Geo-spatial and Geo-temporal Reasoning in Vision-Language and Large Language Models : A Review

Shahriyar Zaman Ridoy[1,2♦†] **and Azmine Toushik Wasi**[1,3♦],
Koushik Ahamed Tonmoy[2], **Md Rizwan Parvez**[4]
[1]Computational Intelligence and Operations Laboratory [2]North South University
[3]Shahjalal University of Science and Technology [4]Qatar Computing Research Institute
[♦]Equal Contribution [†]**Correspondence:** shahriyar.ridoy@northsouth.edu

Abstract

Geo-spatial and geo-temporal reasoning require models to determine not only what information is present but also where and when it applies, which is critical for applications such as crisis response, environmental monitoring, and urban planning. This review surveys recent advances in vision–language models (VLMs) and large language models (LLMs) that address geographically and temporally grounded reasoning across tasks including image geolocation, spatiotemporal question answering, language-guided navigation, and Earth observation. We organize the literature through a structured analysis covering task families, input modalities, model architectures, and reasoning paradigms, highlighting how perception, retrieval, tool use, and intermediate reasoning steps contribute to system performance. The survey further synthesizes existing datasets and benchmarks, identifying common evaluation pitfalls such as spatial granularity mismatch, temporal drift, geographic bias, and potential data leakage. Finally, we outline key research challenges and opportunities for developing robust, verifiable, and time-aware geo-spatial AI systems while addressing broader concerns related to fairness, bias, and location privacy.

1 Introduction

Geo-spatial and geo-temporal reasoning require models to determine not only what is depicted or stated, but also *where* and *when* it applies, capabilities that are essential for applications such as crisis response, environmental monitoring, urban planning, and cultural analysis, where outputs must be geographically and temporally grounded rather than merely plausible in language space (Ji et al., 2025; Schottlander and Shekel, 2025a; Phillips and Matias, 2025; Ning, 2025). Compared with generic captioning or visual question answering, these tasks demand integration of heterogeneous evidence, including street-level and overhead imagery, remote sensing data, maps, graphs, text, and timestamps, alongside structured knowledge of landmarks, spatial relations, and evolving world states (Zhang et al., 2024a; Ou et al., 2025; Jia et al., 2025; Ji et al., 2025; Radford et al., 2021; Vivanco Cepeda et al., 2023).

Recent progress spans image geolocation and place recognition (Hays and Efros, 2008; Vo et al., 2017; Weyand et al., 2016; Astruc et al., 2024), spatiotemporal and knowledge-based question answering (Dai et al., 2024; Chen et al., 2021, 2025), language-guided navigation (Anderson et al., 2018; Chen et al., 2019), and Earth observation with multimodal geo-foundation models (Grainge et al., 2025; Smith et al., 2024; Zhang et al., 2024a; Schottlander and Shekel, 2025a; Roberts et al., 2024; Wang et al., 2024; Jia et al., 2025; Li et al., 2025a). Modern multimodal systems can exploit fine-grained geographic cues from imagery and, in certain settings, rival specialized geolocation approaches (Roberts et al., 2024; Postma and Patin, 2025; Grainge et al., 2025; Jia et al., 2025; Li et al., 2025a). Yet progress remains methodologically fragmented: many studies report end-task accuracy without isolating the contributions of perception, retrieval, tool use, intermediate reasoning, or verification (Yerramilli et al., 2025; Singh et al., 2024; Zhang et al., 2024c; Google Research, 2025; Yao et al., 2022; Schick et al., 2023; Karpas et al., 2022). Unlike prior surveys focused on GeoAI foundations or embodied agents (Jakubik et al., 2023; Ma et al., 2025), this review centers the reasoning layer that links perception to grounded geographic inference.

Geo-spatial tasks further amplify persistent challenges of ambiguity, distribution shift, and world change: visually similar scenes recur across regions, decisive cues may be sparse or occluded, and geographic knowledge, including place names, infrastructure, and points of interest, evolves over time (Grainge et al., 2025; Postma and Patin, 2025; Manvi et al., 2024a; Huang et al., 2025; Jia et al., 2025). Consequently, current systems remain brit-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1377–1411
November 6-10, 2026 ©2026 Association for Computational Linguistics

tle for ambiguous imagery, underrepresented regions, and temporally sensitive queries (Grainge et al., 2025; Postma and Patin, 2025; Jia et al., 2025). These limitations intersect with broader risks, including geographic bias and location privacy concerns (Grainge et al., 2025; Manvi et al., 2024a; Huang et al., 2025).

In this review, we synthesize advances in geo-spatial and geo-temporal reasoning for large language and vision–language models, as detailed in §B. We introduce a structured analysis (§2) of tasks , modalities, architectures, and reasoning paradigms; analyze datasets and evaluation pitfalls (§3); review applications and use cases (§4); and outline open challenges (§5) and future directions (§6) toward robust, verifiable, and time-aware geo-spatial AI.

2 Geo-spatial/-temporal Tasks & Methods

Research on geo-spatial and geo-temporal reasoning spans diverse problems, data sources, and modeling approaches. To structure this space, we organize the literature along four complementary dimensions: the *prediction target*, the *available evidence*, the *modeling interface*, and the *reasoning strategy*. Accordingly, this survey distinguishes four axes: **task families**, which define objectives such as coordinate inference, place identification, navigation, or spatiotemporal fact retrieval; **input modalities**, including images, maps, text, timestamps, and GIS layers; **model architectures**, which govern how these signals are represented and fused; and **reasoning paradigms**, which range from end-to-end prediction to retrieval-, tool-, or verification-based pipelines. Our analysis provides a unified framework for examining work across geolocation, spatiotemporal question answering, navigation, and Earth observation. By separating tasks, inputs, models, and reasoning strategies, it clarifies how different components contribute to system behavior and helps reveal recurring challenges such as geographic ambiguity, distribution shift, temporal drift, and uneven spatial coverage.

2.1 Task Categories

Geo-spatial reasoning tasks differ primarily in *what must be inferred* and *how that inference can be validated*. Across the literature, five recurring families emerge: location inference from perceptual inputs, spatiotemporal question answering over dynamic facts, embodied navigation and mapping, Earth observation analysis, and reasoning over symbolic geographic structures.

■ **Image Geolocation and Place Recognition.** Image geolocation seeks to infer where an image was captured, typically through coordinate regression, spatial-cell classification, or retrieval against large geotagged corpora (Hays and Efros, 2008; Weyand et al., 2016). Closely related place recognition instead identifies a specific known location instance, often via large-scale visual retrieval pipelines (Vivanco Cepeda et al., 2023; Ji et al., 2025). Although both tasks rely on visual cues such as terrain, vegetation, architecture, or signage, recent work increasingly frames geolocation as a multimodal reasoning problem rather than a pure image-to-coordinate mapping (Cheng et al., 2025; Wasi et al., 2026b). The central difficulty lies in the combination of fine spatial granularity and limited observable evidence. Visually similar scenes recur globally, decisive cues are often absent, and geographic coverage in training data is uneven, producing systematic failure under ambiguity or temporal drift (Grainge et al., 2025). As a result, recent benchmarks emphasize evidence-grounded predictions and explanation traces rather than raw coordinate accuracy (Cheng et al., 2025). Correspondingly, model designs increasingly separate local clue extraction (for example text or traffic conventions) from broader contextual inference, reconnecting them through retrieval or external knowledge. In practice, improvements often arise less from larger encoders than from better decomposition and verification of spatial hypotheses.

■ **Spatiotemporal Question Answering.** Spatiotemporal question answering assesses whether models can resolve natural-language queries with explicit spatial and temporal constraints. Geo-spatial QA involves relations such as proximity or containment, while **geo-temporal** QA requires reasoning over facts that vary over time (Dai et al., 2024). Many queries demand multi-hop grounding across maps, points-of-interest databases, schedules, and historical records (Chen et al., 2021, 2025). Benchmarks like POI-QA highlight the compositional nature of these tasks: systems must satisfy location, category, and temporal constraints within urban settings (Han et al., 2025). Current LLMs remain fragile in multi-constraint scenarios, with errors arising from incorrect decomposition or inconsistent execution rather than linguistic misunderstanding (Grainge et al., 2025), and benchmark coverage is skewed toward Global North cities, risk-

ing overestimation of generalization (Manvi et al., 2024a). Consequently, spatiotemporal QA is increasingly framed as a tool-augmented reasoning problem, where models generate structured queries or intermediate programs that interact with map services or databases, emphasizing verifiable execution over generative inference.

■ **Vision-and-Language Navigation and Mapping.** Vision–language navigation (VLN) studies embodied geo-spatial reasoning, requiring agents to interpret natural-language instructions and navigate with partial visual observations. Benchmarks such as **Room-to-Room (R2R)** (Anderson et al., 2018), **CVDN** (Thomason et al., 2020), and **Touchdown** (Chen et al., 2019) illustrate how VLN demands spatial memory, reference resolution, and long-horizon planning, with real-world performance lagging due to sim-to-real gaps. Modern approaches increasingly decouple linguistic planning from low-level control: hierarchical architectures allow LLMs to propose subgoals or route sketches while specialized controllers handle movement, perception, and localization (Driess et al., 2023; Schottlander and Shekel, 2025a; Ma et al., 2025; Wasi et al., 2026a). Related work on **mapping and map augmentation** extends this paradigm, translating multimodal observations into structured geographic updates, such as OpenStreetMap tags from street-level imagery (Juhász et al., 2023). Together, these approaches shift focus from scene understanding to actionable, info-grounded geo-spatial reasoning.

■ **Earth Observation and Remote Sensing Analysis.** Earth observation (EO) extends geo-spatial reasoning to satellite and aerial imagery, often including multispectral or radar inputs, for tasks such as land-cover mapping, object detection, and temporal change analysis. Traditionally addressed with domain-specific pipelines, EO is increasingly framed as a multimodal foundation-model problem supporting natural-language querying and interactive analysis. Models like *EarthGPT* enable question answering and summarization over multisensor inputs (Zhang et al., 2024a; Manvi et al., 2024b), while *GeoPix* couples language queries with dense spatial grounding for pixel-level outputs (Ou et al., 2025). Realizing this requires aligning language supervision with EO annotations, motivating datasets that link imagery with captions and region labels (Manvi et al., 2024b), with early language-facing map pipelines showing similar trends (Juhász et al., 2023). EO highlights both opportunity and challenge: heterogeneous sensors, spectral bands, and extreme scales demand specialized pretraining, while failures from cloud occlusion, spectral shift, or temporal misalignment persist (Grainge et al., 2025). EO thus reshapes the representation problem, requiring reasoning across sensor physics, scale, and time.

■ **Knowledge-based Geo-spatial Reasoning.** A complementary family works over **symbolic** or **database-backed** geographic information, including toponym resolution, route optimization, graph queries, and qualitative spatial reasoning. While classical GIS software and graph algorithms handle these tasks, LLMs are evaluated on how they interface with structured data without sacrificing exactness. Pretrained models encode nontrivial geographic priors, sometimes resembling latent maps (Manvi et al., 2024b), and *GeoLLM* uses prompted context to predict location-indexed indicators, though performance drops in data-sparse regions (Manvi et al., 2024b). Explicit graph interfaces such as *Graph-Instruct* and *GraphGPT* enable LLM interaction with road networks or knowledge graphs (Luo et al., 2024; Tang et al., 2024; Li et al., 2025b), but large-scale traversal and optimization still rely on classical algorithms, and qualitative spatial reasoning over relations like *near* or *between* remains fragile (Ji et al., 2025). Progress is therefore strongest where neural flexibility is coupled with explicit algorithmic back-ends rather than treated as a substitute.

2.2 Input Modalities

Input modality shapes both the evidence available and the types of ambiguity a model faces. Text is expressive but underspecified, images are rich but implicit, maps are precise yet brittle, and multimodal inputs combine these trade-offs. Modality choice is increasingly treated as a reasoning decision rather than a simple data-format issue.

■ **Text (Natural Language).** Textual inputs, questions, captions, or narratives, enable geo-spatial reasoning through LLM-encoded world knowledge, supporting tasks like toponym disambiguation or translating requests into geo-spatial queries (Manvi et al., 2024b; Schottlander and Shekel, 2025a). Yet text is inherently *underspecified*: place names are ambiguous, routes omit constraints, and temporal references are often implicit. Its strength arises when coupled with external resources such as gazetteers, maps, or databases, but this also raises privacy risks, since effective models can infer sensitive locations from casual mentions.

