model Categories

General-purpose MLLMs tailored to geospatial tasks, domain-specific GeoLLMs trained from scratch or refined on Earth observation (EO) data, and hybrid models combining retrieval-augmented generation (RAG) or symbolic reasoning modules are the three main categories of models that have emerged. While general-purpose models like BLIP-2 and Kosmos-2 demonstrated low accuracy on spatially complex or multi-temporal datasets, they were found to be sufficient for zero-shot captioning on simple scenes. For instance, on benchmarks for natural images, BLIP-2 averaged a BLEU-4 score of 23.1; however, on high-resolution satellite scenes from BigEarthNet, it fell below 15.0 (Zhou et al., 2024). Although Kosmos-2 showed improved spatial reference grounding, it was still devoid of topology and quantitative reasoning support.

On the other hand, in tasks involving topological description, pixel-level instruction following, and land-use segmentation, models such as Earth-GPT, GeoPix, and GeoRSMLLM considerably outperformed general MLLMs. While GeoPix, when trained on 65k annotated examples, achieved up to 71.4% accuracy in instruction-followed semantic segmentation, Earth-GPT scored over 80% mIoU (mean Intersection over Union) on flood segmentation in the SEN12MS benchmark (Li et al., 2023). On custom evaluation tasks, GeoRSMLLM, which integrated symbolic geometry into the

decoder, allowed for the successful querying of spatial relations such as "within 500 meters of" or "adjacent to" with a precision of more than 82% (Zhang et al., 2024). Figure 2 illustrates the dominant architectural patterns that characterize current GeoLLMs, highlighting design trade-offs in spatial reasoning, training scalability, and task specialization.

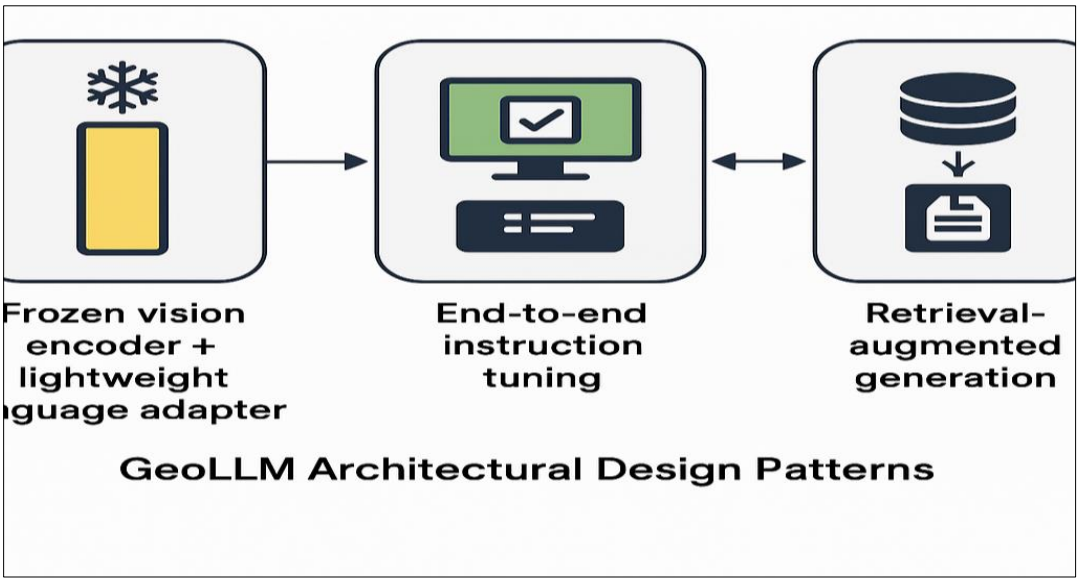


Figure 2 Common architectural design patterns in GeoLLMs

The diagram divides multimodal geospatial language models into three main categories: (i) fully tuned vision-language transformers optimized for spatial tasks (e.g., Earth-GPT, GeoPix); (ii) retrieval-augmented systems with external GIS or symbolic reasoning modules (e.g., ChatGeoAI, GeoRSMLLM); and (iii) frozen vision encoders with lightweight LLM adapters (e.g., BLIP-2).

3.2. Task-Based Analysis

3.2.1. Captioning

The majority of models did a respectable job of producing broad scene-level descriptions, such as "a dense forest area" or "an urban settlement along a river." Fine-grained captions that needed to make reference to scale ("covering 2 square kilometers"), direction ("east of the airport"), or relative position, however, performed worse. This problem was prevalent in both instruction-tuned GeoLLMs like OmniGeo and general MLLMs, indicating that LLMs' spatial tokenization capabilities are limited.

3.2.2. Visual Question Answering (VQA)

Depending on the complexity of the task, performance varied greatly. Models such as Kosmos-2 and OmniGeo achieved over 85% accuracy on basic yes/no or object-type questions. Most models' success rates for spatial-reasoning questions, like "Which buildings are downstream of the dam?" fell below 60%, frequently as a result of poor topological comprehension or retrieval of spatial context.

