# FRIEDA: Benchmarking Multi-Step Cartographic Reasoning in Vision-Language Models

**Anonymous authors**
Paper under double-blind review

## Abstract

Cartographic reasoning is the skill of interpreting geographic relationships by aligning legends, map scales, compass directions, map texts, and geometries across one or more map images. Although essential as a concrete cognitive capability and for critical tasks such as disaster response and urban planning, it remains largely unevaluated. Building on progress in chart and infographic understanding, recent large vision language model (LVLM) studies on map visual question-answering (VQA) often simplify maps as a special case of charts. In contrast, map VQA demands comprehension of layered symbology (e.g., symbols, geometries, and text labels) as well as spatial relations tied to orientation and distance that often span multiple maps and are not captured by chart-style evaluations. To address this gap, we introduce FRIEDA, a benchmark for testing complex open-ended cartographic reasoning in LVLMs. FRIEDA sources real map images from documents and reports in various domains (e.g., geology, urban planning, and environmental assessment) and geographical areas. Following classifications in Geographic Information System (GIS) literature, FRIEDA targets all three categories of spatial relations: *topological* (border, equal, intersect, within), *metric* (distance), and *directional* (orientation). All questions require multi-step inference, and many require cross-map grounding and reasoning. We evaluate eleven state-of-the-art LVLMs under two settings: (1) the *direct* setting, where we provide the maps relevant to the question, and (2) the *contextual* setting, where the model may have to identify the maps relevant to the question before reasoning. Even the strongest models, Gemini-2.5-Pro and GPT-5-Think, achieve only 38.20% and 37.20% accuracy, respectively, far below human performance of 84.87%. These results reveal a persistent gap in multi-step cartographic reasoning, positioning FRIEDA as a rigorous benchmark to drive progress on spatial intelligence in LVLMs.

## 1 Introduction

Recent advances in large vision-language models (LVLMs) have markedly improved multimodal reasoning, with strong results across diverse visual question-answering (VQA) tasks (Dong et al., 2024; Souibgui et al., 2025). Education and cognitive science research characterizes reasoning as a broad capability that spans numeric reasoning, logical deduction (Holyoak & Morrison, 2012), and textual interpretation (Wharton & Kintsch, 1991), as well as interpreting pictures (Mayer, 2020), spatial data (Li et al., 2025), and map images (Goodchild, 2012). Extensive LVLM benchmarks cover many of these facets: visual numeracy in chart and infographics (Lin et al., 2025; Mathew et al., 2022; Masry et al., 2022), document and layout reasoning (Duan et al., 2025; Mathew et al., 2021), multi-image inference (Kazemi et al., 2025; Xia et al., 2025), and even spatial relations in natural images (Shiri et al., 2024). However, reasoning over maps, also a core human competence (Tversky, 2003; Kastens & Ishikawa, 2006; Ishikawa & Newcombe, 2021), which we refer to as **cartographic reasoning**, remains under-examined in LVLMs.

Unlike natural images, maps encode information with an abstract, symbolic visual grammar (e.g., map scales, compass/north arrows, and thematic symbology) (Buckley, 2006), which demands a deeper interpretation than simple pattern recognition. Mastery of these elements must be coupled with the comprehension of spatial relations that are commonly grouped into topological reasoning (e.g., detecting shared boundaries), metric inference (e.g., converting map lengths to real-world dis-

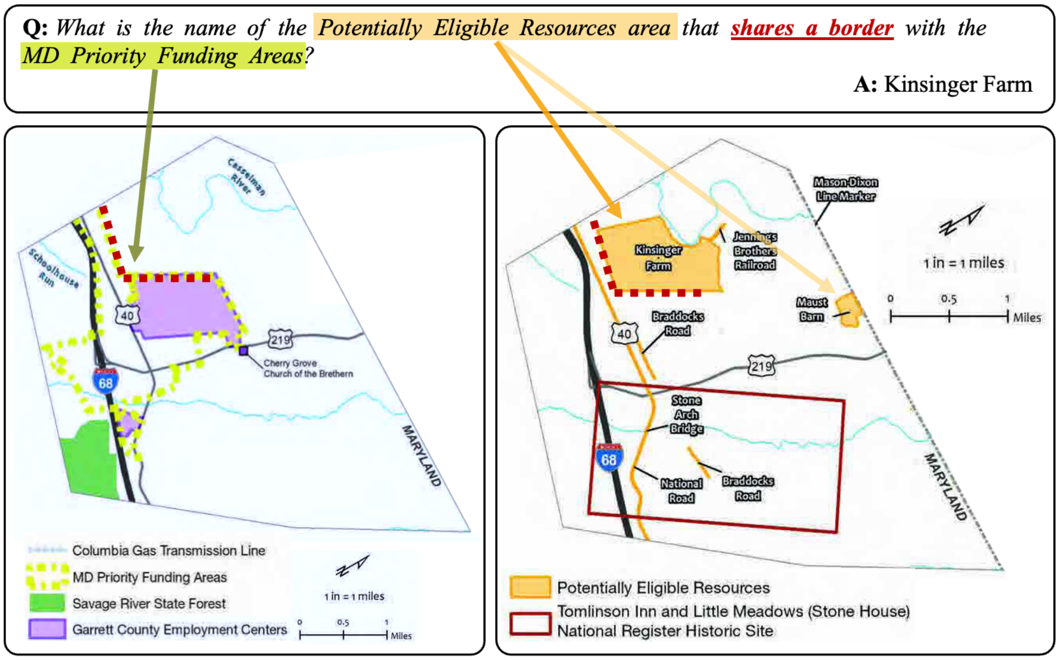



Figure 1: Example of a FRIEDA question requiring multi-map, multi-step cartographic reasoning. To solve the question, the model must (1) use each legend to locate the two referenced regions, (2) evaluate the *border* spatial relation between them, and (3) read the map label of the qualifying feature to answer "Kinsinger Farm."

tances through the map scale), and directional reasoning (using a compass or north arrow) (Clementini et al., 1993; Cohn & Renz, 2007). In addition, human map-reading competencies (Liben et al., 2010; Muir, 1985) frequently require these inferences across multiple maps within a single document. Correctly answering a map question, therefore, draws on map-specific skills (Hegarty & Waller, 2005) such as interpreting map elements, reasoning over spatial relations, and integrating evidence across multiple maps, as well as broader capabilities emphasized in LVLM research that include text grounding (Singh et al., 2019; Sidorov et al., 2020), numeric and logical inference (Lu et al., 2024a; Hu et al., 2023), multi-image integration (Wang et al., 2024a; Xia et al., 2025), and retrieval (Wang et al., 2025a; 2024b). A cartographic reasoning benchmark can therefore probe comprehensive reasoning and provide a clear understanding of the spatial intelligence of LVLMs.