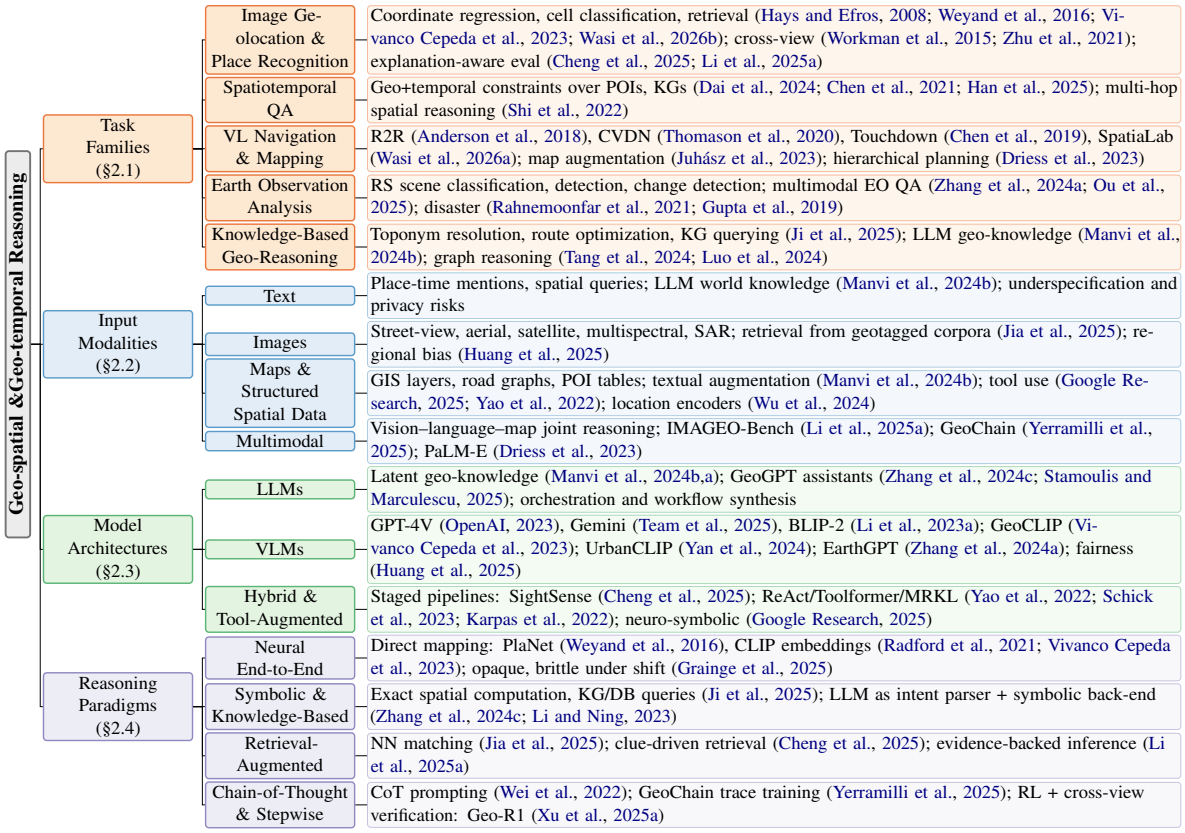


Figure 1: Overview of the geo-spatial/-temporal tasks and methods: task families (§2.1), input modalities (§2.2), model architectures (§2.3), reasoning paradigms (§2.4)

■ **Images (Vision).** Images carry dense but implicit geo-spatial cues, landscape, architecture, signage, climate, and seasonal markers, that can support location and temporal inference. Multimodal models like GPT-4V can combine these cues with world knowledge to generate plausible geolocation hypotheses (Roberts et al., 2024; Postma and Patin, 2025). However, visual reasoning is biased toward well-represented regions, and diverse sources, street-level, panoramic, aerial, satellite, multispectral, SAR, and video, differ drastically in scale, physics, and annotation regimes. Retrieval-based pipelines remain effective but uneven, as sparsely represented areas yield weaker priors and sparse indices (Jia et al., 2025).

■ **Maps and Structured Spatial Data.** Structured data, maps, GIS layers, coordinates, and POI tables, supports precise operations like geocoding, distance computation, and routing, which cannot be approximated by language alone (Manvi et al., 2024a). LLMs interface via textual augmentation of features (Manvi et al., 2024b; Buckmann et al., 2025), tool use with geocoders and routing engines (Ning, 2025; Phillips and Matias, 2025; Yao et al., 2022; Karpas et al., 2022; Schick et al., 2023), or architectural integration with coordinate embeddings (Manvi et al., 2024b; Wu et al., 2024). Each strategy balances precision, flexibility, and cost, but all must contend with interface complexity and coverage bias.

■ **Multimodal Inputs.** Realistic geo-spatial tasks are inherently multimodal, requiring sequential integration of visual, textual, and structured signals. Benchmarks such as IMAGEO-Bench (Li et al., 2025a) and GeoChain (Yerramilli et al., 2025) illustrate challenges where models alternate between perceptual extraction and knowledge reasoning. Embodied settings heighten alignment demands: language, visual streams, and map objects must be co-grounded without letting any modality dominate (Driess et al., 2023). Failures arise from imbalance, hallucination propagation, or regional bias, underscoring that multimodal geo-spatial competence is fundamentally about *shared grounding*, not input concatenation.

2.3 Model Architectures

Model architecture shapes where geo-spatial competence resides, parametric knowledge, perceptual encoders, external tools, or the coordination logic

connecting them. Progress comes from careful combinations rather than dominance of a single architecture.

■ **Large Language Models (LLMs).** Text-only LLMs encode latent geographic and temporal knowledge, supporting coarse spatial reasoning, explanation, and query decomposition (Manvi et al., 2024b). Their limits are structural: no perceptual grounding, unreliable for exact spatial computations, and biased toward well-documented regions. Consequently, LLMs perform best as orchestrators, translating user intent, decomposing tasks, invoking tools, and synthesizing results (Google Research, 2025). GeoGPT-style assistants and prompted LLM place encoders illustrate this pattern (Stamoulis and Marculescu, 2025; Zhang et al., 2024c; Manvi et al., 2024a), highlighting that abstraction and interaction matter more than raw parametric knowledge for geo-spatial tasks.

■ **Vision-Language Models (VLMs).** VLMs provide perceptual grounding for geo-spatial reasoning, linking image inputs to language outputs. General-purpose models such as GPT-4V, InstructBLIP, and LLaVA identify landmarks and generate coarse geolocation hypotheses (Postma and Patin, 2025; Roberts et al., 2024), but struggle on subtle local cues. Fine-tuned or domain-specific models, GeoCLIP-style embeddings, urban scene models, and EO-focused VLMs, improve recognition and retrieval (Jia et al., 2025; Manvi et al., 2024b; Juhász et al., 2023). Benchmarking shows persistent regional biases and uneven transfer, emphasizing that performance gains reflect either genuine cue learning or benchmark-specific exploitation (Huang et al., 2025; Jia et al., 2025). Cross-view modeling with aerial and ground imagery further tests representational generality.

■ **Hybrid and Tool-Augmented Systems.** Top-performing systems combine LLMs or VLMs with external modules, OCR, retrieval, geocoders, routing engines, or GIS operators, to handle tasks requiring exact computation and up-to-date facts (Google Research, 2025). Pipelines often extract clues with vision modules, narrow candidates via retrieval, then finalize hypotheses with LLM reasoning and explanations (Cheng et al., 2025). Iterative tool use or neuro-symbolic execution executes formal queries against reliable back-ends (Juhász et al., 2023; Yao et al., 2022; Karpas et al., 2022; Schick et al., 2023). These designs improve grounding, auditability, and error localization, but add complexity and potential cascade failures. Hybrid architectures thus represent the current practical approach to reconciling fluent generation with precise geo-spatial computation.

2.4 Reasoning Approaches

Geo-spatial reasoning varies in whether inference is implicit, explicit, or externally grounded; the key distinction is whether intermediate claims can be verified against perception, retrieval, or computation.

■ **Neural End-to-End Reasoning.** End-to-end models map inputs directly to outputs without exposing intermediate steps. Classic examples include PlaNet for image geolocation (Weyand et al., 2016) and latent-structure predictors for spatiotemporal tasks (Jia et al., 2025). When data is abundant, this yields effective joint representation and prediction, but models are opaque, brittle under distribution shift, and inherit geographic biases (Grainge et al., 2025). Modern systems therefore wrap end-to-end perception with retrieval, tool use, or explicit verification rather than discard neural inference entirely.

■ **Symbolic and Knowledge-Based Reasoning.** Symbolic methods encode space and time explicitly through coordinates, graphs, or formal logic (Ji et al., 2025), enabling precise containment, distance, and routing computations. LLMs typically handle intent parsing or explanation, while queries execute via SQL, SPARQL, or tool-call sequences (Ji et al., 2025). This is critical for constraint-heavy tasks like route planning, yet governance remains fragile: schema mismatches and tool-invocation errors persist (Grainge et al., 2025). Symbolic execution is strongest when neural flexibility is paired with verifiable computation.

■ **Retrieval-Augmented Reasoning.** Retrieval-augmented models condition on external evidence, geotagged images, POI records, maps, or schedules, rather than relying solely on parametric memory (Cheng et al., 2025). In geolocation, nearest-neighbor matching or clue-driven searches improve accuracy (Jia et al., 2025); in geo-temporal QA, dynamic facts and historical records are consulted (Grainge et al., 2025). This open-book approach mitigates aging and tail-data issues, but introduces failures from sparse coverage, low-quality hits, or geographic bias (Li et al., 2025a).

■ **Chain-of-Thought and Stepwise Reasoning.** CoT methods decompose tasks into explicit intermediate decisions, such as cue extraction, candidate generation, elimination, and verification

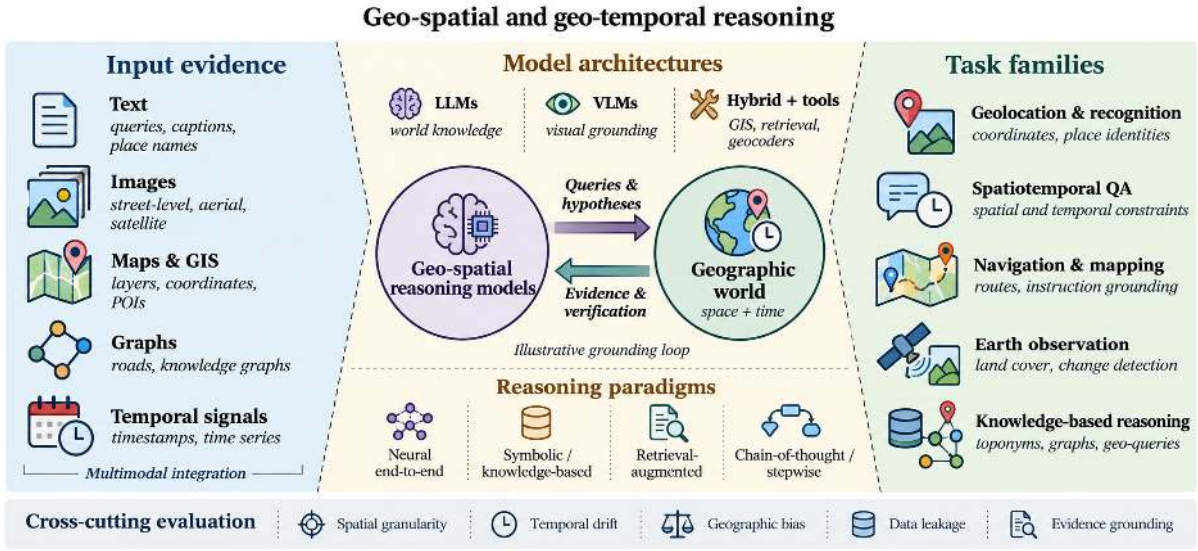


Figure 2: Overview of geo-spatial and geo-temporal reasoning in large language and vision–language models. The figure connects heterogeneous input evidence, model architectures, reasoning paradigms, and five task families. The central loop illustrates evidence grounding and verification in retrieval- and tool-assisted systems, while the bottom band highlights cross-cutting evaluation concerns.

(Wei et al., 2022). Applied to geolocation and geo-temporal QA, they improve accuracy and inspectability (Roberts et al., 2024; Postma and Patin, 2025). GeoChain provides multi-step supervision for Street View reasoning (Yerramilli et al., 2025), while branching deliberation explores multiple candidate paths (Cheng et al., 2025). Crucially, stepwise reasoning is only valuable when intermediate claims are checkable; benchmarks increasingly score explanations against observable evidence rather than accepting fluent narratives as proof (Li et al., 2025a; Cheng et al., 2025).

2.5 Design Choice Effectiveness and Failure Modes

Table 1 synthesizes which architectural and reasoning design choices are most effective under specific task settings, together with the primary failure modes identified across the surveyed corpus. These observations are descriptive: they reflect patterns across the 142 reviewed papers rather than controlled ablation studies.

3 Datasets and Benchmarks

Datasets and benchmarks define both what geospatial models learn and which errors remain invisible. Recent work increasingly evaluates reasoning, grounding, explanation quality, and robustness under geographic and temporal shift rather than raw accuracy (Grainge et al., 2025; Wang et al., 2024; Li et al., 2025a). Benchmark validity is often compromised by pretraining overlap, uneven coverage, temporal staleness, and metric mismatch with deployment needs. Here, we organize key resources by task family (Table 6), emphasizing datasets that have become reference points for LLM/VLM, retrieval-augmented, and tool-assisted systems.

■ **Image Geolocation and Landmark Recognition.** Geotagged corpora such as *YFCC100M* provide foundational training supervision, while evaluation datasets like *Im2GPS* and *Im2GPS3k* maintain continuity across model generations (Thomee et al., 2016; Hays and Efros, 2008; Vo et al., 2017). Landmark identification relies on *Google Landmarks v2*, and cross-view datasets like *TimeSpot* (Wasi et al., 2026b), *CVUSA*, *CVACT*, and *VIGOR* stress robustness across aerial-ground alignment (Workman et al., 2015; Liu and Li, 2019; Zhu et al., 2021). Recent benchmarks such as *LLM-Geo* and *GeoPose3K* highlight multimodal reasoning and fine-grained evaluation (Wang et al., 2024; Brejcha and Čadík, 2017), while *IMAGEO-Bench*, *GeoExplain*, and *GeoGuess* prioritize explanation quality, bias exposure, and evidence-based evaluation (Li et al., 2025a; Huang et al., 2025; Cheng et al., 2025). The field is moving from pure localization metrics toward diagnostic assessment of whether models *localize for the right reasons*.

■ **Spatiotemporal and Knowledge-Based QA.** Generic QA datasets such as *SQuAD* and *TriviaQA*

Table 1: Design effectiveness and dominant failure modes by task family. **Best design** denotes the most consistently effective architectural pattern, while **Failure mode** denotes the most frequently reported limitation.