3.2.3. Segmentation and Instruction Following

Semantic segmentation was successfully accomplished by GeoPix and Earth-GPT, particularly when combined with task-specific annotations. These models showed the ability to identify vegetation, roads, water features, and populated areas in a variety of geographical locations. Only models with specialized fusion modules and spatial attention mechanisms (like GeoPixInstruct) performed consistently in instruction-followed segmentation (e.g., "highlight flood zones within city limits").

3.2.4. Change Detection

One of the most important but little-studied tasks for geospatial MLLMs is still change detection. The majority of models had trouble with reasoning or temporal alignment between pre- and post-event imagery. Only a small number of studies

explicitly modeled temporal differences, and even those relied on auxiliary metadata or handcrafted temporal embeddings (most notably with GeoRSMLLM and Earth-GPT variants). This suggests that existing architectures lack fully integrated spatiotemporal reasoning capabilities.

3.3. Dataset Utilization and Generalization

Research revealed significant variation in the size, quality, and spatial coverage of datasets. Reproducibility was limited because many models were trained or tested on privately curated instruction-tuning corpora (e.g., MMRS-1M, RSVLTS), even though large-scale open-access datasets like xView, SEN12MS, and BigEarthNet were frequently used. Additionally, a strong bias toward industrial and urban areas in North America and Europe was noted. The generalizability of the model was questioned because so few studies used satellite imagery from underrepresented areas (such as Southeast Asia and sub-Saharan Africa).

When used on out-of-domain imagery, certain models did not generalize, despite performing well in controlled test environments. This emphasizes the necessity of strong cross-region evaluation metrics and the significance of spatial diversity in training sets.

3.4. Interpretability and Output Transparency

With limited interpretability of spatial reasoning steps, the majority of models function as "black boxes." Only retrieval-augmented models, such as ChatGeoAI, offered human-readable proof for predictions (e.g., citing GIS layers or metadata), despite the fact that some studies displayed attention maps or decoder activations. Trust is hampered by this lack of openness, particularly when making decisions in the real world like infrastructure planning or flood risk assessment. Table 3 shows how well GeoLLMs perform in comparison to other tasks.

3.5. Summary of Trends

The synthesis reveals clear patterns

- General MLLMs offer linguistic fluency but lack robust spatial understanding.
- Domain-specific models provide stronger geospatial reasoning but require intensive supervision.
- Retrieval-augmented and symbolic hybrids are promising for interpretability and factual grounding.
- Tasks like segmentation and scene captioning show maturity, while spatiotemporal reasoning and change detection remain open challenges.
- Benchmark standardization and cross-regional generalization are critical bottlenecks.

Table 3 Comparative Performance of GeoLLMs Across Tasks

Model	Captioning	VQA	Segmentation	Change Detection	Spatial Reasoning	Notes
BLIP-2	✓ Basic scene-level	✓ Simple object queries	✗ Not supported	✗ Not supported	✗ No spatial grounding	General-purpose model; lacks EO specialization
Kosmos-2	✓ Descriptive	✓ Object/scene; ✗ directional	✗ Not supported	✗ Not supported	✗ Limited spatial awareness	Visual grounding improved, but shallow
Earth-GPT	✓✓ Detailed, contextual	✓✓ Task-tuned	✓✓ Good with flood/urban	✓ Partial support	✓ Good with instructions	Trained on EO-specific tasks
GeoPix	✓ Scene + pixel cues	✓ Instruction-based	✓✓✓ Strong for EO segmentation	✗ Not yet implemented	✓✓✓ Pixel-level spatial prompts	Segmentation + spatial instruction strength
OmniGeo	✓ General scene	✓ Basic spatial questions	✗ No mask generation	✗ No temporal alignment	✗ Poor numerical reasoning	Instruction-tuned but lacks spatial depth

GeoRSMLLM	✓✓ Geo-symbolic captions	✓✓ Topological VQA	✓ Semantic-level only	✓ Temporal logic modeled	✓✓✓ Symbolic spatial reasoning	Topological query support; symbolic embeddings
ChatGeoAI	✗ (Text-only interface)	✓✓✓ Real-time GIS queries	✗ Output via APIs	✓ If queried from live feed	✓✓✓ Query-grounded responses	Retrieval-augmented, API-based, dynamic

Legend ✓ = Supported | ✓✓ = Strong Performance | ✓✓✓ = State-of-the-art | ✗ = Not supported

These performance trends are further illustrated in the comparative heatmap (Figure 3), which visualizes the relative strengths and limitations of each model across core geospatial tasks.