A growing line of work began to evaluate LVLMs on map VQA, yet these benchmarks do not fully assess cartographic reasoning. Earlier datasets pose chart-style questions that can be answered without interpreting spatial relations, which bypasses the topological, metric, and directional inferences that are central to map comprehension (Koukouraki et al., 2025; Chang et al., 2022). Other efforts cover only a subset of relations as they target specific tasks such as navigation (Feng et al., 2025; Kazemi et al., 2025) or entity identification (Dihan et al., 2025). While suitable for those objectives, such coverage is insufficient for evaluating human-like map understanding (Liben et al., 2010). Many benchmarks also restrict the stylistic variability of maps. Some focuses on choropleths (Koukouraki et al., 2025; Chang et al., 2022; Mukhopadhyay et al., 2025), others rely on maps created with map-coloring tools (Srivastava et al., 2025) or common web basemaps (Kazemi et al., 2025; Dihan et al., 2025). Several further focus on limited thematic domains (e.g., geology (Huang et al., 2025)) or restricted geography coverage (Chang et al., 2022; Srivastava et al., 2025). These constraints overlook the heterogeneity in toponyms, labeling conventions, projections, and symbology that real-world cartography demands (Slocum et al., 2022; Robinson et al., 1978). Multi-map reasoning is rarely evaluated, with limited exceptions (Kazemi et al., 2025), even though practical cases often require integrating evidence across multiple maps (e.g., reconciling transit maps with future land-use maps for urban planning) and aligning overlapping information (Lupien & Moreland, 1987). Moreover, although document-level multimodal understanding is emphasized in other LVLM benchmarks, existing map VQA benchmarks seldom require selecting the correct map among many images in long reports, despite government documents and technical documents containing numerous, visually similar, context-dependent maps (Federal Emergency Management Agency, 2025; U.S. Environmental Protection Agency, 2025; SEDAR+, 2025). As a result, current map VQA settings underestimate the demands of comprehensive map understanding, leaving it unclear whether LVLMs possess human-like map-reading competencies. Full cartographic reasoning remains beyond the scope of what existing VQA benchmarks assess.

We introduce FRIEDA, a benchmark designed for evaluating **multi-map, multi-step, comprehensive cartographic reasoning** in LVLMs. We curate maps from public documents of various thematic domains (e.g., geological surveys, planning reports, environmental studies) to develop questions that require models to interpret maps as they appear in reports, mirroring practical scenarios in which a reader must synthesize evidence from maps embedded in a document (see Figure 1). The collection encompasses a diverse range of styles, projections, and scales. We create each question such that it requires (1) reasoning over topological, metric, and directional relations, (2) interpreting map elements and their semantics (e.g., legends, map scales, and north arrows), and, when applicable, (3) integrating information across multiple maps, and (4) selecting the appropriate map(s) from a document to answer the query. To probe genuine reasoning rather than random guessing, the answers are in a free-form (not multiple-choice) format. The benchmark evaluation includes two settings: a *direct* setting, which provides the relevant map images with the question to focus on evaluating map comprehension, and a *contextual* setting, where the model must first retrieve the correct maps from a broader within-document collection before answering. We score outputs using a unified, task-aware protocol aligned to the three spatial-relation categories. We evaluate textual responses (topological and semantic labels) with LLM-as-Judge (Gu et al., 2025), distance responses (numeric values with units) with unit-aware parsing and mean absolute percentage error (MAPE), and directional responses (cardinal directions for relative position) with angular tolerance over the eight directions. We compare the result against the human upper bound derived from multi-annotator agreement to contextualize LVLM performance. By aligning our tasks with the competencies expected of human map-readers (Goodchild, 2012; Liben et al., 2010) and explicitly targeting compositional cross-image inference that is largely absent from prior map VQA work, FRIEDA fills a crucial gap in state-of-the-art LVLM evaluation.

Across 11 LVLMs (both proprietary and open source), we find that even state-of-the-art models struggle with multi-step cartographic reasoning. With FRIEDA-direct, where the relevant maps are provided, the best-performing model (Gemini-2.5-Pro) correctly answers fewer than 40% of the questions, far below human performance ($> 80\%$). Overall accuracy remains essentially unchanged in the contextual setting, indicating that retrieval and disambiguation are not the primary bottlenecks; the core difficulty lies in cartographic reasoning itself. Our error analysis highlights recurring failures, such as misreading legends (confusing symbol shapes and colors) and misaligning information across maps when map styles, projections, or map scales differ. We also observe heterogeneous strengths across models (e.g., GPT-5-Think on multi-map questions and Claude-Sonnet-4 on distance queries). However, overall accuracy remains low, highlighting the gap between current LVLMs and the multi-step, cross-image cartographic reasoning skills required.

We organize the remainder of the paper as follows. Section 2 formalizes the tasks and core skills of cartographic reasoning; Section 3 describes the benchmark design and dataset statistics; Section 4 details the models, experimental setup, and evaluation protocol, and reports the results; Section 5 presents ablations and error analyses.

## 2 TASK DEFINITION

Cartographic reasoning is the ability to interpret maps and draw justified inferences from them. In FRIEDA, we design questions to assess core map-reading competence while mirroring realistic document use, where a reader may need to navigate a document to locate the relevant map(s). All questions require (1) reasoning over *spatial relations*, (2) interpreting heterogeneous *map elements*, and (3) integrating evidence across *multiple maps* when necessary. We also include a (4) *contextual setting* in which additional maps are provided, requiring the model to identify relevant map(s) before performing the reasoning. We detail these categories and the accompanying taxonomy below.

**Spatial Relation** Spatial relations describe how geographic features relate in space (Carlson & Logan, 2001), how they are positioned in space (Majic et al., 2021), and how their geometries interact (Renzhong, 1998). In geographic information systems (GIS) and spatial cognition, these relations are often grouped into three categories: topological, metric, and directional (Cohn & Renz, 2007; Clementini et al., 1993). To make these abilities measurable and comparable, FRIEDA separates questions by spatial relation type and grounds the topological portion in the 9-intersection model (Clementini et al., 1993). We consolidate finer-grained subtypes into their broader categories (e.g., *cross* classified as *intersect*, and *contain* classified as *within*), yielding four topological classes: *border* (shared boundary between regions), *equal* (coincident geometries), *intersect* (crossing or overlap of features), and *within* (containment or inclusion of one area inside another). We complement these with one metric primitive, *distance*, and one directional primitive, *orientation*. Together, these six relations maintain the expressiveness of spatial queries while aligning with users' intuitive spatial reasoning.