Task Family	Most Effective Design Choice	Primary Failure Mode	Representative Evidence
Image Geolocation	Retrieval + VLM re-ranking; staged clue extraction over single-shot prediction	Regional coverage gap; near-duplicate leakage	(Cheng et al., 2025; Jia et al., 2025; Li et al., 2025a)
Spatiotemporal QA	Tool-augmented LLMs querying structured databases; CoT decomposition	Temporal staleness; incorrect multi-hop decomposition	(Dai et al., 2024; Han et al., 2025; Chen et al., 2021)
VL Navigation	Hierarchical planning (LLM subgoal + specialist controller)	Sim-to-real gap; direction error propagation	(Anderson et al., 2018; Driess et al., 2023; Ma et al., 2025)
Earth Observation	Domain-specific pretraining (EO-aligned corpora); pixel-level grounding	Spectral/sensor domain shift; temporal misalignment	(Zhang et al., 2024a; Ou et al., 2025; Smith et al., 2024)
Knowledge-Based Geo-Reasoning	LLM as intent parser + symbolic/graph back-end execution	Schema mismatch; qualitative spatial relation failures	(Ji et al., 2025; Zhang et al., 2024c; Li and Ning, 2023)

remain useful for coarse factual testing (Rajpurkar et al., 2016; Joshi et al., 2017), but spatiotemporal QA demands structured reasoning over location, relation, and time. *POI-QA* requires joint resolution of points of interest and temporal constraints (Grainge et al., 2025; Han et al., 2025), while *ChronoQA* and *TimeQA* isolate temporal reasoning (Chen et al., 2025, 2021), and *StepGame* targets multi-hop qualitative spatial inference (Shi et al., 2022). *STKGQA* frames QA over spatiotemporal knowledge graphs, demanding explicit query execution (Dai et al., 2024). Tool-augmented evaluations further emphasize evidence retrieval and platform operations (Singh et al., 2024), reflecting that robust spatiotemporal QA hinges more on grounding and execution than linguistic fluency.

■ **Navigation and Embodied Benchmarks.** Navigation benchmarks probe sequential geo-spatial reasoning. Indoor suites like *R2R*, *RxR*, *CVDN*, and *ALFRED* test instruction grounding, memory, and planning (Anderson et al., 2018; Ku et al., 2020; Thomason et al., 2020; Shridhar et al., 2020), while outdoor datasets such as *Touchdown* and *StreetNav* add urban complexity and localization (Chen et al., 2019; Hermann et al., 2020). Embodied QA and dialog navigation (*EQA*, *NDH/CVDN*) extend these challenges by requiring evidence gathering (Das et al., 2018; Thomason et al., 2020). Mapping proxies like *Talk2Nav* and OSM-completion tasks test structured geographic updates (Vasudevan et al., 2021; Juhász et al., 2023). Evaluation increasingly favors trajectory-faithfulness and interactive competence over static correctness (Magalhaes et al., 2019).

■ **Remote Sensing and Earth Science.** Remote-sensing datasets remain fragmented, reflecting domain-specific origins. Large-scale detection is benchmarked via *xView* and *DOTA* (Lam et al., 2018; Xia et al., 2018), while caption datasets such as *RSICD* and *UCM Captions* pair imagery with text, and *RSVQA-Bench* supports QA over overhead scenes (Lu et al., 2017; Qu et al., 2016; Lobry et al., 2021). Disaster and temporal datasets like *FloodNet* and *xBD* enable change detection and assessment tasks (Rahnemoonfar et al., 2021; Gupta et al., 2019). Multimodal time-series resources such as Google's *PDFM* and spatial representation benchmarks like *OBSR* promote evaluation of representation transfer and reasoning, not just classification (Agarwal et al., 2025; Moska et al., 2025). Together, these benchmarks shift Earth observation evaluation from isolated tasks toward integrated perception, grounding, and temporal reasoning.

4 Applications and Use Cases

Geo-spatial LLM/VLM systems are most effective as interface and triage layers over existing pipelines, helping users query heterogeneous data, summarize outputs, and surface candidate actions while leaving verification to classical tools and humans (Table 5). Across domains, AI primarily accelerates access and synthesis rather than replacing expertise.

■ **Earth Observation & Environmental Monitoring.** EO systems combine imagery, spatial layers, and time series. Models such as *EarthGPT* and *GeoPix* support instruction-following over imagery and spatially grounded outputs for monitoring and disaster triage (Zhang et al., 2024a; Ou et al., 2025; Rahnemoonfar et al., 2021; Gupta et al., 2019). Text-oriented systems like *ClimateBERT* and *ClimateGPT* assist climate-document synthesis (Webersinke et al., 2021; Thulke et al., 2024),

while multimodal models such as *PDFM* learn transferable population–environment representations (Agarwal et al., 2025). In practice, these systems reorganize workflows around retrieval, triage, and evidence-linked summaries (Schottlander and Shekel, 2025b).

■ **Crisis Response & Humanitarian Aid.** AI mainly serves as a decision-support layer in disasters. LLMs summarize pre/post imagery for localized damage assessment on benchmarks such as *xBD* and *xView2* (Gupta et al., 2019; Singh et al., 2024), while SAR systems convert reports into structured hypotheses without replacing probabilistic planning (Ji et al., 2025; Google Research, 2025; Abi-Zeid et al., 2019). Humanitarian mapping accelerates candidate generation for *Missing Maps* and OSM while preserving human validation (Missing Maps; UN Maps; OpenStreetMap Wiki contributors; Juhász et al., 2023). Resource-allocation systems similarly rank options under explicit assumptions rather than issuing autonomous recommendations (Ji et al., 2025; Pizzi et al., 2020; Gella et al., 2023).

■ **Navigation & Transportation.** LLMs translate underspecified goals into routing or POI queries and explain trade-offs in natural language (Ji et al., 2025; Schottlander and Shekel, 2025b; Google Research, 2025; Daniel, 2024). They also augment traffic forecasting with contextual signals (Zhao et al., 2025), though classical solvers remain essential for safety and feasibility. Similar constraints apply to delivery robotics and drones, where language-guided planning must remain grounded in perception and action models (Driess et al., 2023; Ma et al., 2025).

■ **Smart Cities & Urban Planning.** Conversational GIS is a strong near-term application, translating natural-language queries into geo-spatial retrieval and analysis (Google Research, 2025; Li and Ning, 2023; Akinboyewa et al., 2025; Ning et al., 2025). Generative visualizations aid rapid "what-if" communication (Hua et al., 2025), but outputs remain assistive rather than policy recommendations (Li and Ning, 2023; Zhang et al., 2025b). The main benefit is shortening analytic loops and improving accessibility.

■ **Social Media Geolocation & OSINT.** LLMs and VLMs can infer candidate locations from images or videos and highlight supporting cues, reducing verification effort while leaving final judgment to humans (Postma and Patin, 2025). These capabilities accelerate OSINT workflows but also raise privacy concerns because even non-obvious imagery may be geolocated (Grainge et al., 2025). OSINT therefore highlights both the utility and dual-use risks of geo-spatial AI.

Across domains, the dominant pattern is human–AI collaboration: models accelerate retrieval, synthesis, and preliminary interpretation, while experts remain responsible for verification and high-stakes decisions.

5 Open Challenges

Despite rapid progress, geo-spatial and geo-temporal AI systems remain limited by **temporally stale world models**, **weak multimodal grounding**, **uneven geographic coverage**, costly deployment, and poor interoperability with trusted GIS workflows. **Temporal reasoning** is a major weakness because models often fail to account for changing world states, motivating timestamp-aware training, retrieval over time-indexed knowledge bases, and evaluation protocols with time-sliced and recency-aware splits (Han et al., 2025; Chen et al., 2021, 2025; Dai et al., 2024; Smith et al., 2024). **Multimodal alignment** remains unreliable across geometry, cross-view consistency, modality arbitration, and fine-grained grounding, highlighting the need for tool-based verification, pixel-level grounding, and explicit auditing frameworks (OpenAI, 2023; Ji et al., 2025; Xu et al., 2025a; Liu and Li, 2019; Zhu et al., 2021; Driess et al., 2023; Ou et al., 2025; Zhang et al., 2024b; Chen et al., 2024; Liu et al., 2024). **Data bias and coverage gaps** further create systematic disparities across regions, languages, time periods, and capture conditions, reducing robustness in underrepresented settings (Huang et al., 2025; Li et al., 2025a; Jia et al., 2025). Addressing these challenges will require broader geo-spatial datasets, multilingual and region-aware retrieval, calibrated abstention, and geographically stratified evaluation rather than reliance on single aggregate benchmarks (Astruc et al., 2024). More details are available in Appendix D.

6 Future Directions

The directions below reflect the authors' prioritized recommendations based on the limitations identified in §3–5, rather than descriptive findings from the corpus. Future geo-spatial AI should move toward **unified 4D foundation models** integrating imagery, vector geometries, text, place graphs, and temporal signals. These systems should support

multi-resolution spatial hierarchies, **world-state versioning**, and **calibrated uncertainty** for probabilistic spatial reasoning (Dai et al., 2024).

✓ **Interactive and verifiable agents** should translate natural-language instructions into GIS, remote-sensing, and knowledge-graph workflows, returning auditable intermediate artifacts like buffers, joins, and route constraints. Closed-loop verification over authoritative sources ensures geometric and topological consistency. Tool-augmented architectures (ReAct, Toolformer, MRKL) and retrieval-augmented generation enable precise, current-data grounded reasoning (Yao et al., 2022; Schick et al., 2023; Karpas et al., 2022).

✓ **Human-in-the-loop mapping** extends coverage to under-mapped or crisis regions. Active spatial learning prioritizes high-uncertainty areas, while provenance tracking records contributors, timestamps, and evidence. Integration with crowdsourced platforms improves calibration and prevents overconfident predictions (Missing Maps; UN Maps; Juhász et al., 2023).

✓ **Governance, privacy, and dual-use safety** require red-teaming, bias audits, access control, anonymization, and query logging. Sensitive inferences must be restricted, with careful release management to mitigate misuse.

Overall, near-term progress favors *grounded, retrievable, tool-augmented, auditable, and human-centered systems* rather than monolithic *geo-spatial AGI*, emphasizing *robustness, interpretability, ethical deployment, and verified reasoning*. More details are available in Appendix E.

7 Conclusion

Geo-spatial and geo-temporal AI advances depend on aligning tasks, modalities, architectures, and reasoning strategies. Success increasingly relies on modular, tool-augmented, or stepwise approaches that combine neural flexibility with explicit verification. Multimodal integration, retrieval, and symbolic grounding help mitigate ambiguity, temporal drift, and geographic bias, while benchmarks now emphasize evidence-backed predictions. Robust systems reason over heterogeneous inputs, decompose complex queries, and produce checkable intermediate claims, highlighting the need for careful design, evaluation, and governance.

Limitations

This review, while comprehensive in surveying LLMs and VLMs for geo-spatial and geo-temporal reasoning, inevitably cannot capture every recent development due to the rapid pace of AI research. It emphasizes conceptual synthesis and cross-domain comparison rather than presenting new experiments, empirical validation, or system-level benchmarks. As such, quantitative performance evaluation, implementation details, and comparative analyses of specific models fall outside its scope. The focus is on providing a structured framework and highlighting trends, trade-offs, and open challenges, rather than introducing new datasets, algorithms, or technical contributions. Consequently, readers should interpret the survey as a high-level analytic perspective rather than a source of empirical conclusions. Several exclusion decisions may affect corpus representativeness. Non-English papers and workshop abstracts without full technical content were removed to maintain analytical consistency, but this may underrepresent geo-spatial AI work published in Chinese, Japanese, or other languages, and early-stage contributions presented at workshops. The exclusion of arXiv preprints from quality filtering is addressed in Appendix B.10. The fast-moving preprint landscape means the survey reflects a snapshot of the field as of early 2025, and studies published after the curation cutoff are not covered.

References

Irène Abi-Zeid, Michael Morin, and Oscar Nilo. 2019. Decision support for planning maritime search and rescue operations in canada.

Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, Swapnil Vispute, Atul Kumar, Yael Mayer, David Fork, Sheila de Guia, Jamie McPike, Adam Boulanger, Tomer Shekel, David Schottlander, Yao Xiao, Manjit Chakravarthy Manukonda, Yun Liu, Neslihan Bulut, Sami Abu el haija, Bryan Perozzi, Monica Bharel, Von Nguyen, Luke Barrington, Niv Efron, Yossi Matias, Greg Corrado, Krish Eswaran, Shruthi Prabhakara, Shravya Shetty, and Gautam Prasad. 2025. General geospatial inference with a population dynamics foundation model. *Preprint*, arXiv:2411.07207.

Temitope Akinboyewa, Zhenlong Li, Huan Ning, and M Naser Lessani. 2025. Gis copilot: Towards an autonomous gis agent for spatial analysis. *International Journal of Digital Earth*, 18(1):2497489.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao XU, Hongyu Zhou, and Loic Landrieu. 2024. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Jan Brejcha and Martin Čadík. 2017. Geopose3k: Mountain landscape dataset for camera pose estimation in outdoor environments. *Image and Vision Computing*, 66:1–14.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Marcus Buckmann, Quynh Anh Nguyen, and Edward Hill. 2025. Revealing economic facts: Llms know more than they say. *Preprint*, arXiv:2505.08662.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *Preprint*, arXiv:2108.06314.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. 2024. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. 2025. A question answering dataset for temporal-sensitive retrieval-augmented generation. *Preprint*, arXiv:2508.12282.

Fenghua Cheng, Jinxiang Wang, Sen Wang, Zi Huang, and Xue Li. 2025. Geoguess: Multimodal reasoning based on hierarchy of visual information in street view. *Preprint*, arXiv:2506.16633.

Xinbang Dai, Huiying Li, and Guilin Qi. 2024. Question answering over spatio-temporal knowledge graph. *Preprint*, arXiv:2402.11542.