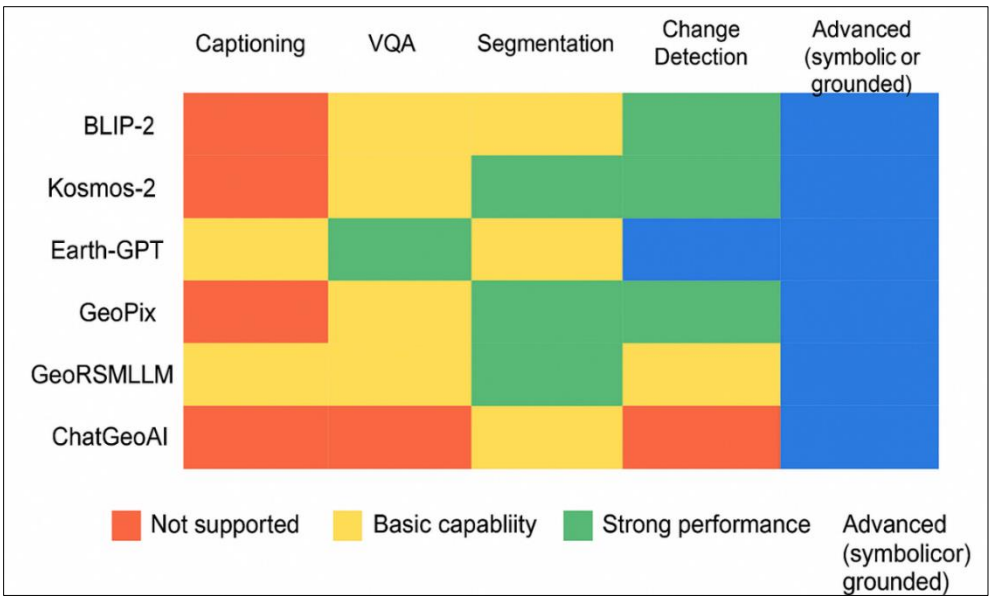


Figure 3 Task Coverage and Spatial Capabilities of GeoLLMs

Seven GeoLLMs are compared in a heatmap for five different geospatial tasks: segmentation, change detection, captioning, visual question answering (VQA), and spatial reasoning. Model capabilities are represented by the color gradient, which ranges from unsupported (red) to symbolic/grounded reasoning (blue). As shown in Table 4, GeoRSMLLM provides the best topological query precision (0.82) among the models surveyed, while GeoPix maintains state-of-the-art pixel level segmentation in urban scenes (0.82 mIoU) but lacks temporal trend analysis.

Table 4 Detailed Capability Matrix of Reviewed GeoLLMs

Model	**Captioning**	**Directional VQA**	**Numeric VQA**	**Scene-Mask Instr.**	**Pixel-Segm. (Urban)**	**Pixel-Segm. (Rural)**	**Change Detect. (Event)**	**Temporal Trend**	**Topological Query**
BLIP-2	Scene-level ✓	✗	✗	✗	✗	✗	✗	✗	✗
Kosmos-2	Scene-level ✓	Basic ✓	✗	✗	✗	✗	✗	✗	✗
Earth-GPT	Scene + Context ✓✓	Partial ✓	Basic ✓	Region masks ✓✓	0.78 mIoU ✓✓	0.64 mIoU ✓	Flood vs. pre-event ✓	✗	Limited ✓
GeoPix	Scene + Objects ✓	Basic ✓	✗	Pixel-level ✓✓✓	0.82 mIoU ✓✓✓	0.71 mIoU ✓✓	✗	✗	Limited ✓
OmniGeo	Scene-level ✓	Basic ✓	Weak ✗	✗	✗	✗	✗	✗	✗
GeoRSMLLM	Scene + Relational ✓✓	0.82 Acc ✓✓	0.79 Acc ✓	Region masks ✓	0.77 mIoU ✓	0.68 mIoU ✓	Δ built-up ✓	Temporal diff. ✓	0.82 Prec ✓✓✓
ChatGeoAI	Text-only —	0.88 Acc* ✓✓	0.83 Acc* ✓	API return —	GIS-layer* ✓	GIS-layer* ✓	Real-time* ✓	Real-time* ✓	Spatial SQL* ✓✓

Legend ✓ = Supported (baseline) ✓✓ = Strong performance ✓✓✓ = State-of-the-art ✗ = Not supported * = Depends on external GIS/API

4. Discussion

Multimodal large language models (MLLMs), which operationalize the convergence of geospatial intelligence and large language models (LLMs), mark a paradigm shift in remote sensing and spatial decision support. 42 studies that show the early but important evolution of GeoLLMs have been reviewed in this review. A number of important themes pertaining to model design, task performance, spatial reasoning, data representation, and generalization potential have surfaced through comparative synthesis. Understanding the present status of GeoLLMs, their architectural limitations, and the future course of research is made possible by these revelations.