**Map Elements** Maps are symbolic representations that encode spatial information through abstract conventions (Slocum et al., 2022). Therefore, interpreting map elements is a distinct skill central to cartographic reasoning. The key elements we target are *map text* (place and feature names), *legends* (mappings from color, icons, and patterns to semantic classes), *map scales* (measurements that convert the map distance to the real-world distance), and the *compass* (ESRI, 2021). The styles of these components vary widely across maps: map texts may use different typography or placement rules (Monmonier, 2015), legends may use continuous color ramps or discrete pictograms (Slocum et al., 2022), map scales may appear as bars or frames around the map (Robinson, 1995), and the compass may be a compass rose or a north arrow (Slocum et al., 2022). Practical map interpretation requires grasping the concepts of map elements rather than simply recognizing their shapes. Consequently, our design includes questions that require reading map texts, decoding legends, using the map scale, and applying orientation to demonstrate true map literacy by linking abstract visual encodings to their underlying semantics.

**Multi-Map Reasoning** Beyond interpreting spatial relations and map elements, practitioners regularly perform cross-map comparison and fusion to synthesize multiple map editions or thematic layers (Lupien & Moreland, 1987). Our multi-map setting reflects this practice: we curate questions that present two or more maps together and require the model to integrate evidence by aligning shared symbols, reconciling differences in labels, map scales, and orientation, and identifying co-referent regions or features (Foody, 2007). Extracting distributions and patterns is widely recognized as a core capability (Ishikawa, 2016; Rexigel et al., 2024; Morita & Fukuya, 2025). By testing this setting, we move beyond isolated spatial computation to evaluate deeper cartographic reasoning across varied depictions of the same space.

**Contextual Setting** To mirror practical workflows (Mathew et al., 2021; Tanaka et al., 2023), we evaluate a contextual setting (FRIEDA-contextual), where a model must identify the relevant map before answering a question. In this scenario, we provide the model with multiple maps from the same source (i.e., a document), and the model must perform within-document retrieval using cues in the map, such as titles, legends, or labels. By evaluating model performance on FRIEDA-contextual, we capture a core aspect of real map use: the model must understand how map elements encode meaning and leverage that understanding to select the required map from thematically related alternatives that vary in data layers, geographic extent, or purpose (Ishikawa, 2016).

## 3 FRIEDA

We present FRIEDA, a benchmark for assessing LVLM's comprehensive cartographic reasoning, with an emphasis on cross-map (i.e., multi-image) scenarios. This section summarizes the benchmark statistics and details the dataset curation procedure.

### 3.1 Benchmark Statistics

Table 1 shows that FRIEDA comprises 17,030 map images drawn from 210 documents and a total of 500 questions. To capture real-world variability, the maps span diverse geographies (32 countries) and thematic domains (six domains), exhibiting heterogeneous styles, such as varied color palettes, legends, and symbol conventions.

| Statistics | Number |
|---|---|
| Total questions | 500 |
| Textual answers | 372 (74.4%) |
| Metric answers | 45 (9.0%) |
| Directional answers | 83 (16.6%) |
| Single-map | 202 (40.4%) |
| Multi-map | 298 (59.6%) |
| Total number of documents | 210 |
| Total number of images | 17,030 |
| Maximum question length | 60 |
| Average question length | 24.9 |

Table 1: Key statistics.

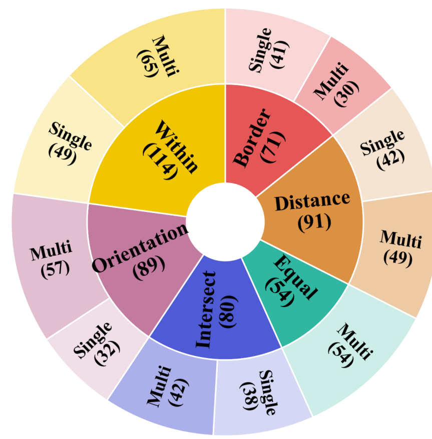



Figure 2: Question distribution by spatial relation (inner) and map count (outer). Sizes are proportional to the number of questions in each category.

Figure 2 reports the question distribution for each spatial relation, further broken down by single- and multi-map questions.[1] The distribution is roughly balanced across relations and settings. We provide the detailed counts by spatial relation and by setting (single- or multi-map) in Appendix C.1 and include a representative example of each in Appendix D.

### 3.2 Benchmark Construction

This section describes the FRIEDA benchmark construction process. The construction phase consists of three stages: map image collection, question generation, and question validity verification.

**Map Image Collection** To capture both stylistic and geographic diversity, we curate maps from publicly available government and multilateral reports across multiple domains: geology (SEDAR+, 2025), national park management (National Park Service, 2025), environmental assessments (U.S. Environmental Protection Agency, 2025; Environmental Protection Agency, 2025; Ministry of Agriculture Climate Change and Environment, 2025), disaster response (Federal Emergency Management Agency, 2025), urban planning (Seattle Planning and Community Development, 2024; Department of Community Development, 2025; Urban Redevelopment Authority, 2025; City of Cape Town, 2025), and infrastructure investment (AIIB, 2025). We limit the sources to documents with Latin-script text to focus on cartographic reasoning rather than transliteration or translation. Since these documents also contain non-map graphics (e.g., photographs, logos, charts), we use Idefics3-8B (Laurençon et al., 2024) with a tailored prompt (Appendix B.1) to flag map images. Also, we manually verify that each extracted set contains only cartographic maps (examples of excluded non-maps appear in Appendix B.1.1). To support FRIEDA-contextual, we retain only documents with at least four maps and construct the contextual map set for each target image based on the proximity of the page and location within the document. We then randomly shuffle the order of the maps in the contextual set to prevent LVLM from using positional cues to identify the target maps.

**Question Generation** For each collected map, we use GPT-4 and GPT-o3 (Appendix B.6) to propose candidate questions, the targeted spatial relations, and a reference answer. We do not impose templates and accept any phrasing that unambiguously represents the target relation as valid to reflect various forms of paraphrases of spatial relations (e.g., "Is A within B?" vs. "Does B contain A"). We then manually review all candidate questions to verify correctness, refine wording, ensure clarity, and confirm the reference answers (i.e., gold answer). To evaluate cartographic reasoning rather than memorization or web retrieval ability, we screen each question for searchability: using GPT with web search enabled, we discard any questions that can be answered without inspecting the map image. Finally, we categorize answer types as textual (short span or categorical label), distance

[1] As *equal* denotes identical geometry (i.e., identical location and size), the benchmark contains no single-map *equal* questions.