Miriam Daniel. 2024. New in maps: Inspiration curated with gemini, enhanced navigation and more. Accessed: 2026-01-03.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. *Preprint*, arXiv:2303.03378.

Emilie Durrieu, Christophe Hurter, Philippe Muller, and Victor Boutin. 2026. Object-level explanations for image geolocation models: a geoguessr use-case. *Preprint*, arXiv:2605.00912.

European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Accessed: 2026-03-04.

Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, Songwei Li, Yunke Zhang, Yuming Lin, Tong Li, Jingtao Ding, Chen Gao, Fengli Xu, and Yong Li. 2025. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *Preprint*, arXiv:2504.09848.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.

Getachew Workineh Gella, Dirk Tiede, Stefan Lang, Lorenz Wendit, and Yunya Gao. 2023. Spatially transferable dwelling extraction from multi-sensor imagery in idp/refugee settlements: A meta-learning approach. *International Journal of Applied Earth Observation and Geoinformation*, 117:103210.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Google Research. 2025. Introducing geospatial reasoning: Unlocking insights with generative ai and multiple foundation models. Accessed: 2025-10-25.

Oliver Grainge, Sania Waheed, Jack Stilgoe, Michael Milford, and Shoaib Ehsan. 2025. Assessing the geolocation capabilities, limitations and societal risks of generative vision-language models. *Preprint*, arXiv:2508.19967.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. 2019. xbd: A dataset for assessing building damage from satellite imagery. *Preprint*, arXiv:1911.09296.

Xiao Han, Dayan Pan, Xiangyu Zhao, Xuyuan Hu, Zhaolin Deng, Xiangjie Kong, and Guojiang Shen. 2025. A dataset for spatiotemporal-sensitive poi question answering. *Preprint*, arXiv:2505.10928.

James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Tongyan Hua, Lutao Jiang, Ying-Cong Chen, and Wufan Zhao. 2025. Sat2city: 3d city generation from a single satellite image with cascaded latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27978–27988.

Jingyuan Huang, Jen tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. 2025. Ai sees your location, but with a bias toward the wealthy world. *Preprint*, arXiv:2502.11163.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. 2023. Foundation models for generalist geospatial artificial intelligence. *Preprint*, arXiv:2310.18660.

Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. 2025. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science*, pages 1–38.

Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Sharon Li. 2025. Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization. *Preprint*, arXiv:2509.04334.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Levente Juhász, Peter Mooney, Hartwig H Hochmair, and Boyuan Guan. 2023. Chatgpt as a mapping assistant: A novel method to enrich maps with generative ai and content derived from street-level photographs. *arXiv preprint arXiv:2306.03204*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Andreas Kamilaris and Francesc X Prenafeta-Boldú. 2018. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *Preprint*, arXiv:2205.00445.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. 2023. Towards global, general-purpose pretrained geographic location encoders. In *NeurIPS 2023 Workshop on Distribution Shifts (DistShift)*. Poster; OpenReview.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *Preprint*, arXiv:2010.07954.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. 2018. xview: Objects in context in overhead imagery. *Preprint*, arXiv:1802.07856.

Jett Legacion, Mallory Delgadillo, Nicholas Giner, and Priscilla Kim. 2025. What's new in ai assistants (october 2025). ArcGIS Blog. Accessed: 2026-03-04.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025a. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *Preprint*, arXiv:2508.01608.

Xin Sky Li, Qizhi Chu, Yubin Chen, Yang Liu, Yaoqi Liu, Zekai Yu, Weize Chen, Chen Qian, Chuan Shi, and Cheng Yang. 2025b. Graphteam: Facilitating large language model-based graph analysis via multi-agent collaboration. *Preprint*, arXiv:2410.18032.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023b. Geolm: Empowering language models for geospatially grounded language understanding. *Preprint*, arXiv:2310.14478.

Zhenlong Li and Huan Ning. 2023. Autonomous gis: the next-generation ai-powered gis. *International Journal of Digital Earth*, 16(2):4668–4686.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Liu Liu and Hongdong Li. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. *Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection*, page 38–55. Springer Nature Switzerland.

Sylvain Lobry, Begüm Demir, and Devis Tuia. 2021. Rsvqa meets bigearthnet: a new, large-scale, visual question answering dataset for remote sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1218–1221. IEEE.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

Zihan Luo, Xiran Song, Hong Huang, Jianxun Lian, Chenhao Zhang, Jinqi Jiang, and Xing Xie. 2024. Graphinstruct: Empowering large language models with graph understanding and reasoning capability. *Preprint*, arXiv:2403.04483.

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2025. A survey on vision-language-action models for embodied ai. *Preprint*, arXiv:2405.14093.

Gabriel Ilharco Magalhaes, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. In *NeurIPS Visually Grounded Interaction and Language (ViGIL) Workshop*, volume 1.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2024. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Trans. Spatial Algorithms Syst.*, 10(2).

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024a. Large language models are geographically biased. *Preprint*, arXiv:2402.02680.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2024b. Geollm: Extracting geospatial knowledge from large language models. *Preprint*, arXiv:2310.06213.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. 2023. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Missing Maps. Missing maps. Accessed: 2026-01-03.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Medicine*, 6(7):e1000097.

Marco Monti, Oliver Kutz, Nicolas Troquard, Guendalina Righetti, and Nicolas Troquard. 2024. Improving the accuracy of black-box language models with ontologies: A preliminary roadmap. In *JOWO*.

Julia Moska, Oleksii Furman, Kacper Kozaczko, Szymon Leszkiewicz, Jakub Polczyk, Piotr Gramacki, and Piotr Szymański. 2025. Obsr: Open benchmark for spatial representations. *Preprint*, arXiv:2510.05879.

Zachary Munn, Micah D J Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(1):143.

Huan Ning. 2025. Geospatial reasoning by google: A leap towards autonomous gis. *Penn State GeoInformation and Big Data Research Laboratory*. Accessed: 2025-10-25.

Huan Ning, Zhenlong Li, Temitope Akinboyewa, and M Naser Lessani. 2025. An autonomous gis agent framework for geospatial data retrieval. *International Journal of Digital Earth*, 18(1):2458688.

OpenAI. 2023. Gpt-4v(ision) system card. Accessed: 2026-01-03.

OpenStreetMap Wiki contributors. Facebook ai-assisted road tracing. Accessed: 2026-01-03.

Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. 2025. Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*.

Chris Phillips and Yossi Matias. 2025. Google earth ai: Our state-of-the-art geospatial ai models. Accessed: 2025-10-27.

Michael Pizzi, Mila Romanoff, and Tim Engelhardt. 2020. Ai for humanitarian action: Human rights and ethics. *International Review of the Red Cross*, 102(913):145–180.

Foeke Postma and Nathan Patin. 2025. Have llms finally mastered geolocation? *Bellingcat*. Accessed: 2025-10-25.

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, pages 1–5. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. 2021. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2024. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–563.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

David Schottlander and Tomer Shekel. 2025a. Geospatial reasoning: unlocking insights with generative ai and multiple foundation models.

David Schottlander and Tomer Shekel. 2025b. Geospatial reasoning: Unlocking insights with generative ai and multiple foundation models. Accessed: 2025-10-27.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, Zhenzhong Kuang, and Jiajun Ding. 2025. Imp: Highly capable large multimodal models for mobile devices. *IEEE Transactions on Multimedia*.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11321–11329.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Simranjit Singh, Michael Fore, and Dimitrios Stamoulis. 2024. Evaluating tool-augmented agents in remote sensing platforms. *arXiv preprint arXiv:2405.00709*.

Michael J. Smith, Luke Fleming, and James E. Geach. 2024. Earthpt: a time series foundation model for earth observation. *Preprint*, arXiv:2309.07207.

Dimitrios Stamoulis and Diana Marculescu. 2025. Geoolm: Enabling sustainable earth observation studies with cost-efficient open language models & state-driven workflows. In *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 608–619.

Jian Sun, Kangdao Liu, Chi Zhang, Chuangquan Chen, Junge Shen, and Chi-Man Vong. 2025. Cross-view uav geo-localization with precision-focused efficient design: A hierarchical distillation approach with

multi-view refinement. *arXiv e-prints*, pages arXiv–2510.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. *Preprint*, arXiv:2310.13023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jin-

wei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Ray-

nald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer,

Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *Preprint*, arXiv:2401.09646.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

UN Maps. The community | un mappers. Accessed: 2026-01-03.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1):246–266.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701.

Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

W3C SPARQL Working Group. 2013. Sparql 1.1 query language. W3C Recommendation; Accessed: 2026-03-04.

Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700.

Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. 2024. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *Preprint*, arXiv:2405.20363.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. 2026a. Spatialab: Can vision-language models perform spatial reasoning in the wild? *Preprint*, arXiv:2602.03916.

Azmine Toushik Wasi, Shahriyar Zaman Ridoy, Koushik Ahamed Tonmoy, Kinga Tshering, S. M. Muhtasimul Hasan, Wahid Faisal, Tasnim Mohiuddin, and Md Rizwan Parvez. 2026b. Timespot: Benchmarking geo-temporal understanding in vision-language models in real-world settings. *Preprint*, arXiv:2603.06687.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, pages 37–55. Springer.

Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969.

Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, X. Angela Yao, Hongxu Ma, Lan Mu, Stefano Ermon, Tanuja Ganu, Akshay Nambi, Ni Lao, and Gengchen Mai. 2024. Torchspatial: A location encoding framework and benchmark for spatial representation learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Chenhui Xu, Fuxun Yu, Michael J. Bianco, Jacob Kovarskiy, Raphael Tang, Qi Zhang, Zirui Xu, Will LeVine, Brandon Dubbs, Heming Liao, Cassandra Burgess, Suvam Bag, Jay Patravali, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Jinjun Xiong. 2025a. Geo-r1: Unlocking vlm geospatial reasoning with cross-view reinforcement learning. *Preprint*, arXiv:2510.00072.

Daliang Xu, Hao Zhang, Liming Yang, Ruiqi Liu, Gang Huang, Mengwei Xu, and Xuanzhe Liu. 2025b. Fast on-device llm inference with npus. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 445–462.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*, pages 4006–4017.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. 2025. Geochain: Multimodal chain-of-thought for geographic reasoning. In *EMNLP (Findings)*, pages 23624–23639.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025a. Mapnav: A novel memory representation via annotated semantic maps for vision-and-language navigation. *Preprint*, arXiv:2502.13451.

Qianheng Zhang, Song Gao, Chen Wei, Yibo Zhao, Ying Nie, Ziru Chen, Shijie Chen, Yu Su, and Huan Sun. 2025b. Geoanalystbench: A geoai benchmark for assessing large language models for spatial analysis workflow and code generation. *Transactions in GIS*, 29(7):e70135.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. 2024a. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024b. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238.

Yifan Zhang, Cheng Wei, Zhengting He, and Wenhao Yu. 2024c. Geogpt: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131:103976.

Yusheng Zhao, Xiao Luo, Haomin Wen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025. Embracing large language models in traffic flow forecasting. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8108–8123.

Sijie Zhu, Taojiannan Yang, and Chen Chen. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36.

A Related Work

A number of surveys and reviews have examined geo-spatial AI, remote sensing, and spatiotemporal modeling, often with a focus on specific data types, applications, or methodological families. Prior work in remote sensing has extensively covered deep learning for satellite and aerial imagery analysis, including land use classification, object detection, and change detection (Zhu et al., 2017; Kamilaris and Prenafeta-Boldú, 2018), as well as convolutional network architectures for high-resolution imagery (Ma et al., 2019) and multimodal fusion approaches combining imagery with GIS databases (Wang et al., 2020). Reviews on spatiotemporal modeling and geo-spatial analytics have emphasized trajectory prediction, spatial networks, and urban computing (Feng et al., 2025), while GIS-focused surveys have examined workflow integration, crowdsourced mapping, and participatory sensing (Mai et al., 2024).

Despite their breadth, these prior surveys have several limitations. Most focus on narrow modalities such as imagery or vector data rather than unified 4D representations that jointly encode space and time. They tend to emphasize benchmark performance over methodological generalization, verification, or uncertainty calibration. Human-in-the-loop mapping, tool-augmented pipelines, and governance considerations, including privacy and dual-use safety, are rarely addressed. Furthermore, systematic curation of foundation-model-based approaches that integrate LLMs, VLMs, and GIS tools for spatiotemporal reasoning is largely absent. As a result, existing surveys provide only partial insight into the design patterns, trade-offs, and open challenges of next-generation geo-spatial AI systems.

Our work fills these gaps by employing a structured, multi-stage methodology across journals, conferences, and databases including arXiv, IEEE Xplore, Scopus, and ACM DBLP, resulting in a curated corpus of 142 papers. Unlike prior surveys, we classify contributions along a taxonomic framework that incorporates representation type, tool-augmented workflows, and human-in-the-loop methods. We highlight approaches for verification, uncertainty quantification, provenance tracking, and ethical deployment, topics largely missing from previous reviews. By synthesizing findings across modalities, temporal scales, and application domains such as urban planning, disaster response, navigation, environmental monitoring, and social simulation, this survey provides a comprehensive roadmap for future research in grounded, retrievable, tool-augmented, and human-centered geo-spatial AI.

B Survey Methodology

This survey is a **structured scoping review** (Munn et al., 2018), not a PRISMA-compliant systematic review (Moher et al., 2009). We do not claim to be exhaustive across all databases or to have applied formal inter-rater reliability statistics. Instead, we follow a transparent, multi-stage curation protocol designed to maximize coverage of methodologically novel contributions in geo-spatial and geo-temporal AI, while remaining reproducible in its documented decisions. The choice of a scoping-review design reflects the breadth and pace of the target literature: PRISMA's emphasis on exhaustive retrieval and risk-of-bias appraisal is optimized for clinical evidence synthesis, whereas our goal is to map a rapidly evolving landscape across AI, computer vision, NLP, remote sensing, and GIS. This design aligns with established practice in AI/NLP survey papers (Bender et al., 2021; Bommasani et al., 2022), where structured narrative or scoping approaches are standard. The objective is to construct a representative, thematically coherent corpus, classify contributions, and synthesize findings along methodological axes.