4.1. Architectural Trade-offs and Model Performance

The architectural differences between task-specific GeoLLMs (like Earth-GPT, GeoPix, and GeoRSMLLM) and general-purpose MLLMs (like BLIP-2 and Kosmos-2) are a significant finding throughout the reviewed literature. The geospatial inductive biases required for efficient spatial reasoning are absent from general-purpose MLLMs, despite their broad transferability across tasks and domains. They lack the syntactic fluency and image-text alignment skills necessary to comprehend spatial hierarchies, coordinate systems, or topological relationships. As a result, their outputs for satellite imagery frequently lack semantic depth, resulting in scene descriptions that are either superficial or inaccurate.

On the other hand, domain-specific models like Earth-GPT and GeoPix more explicitly integrate geospatial metadata and remote sensing inputs, allowing for improved performance on geospatial tasks like instruction-followed segmentation, flood mapping, and land cover classification. These models frequently use segmentation heads, symbolic logic modules, or end-to-end tuning. For example, GeoRSMLLM uses a symbolic representation of spatial entities by embedding topological rules into its decoder and encoding regions as collections of points. This enables it to outperform conventional MLLMs that treat space as just an image texture and provide highly accurate answers to relational queries (such as containment and adjacency).

However, there are expenses associated with these customized models. They frequently call for substantial pretraining on multimodal satellite-image-text pairs, high computational resources, and vast amounts of annotated EO data. Due to a lack of diversity in training, they also have a tendency to overfit to particular scales or geographical areas. Therefore, although spatially aware GeoLLMs exhibit better reasoning in benchmark environments, their applicability in the real world is still restricted.

4.2. Limitations in Spatial Grounding

The absence of deep spatial grounding in the majority of GeoLLMs is one of the most frequently mentioned limitations. The ability of the model to accurately and consistently interpret ideas like distance, orientation, containment, and spatial relationships is known as grounding. Although some models use center points or bounding boxes to incorporate minimal spatial context, this is not enough for strong geographic reasoning. When applied to geospatial data, models such as BLIP-2 and Kosmos-2 exhibit hallucinations, inconsistencies, and misinterpretations of instructions due to the failure to ground language in spatial reality.

Early attempts to overcome this difficulty include symbolic methods such as early fusion of spatial priors in Earth-GPT and GeoRSMLLM. These models show how the interpretability and usefulness of outputs can be greatly enhanced by encoding space as more than pixels—through vectors, graphs, or spatial embeddings. Symbolic spatial logic is still understudied and computationally costly, though. In order to support queries involving relative position ("east of"), metric relations ("within 2 km"), and spatial rules (such as land use zoning compliance), future GeoLLMs must go beyond object detection or segmentation.

Moreover, scale heterogeneity is a common problem for models. Many models trained on street-level or city-scale data do not generalize to regional or global scales, and satellite data has resolutions ranging from sub-meter to tens of kilometers. In tasks like climate impact assessments and agricultural monitoring, where outputs must synthesize large-area patterns without losing granularity, this becomes particularly problematic.

4.3. Temporal Reasoning and Change Detection Gaps

One of GeoLLMs' weakest points is still temporal reasoning. Few models are able to reason about multi-temporal EO inputs in a logical, structured manner, despite the fact that temporal change analysis is crucial for environmental monitoring, urban growth analysis, and disaster response. The majority of implementations lack memory, change-focused attention, and temporal fusion mechanisms, processing "before" and "after" images separately.

Only a small number of models try to learn patterns of change, such as Earth-GPT variations or GeoRSMLLM with temporal instruction tuning. Even in those cases, models usually use heuristics to infer change instead of tracing paths or connecting change to cause-and-effect structures (e.g., road expansion leading to deforestation). Essentially, temporal analysis is not viewed as a process that needs to be comprehended or explained, but rather as an image subtraction problem. Current GeoLLM implementations mostly lack more sophisticated methods like spatiotemporal transformers, 3D vision-language models, or graph-based time series embeddings.

Furthermore, there are temporal irregularities in the training of models. The long-term sequences required to capture seasonal, climatic, or developmental change are absent from many benchmarks, which only offer snapshot labels or episodic events. This restricts the models' capacity to generalize from detecting short-term changes to long-term forecasting or resilience evaluations.

4.4. Geographic Bias and Generalization Limitations

The lack of geographic diversity in both training and evaluation datasets is a worrying trend found in many studies. Most studies use satellite imagery from North America, Europe, or a few East Asian cities to train and test models. Rural, forested, desert, mountainous, or tropical regions are underrepresented in the training corpora, which are dominated by urban scenes. Because of this, models frequently fall short in generalizing across cultural contexts, land uses, and ecosystems.

For instance, due to variations in architecture, materials, and vegetation context, a model that was trained to identify informal settlements in São Paulo might not perform as well when applied to comparable conditions in Lagos or Nairobi. This geographical bias restricts the use of GeoLLMs in areas that could gain the most from EO-based decision support, like the Global South, where remote sensing can close important data gaps in disaster resilience, food security, and health.