(numeric with units), or direction, to allow analysis by response modality. These steps ensure that FRIEDA consists only high-quality, unambiguous questions with reliable gold answers.

**Annotation Pipeline** We validate every question with annotations from 11 Ph.D. researchers in computer science (eight with expertise in maps). We collect annotations for four weeks, with four batches of question sets assigned to each annotator. Annotators confirm that each question is answerable from the provided map(s), and for multi-map questions, they verify that all maps are required to answer the question. To avoid bias, the researcher who verified and edited a given question did not participate in its validation. As each question is independently labeled by three annotators, we only keep the question if a majority agrees with the gold answer; otherwise, it is discarded. In total, we remove 61 questions that fail to reach an agreement $\geq$2/3. Appendix B.3 details the instruction prompt provided to the annotators, and Appendix B.5 shows the annotation interface.

# 4 Experiments

This section details the experimental setup, baselines, and evaluation metrics, and then presents the main result, showing that FRIEDA is a challenging benchmark even for the strongest LVLMs.

## 4.1 Experimental Setup

**Models** We evaluate 11 LVLMs with multi-image support on FRIEDA. For proprietary models, we test three models: Gemini-2.5-Pro (Gemini Team, 2025), GPT-5-Think (OpenAI, 2025), and Claude-Sonnet-4 (Anthropic, 2025). For open source models, we consider eight model families and evaluate the largest available model from each family: LLaVA-NeXT-110B (Li et al., 2024b), GLM4.5V-108B (Team et al., 2025), InternVL3-78B (Chen et al., 2024b), LLaVA-OneVision-72B (Li et al., 2024a), Qwen2.5VL-72B (Bai et al., 2025), InternVL3.5-38B (Wang et al., 2025b),[2] Ovis2-34B (Lu et al., 2024b), and Ovis2.5-9B (Lu et al., 2025).

To enforce determinism in open-source models, we set `do_sample=False` and `temperature=0`. For proprietary models, we use the default settings of each model with maximum reasoning enabled (e.g., `reasoning=high` for GPT-5-Think) and append the instruction "Do not use search" to turn off external retrieval. All models receive the same set of instructions that human annotators receive (Appendix B.4).

**Evaluation metrics** Answers in FRIEDA fall into three categories: textual, distance, and direction. For textual answers, we employ an LLM-as-Judge (Gu et al., 2025) method, utilizing Mistral Small 3.1 (Mistral AI, 2024) as the evaluator.[3] The full judge prompt appears in Appendix E.1. This setup handles minor variation (e.g., 'Cypress Creek' vs. 'Cypress') by matching semantics rather than identifying exact string equality. For distance-based answer, we report mean absolute percentage error (MAPE) and consider predictions within 20% error as correct, following Lewis (1982). For directional answers, we mark a response correct if it matches the target cardinal direction within one adjacent label (e.g., if the gold answer is North, accept North, North West, and North East), reflecting the perceptual nature of the labels. We validate the reliability of the evaluation method against manual annotations, achieving a Cohen's $\kappa$ of 0.9028 across all judged questions, which supports its suitability for evaluation.

## 4.2 Evaluation Results

Figure 3 summarizes the overall performance, and Table 2 reports accuracy by spatial relation. As FRIEDA retains questions with at least 2/3 annotator agreeing on the gold answer, we report accuracy for two subsets: *All-Agree*, where all three annotators agreed, and *Partial-Agree*, where 2/3 annotators agreed. *All-Agree* items serve as an indirect indicator of questions that are easier and less ambiguous for the annotators under our task and instructions, whereas *Partial-Agree* items may be considered as intrinsically more difficult or ambiguous to answer correctly. We also report

[2] We evaluate the 38B variant rather than the 241BA28B variant as the latter activates only 28B parameters during inference. We report the results for the 241BA28B setting in Appendix F.2.

[3] Mistral is not the language backbone of any tested LVLM, thereby reducing potential bias.

the *Overall Accuracy*, which aggregates both subsets. Even the strongest LVLM (Gemini-2.5-Pro) remains below 40% overall accuracy, well behind human performance at 84%. The best open source result (Ovis2.5-9B-Think) achieves 24% overall accuracy, underperforming proprietary systems and far below humans. We find no clear relationship between model size and performance, suggesting that training data, training objectives, and explicit reasoning mechanisms matter more than scale for cartographic reasoning.

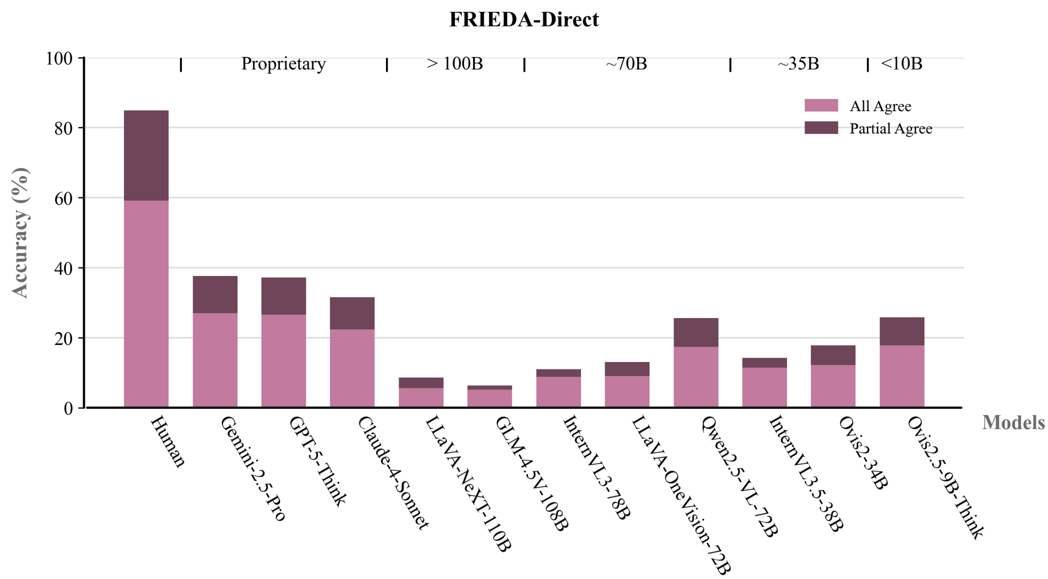



Figure 3: Overall accuracy of different models on the FRIEDA-direct benchmark.