B.1 Stage 1: Data Collection

Candidate papers were collected from preprint servers, top-tier journals, leading conferences, and major academic databases in AI, computer vision, robotics, remote sensing, and GIS. Table 3 lists the primary venues and databases considered.

Searches were conducted using combinations of keywords including: `geo-spatial AI`, `geo-temporal modeling`, `spatial reasoning`, `place graphs`, `4D geo-spatial foundation models`, `raster-vector fusion`, `spatiotemporal predictive modeling`, `map reasoning`, `GIS tool integration`, `uncertainty calibration`, `human-in-the-loop mapping`, `crowdsourced mapping`, `dual-use safety`, `spatial knowledge graphs`, `multimodal geo-spatial representation`, `tool-augmented AI`, `spatiotemporal retrieval`.

The initial query phase across arXiv, IEEE

Table 2: Comparison of this survey with closely related prior surveys. **Cov.** = coverage scope; **Tax.** = structured taxonomy; **Bench.** = benchmark/evaluation analysis; **Tool** = tool-augmented/agentic methods; **HITL** = human-in-the-loop; **Gov.** = governance/privacy/bias; **Method.** = explicit curation methodology.

Survey	Year	Cov.	Tax.	Bench.	Tool	HITL	Gov.	Method.
Zhu et al. (Zhu et al., 2017)	2017	RS imagery	Partial	✓	–	–	–	Narrative
Kamilaris & Prenafeta-Boldú (Kamilaris and Prenafeta-Boldú, 2018)	2018	RS/Agri	Partial	✓	–	–	–	Narrative
Ma et al. (Ma et al., 2019)	2019	RS/CV	Partial	✓	–	–	–	Narrative
Wang et al. (Wang et al., 2020)	2020	RS+GIS	Partial	✓	–	–	–	Narrative
Mai et al. (Mai et al., 2024)	2023	GeoAI	Partial	–	Partial	–	Partial	Narrative
Jakubik et al. (Jakubik et al., 2023)	2023	EO/FM	Partial	✓	–	–	–	Narrative
Feng et al. (Feng et al., 2025)	2025	Spatiotemporal	Partial	✓	Partial	–	–	Narrative
Ma et al. (Ma et al., 2025)	2025	VLA/Embodied	Partial	✓	✓	–	–	Narrative
This survey	**2026**	**LLM/VLM geo**	✓	✓	✓	✓	✓	**Multi-stage (8 stages, n=142)**

Table 3: Primary journals, conferences, and databases surveyed.

Journals / Databases	Conferences / Repositories
IEEE Xplore	CVPR, ICCV, ECCV
Scopus	NeurIPS, ICML, ICLR
ACM Digital Library / DBLP	ACL, EMNLP, NAACL
IEEE TGRS, ISPRS Journal of Photogrammetry and Remote Sensing	GIScience, ACM SIGSPATIAL, COSIT
IJGIS, Remote Sensing of Environment, JOSIS	RSS, CoRL, AAMAS
Computers, Environment and Urban Systems	AAAI, IJCAI

Xplore, Scopus, ACM DBLP, and the listed journals/conferences yielded approximately 1,872 candidate papers.

B.2 Stage 2: Deduplication and Metadata Cleaning

After collecting candidate papers from multiple sources, including arXiv, IEEE Xplore, Scopus, ACM DBLP, and leading AI and geo-spatial conferences, all records were merged and standardized to ensure uniformity in metadata. Papers were excluded if they were duplicates, had multiple preprint versions, were withdrawn, or were otherwise inaccessible. Non-English papers and workshop abstracts lacking full technical content were also removed to maintain analytical rigor. This cleaning process ensured that each entry represented a distinct, citable contribution with complete information. Metadata fields such as title, authors, venue, year, DOI, and keywords were verified for consistency and completeness. After this stage, a total of 1,305 papers remained, forming the basis for the subsequent relevance screening. This step provided a reproducible foundation and minimized noise in the final corpus.

B.3 Stage 3: Title Screening

Titles were systematically screened to assess explicit relevance to geo-spatial representation learning, spatiotemporal reasoning, GIS integration, predictive modeling, or tool-augmented geo-spatial workflows. Works that appeared clearly unrelated to these methodological goals, including those focused solely on downstream applications without novel modeling contributions, were excluded. This stage emphasized precision over recall to reduce the workload for later, more intensive reviews. By carefully considering semantic cues in titles, we filtered out research that did not align with the survey's core objectives. Each remaining title was cross-checked against keywords in the search query to ensure coverage of emerging topics. After this initial relevance assessment, 612 papers were retained for abstract-level screening. This process balanced breadth and specificity while preserving potentially important methodological contributions.

B.4 Stage 4: Abstract Screening

Abstracts were examined to assess the presence of either (a) a methodological contribution—explicit spatial or temporal modeling, representation learning, multimodal data integration, or verification pipelines—or (b) a substantive application contribution that provides evidence about deployment, failure modes, governance, or human-in-the-loop use. Studies lacking either type of contribution were excluded. This dual inclusion criterion was adopted because the survey covers not only model design but also the practical consequences of deploying

geo-spatial AI, including operational reliability, geographic bias, and dual-use risk—topics for which evidence arises primarily from application-oriented studies. Special attention was given to studies that combined multiple data modalities, incorporated temporal dynamics, or proposed human-in-the-loop approaches. Ambiguous cases were flagged for full-text assessment to avoid prematurely discarding potentially relevant contributions. After this stage, 314 papers remained, forming a more focused set for in-depth evaluation.

B.5 Stage 5: Full-Text Eligibility Assessment

Full-text review was conducted to evaluate whether each paper met inclusion criteria requiring explicit modeling of geo-spatial state, dynamics, or structured reasoning, beyond simple task performance. Studies were retained if they integrated raster, vector, or temporal data in a computational model and considered elements such as uncertainty, provenance, or human-in-the-loop workflows. Papers lacking formal methodological contributions, or that did not address multimodal geo-spatial reasoning, were excluded. This stage also assessed technical clarity, reproducibility, and whether sufficient methodological detail was provided to support coding and classification. Each work was examined for representational fidelity, temporal reasoning capabilities, and alignment with geo-spatial modeling principles. By the end of this stage, 172 papers were selected for further relevance verification.

B.6 Stage 6: Geo-spatial AI Relevance Verification

To ensure alignment with geo-spatial AI and geotemporal modeling, we applied an additional relevance check focused on methodological scope. Papers were retained only if they explicitly involved multimodal fusion of imagery, maps, graphs, or sensor data, or leveraged foundation models for geospatial tasks and spatiotemporal reasoning. Tool-augmented workflows, including GIS, remote sensing, and knowledge-graph integration, were also a criterion for inclusion. This stage was essential to ensure conceptual coherence with the survey's focus on next-generation geo-spatial AI. Studies that did not leverage foundation models, tool-assisted pipelines, or spatiotemporal reasoning were excluded to avoid diluting the analysis.

After applying these strict criteria, 142 papers remained for the final corpus. This final set represents a carefully curated and methodologically coherent body of work for subsequent coding and analysis. Our year distribution shows the survey's emphasis on recent work: 34 references from 2025, 23 from 2024, and 22 from 2023, so roughly 56% of all citations are from the 2023–2025 window, consistent with the paper's stated focus on 2021–2025.

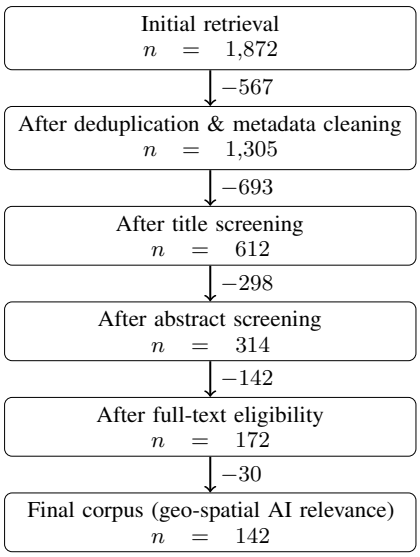


Figure 3: Paper selection flow. Numbers indicate papers removed at each stage.

B.7 Stage 7: Theoretical Coding

Each paper was systematically coded according to its primary methodological contribution, rather than the application domain. Contributions were classified as raster-based predictive models, vector and graph-based models, hybrid 4D geo-spatial models integrating temporal and multimodal data, tool-augmented reasoning pipelines, or human-in-the-loop participatory mapping frameworks. Classification considered the main representational strategy, workflow architecture, and computational innovations. Ambiguous or multi-strategy papers were coded according to the dominant contribution to preserve analytical clarity. This coding framework enabled structured comparison across diverse geo-spatial AI approaches and highlighted emerging patterns and trade-offs. The goal was to create a taxonomy that supports both conceptual synthesis and quantitative summary of research trends.

B.8 Stage 8: Application Mapping

In addition to theoretical coding, each paper was mapped to one or more application domains, including urban planning, smart cities, disaster response, humanitarian mapping, autonomous navigation, robotics, environmental monitoring, climate

modeling, and social or agentic simulation. Papers spanning multiple domains were assigned all relevant contexts to maintain analytical completeness and avoid underrepresenting cross-domain contributions. This mapping allowed the survey to assess not only methodological trends but also the practical areas where geo-spatial AI is being applied. By linking representation strategies to domain-specific challenges, we were able to identify both coverage gaps and opportunities for tool-augmented, human-in-the-loop approaches.

B.9 Curatorial Principles

Papers were selected based on **methodological contribution** rather than benchmark performance. Priority was given to works introducing new representations, transition operators, uncertainty modeling, or evaluation protocols. Application-oriented studies providing evidence on deployment reliability, operational failure modes, geographic bias, governance, or human-in-the-loop workflows were also retained, consistent with the survey's coverage of real-world use cases (§4). Disagreements in categorization were resolved via consensus review.

B.10 Rationale for Including Preprints

The corpus includes papers from arXiv alongside peer-reviewed publications indexed in IEEE Xplore, Scopus, and ACM DBLP. We include preprints for two reasons. First, the target domain (LLM/VLM applications to geo-spatial reasoning) is fast-moving: many of the most-cited architectures, benchmarks, and evaluation results appear first as arXiv preprints and are subsequently published or widely adopted before formal journal indexing. Excluding preprints entirely would introduce a systematic recency bias against the period most relevant to this survey (2023–2025). Second, we apply quality filtering regardless of venue: inclusion required an explicit methodological contribution (new representation, reasoning paradigm, benchmark, or evaluation protocol), not merely reported results (Stage 5, §B). Of the 142 final papers, approximately 52 appear as preprints or arXiv-only entries; the remainder are published in peer-reviewed venues. Preprint-only papers are cited as such in the reference list and readers should treat their specific empirical claims as provisional.

B.11 Synthesis Strategy

Analysis proceeded along dual axes: representation methodology and application domain. This enabled comparison across architectural families while evaluating domain-specific constraints independently. Quantitative trends were summarized descriptively, and qualitative insights were synthesized to highlight common design patterns, trade-offs, and open challenges in next-generation geo-spatial AI.

Table 4: Summary of paper selection process.

Stage	Papers Remaining
Initial retrieval	1,872
Deduplication and metadata cleaning	1,305
Title screening	612
Abstract screening	314
Full-text eligibility	172
Geo-spatial AI relevance verification	142

C Key Methods and Models

Recent progress has not come from a single dominant architecture. Instead, the field has advanced through four recurrent design moves: stronger retrieval/classification backbones for candidate generation, multi-stage pipelines that expose intermediate clues, training schemes that optimize decomposed reasoning rather than single-shot prediction, and pretraining strategies that inject geographic structure directly into model representations. We organize the methods literature around these design patterns and then use representative systems to illustrate the main trade-offs.

C.1 Global Image Retrieval & Classification Approaches

Global image geolocation still rests on two basic backbones: *classification* over discretized geographic regions and *retrieval* against large geotagged corpora. **PlaNet** is the canonical classification formulation, casting localization as prediction over Earth cells and using the penultimate representation as a location-aware embedding (Weyand et al., 2016). Later work refined this recipe with hierarchical prediction and domain-specific classifiers, but the core assumption remained the same: geography is approximated by a learned partition.