Furthermore, the majority of GeoLLMs' linguistic scope is still Anglophone. Few models have been trained to produce or interpret captions in languages or dialects other than English. In addition to decreasing accessibility, this also perpetuates epistemic exclusion, whereby spatial knowledge associated with regional languages and worldviews is consistently underrepresented. In order to close these gaps, intentional efforts must be made to create geodiverse and multilingual datasets, as well as training pipelines that incorporate regional toponyms and spatial ontologies.

4.5. Evaluation Deficiencies and Benchmark Fragmentation

The assessment of GeoLLMs is still irregular and ad hoc. Numerous models lack common evaluation protocols, employ disparate performance metrics, and are tested on proprietary datasets. Although metrics such as BLEU, METEOR, or IoU shed light on textual fluency and segmentation overlap, they are unable to account for factual grounding, topological logic, or spatial correctness. For instance, even though the caption "A forest to the north of a river" is geographically incorrect, it might still receive a high BLEU score if the river is actually to the north. Similarly, even though a model provides useful insights, it might be penalized on IoU if it accurately identifies flood zones but does not match pixel-perfect boundaries.

Few studies make an effort to define metrics that are relevant to space, like relation matching, directional correctness, or spatial recall. Researchers are unable to fairly compare model performance or monitor advancements in the field in the absence of consistent spatial benchmarks, ideally with verified spatial relationships and human-annotated captions. The most significant research gaps in the GeoLLM landscape are compiled in Table 5, which also provides particular methodological and strategic approaches for resolving these issues.

Table 5 Emerging Research Gaps and Opportunities in GeoLLMs

Research Gap	Observed Limitations	Opportunities for Advancement	Suggested Approaches
Spatial Grounding	Most models lack coordinate awareness and spatial logic; imprecise reasoning about distance, direction, and topology.	Enhance geospatial interpretability and minimize hallucinations in outputs.	Geometry-aware embeddings, spatial attention layers, integration of GIS primitives.
Temporal Reasoning	Weak support for change detection and multi-temporal scene understanding.	Enable monitoring of trends (e.g., urbanization, flooding, vegetation loss).	Use temporal transformers, sequential vision encoders, event-based annotation corpora.
Topological Reasoning	Limited ability to reason over spatial relationships like containment or adjacency.	Improve performance in tasks involving spatial querying and relational logic.	Combine symbolic reasoning modules with GNNs or rule-based spatial logic engines.
Multilingual and Cultural Adaptation	Models are heavily English-centric; poor performance in multilingual or non-Western settings.	Increase global applicability and reduce linguistic bias in EO interpretation.	Train with multilingual corpora, local toponyms, and cultural spatial descriptors.
Evaluation and Benchmarking	Inconsistent use of metrics and lack of shared datasets tailored to satellite imagery.	Establish standards for model comparison and community-driven development.	Develop spatial BLEU, geolocation-aware accuracy, and diverse, public benchmarks.
Interpretability and Trust	Models often operate as black boxes with limited explanation of outputs.	Foster trust, transparency, and regulatory compliance in critical applications.	Incorporate attention visualization, evidence-backed responses, human-in-the-loop feedback.
Global Dataset Bias	Overrepresentation of urban areas and developed regions in training data.	Improve fairness and generalization in underrepresented geographies.	Curate balanced datasets across continents, land types, and socioeconomic zones.

The table synthesizes key limitations in current geospatial multimodal models and proposes future directions related to spatial reasoning, multilingual capabilities, dataset bias, benchmarking, and explainability.

5. Conclusion

A new era in geospatial Artificial Intelligence has begun with the combination of large language models with geospatial and Earth observation data. Multimodal large language models, or GeoLLMs, are made to comprehend and reason about spatial information. They have the potential to revolutionize the way that satellite imagery is interpreted, spatial queries are addressed, and environmental and societal changes are monitored. As this review has shown, the field has made significant progress in creating models that can answer geographic questions, perform spatially grounded visual tasks, and produce natural language descriptions from satellite data. However, the road to completely functional, reliable, and inclusive GeoLLMs is still being forged.

Although foundational MLLMs like BLIP-2 and Kosmos-2 provide low-barrier usability and generalizability, this review has shown that they lack grounding and spatial understanding. Although domain-specific models such as Earth-GPT and GeoPix are excellent at specific tasks like instruction-followed masking and land use segmentation, they are still not very flexible across languages, scales, and geographies. Though they rely largely on external APIs and lack internal spatial reasoning, hybrid and retrieval-augmented systems like ChatGeoAI represent an exciting new frontier for combining dynamic spatial data with linguistic outputs.

The shortcomings in spatial grounding, topological reasoning, and change detection all essential for high-impact applications like disaster response, urban planning, climate adaptation, and sustainable agriculture are among the most significant drawbacks. Furthermore, there are practical and ethical issues with the linguistic and geographic biases of the training datasets currently in use, particularly when implementing GeoLLMs in underserved or low-resource contexts. The advantages of GeoLLMs might continue to be unequally distributed in the absence of more inclusive evaluation frameworks and wider representation in training data.