| | Overall (500) | Border (71) | Distance (91) | Equal (54) | Intersect (80) | Orientation (89) | Within (115) |
|---|---|---|---|---|---|---|---|
| Human Average | **84.87** | **89.00** | **78.28** | **89.10** | **85.53** | **91.80** | **88.08** |
| ***Proprietary LVLMs*** | | | | | | | |
| Gemini-2.5-Pro | **38.20** | <u>32.39</u> | <u>25.27</u> | 33.33 | <u>28.75</u> | **71.59** | **35.34** |
| GPT-5-Think | <u>37.20</u> | 25.35 | **27.47** | **44.44** | **31.25** | <u>69.32</u> | <u>28.45</u> |
| Claude-Sonnet-4 | 31.60 | **33.80** | 23.08 | <u>37.04</u> | 22.50 | 56.82 | 21.55 |
| ***Open Source LVLMs*** | | | | | | | |
| LLaVA-NeXT-110B | 8.60 | 4.23 | 10.99 | 11.11 | 16.25 | 0.00 | 9.48 |
| GLM-4.5V-108B | 6.40 | 5.41 | 2.15 | 21.57 | 6.17 | 1.16 | 7.83 |
| InternVL3-78B | 11.00 | 1.41 | 4.40 | 12.96 | 5.00 | 34.09 | 7.76 |
| LLaVA-OneVision-72B | 13.00 | 9.86 | 10.99 | 5.56 | 8.75 | 29.55 | 10.34 |
| Qwen2.5-VL-72B | 25.60 | 11.27 | 14.29 | 25.93 | 17.50 | 55.68 | 25.86 |
| InternVL3.5-38B | 14.20 | 11.27 | 8.79 | 14.81 | 2.50 | 36.36 | 11.21 |
| Ovis2-34B | 17.80 | 25.35 | 13.19 | 25.93 | 26.25 | 2.27 | 18.97 |
| Ovis2.5-9B-Think | 25.80 | 12.68 | 20.88 | 24.07 | 22.50 | 51.14 | 21.55 |

Table 2: Overall and per spatial relation accuracy of human and LVLMs on FRIEDA-direct.

## 5 Analysis

**Error analysis on Gemini Pro** To pinpoint where LVLMs fail, we analyze Gemini-2.5-Pro on the *All-Agree* subset (in total, 167 questions). This ensures that our analysis targets distinct model failures on questions that humans find straightforward. We assign each incorrect answer to a single primary error category. When multiple issues co-occur, we prioritize errors that occur earlier in the reasoning pipeline that propagate to downstream steps. The largest source of error involves the misinterpretation of legends (25.61%): cases where the model assigns colors or symbols to the wrong class. The remaining 23.78% is due to cross-map interpretation failures, which reflect difficulties in aligning the map scales and shared features across maps, and 16.46% is due to spatial-relation semantics error, which arises when the model mixes up spatial relations (e.g., labeling region B *within* A when it only *touches* A at the boundary). Map-element misunderstandings include mistakes with the map scale (9.76%; unit or ratio errors), map text (8.93%; selecting the wrong place or feature from labels), geometry or shape reference (3.66%; pointing to the wrong area on the map), and orientation (3.05%; ignoring a tilted compass). Finally, we observe generic VQA errors not specific to cartography, such as miscounting (6.71%), subject-object confusion (1.82%; referring 'A relative to B' as 'B relative to A'), and hallucination (1.20%). For the top three error categories, we provide examples and rationales returned by the three proprietary models in Appendix F.1.

**Performance by spatial relation** Figure 4 reports per-spatial relation accuracy for human annotators and the three proprietary models. LVLM performance broadly tracks the human baseline: both are most accurate on orientation and struggle most with distance. On questions where an annotator answers incorrectly, LVLMs are also incorrect 84.53% of the time. While GPT-5-Think and Gemini-2.5-Pro achieve comparable overall accuracy, GPT-5-Think is stronger on tasks that require multi-map reasoning (Table 8), indicating better integration of evidence across maps. This is most evident in the equal relation questions, a multi-map exclusive task, where GPT-5-Think's accuracy is nearly 13% higher compared to Gemini-2.5-Pro. Notably, Claude-Sonnet-4 is the strongest on distance questions, particularly those requiring interpretation of the map scale to compute exact distances.

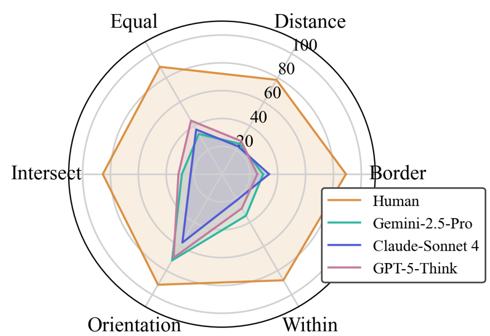



Figure 4: Per spatial relation accuracy (%) of human annotators and three proprietary LVLMs (Gemini-2.5-Pro, Claude-Sonnet-4, and GPT-5-Think) on FRIEDA-direct.

**Performance on contextual setting** We observe a minimal difference in accuracy between the FRIEDA-direct (Figure 3 and Table 2) and FRIEDA-contextual (Appendix E.2). To verify that this is not an artifact of the accuracy metric, we directly compare the per-question performance of the eight open-source models under deterministic settings (i.e., `do_sample=False` and `temperature=0`). We observe 88.03% per-question performance agreement between the direct and contextual settings, indicating that contextual images (maps from the same document that are not required to answer the question) rarely affect the model's prediction.

| Model | Accuracy (%) |
|---|---|
| Ovis2.5-9B | 19.00 |
| Ovis2.5-9B-Think | **24.80** |

Table 3: Performance of Ovis2.5 model on FRIEDA-direct.