Retrieval-based methods instead defer precision to the reference set. The success of CLIP-style representation learning made this especially attractive because images, text, and location descriptors can be aligned in a shared embedding space (Radford et al., 2021). **GeoCLIP** makes this logic explicit by aligning image features with learnable GPS representations, effectively turning localization into

Table 5: Geo-spatial LLM/VLM Applications and Representative References

Application Area	Description and References
Earth Observation & Environmental Monitoring	Instruction-following over imagery and spatial layers (e.g., *EarthGPT*, *GeoPix*) for analyst-facing monitoring and disaster triage without replacing pipelines (Zhang et al., 2024a; Ou et al., 2025; Rahnemoonfar et al., 2021; Gupta et al., 2019). Text-focused systems (*ClimateBERT*, *ClimateGPT*) support document synthesis (Webersinke et al., 2021; Thulke et al., 2024). Multimodal backbones (*PDFM*) enable population-environment representations (Agarwal et al., 2025). EO LLMs reorganize workflows around retrieval, triage, and evidence-linked summaries (Schottlander and Shekel, 2025b).
Crisis Response & Humanitarian Aid	AI accelerates triage and communication. Damage assessment LLMs synthesize pre/post imagery over *xBD* and *xView2* (Gupta et al., 2019; Singh et al., 2024). SAR applications structure narrative reports (Ji et al., 2025; Google Research, 2025; Abi-Zeid et al., 2019). Humanitarian mapping speeds candidate generation (e.g., *Missing Maps*, OSM) while validation remains human-driven (Missing Maps; UN Maps; OpenStreetMap Wiki contributors; Juhász et al., 2023). Resource allocation models rank options under assumptions (Ji et al., 2025; Schottlander and Shekel, 2025b; Pizzi et al., 2020; Gella et al., 2023).
Navigation & Transportation	LLMs translate user goals into structured routing or POI queries and explain trade-offs (Ji et al., 2025; Schottlander and Shekel, 2025b; Google Research, 2025; Daniel, 2024). Contextual signals augment traffic forecasting (Zhao et al., 2025), but classical solvers remain essential. Language-guided planning in robotics/drones requires grounding in perception and constrained action models (Driess et al., 2023; Ma et al., 2025).
Smart Cities & Urban Planning	Conversational GIS enables natural-language queries to geo-spatial retrieval, analysis, and summaries (Google Research, 2025; Li and Ning, 2023; Akinboyewa et al., 2025; Ning et al., 2025). Generative visualizations support "what-if" scenario communication (Hua et al., 2025). Outputs accelerate analysis but are assistive, not validated policy recommendations (Li and Ning, 2023; Zhang et al., 2025b).
Social Media Geolocation & OSINT	LLMs/VLMs propose candidate locations from images/videos, highlight supporting cues, and reduce verification effort while humans retain responsibility (Postma and Patin, 2025). Capabilities accelerate OSINT workflows but raise privacy risks (Grainge et al., 2025).

Table 6: Key Geo-spatial Datasets and Benchmarks by Task Family

Task Family	Representative Datasets / References	Evaluation Focus / Notes
Image Geolocation and Landmark Recognition	YFCC100M (Thomee et al., 2016), Im2GPS (Hays and Efros, 2008), Im2GPS3k (Vo et al., 2017), Google Landmarks v2, TimeSpot (Wasi et al., 2026b), CVUSA (Workman et al., 2015), CVACT (Liu and Li, 2019), VIGOR (Zhu et al., 2021), LLM-Geo (Wang et al., 2024), GeoPose3K (Brejcha and Čadík, 2017), IMAGEO-Bench (Li et al., 2025a), GeoExplain (Huang et al., 2025), GeoGuess (Cheng et al., 2025)	Geotagging accuracy, cross-view robustness, landmark recognition, multimodal reasoning, explanation quality, bias detection
Spatiotemporal and Knowledge-Based QA	SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), POI-QA (Grainge et al., 2025; Han et al., 2025), ChronoQA (Chen et al., 2025), TimeQA (Chen et al., 2021), StepGame (Shi et al., 2022), STKGQA (Dai et al., 2024), tool-augmented evaluation frameworks (Singh et al., 2024)	Structured spatiotemporal reasoning, temporal and spatial multi-hop QA, evidence retrieval, grounding, query execution over knowledge graphs
Navigation and Embodied Benchmarks	R2R (Anderson et al., 2018), RxR (Ku et al., 2020), CVDN (Thomason et al., 2020), ALFRED (Shridhar et al., 2020), Touchdown (Chen et al., 2019), StreetNav (Hermann et al., 2020), EQA (Das et al., 2018), NDH/CVDN (Thomason et al., 2020), Talk2Nav (Vasudevan et al., 2021), OSM-completion tasks (Juhász et al., 2023)	Instruction grounding, sequential planning, trajectory accuracy, memory and reasoning in indoor and outdoor navigation, interactive competence, evidence gathering, mapping updates
Remote Sensing and Earth Science	xView (Lam et al., 2018), DOTA (Xia et al., 2018), RSICD (Lu et al., 2017), UCM Captions (Qu et al., 2016), RSVQA-Bench (Lobry et al., 2021), FloodNet (Rahnemoonfar et al., 2021), xBD (Gupta et al., 2019), PDFM (Agarwal et al., 2025), OBSR (Moska et al., 2025)	Large-scale object detection, image-captioning alignment, QA over overhead imagery, disaster and change detection, spatiotemporal reasoning, representation transfer, multimodal time-series evaluation

nearest-neighbor search in a continuous coordinate-aware latent space (Vivanco Cepeda et al., 2023; Jia et al., 2025). In practice, current systems often combine the two views: a fast backbone produces a coarse candidate set, then a slower reasoner reranks or verifies it (Jia et al., 2025; Postma and Patin, 2025).

The strength of these approaches is efficiency and scale. Their weakness is that they plateau once scenes become rare, ambiguous, or poorly covered by the index. Large reference sets such as **OSV5M** reduce this problem but do not remove it (Astruc et al., 2024; Jia et al., 2025). Retrieval-heavy systems also make it difficult to separate genuine transfer from database dependence, especially in GeoGuessr-style settings where near-duplicate or highly similar imagery can quietly dominate performance (Durrieu et al., 2026; Postma and Patin, 2025). This is precisely why benchmarks such as **GeoArena** matter: they stress fresh, user-provided imagery to better distinguish robust localization from memorization or leakage (Jia et al., 2025). The current frontier is therefore not replacing retrieval/classification backbones, but constraining them with higher-level reasoning when the candidate set remains ambiguous.

C.2 Hierarchical Reasoning & Multi Stage Pipelines

As benchmarks shifted from coordinate prediction toward evidence-backed explanation, monolithic models gave way to *multi-stage pipelines*. The

central idea is simple: decompose the task into subproblems that are easier to validate than the final answer. In geolocation, this usually means separating *local* cues such as signage, plates, objects, or plant species from *global* cues such as landscape, road geometry, architecture, and settlement density (Cheng et al., 2025). Specialized modules extract these clues, convert them into symbolic observations, and pass them to a fusion stage that produces the final hypothesis.

This design improves controllability for two reasons. First, intermediate outputs are inspectable; second, individual modules can be upgraded without retraining the entire system. **SightSense** is a representative example, but the pattern is broader than any one model (Cheng et al., 2025). The downside is equally structural: early-stage errors propagate. OCR mistakes, spurious detections, or overconfident scene summaries can anchor the downstream reasoner on the wrong region. Systems therefore increasingly propagate confidence scores, keep multiple hypotheses alive, or insert verification stages before synthesis (Cheng et al., 2025).

A closely related decomposition appears in geospatial QA. Rather than answering directly, the model first emits a plan—for example as JSON, pseudo-SQL, or a tool-call sequence—then retrieval or computation fills in the required facts, and only then is a final response generated. Structured-output protocols such as those used in **IMAGEO-Bench** make this design easier to evaluate because observations, evidence, and conclusions can be separated explicitly (Li et al., 2025a). More agentic versions couple this with iterative tool use and revision loops, following the broader logic of ReAct-, Toolformer-, and MRKL-style systems, but instantiated over geo-spatial APIs and databases rather than generic tools (Google Research, 2025; Yao et al., 2022; Schick et al., 2023; Karpas et al., 2022). The appeal of these pipelines is not only better accuracy. It is that they expose *where* the system failed, which matters more in geo-spatial reasoning than a polished wrong answer.

C.3 Chain-of-Thought Training and Reinforcement Learning

Chain-of-thought (CoT) is no longer just an inference-time prompt trick; it has become a training target (Wei et al., 2022). In geo-spatial reasoning, this matters because many failures stem from poor decomposition rather than missing raw knowledge. **GeoChain** addresses the scarcity of expert-authored rationales by generating large-scale multimodal reasoning traces and filtering them for correctness, allowing smaller models to be fine-tuned on explicit geographic search procedures rather than only final labels (Yerramilli et al., 2025). This makes the model learn not just *what* answer is correct, but *how* to search for it.

A common recipe is teacher distillation: a stronger model produces candidate reasoning traces, and only those whose final answers match ground truth are kept for supervised fine-tuning (Yerramilli et al., 2025). This improves sample efficiency, but it does not solve the faithfulness problem. Correct final answers can still be paired with brittle or post hoc reasoning, so CoT supervision helps most when intermediate steps are externally checkable.

Reinforcement learning pushes this further by rewarding not only final accuracy but also *verifiable intermediate behavior*. **Geo-R1** is the clearest example: it optimizes geolocation with rewards tied to both final localization and intermediate cross-view consistency checks, encouraging the model to prefer evidence that can be corroborated across modalities (Xu et al., 2025a; Cheng et al., 2025). The same logic extends naturally to tool use, where the agent is rewarded for invoking computation or lookup when the task demands it, rather than freewheeling through plausible nonsense. Overall, the shift is from prompted reasoning to *optimized reasoning policies*. That is a substantive methodological change, not just nicer prompting.

C.4 Spatially-Grounded Pretraining & Representation Learning

A parallel line of work tries to make models spatially competent *before* task-specific fine-tuning. In language, **GeoLM** injects explicit geographic supervision during pretraining so that location is represented as more than incidental text co-occurrence (Li et al., 2023b). **GeoLLM** explores a related but more pragmatic direction: prompt an LLM with coordinate-indexed descriptors and use it as a predictor or feature extractor for downstream socioeconomic tasks (Manvi et al., 2024b). The results are striking, but they also reveal the central limitation of this paradigm: implicit geographic knowledge is useful yet regionally uneven, and performance degrades where textual coverage is sparse or systematically biased (Manvi et al., 2024a).

On the vision side, contrastive learning remains the main workhorse. **GeoCLIP** aligns images

with location representations for geolocalization (Vivanco Cepeda et al., 2023), **UrbanCLIP** specializes contrastive pretraining for urban imagery and place semantics (Yan et al., 2024), and **SatCLIP** adapts the same principle to multispectral satellite data to produce more transferable location encoders under geographic shift (Klemmer et al., 2023). These models matter because they move geographic structure from the downstream loss into the representation itself.

Coordinate encoding is the complementary problem. Latitude and longitude are continuous, periodic, and geometrically awkward, so the way they are embedded has nontrivial effects on generalization. **TorchSpatial** makes this explicit by benchmarking continuous and discrete encoders and showing trade-offs between wrap-around behavior, boundary artifacts, and transfer across tasks (Wu et al., 2024). The broader conclusion is sobering: spatial pretraining helps, but no single representation yet provides globally reliable geographic priors across scales, regions, and modalities (Li et al., 2025a). Representation learning has improved the starting point; it has not removed the need for retrieval, tools, or verification.

C.5 Notable Recent Models and Systems

The models below are best read as *design exemplars* rather than isolated winners. Each is influential because it operationalizes a different answer to the same question: where should geo-spatial competence live—in a general multimodal foundation model, in a domain-specific backbone, or in a composite system?

■ **GPT-4V (2023) and Gemini (2023–).** Closed multimodal models have functioned as strong reference baselines for geo-spatial reasoning. **GPT-4V** extends GPT-4 with visual input and has shown competitive performance on geospatially flavored map- and image-based tasks (OpenAI, 2023; Roberts et al., 2024). **Gemini** plays a similar role on the Google side as a general-purpose multimodal reasoner rather than a geo-spatial specialist (Team et al., 2025). Practitioner-oriented evaluations suggest that rankings among these systems are sensitive to prompting, tool access, and task format (Postma and Patin, 2025). Their importance in the literature is therefore less about architectural transparency than about setting a moving upper bound that open systems try to match.

■ **BLIP-2 (2023). BLIP-2** is influential less as a geo-spatial model than as a *modular interface* between perception and language. Its Q-Former architecture allows a frozen visual encoder and a frozen LLM to be coupled cheaply, making it a convenient front-end for downstream captioning and VQA-style pipelines (Li et al., 2023a). In geospatial workflows, this has made BLIP-2 useful as an "artificial analyst" that turns street-level imagery into textual descriptions, which can then feed downstream spatial reasoning or map-editing systems (Juhász et al., 2023). The key lesson is architectural: modular multimodal front-ends can be more reusable than end-to-end geo-spatial specialists, especially in low-resource application settings.

■ **SightSense (Cheng et al., 2025). SightSense** is one of the clearest demonstrations that deliberate decomposition can outperform monolithic inference on explanation-centric geolocation. It combines open-vocabulary grounding, CLIP-style retrieval, and VLM-based synthesis into a staged pipeline designed around **GeoExplain** and **GeoGuess** (Cheng et al., 2025). Concretely, it uses **GroundingDINO** for clue detection, CLIP-style retrieval for evidence gathering, and **Qwen-VL-7B** for final localization and rationale generation (Liu et al., 2024; Radford et al., 2021; Bai et al., 2023; Cheng et al., 2025). The reported ablations are as important as the headline numbers: removing retrieval hurts substantially, which suggests that even strong multimodal backbones still benefit from explicit evidence access (Cheng et al., 2025). SightSense matters because it demonstrates that targeted system design can beat brute-force generality when the benchmark explicitly rewards grounded explanation.

■ **EarthGPT (2024). EarthGPT** represents the opposite design choice: build a domain-specific multimodal model for Earth observation rather than retrofitting a web-trained VLM (Zhang et al., 2024a). Its contribution is not only architectural, but also data-centric through **MMRS-1M**, which supports instruction tuning across scene classification, captioning, VQA, grounding, and detection in remote sensing (Zhang et al., 2024a). The broader significance is that EarthGPT treats geo-spatial perception as a first-class pretraining problem rather than as a narrow downstream adaptation. This line sits alongside related Earth-scale backbones such as EarthPT and broader geo-spatial reasoning systems, and it helps explain why remote-sensing performance increasingly depends on domain-aligned pretraining rather than generic multimodal scale alone (Smith et al., 2024; Agarwal et al., 2025;

Schottlander and Shekel, 2025b).