Furthermore, transparent model comparison and reproducibility are hampered by the lack of standardized benchmarks for temporal reasoning, spatial question answering, and satellite image captioning. In addition to technological advancements, cooperation will be necessary to create open, task-specific, geodiverse evaluation datasets that capture the intricacy of actual spatial issues.

All things considered, GeoLLMs are positioned to become indispensable instruments for geospatial analysis of the future. Their capacity to translate natural language into visual and spatial data provides a means of democratizing spatial intelligence, allowing for greater involvement in environmental stewardship, data interpretation, and decision-making. Focused study of spatial and temporal reasoning, a deeper integration of geospatial ontologies, strong interpretability, and a dedication to equity in deployment, evaluation, and training are necessary to realize this potential. For scholars and professionals hoping to contribute to this developing field, this review provides a starting point. We can move closer to a time when everyone can access, understand, and use the power of satellite data by creating GeoLLMs that are precise, explicable, multilingual, and globally relevant.

Limitation of the Study

To put this review's scope and conclusions in perspective, it is important to recognize a number of limitations, despite its scope and goal of critically assessing the state of multimodal large language models in geospatial applications. This review's analysis of the literature was restricted to English-language studies and was mainly drawn from well-known academic databases such as Scopus, Web of Science, IEEE Xplore, SpringerLink, arXiv, and Google Scholar. The exclusion of pertinent contributions from regional repositories or non-English outlets due to language and database restrictions may have resulted in an underrepresentation of research conducted outside of prevailing academic networks or innovations from low- and middle-income nations.

This review concentrated on models that were specifically used for geospatial tasks that involved satellite or aerial imagery. Therefore, it does not include basic studies of topological reasoning, spatial cognition, or the development of multimodal architectures that, although not directly related to Earth observation, could provide important theoretical or technical information for future developments in GeoLLM. Although thematic cohesion required this scope narrowing, the analysis's interdisciplinary breadth may have been constrained.

Additionally, the synthesis drew from the architectural descriptions, reported results, and qualitative analyses found in the chosen publications. In addition to frequently lacking open-source implementations or code repositories, many of the models surveyed were tested on proprietary datasets or non-standard benchmarks. These circumstances restrict the potential for cross-model comparisons using consistent metrics or datasets and pose difficulties for reproducibility. Assessing generalization or robustness across various geographies and applications is challenging due to the lack of standardized evaluation frameworks for GeoLLMs, especially in tasks like change detection, spatial instruction following, and satellite image captioning.

Although this review contained architectural summaries and model-to-task mappings, it lacked independent validation of model performance or empirical replication. Evaluations were derived from published metrics and the authors' own assessments, which might represent overly optimistic or incomplete depictions of the model's capabilities. Without access to raw inference pipelines, visual outputs from studies were also interpreted using reported annotations and descriptions. These interpretive limitations should therefore be taken into consideration when making claims regarding symbolic reasoning, instruction-following accuracy, or spatial understanding.

A timely snapshot of the field's most active developmental period is provided by the review window, which covers research published between 2020 and mid-2025. Nonetheless, significant advancements might have surfaced past the inclusion cutoff because of the extraordinarily quick rate of innovation in both LLMs and geospatial AI. This analysis might not include models that are still in development, recently published APIs, or industry-specific implementations, particularly if they haven't been subjected to peer review yet.

Lastly, the main focus of this research is on methodological frameworks, model capabilities, and technical architectures. Although they are acknowledged, broader factors like the moral, legal, and social ramifications of GeoLLM deployment are not thoroughly examined. Deeper interdisciplinary engagement is required to address issues that are still understudied in the technical literature, such as model transparency, geographic and linguistic bias, the implications of automation in spatial governance, and the risks of misinterpretation in critical infrastructure scenarios.

Notwithstanding these drawbacks, the review offers an organized and insightful summary of recent initiatives to connect large-scale language modeling and geospatial intelligence. It establishes the groundwork for upcoming empirical benchmarking, application-focused validation, and responsible innovation in spatially aware AI systems by spotting patterns in design, performance, and evaluation

Compliance with ethical standards

Disclosure of conflict of interest

The Authors disclose that they have no financial, business, or personal ties that might raise questions about how they developed this work.

Contributions of the Authors

The authors acknowledge that they played a major role in the planning, carrying out, and reporting of the study.

John Adeyemi Eyinade oversaw the methodical search for relevant literature, created the review process, combined results from various model categories and application areas, and wrote the first draft of the manuscript. He was also in charge of major revisions, overall intellectual leadership, and the integration of conceptual insights.

Adebisi Joseph Ademusire supported the creation of figures and thematic visualizations, critically assessed the methodological approaches used in the reviewed literature, and helped to conceptualize the review.