**Impact of reasoning (think) on cartographic question types** Despite being the smallest model tested with FRIEDA, Ovis2.5-9B-Think attains strong results (4th overall and 1st among open source models). To identify what drives this performance, we further evaluate Ovis2.5-9B with explicit reasoning (i.e., *Think*) disabled (Table 3). The overall accuracy of Ovis2.5-9B remains above the open source average, indicating that model characteristics (e.g., architecture, training data) contribute to the model's strong results. Enabling *Think* adds an additional 5% performance gain. To identify which question types benefit from explicit reasoning and whether it improves cartographic performance, we manually analyze the 60 questions that only the *Think* variant answers correctly. Reasoning helps mostly with cardinal-direction questions, where north faces the top of the image (48.33%), followed by multi-map alignment (23.33%). Additional improvements come from correctly reading map text (15%), interpreting the map scale (5%), associating legend with symbol (5%), and counting (3%). Together, these patterns suggest that explicit reasoning primarily strengthens orientation-related and multi-map questions, which are central to carographic reasoning, while yielding smaller gains in symbol and map scale interpretation.[4]

[4] We further evaluate the association between performance and model size in Appendix F.2

# 6 Related Work

**Document & Infographic/Chart VQA** With the advancement of LVLMs, researchers proposed various benchmarks to test LVLM's ability to perform visual reasoning over documents and designed graphics. DocVQA (Mathew et al., 2021) introduced a large-scale question-answering (QA) dataset over real forms and reports, and DocVXQA (Souibgui et al., 2025) builds upon these benchmarks to design a self-explanatory framework that produces interpretable rationales of LVLMs. DocoPilot (Duan et al., 2025) evaluates LVLMs on scientific articles, which not only test text understanding but also interpretation of embedded figures such as charts. For graphics, InfographicsVQA (Mathew et al., 2022) tests joint reasoning over text, layout, and pictorial elements in visually rich infographics. InfoChartQA (Lin et al., 2025) pairs plain charts and infographics to identify design elements that degrade LVLMs' performance. In general, VQA evaluation on frontier LVLMs reveals a trend of being competent at high-level patterns, such as trends and extrema, but struggles with precise value extraction and, in general, robustness. FRIEDA evaluates these shortcomings in a cartographic setting where layout, symbols, legends, scales, and compass orientation interact tightly to measure how well LVLMs integrate these signals to answer map-based questions.

**Map VQA and Spatial Reasoning** LVLMs continue to struggle with spatial and symbolic logic, which are essential to cartographic reasoning. Recent map VQA benchmarks have made progress but remain narrow in scope. MapQA (Chang et al., 2022) evaluates question answering on choropleth maps and shows that general VQA and ChartVQA systems underperform on maps. MapWise (Mukhopadhyay et al., 2025) broadens the geographic coverage and uses 43 question templates targeting spatial relations and map features. MapIQ (Srivastava et al., 2025) extends the map type coverage to include cartograms and proportional-symbol maps and tests analytic tasks under different color schemes. MapEval (Dihan et al., 2025) assesses geospatial reasoning across various cities and reports consistent human-LVLMs performance gaps. There also exists map VQAs on domain-specific datasets such as PEACE (Huang et al., 2025) on geology maps and ReasonMap (Feng et al., 2025) on navigation tasks. Overall, prior work has mainly assessed single-map reasoning under constrained map styles and geographic diversity, as well as limited types of spatial relations. Multi-image understanding on heterogeneous maps remains underexplored. Cross-image benchmarks, such as ReMI (Kazemi et al., 2025), likewise reveal significant human-model gaps. We detail key differences between prior map VQA benchmarks and FRIEDA in Appendix G.

Spatial reasoning benchmarks, such as SpatialVLM (Chen et al., 2024a) and SpatialRGPT (Cheng et al., 2024), have advanced 2D and 3D spatial perception and reasoning on natural images. However, their work does not engage with symbolic conventions unique to maps (i.e., legends, scales, compasses, and map texts). In contrast, our benchmark closes this gap by evaluating multi-step cartographic reasoning over heterogeneous, real-document maps, which requires models to integrate evidence across multiple figures and align legends, scales, and orientation to infer key spatial relations (i.e., border, distance, equal, intersect, orientation, and within).

# 7 Conclusion

We present FRIEDA, a benchmark for evaluating multi-step cartographic reasoning across six spatial relations, often requiring multi-image alignment. Our evaluation across 11 state-of-the-art LVLMs demonstrates a substantial gap between current performance and the proficiency required for robust map understanding. Analysis reveals that these failures extend beyond issues observed in prior VQA datasets, highlighting the need for novel architectures and effective training methods that incorporate cartographic priors and explicit reasoning over map elements. We will release the error taxonomy and baseline results, alongside FRIEDA, to catalyze progress. We encourage the community to build on FRIEDA with methods that explicitly integrate text, symbology, and geospatial structure, toward LVLMs that reason reliably over real-world maps.

## Ethics Statement

We introduce a benchmark for evaluating cartographic reasoning in large vision-language models. We curate maps from publicly available documents (e.g., government reports, planning, and environmental studies) and retain only the figures necessary for research purposes. To the best of our knowledge, we use all materials under terms that permit research and non-commercial distribution.

All annotators provided informed consent. We collected no personal data about annotators beyond task performance. Our institution's IRB reviewed the annotation protocol and determined that the project does not constitute human subjects research; therefore, no further IRB review was required.

The benchmark inevitably reflects the patterns in the source documents and may exhibit representation bias, including uneven geographic coverage and map types, English-language focus, and unequal representation across regions and themes. We document these limitations and their potential impact in the dataset card (Appendix A) to aid transparency and interpretation.

### Reproducibility Statement

Upon the end of the anonymity period, we plan to release: (1) the benchmark (images, QA JSON, taxonomy, and provenance), (2) code for data loading, inference, evaluation, and table/figure generation, (3) code to replicate the annotation interface, and (4) all prompts and configuration files used for annotation and inference. In the meantime, we provide all details needed to reproduce our results in the main text and appendices: Section 3 describes dataset construction, Section 4 specifies models and inference settings, and Section 5 reports ablations and error analyses.

## References


AIIB. Our projects, 2025. URL `https://www.aiib.org/en/projects/list/index.html`.

Anthropic. System card: Claude opus 4 & claude sonnet 4, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL `https://aclanthology.org/Q18-1041/`.

Aileen Buckley. Make maps people want to look at: five primary design principles for cartography. *ArcNews Online*, 2006.

L A Carlson and G D Logan. Using spatial terms to select an object. *Mem. Cognit.*, 29(6):883–892, September 2001.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*, 2022. URL `https://openreview.net/forum?id=znKbVjeR0yI`.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024.

City of Cape Town. Document centre, 2025. URL `https://www.capetown.gov.za/Document-centre`.

Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In David Abel and Beng Chin Ooi (eds.), *Advances in Spatial Databases*, pp. 277–295, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. ISBN 978-3-540-47765-5.