■ **GeoLLM. GeoLLM** is better understood as a methodology than as a new backbone. It probes and exploits latent geographic knowledge in general-purpose LLMs by conditioning them on structured local descriptors and predicting downstream socioeconomic targets (Manvi et al., 2024b). What makes the result interesting is the tension it exposes: models such as GPT-3.5 and Llama 2 can be surprisingly competitive with limited supervision, yet that success depends heavily on prompt-accessible local context and on the geographic regularities already encoded in the base model (Manvi et al., 2024b; Touvron et al., 2023). GeoLLM therefore shows both the promise and the danger of "mental maps": these models do contain useful spatial priors, but those priors are neither uniformly distributed nor inherently trustworthy (Manvi et al., 2024a).

■ **GraphGPT and GraphInstruct (2024). GraphGPT** and **GraphInstruct** matter because many geo-spatial problems are, underneath the prose, graph problems: routing, connectivity, constrained traversal, and spatiotemporal KG queries (Tang et al., 2024; Luo et al., 2024). These models instruction-tune LLMs to interpret graph-structured inputs and produce human-readable rationales over them. **GraphTeam** extends this line toward collaborative graph analysis (Li et al., 2025b). Their value for geo-spatial AI is not that they replace exact solvers, but that they make graph computation accessible through conversational interfaces while preserving more structure than free-form language alone.

■ **PaLM-E (2023). PaLM-E** is central to the embodied branch of this literature because it treats sensory observations as tokens that condition action generation (Driess et al., 2023). Although introduced in robotics rather than outdoor GIS settings, the design is directly relevant to geo-spatial agents: it provides a template for models that must ground language in spatial state and then act. That connection becomes clearer when read against navigation benchmarks such as **R2R** and **Touchdown**, and against later surveys of vision-language-action models (Anderson et al., 2018; Chen et al., 2019; Ma et al., 2025). Follow-on work on map-like memories further underscores the same point: for long-horizon navigation, explicit spatial state representations can outperform raw history replay (Zhang et al., 2025a).

■ **Environmental and Social Science Models.** A final trend is the rise of *domain-aligned* geo-spatial foundation models. **ClimateGPT** targets climate-oriented synthesis and QA (Thulke et al., 2024); **PDFM** provides a reusable backbone for population and human–environment inference (Agarwal et al., 2025); and generative systems such as **Sat2City** extend the space toward planning-oriented scenario generation from satellite imagery (Hua et al., 2025). These systems are narrower than general multimodal models, but that narrowness is often an advantage: it allows training data, objectives, and outputs to be aligned with real domain workflows.

Taken together, recent systems increasingly combine perception-centric geo-spatial backbones such as **EarthGPT**, **GeoPix**, and related foundation models (Zhang et al., 2024a; Ou et al., 2025; Schottlander and Shekel, 2025a), language-first extraction and reasoning frameworks such as **GeoLLM** and **GraphGPT** (Manvi et al., 2024b; Tang et al., 2024), and specialized urban or spatial pretraining such as **UrbanCLIP** (Yan et al., 2024). The field has therefore moved from isolated task models toward composite, reusable systems with stronger priors and more explicit reasoning interfaces (Jakubik et al., 2023; Mendieta et al., 2023). The unresolved question is no longer whether larger models help. It is where scaling ends and explicit geographic structure has to begin.

D Extended Discussion on Open Challenges

Despite rapid progress, current geo-spatial and geo-temporal AI remains constrained by a small set of recurring failures: temporally stale world models, weak cross-modal grounding, uneven geographic coverage, evaluation protocols that overstate real-world readiness, high computational cost, and limited interoperability with trusted GIS workflows. These are not peripheral defects. They determine whether a model can answer with the correct world state, align language with geometry, calibrate uncertainty, and remain auditable in deployment. We organize the remaining challenges around these pressure points and then outline research directions that appear most consequential.

D.1 Temporal Generalization and World Change

Temporal validity remains a foundational weakness. Most LLM/VLM systems internalize a largely

static world, yet geo-spatial facts change continuously: businesses open and close, transit lines are added, administrative boundaries shift, and the same coordinates may denote different entities at different times. Recent benchmarks make this failure mode explicit. **POI-QA** requires models to satisfy temporal as well as spatial constraints over points of interest (Han et al., 2025), and text-based temporal QA benchmarks show that models often ignore temporal scope unless the task and evidence are explicitly time-conditioned (Chen et al., 2021, 2025). The same lesson appears in spatiotemporal KG settings, where correct answers depend on querying the right world state rather than recalling a timeless fact (Dai et al., 2024).

The methodological implication is straightforward: temporal reasoning should be treated as a first-class modeling and evaluation problem rather than an edge case. Promising directions include timestamp-aware pretraining and fine-tuning, retrieval over time-indexed corpora or versioned spatiotemporal knowledge bases, and time-series pretraining for periodic or seasonal dynamics where repeated observation is available (Dai et al., 2024; Smith et al., 2024). Evaluation should likewise move beyond static test sets toward time-sliced splits, recency-stratified reporting, and explicit separation between historical reasoning, current-state reasoning, and forecasting.

D.2 Multimodal Grounding and Alignment

A second bottleneck is multimodal misalignment. Current systems can produce fluent spatial narratives that are only weakly tied to the actual image, map, or database evidence. In geo-spatial settings, this is unusually costly: a small grounding error can flip direction, cross a boundary, or attach a claim to the wrong intersection or parcel (OpenAI, 2023). The problem is best understood not as one monolithic hallucination issue, but as several distinct alignment failures.

■ **Geometric alignment.** Geo-spatial language must preserve geometry and topology, yet recent evaluations show that LLM-style systems remain unreliable on explicit spatial predicates such as containment, adjacency, and direction (Ji et al., 2025). A practical mitigation is to separate generation from geometric verification: let the model propose claims, but check them against coordinates, polygons, or network structure before surfacing them to the user.

■ **Cross-view consistency.** Cross-view geolocalization exposes another persistent failure mode: a model may extract plausible street-level cues yet fail to reconcile them with overhead context. Recent results from **Geo-R1** suggest that reinforcement learning with verifiable cross-view consistency signals can improve this alignment (Xu et al., 2025a). Classical cross-view benchmarks remain useful here because they isolate the view-translation problem more cleanly than broader multimodal QA settings (Liu and Li, 2019; Zhu et al., 2021).

■ **Modality arbitration.** Many geo-spatial tasks are multimodal not because more modalities are always better, but because different subtasks belong to different evidence sources. Images support visual attributes, maps support geometry and topology, and external databases support demographics or schedules. Current models often arbitrate poorly, for example inferring quantitative or administrative facts from imagery when a structured source should have been consulted. Tool-using multimodal systems point toward a better pattern: route subtasks to the modality or tool that can answer them verifiably rather than forcing every subproblem through free-form generation (Driess et al., 2023).

■ **Fine-grained spatial grounding.** Coarse image captions are rarely enough for reliable geo-spatial explanation. What is often needed is token-to-region or phrase-to-mask grounding, so that textual claims can be attached to specific image regions or objects. **GeoPix** is one example of this direction in remote sensing (Ou et al., 2025). More general multimodal grounding work, including **GROUNDHOG** and **SAM4MLLM**, points to the same conclusion: pixel-level alignment can reduce hallucination and produce more verifiable explanations than box-level or caption-level grounding alone (Zhang et al., 2024b; Chen et al., 2024).

■ **Evaluation of alignment.** Alignment should be evaluated directly rather than inferred from answer quality. One route is automated evidence checking using grounding and open-vocabulary detection models that verify whether mentioned objects or attributes are actually present (Liu et al., 2024). More generally, deployment studies of multimodal systems continue to show that surface fluency and visual faithfulness can diverge sharply (OpenAI, 2023). The open problem is therefore not just better multimodal generation, but better multimodal auditing.

D.3 Data Bias and Coverage

Geo-spatial data is not merely imbalanced; it is structurally uneven across regions, languages, time periods, and capture conditions. Models therefore inherit systematic performance disparities and representational skew (Huang et al., 2025). This is a core scientific issue because geo-spatial reasoning is often deployed exactly where coverage is weakest.

■ **Accuracy bias.** Region-stratified evaluations already show that performance can vary sharply by geography. **IMAGEO-Bench** reports substantial gaps between higher-resource regions and lower-resource ones in image geolocalization (Li et al., 2025a), and similar disparities appear in other large-scale evaluations (Jia et al., 2025; Huang et al., 2025). Aggregate accuracy therefore hides geographically concentrated failure.

■ **Content bias.** Training corpora disproportionately feature tourist landmarks, wealthy urban centers, and visually distinctive canonical scenes. This biases models toward recognizable postcard geography while weakening performance on ordinary, rural, or crisis-specific imagery. The consequence is not only lower recall in under-photographed places, but also distorted model priors about what a region "should" look like.

■ **Language and knowledge bias.** Text-centric models typically encode much richer knowledge about places that are well documented in major languages. For less-covered regions, parametric recall becomes sparse and hallucination risk increases. Retrieval from local gazetteers, multilingual corpora, and region-specific knowledge sources is therefore not an optional enhancement but a fairness intervention.

■ **Temporal and capture-condition bias.** Coverage varies over time and acquisition conditions as well. Many datasets are dominated by daytime, clear-weather, and relatively recent imagery, leaving night scenes, winter conditions, cloud cover, and older infrastructure underrepresented. Robustness under these conditions cannot be assumed from headline benchmark performance.

Mitigation has to operate at multiple levels. On the data side, broader street-level resources such as **OpenStreetView-5M** can improve geographic coverage for both training and evaluation (Astruc et al., 2024). On the modeling side, geographic adaptation, multilingual retrieval, and calibrated abstention can reduce confident guessing out of distribution. On the evaluation side, reporting should be stratified by region, language, urbanicity, time, and capture condition rather than collapsed into a single global score.

D.4 Evaluation and Benchmarking Challenges

Evaluation remains one of the least settled parts of the field. Geo-spatial outputs are often partially correct, multi-granular, time-dependent, and supported by heterogeneous evidence. A single metric therefore rarely captures the actual failure mode.

■ **Multiple Correct Answers and Graded Correctness.** Many geo-spatial tasks admit answers at multiple acceptable granularities. In image geolocation, for example, a country-level answer may be defensible even when exact coordinates are wrong. Strong benchmarks therefore combine hierarchical administrative accuracy with continuous distance error and region-stratified reporting (Cheng et al., 2025; Grainge et al., 2025; Li et al., 2025a). Both views are necessary because they reveal different failure modes.

■ **Evaluating Explanations and Reasoning Traces.** Reasoning traces are useful for debugging and trust, but they are difficult to score. Text-overlap metrics do not measure evidential grounding, and human evaluation does not scale cleanly (Cheng et al., 2025). Two more promising directions are emerging: preference-based evaluation over paired responses (Jia et al., 2025), and evidence verification using detectors, grounders, or GIS constraints that check whether claimed observations and relations actually hold (Liu et al., 2024; Ji et al., 2025). The common shift is from surface similarity to factual and geometric consistency.

■ **Dynamic Benchmarks and Contamination.** Static benchmarks are vulnerable to contamination because pretraining corpora may contain overlapping images, near-duplicates, or benchmark-adjacent content. Dynamic, in-the-wild evaluation reduces this risk by continuously introducing fresh queries, but at the cost of reproducibility (Jia et al., 2025). A useful compromise is snapshot-based evaluation, where fresh benchmarks are periodically frozen and released for offline comparison.

■ **Holistic Task Evaluation for Tool-Using Agents.** For tool-using systems, final-answer accuracy is too narrow. Evaluation should also measure tool selection, intermediate artifact validity, efficiency, and recovery from earlier mistakes. This is already familiar in embodied navigation, where

success and path-quality metrics matter alongside task completion (Anderson et al., 2018; Chen et al., 2019; Shridhar et al., 2020). In remote-sensing and GIS-agent settings, the same logic extends to whether the system queried the correct layer, produced valid intermediate outputs, and propagated errors across steps (Singh et al., 2024).

■ **Human Preferences and Helpfulness.** Assistant-style systems must also be helpful, not merely correct. **GeoArena** captures this through pairwise preference judgments aggregated into Elo-style rankings (Jia et al., 2025). This is useful, but only if paired with objective correctness checks; otherwise, fluent overconfident answers can win on style.

■ **Edge Cases, Ambiguity, and Robustness.** Strong evaluations should stress ambiguity rather than quietly filtering it away. Border regions, visually similar places, misleading cues, and low-signal scenes should be first-class slices, and systems should be rewarded for calibrated uncertainty rather than punished for not pretending certainty. Robustness reporting across region, weather, day/night, and rural/urban splits is therefore essential (Li et al., 2025a).

■ **Reproducibility and Open Data.** Reproducibility is also constrained by access. Many geo-spatial tasks depend on proprietary imagery or platform-specific data, which weakens comparability across studies. Open resources such as **OpenStreetView-5M** and **IMAGEO-Bench** improve this situation by enabling broader participation and more transparent evaluation (Astruc et al., 2024; Li et al., 2025a). More generally, benchmark design should be treated as a research problem in its own right, not as bookkeeping.

D.5 Model Size and Efficiency

Accuracy alone is not enough if deployment requires large multimodal backbones with prohibitive latency, memory, or energy cost. This is especially acute in edge and resource-constrained settings such as on-device navigation, robotics, AR, and disaster response. Efficiency should therefore be treated as a first-class evaluation axis rather than an implementation afterthought (Schwartz et al., 2020).

■ **Distillation.** Knowledge distillation remains a practical route to deployment, especially when a compact student can absorb not only final outputs but also intermediate representations or structured traces from a stronger teacher (Hinton et al., 2015). Recent cross-view geolocalization work shows that hierarchical distillation can compress multi-branch models into edge-friendly systems without catastrophic accuracy loss (Sun et al., 2025). For geospatial agents, distilling retrieval decisions and tool traces may be as important as distilling logits.