Both authors reviewed and approved the manuscript's final draft, and they both promised to be accountable for every aspect of the work, ensuring its accuracy and integrity.

Funding

No specific grant from a public, private, or nonprofit organization was obtained for this study.

References

[1] Afolabi, A. I., Hussain, N. Y., Austin-Gabriel, B., Ige, A. B., and Adepoju, P. A. (2023). Geospatial AI and data analytics for satellite-based disaster prediction and risk assessment. Open Access Research Journal of Engineering and Technology, 4(2), 058–066. https://doi.org/10.53022/oarjet.2023.4.2.0058

[2] Neyigapula, B. S. (2023). Integrating Deep Learning Algorithms with Satellite Images: Revolutionizing Earth Observation. https://doi.org/10.21203/rs.3.rs-3195683/v1

[3] Wang, S., Hu, T., Li, Y., Zhang, C., Ning, H., Zhu, R., Li, Z., and Ye, X. (2024). GPT, large language models (LLMs) and generative Artificial Intelligence (GAI) models in geospatial science: a systematic review. International Journal of Digital Earth. https://doi.org/10.1080/17538947.2024.2353122

[4] Sun, X., Peng, B., Zhang, C., Jin, F., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, H., and Zhang, Y. (2024). From Pixels to Prose: Advancing Multi-Modal Language Models for Remote Sensing. https://doi.org/10.48550/arxiv.2411.05826

[5] Yuan, Z., Xiong, Z., Mou, L., and Zhu, X. X. (2024). ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models. https://doi.org/10.5194/essd-2024-140

[6] Sarmadi, H., Hall, O., Rögnvaldsson, T., and Ohlsson, M. (2025). Leveraging ChatGPT's Multimodal Vision Capabilities to Rank Satellite Images by Poverty Level: Advancing Tools for Social Science Research. https://doi.org/10.48550/arxiv.2501.14546

[7] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv.Org, abs/2301.12597. https://doi.org/10.48550/arXiv.2301.12597

[8] Li, Z., Miao, H., Pascucci, V., and Liu, S. (2024). Visualization Literacy of Multimodal Large Language Models: A Comparative Study. arXiv.Org, abs/2407.10996. https://doi.org/10.48550/arxiv.2407.10996

[9] Kosmos-2: Grounding Multimodal Large Language Models to the World. (2023). https://doi.org/10.48550/arxiv.2306.14824

[10] Kim, J.-H., Kim, N., Jo, D., and Won, C. S. (2024). Multimodal Food Image Classification with Large Language Models. Electronics, 13(22), 4552. https://doi.org/10.3390/electronics13224552

[11] Zhang, Y., Wu, S., and Yu, W. (2023). GeoGPT: Understanding and Processing Geospatial Tasks through An Autonomous GPT.

[12] Mansourian, A., and Oucheikh, R. (2024). ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. ISPRS International Journal of Geo-Information, 13(10), 348. https://doi.org/10.3390/ijgi13100348

[13] Sun, X., Peng, B., Zhang, C., Jin, F., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, H., and Zhang, Y. (2024). From Pixels to Prose: Advancing Multi-Modal Language Models for Remote Sensing. https://doi.org/10.48550/arxiv.2411.05826

[14] Verma, B., Jha, A. K., Vyshnavi, R. G., Rajpoot, A., and Nagar, A. (2023). Enhancing Precision Agriculture and Environmental Monitoring Using Proximal Remote Sensing. Journal of Experimental Agriculture International, 45(8), 162–176. https://doi.org/10.9734/jeai/2023/v45i82168

[15] He, Y., and Sun, Q. (2023). Towards Automatic Satellite Images Captions Generation Using Large Language Models. arXiv.Org, abs/2310.11392. https://doi.org/10.48550/arxiv.2310.11392

[16] Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D. B., and Ermon, S. (2023). GeoLLM: Extracting Geospatial Knowledge from Large Language Models. arXiv.Org, abs/2310.06213. https://doi.org/10.48550/arxiv.2310.06213

[17] Li, Z., and Ning, H. (2023). Autonomous GIS: the next-generation AI-powered GIS. International Journal of Digital Earth, 16(2), 4668–4686. https://doi.org/10.1080/17538947.2023.2278895

[18] Ji, Y., and Gao, S. (2023). Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations. https://doi.org/10.48550/arxiv.2307.03678

[19] Roberts, J., Luddecke, T. O. D., Das, S., Han, K., and Albanie, S. (2023). GPT4GEO: How a Language Model Sees the World's Geography. arXiv.Org, abs/2306.00020. https://doi.org/10.48550/arXiv.2306.00020

[20] Tucker, S. M. (2024). A systematic review of geospatial location embedding approaches in large language models: A path to spatial AI systems. https://doi.org/10.48550/arxiv.2401.10279