Anthony G. Cohn and Jochen Renz. *Handbook of Knowledge Representation*. Elsevier Science, San Diego, CA, USA, 2007. ISBN 0444522115.

Department of Community Development. Media centre, 2025. URL https://addcd.gov.ae/.

Mahir Labib Dihan, MD Tanvir Hassan, MD TANVIR PARVEZ, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hS2Ed5XYRq.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, et al. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4026–4037, 2025.

Environmental Protection Agency. Publications, 2025. URL https://www.epa.ie/publications/.

ESRI. Map elements, 2021. URL https://desktop.arcgis.com/en/arcmap/latest/map/page-layouts/map-elements.htm.

Federal Emergency Management Agency. Fema, 2025. URL https://www.fema.gov/.

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

Giles M Foody. Map comparison in GIS. *Prog. Phys. Geogr.*, 31(4):439–445, August 2007.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, November 2021. ISSN 0001-0782. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Google Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Michael F. Goodchild. The fourth r? rethinking gis education. *ArcUser Fall*, pp. 46–51, 2012.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Mary Hegarty and David A. Waller. *Individual Differences in Spatial Abilities*, pp. 121–169. Cambridge Handbooks in Psychology. Cambridge University Press, 2005.

Keith J. Holyoak and Robert G. Morrison. *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, 03 2012. ISBN 9780199734689. doi: 10.1093/oxfordhb/9780199734689.001.0001. URL https://doi.org/10.1093/oxfordhb/9780199734689.001.0001.

Pengbo Hu, Jingxian Qi, Xingyu Li, Hong Li, Xinqi Wang, Bing Quan, Ruiyu Wang, and Yi Zhou. Tree-of-mixed-thought: Combining fast and slow thinking for multi-hop visual reasoning. *ArXiv*, abs/2308.09658, 2023. URL https://api.semanticscholar.org/CorpusID:261031742.

Yangyu Huang, Tianyi Gao, Haoran Xu, Qihao Zhao, Yang Song, Zhipeng Gui, Tengchao Lv, Hao Chen, Lei Cui, Scarlett Li, et al. Peace: Empowering geologic map holistic understanding with mllms. *arXiv preprint arXiv:2501.06184*, 2025.

Toru Ishikawa. Spatial thinking in geographic information science: Students' geospatial conceptions, map-based reasoning, and spatial visualization ability. *Ann. Am. Assoc. Geogr.*, 106(1): 76–95, January 2016.

Toru Ishikawa and Nora S Newcombe. Why spatial is special in education, learning, and everyday activities. *Cogn. Res. Princ. Implic.*, 6(1):20, March 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Kim A Kastens and Toru Ishikawa. Spatial thinking in the geosciences and cognitive sciences: A cross-disciplinary look at the intersection of the two fields. In *Earth and Mind: How Geologists Think and Learn about the Earth*, pp. 53–76. Geological Society of America, 2006.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, and Ahmed Qureshi. Remi: a dataset for reasoning with multiple images. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Eftychia Koukouraki, Auriol Degbelo, and Christian Kray. Assessing Map Reproducibility with Visual Question-Answering: An Empirical Evaluation. In Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan (eds.), *13th International Conference on Geographic Information Science (GIScience 2025)*, volume 346 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 13:1–13:12, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-378-2. doi: 10.4230/LIPIcs.GIScience.2025.13. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.GIScience.2025.13.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.

C.D. Lewis. *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*. Butterworth scientific. Butterworth Scientific, 1982. ISBN 9780408005593. URL https://books.google.com/books?id=t8W4AAAAIAAJ.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Zekun Li, Malcolm Grossman, Mihir Kulkarni, Muhao Chen, Yao-Yi Chiang, et al. Mapqa: Open-domain geospatial question answering on map data. *arXiv preprint arXiv:2503.07871*, 2025.

Lynn S. Liben, Lauren J. Myers, and Adam E. Christensen. Identifying locations and directions on field and representational mapping tasks: Predictors of success. *Spatial Cognition & Computation*, 10(2-3):105–134, 2010. doi: 10.1080/13875860903568550. URL https://doi.org/10.1080/13875860903568550.

Minzhi Lin, Tianchi Xie, Mengchen Liu, Yilin Ye, Changjian Chen, and Shixia Liu. Infochartqa: A benchmark for multimodal question answering on infographic charts, 2025. URL https://arxiv.org/abs/2505.19028.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024a.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024b.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report. *arXiv:2508.11737*, 2025.

Anthony E Lupien and William H Moreland. A general approach to map conflation. In *Proceedings of 8th International Symposium on Computer Assisted Cartography (AutoCarto 8)*, pp. 630–639, 1987.

Ivan Majic, Elham Naghizade, Stephan Winter, and Martin Tomko. RIM: a ray intersection model for the analysis of the between relationship of spatial objects in a 2D plane. *Geogr. Inf. Syst.*, 35 (5):893–918, May 2021.

Ahmedand Masry, Doand Long, Jia Qingand Tan, Shafiqand Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2022. doi: 10.1109/WACV51458.2022.00264.

Richard E Mayer. *Multimedia learning*. Cambridge University Press, Cambridge, England, 3 edition, July 2020.

Ministry of Agriculture Climate Change and Environment. Resources, 2025. URL https://macce.gov.sc/resources/.

Mistral AI. Mistral-small-instruct-2409. https://huggingface.co/mistralai/Mistral-Small-Instruct-2409, 2024. Hugging Face model card. License: Mistral AI Research License (MRL).

M. Monmonier. *The History of Cartography, Volume Six: Cartography in the Twentieth Century*. University of Chicago Press, 2015. ISBN 9780226152127. URL https://books.google.td/books?id=BZRfEAAAQBAJ.

Aiko Morita and Izumi Fukuya. Integrative processing of text and multiple maps in multimedia learning: an eye-tracking study. *Front. Psychol.*, 16(1487439):1487439, August 2025.

Sharon Pray Muir. Understanding and improving students' map reading skills. *Elem. Sch. J.*, 86(2): 207–216, November 1985.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. MAPWise: Evaluating vision-language models for advanced map queries. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9348–9378, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.473. URL https://aclanthology.org/2025.naacl-long.473/.

National Park Service. Publications, 2025. URL https://www.nps.gov/aboutus/publications.htm.

OpenAI. GPT-5 system card, 2025.

Guo Renzhong. SPATIAL OBJECTS AND SPATIAL RELATIONSHIPS. *Geo Spat. Inf. Sci.*, 1(1): 38–42, October 1998.