■ **Optimized Architectures and Spatial Inductive Bias.** Generic backbones are often statistically inefficient for coordinate-conditioned or geometry-heavy functions. Injecting spatial inductive bias can improve sample efficiency and reduce the burden on scale alone. Fourier feature mappings are one well-established example for coordinate-based learning (Tancik et al., 2020). More broadly, architectures that represent coordinates explicitly, separate symbolic geometry from language generation, and offload fast-changing facts to external memory are likely to be both smaller and more reliable.

■ **Split Computing and On-Device Offloading.** A practical systems pattern is split execution: keep lightweight perception or privacy-sensitive processing local, while selectively offloading heavier reasoning. Transformer offloading work already shows that careful scheduling across device accelerators can reduce latency and energy use substantially (Xu et al., 2025b). In geo-spatial settings this aligns naturally with privacy constraints, for example by keeping raw imagery on-device while sending only abstracted features or tool queries to the cloud.

■ **Quantization and Low-Bit Inference.** Quantization is an obvious lever because it directly targets memory and bandwidth. Post-training low-bit methods such as GPTQ and AWQ, together with low-rank/low-bit adaptation strategies such as QLoRA, make it increasingly feasible to deploy large transformer backbones more cheaply (Frantar et al., 2022; Lin et al., 2024; Dettmers et al., 2023). Mobile-oriented multimodal systems further suggest that lightweight LMM deployment on commodity accelerators is becoming practical (Shao et al., 2025). The open question is how robust these compressed systems remain under geo-spatial distribution shift, where small numerical changes can propagate through long reasoning chains.

■ **Energy and Carbon Footprint.** The environmental cost of large-model training and serving is also material. The *Green AI* perspective is especially relevant here because geo-spatial AI is often justified in the language of sustainability and public benefit (Schwartz et al., 2020). Efficient reuse of pretrained backbones, parameter-efficient tuning,

and modular system design are therefore not merely engineering conveniences; they affect accessibility and who can actually deploy the models.

■ **Real-Time Constraints and Systems Optimizations.** Many geo-spatial systems are interactive or safety-critical, so mean latency is insufficient. Tail latency, predictable response time, and graceful failure matter as much as throughput. Systems optimizations such as FlashAttention show how much can be gained without changing the high-level model design (Dao et al., 2022). Future geo-spatial evaluations should report accuracy–latency trade-offs explicitly.

■ **Modular and Specialist Systems.** A final efficiency direction is modularity. Replacing one monolithic model with smaller specialists for perception, geocoding, geometry, retrieval, and language can reduce parameter count, simplify updates, and strengthen verification. Existing geo-spatial pipelines already move in this direction (Cheng et al., 2025). The remaining challenge is standardizing interfaces and evaluation so modular systems remain comparable to end-to-end baselines.

D.6 Interpretability and Trustworthiness

In high-stakes geo-spatial settings, trust cannot be reduced to point accuracy. Users need to know why a system produced a location, recommendation, or narrative, and whether that explanation is supported by checkable evidence. This matters because geo-spatial errors are often small in form but large in consequence.

■ **Faithful rationales and evidence attribution.** Chain-of-thought prompting can improve task performance (Wei et al., 2022), but generated rationales are not guaranteed to be faithful accounts of the model's actual decision process (Jacovi and Goldberg, 2020; Huang et al., 2023; OpenAI, 2023). A more defensible design is evidence attribution: decompose the answer into atomic claims and attach OCR snippets, retrieved facts, map entities, or geometric predicates to each one (Min et al., 2023). Retrieval-and-revision pipelines such as RARR are useful precisely because they revise unsupported statements into attributable ones (Gao et al., 2023). In spatial settings, the same logic extends naturally to geometric verification (Ji et al., 2025).

■ **Verification loops and self-checking.** Explicit verification loops are a promising way to reduce geo-spatial hallucination: draft an answer, propose verification queries, check them independently, then revise. **CoVe** provides one generic template for this pattern (Dhuliawala et al., 2024), while sampling-based approaches such as **SelfCheck-GPT** can detect unstable or weakly supported claims even in black-box settings (Manakul et al., 2023). These mechanisms are especially useful when answers mix imagery, maps, text, and tools.

■ **Uncertainty, calibration, and abstention.** Ambiguity is intrinsic to many geo-spatial tasks, yet neural confidence is often miscalibrated (Guo et al., 2017). Approximate uncertainty estimation is possible through Monte Carlo dropout, deep ensembles, and related methods (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017), and LLMs can sometimes provide useful self-estimates of correctness (Kadavath et al., 2022). But the more important design choice is often selective prediction: systems should abstain or widen the hypothesis set when evidence is weak (Geifman and El-Yaniv, 2019). For geo-spatial interfaces, this often means heatmaps, ranked candidate regions, or time-scoped validity rather than a single overconfident pin.

■ **Ambiguity and incomplete information.** Trustworthy systems should recognize when the input is genuinely underdetermined. Featureless roads, generic forests, and occluded satellite tiles should trigger abstention or targeted requests for additional views, landmarks, or timestamps rather than forced guesses. Evaluation should reward that behavior.

■ **Robustness to manipulation and adversaries.** Geo-spatial systems are also exposed to adversarial conditions: misleading signs, doctored imagery, manipulated map layers, and coordinated misinformation. Vision models are already known to be vulnerable to adversarial perturbation (Goodfellow et al., 2014; Brown et al., 2017). Geo-spatial pipelines therefore need provenance tracking, adversarially informed stress tests, and conservative behavior under suspected manipulation.

■ **Auditability, bias, and governance.** Interpretability is inseparable from governance because bias cannot be diagnosed if the evidence trail is opaque. Documentation frameworks such as *Model Cards* and *Datasheets* provide a starting point for reporting intended use, subgroup performance, and known failure modes (Mitchell et al., 2019; Gebru et al., 2021). This aligns with emerging regulatory pressure for transparency and human oversight in higher-risk systems (European Union, 2024).

■ **Human–AI collaboration.** Human oversight is

most effective when experts can intervene directly: correct an extracted cue, constrain a search region, request counter-hypotheses, or inspect provenance. Preference-based evaluation in systems such as **GeoArena** already hints at this broader interaction model (Jia et al., 2025). Mapping workflows provide another concrete case, where LLMs are useful as assistants only when their suggestions remain reviewable and attributable (Juhász et al., 2023).

D.7 Integration with GIS and Domain Expertise

A central deployment barrier is that geo-spatial work already has mature execution environments: GIS software, spatial databases, and domain-specific simulators. Experts usually trust deterministic geoprocessing operators more than fluent end-to-end generation. The most plausible integration pattern is therefore hybrid: use LLMs and VLMs for intent understanding, workflow synthesis, and interface translation, but delegate execution and verification to GIS engines and authoritative data.

■ **Natural language to geoprocessing workflows.** Recent systems increasingly operationalize this design. **GeoGPT** frames GIS automation as an agentic planning loop over spatial tools (Zhang et al., 2024c). The broader *Autonomous GIS* agenda makes the same move more explicitly, treating the LLM as a workflow generator and self-checker (Li and Ning, 2023). **GIS Copilot** and related QGIS-integrated assistants show that this is no longer purely conceptual: natural-language requests can be turned into executable spatial workflows and code, albeit with nontrivial failure rates that still require expert review (Akinboyewa et al., 2025; Ning et al., 2025).

■ **Tool-using LLM architectures for GIS.** The technical requirement underneath all of this is reliable tool use. ReAct, Toolformer, and MRKL-style architectures provide the general template: interleave reasoning with external API calls and keep symbolic computation separate from free-form language when exactness matters (Yao et al., 2022; Schick et al., 2023; Karpas et al., 2022). For GIS, this means evaluation should score tool choice, parameter validity, CRS and unit handling, and reproducibility of the generated workflow.

■ **Data access, provenance, and knowledge graphs.** Practical systems also need first-class access to live repositories rather than static training memories. Spatial databases, catalogs, and knowledge graphs such as Wikidata and SPARQL-accessible endpoints are obvious building blocks here (Vrandečić and Krötzsch, 2014; W3C SPARQL Working Group, 2013). The key requirement is provenance: the system should expose dataset version, query, spatial extent, and transformation history rather than silently surfacing stale or contextless facts.

■ **Domain constraints and ontology-guided consistency.** Many domain decisions depend on hard constraints: zoning rules, geologic plausibility, network connectivity, topological validity. These can be encoded as checkable predicates over intermediate outputs, with symbolic revision or rejection when a claim violates the ontology or rule base (Monti et al., 2024). This kind of constraint-guided consistency is one of the clearest places where domain expertise should shape model behavior directly.

■ **Interoperability and benchmarking.** Interoperability matters because public-health, transport, environmental, and planning agencies rarely operate on a single unified stack. Multi-model orchestration is therefore more realistic than a universal assistant, and recent geo-spatial-reasoning initiatives increasingly adopt that framing (Google Research, 2025). Benchmarks such as **GeoAnalystBench** are valuable because they evaluate LLM-generated geoprocessing workflows against validated task suites rather than against vague demos (Zhang et al., 2025b).

■ **From prototypes to practice.** Industry platforms are beginning to expose similar affordances. Esri, for example, now describes AI assistants across ArcGIS products for code generation, question answering over spatial data, and map-linked interaction (Legacion et al., 2025). The unresolved issue is no longer whether such systems can demo, but whether they can remain deterministic, provenance-first, and overrideable when used in operational settings.

E Extended Future Directions

■ **Unified 4D Geo-spatial Foundation Models.** Next-generation models should jointly represent **raster imagery, vector geometries, place graphs, text, and irregular time series**, capturing both spatial and temporal structure. Explicit **multi-resolution hierarchies** allow fine-grained reasoning at street, city, and regional scales, while **world-**

state versioning tracks temporal changes such as POI updates, infrastructure development, and seasonal dynamics. **Calibrated uncertainty** should produce probabilistic heatmaps, spatiotemporal risk surfaces, or scenario-based forecasts, enabling robust decision-making under incomplete or conflicting evidence (Dai et al., 2024). Integration of **coordinate-conditioned attention** or Fourier features can improve geometric fidelity, and **hierarchical graph representations** enable scalable reasoning over large spatial networks.

■ **Interactive and Verifiable Geo-spatial Agents.** Agents should **translate natural language instructions into multi-step workflows** across GIS tools, remote-sensing analysis, and knowledge graphs. They must return **auditable intermediate artifacts**—buffers, joins, route constraints, and uncertainty maps—that can be independently verified. **Closed-loop verification** involves querying authoritative sources, running simulations, or applying topological/geometric consistency checks before returning outputs. Tool-augmented architectures (e.g., **ReAct, Toolformer, MRKL**) enable precise reasoning grounded in current data (Yao et al., 2022; Schick et al., 2023; Karpas et al., 2022). Retrieval-augmented generation from live geo-spatial databases, open-source satellite imagery, and temporally indexed corpora ensures evidence-based outputs. Multi-agent or collaborative pipelines could decompose large tasks, verifying intermediate steps across agents.

■ **Human-in-the-Loop and Participatory Mapping. Active spatial learning** prioritizes regions with high uncertainty, low coverage, or high expected impact. Crowdsourced contributions from platforms like *Missing Maps* and *UnMappers* can improve calibration, but require careful **provenance tracking** (timestamps, contributor IDs, and evidence sources) to prevent overconfident or spurious predictions (Missing Maps; UN Maps; Juhász et al., 2023). Hybrid human-AI pipelines can implement **iterative verification loops** where human reviewers check AI-generated predictions before final integration. Participatory workflows can also flag domain-specific constraints, such as zoning or environmental regulations, for automated compliance checks.

■ **Governance, Privacy, and Dual-Use Safety.** Geo-spatial AI is inherently dual-use: improved location inference or trajectory modeling can aid disaster response but also surveillance. Systems must embed **red-teaming, bias audits, anonymization, access control, and query logging** into design. Sensitive inferences—e.g., private addresses or vulnerable populations—require **granular release policies**, differential privacy, or encrypted on-device computation. Governance frameworks should track dataset versioning, transformations, and limitations, enabling external auditing. Ethical safeguards include limiting high-resolution outputs in sensitive areas and flagging high-risk predictions automatically.

■ **Modular and Adaptive System Design.** Rather than monolithic geo-spatial AGI, near-term progress favors **grounded, retrievable, tool-augmented, auditable, and human-centered systems**. Modular architectures separate perception, geocoding, reasoning, and tool selection, allowing independent updates and verification. Adaptive pipelines can switch between models or tools depending on region, modality, or temporal relevance. Combined with retrieval over current sources, explicit reasoning traces, and uncertainty-aware outputs, these systems can remain robust under real-world changes.

■ **Integration with Domain Knowledge and Ontologies.** Spatial reasoning should enforce **domain constraints** such as topology, network connectivity, and geologic plausibility. Ontology-guided validation ensures outputs comply with rules and allows symbolic revision of inconsistent claims (Monti et al., 2024). Linking outputs to structured knowledge graphs, metadata, and authoritative databases ensures provenance, verifiability, and traceability of every claim.

■ **Advanced Evaluation and Benchmarking.** Future benchmarks should combine **hierarchical spatial accuracy, temporal correctness, uncertainty calibration, and tool-selection fidelity**. Dynamic, in-the-wild evaluation datasets, along with **stratified metrics** across regions, languages, and capture conditions, will reveal hidden failure modes. Human-AI preference judgments, coupled with evidence-backed scoring, measure both helpfulness and correctness.

Overall, the strongest near-term path combines **multi-resolution 4D representations, modular verification pipelines, human-in-the-loop workflows, tool augmentation, and rigorous governance** to produce **robust, interpretable, ethical, and auditable geo-spatial AI** capable of adapting to a changing world.