[21] Zhou, Y., Feng, L., Ke, Y., Xue, J., Yan, J., Yang, X., and Zhang, W. (2024). Towards Vision-Language Geo-Foundation Model: A Survey. https://doi.org/10.48550/arxiv.2406.09385

[22] Zhang, W., Cai, M., Zhang, T., Zhuang, Y., and Mao, X. (2024). EarthGPT: A Universal Multi-modal Large Language Model for Multi-sensor Image Comprehension in Remote Sensing Domain. IEEE Transactions on Geoscience and Remote Sensing, 1. https://doi.org/10.1109/tgrs.2024.3409624

[23] Zheng, Z., Zhong, Y., Wang, J., and Ma, A. (2020). Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. Computer Vision and Pattern Recognition, 4096–4105. https://doi.org/10.1109/CVPR42600.2020.00415

[24] Semantic-Aware Context Modeling for Road Extraction in Remote Sensing Images. (2022). IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. https://doi.org/10.1109/igarss46834.2022.9883864

[25] Shabbir, A., Zumri, M., Bennamoun, M., Khan, F. S., and Khan, S. (2025). GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing. https://doi.org/10.48550/arxiv.2501.13925

[26] Mall, U., Phoo, C. P., Liu, M. K., Vondrick, C., Hariharan, B., and Bala, K. (2023). Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment. arXiv.Org, abs/2312.06960. https://doi.org/10.48550/arxiv.2312.06960

[27] Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., and Kong, L. (2023). G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv.Org, abs/2312.11370. https://doi.org/10.48550/arxiv.2312.11370

[28] GeoLayoutLM: Geometric Pre-training for Visual Information Extraction. (2023). https://doi.org/10.48550/arxiv.2304.10759

[29] Shi, Y., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., and Sun, M. (2024). VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. https://doi.org/10.48550/arxiv.2410.10594

[30] Wang, Z., Ge, X., Jose, J. M., Yu, H.-T., Ma, W., Ren, Z., and Xin, X. (2024). R 3 AG: First Workshop on Refined and Reliable Retrieval Augmented Generation. 307–310. https://doi.org/10.1145/3673791.3698435

[31] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., and Shou, M. Z. (2024). Hallucination of Multimodal Large Language Models: A Survey. https://doi.org/10.48550/arxiv.2404.18930

[32] Banerjee, S., Agarwal, A., and Singla, S. (2024). LLMs Will Always Hallucinate, and We Need to Live With This. https://doi.org/10.48550/arxiv.2409.05746

[33] Villa, A., León Alcázar, J., Alfarra, M., Araujo, V., Soto, Á., and Ghanem, B. (2025). EAGLE: Enhanced Visual Grounding Minimizes Hallucinations in Instructional Multimodal Models. https://doi.org/10.48550/arxiv.2501.02699

[34] Sun, X., Peng, B., Zhang, C., Jin, F., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, H., and Zhang, Y. (2024). From Pixels to Prose: Advancing Multi-Modal Language Models for Remote Sensing. https://doi.org/10.48550/arxiv.2411.05826

[35] Imam, M. F., Lyu, C., and Aji, A. F. (2025). Can Multimodal LLMs do Visual Temporal Understanding and Reasoning? The answer is No! https://doi.org/10.48550/arxiv.2501.10674

[36] Nathvani, R., Vishwanath, D., Clark, S., Alli, A. S., Muller, E., Coste, H., Bennett, J. E., Nimo, J., Moses, J. B., Baah, S., Hughes, A. F., Suel, E., Metzler, A. B., Rashid, T., Brauer, M., Baumgartner, J., Owusu, G., Agyei-Mensah, S., Arku, R. E., and Ezzati, M. (2023). Beyond here and now: Evaluating pollution estimation across space and time from street view images with deep learning. Science of The Total Environment, 166168. https://doi.org/10.1016/j.scitotenv.2023.166168

[37] Wang, H., Chen, J., Pan, T., Dong, Z., Zhang, L., Jiang, R., and Song, X. (2024). Evaluating the Generalization Ability of Spatiotemporal Model in Urban Scenario. https://doi.org/10.48550/arxiv.2410.04740

[38] Marsocci, V., Jia, Y., Le Bellier, G., Kerekes, D., Zeng, L., Häfner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., and Nascetti, A. (2024). PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. https://doi.org/10.48550/arxiv.2412.04204

[39] Rane, J., Kaya, Ö., Mallick, S. K., and Rane, N. (2024). Artificial Intelligence-powered spatial analysis and ChatGPT-driven interpretation of remote sensing and GIS data. https://doi.org/10.70593/978-81-981271-7-4_5

[40] Zhu, J., Dang, P., Cao, Y., Lai, J., Guo, Y., Wang, P., and Li, W. (2024). A flood knowledge-constrained large language model interactable with GIS: enhancing public risk perception of floods. International Journal of Geographical Information Science. https://doi.org/10.1080/13658816.2024.2306167