Eva Rexigel, Jochen Kuhn, Sebastian Becker, and Sarah Malone. The more the better? a systematic review and meta-analysis of the benefits of more than two external representations in STEM education. *Educ. Psychol. Rev.*, 36(4), December 2024.

A.H. Robinson. *Elements of Cartography*. Wiley, 1995. ISBN 9780471555797. URL https://books.google.com/books?id=mUyAAAAAMAAJ.

A.H. Robinson, R.D. Sale, and J.L. Morrison. *Elements of Cartography*. Wiley, 1978. ISBN 9780471017813. URL https://books.google.com/books?id=QknctEDueRcC.

Seattle Planning and Community Development. Current projects, 2024. URL https://www.seattle.gov/opcd/current-projects.

SEDAR+. Sedar archive, 2025. URL https://www.sedarplus.ca:5443/t/legacysedardata/views/LegacySedarReportFinalPublic/Home?%3Aembed=y&%3AisGuestRedirectFromVizportal=y.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1195. URL https://aclanthology.org/2024.emnlp-main.1195/.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. 2020.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. *Thematic cartography and geovisualization, fourth edition*. CRC Press, London, England, 4 edition, August 2022.

Mohamed Ali Souibgui, Changkyu Choi, Andrey Barsky, Kangsoo Jung, Ernest Valveny, and Dimosthenis Karatzas. DocVXQA: Context-aware visual explanations for document question answering. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wex0vL4c2Y.

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. MapIQ: Evaluating multimodal large language models for map question answering. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=VSwRuGtB5n.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: a dataset for document visual question answering on multiple images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on*

*Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i11.26598. URL https://doi.org/10.1609/aaai.v37i11.26598.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Barbara Tversky. Navigating by mind and by body. In Christian Freksa, Wilfried Brauer, Christopher Habel, and Karl F. Wender (eds.), *Spatial Cognition III*, pp. 1–10, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45004-7.

Urban Redevelopment Authority. Master plan, 2025. URL https://www.ura.gov.sg/Corporate/Planning/Master-Plan.

U.S. Environmental Protection Agency. Environmental impact statement (eis) database, 2025. URL https://www.aiib.org/en/projects/list/index.html.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.166. URL https://aclanthology.org/2025.naacl-long.166/.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024b.

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

Cathleen Wharton and Walter Kintsch. An overview of construction-integration model: a theory of comprehension as a foundation for a new cognitive architecture. *SIGART Bull.*, 2(4):169–173, July 1991. ISSN 0163-5719. doi: 10.1145/122344.122379. URL https://doi.org/10.1145/122344.122379.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. MMIE: Massive multimodal interleaved comprehension benchmark for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HnhNRrLPwm.

# A DATACARD

We adopt the data statement framework of Bender & Friedman (2018) and integrate complementary fields from Datasheets for Datasets (Gebru et al., 2021) to centralize key information for the analysis, reuse, and deployment of FRIEDA.

## A.1 CURATION RATIONALE

We design FRIEDA to evaluate carographic reasoning (the ability to interpret map-specific symbols, comprehend spatial relations, and integrate evidence across one or more maps). We source questions from public documents to reflect map-reading tasks encountered in practice (e.g., planning, hazard assessment, and geology). High-level goals, task definitions, and design choices appear in the main text (Section 2 and Section 3). We further expand on the benchmark curation process in Appendix B.

## A.2 BENCHMARK COMPOSITION

- **Total size**: 500 validated questions; each question with 1 gold answer
- **Agreement**: Each question is annotated by three annotators; we record the problem-level agreement and mark items with unanimous agreement on the gold answer as *All-Agree*, and those with 2/3 agreement as *Partial-Agree*.
- **Modalities**: Every question involves one or multiple map image(s) and associated question text.
- **Spatial relations (6)**: Border, Equal, Intersect, Within, Distance, Orientation
- **Answer types (3)**: Textual (short text), Distance, and Direction
- **Provenance**: Public documents from 32 countries across six continents. Documents are from six domains (urban planning, environmental assessment, national park management, geologic reports, disaster and hazard reports, infrastructure and investment reports). Sources are detailed further in Section 3 and Appendix C.3.
- **Languages**: Questions and instructions are in English (en-US); source maps primarily use English labels but may include other languages written in the Latin script.

## A.3 DATA COLLECTION PROCESS

- **Acquisition**: We collected maps from public reports, then filtered for reading map elements and task suitability.
- **Question creation**: Curators wrote questions that required reading the legend, scale, and compass, and reasoning over one or more spatial relations; questions were rejected if (1) they were solvable without using any maps or (2) if question ambiguity could not be resolved by manual editing.

## A.4 ANNOTATOR DEMOGRAPHIC

We share the annotator demographics to contextualize potential biases while preventing re-identification.

- **Count**: 11 Annotators in total (2 also served as question curator)
- **Academic background**: Ph.D. Researchers [100%]
- **GIS/cartography background**: $\leq$1 year: [27%]; 1–3 years: [27%]; 3–5 years: [18%]; 5+ years: [27%].
- **Language**: All authoring and communication used American English (en-US). As the task focuses on cartographic symbols and spatial relations (not dialect), we do not report individual annotator nationalities. Non-native participation may introduce minor phrasing variance. We standardized qustion phrasing during review and removed questions flagged as ambiguous by $\geq$ 2/3 annotators.

### A.5 EVALUATION & METRICS

- **Primary metric**: Accuracy
- **Textual (LLM-as-Judge)**: After attempting exact string match, we use an LLM-as-Judge to compare model outputs to gold answers. Appendix E.1 provides the judging prompt for reproducibility.
- **Distance (MAPE)**: We apply mean absolute error (MAPE) and unit-aware parsing and consider all distance answers with in 20% as correct.
- **Direction**: We canonicalize directional answers to the eight cardinal directions and consider all cardinal direction within one adjacent unit as correct.

### A.6 KNOWN LIMITATIONS & BIASES

- **Regional representation bias**: As FRIEDA uses only English-language documents, regions where English is a dominant language are overrepresented, and non-English conventions and locales are not covered.
- **Domain skew**: The corpus emphasizes planning, environmental, and government reports with less coverage on other types of maps, such as nautical or military charts.

## B DETAILED BENCHMARK CONSTRUCTION

### B.1 MAP IMAGE FILTERING

We use Idefics3-8B (Laurençon et al., 2024) to filter map images from the document. To produce a strict Yes/No decision, we prompt the model:
`Is this a cartographic map? Answer only with Yes or No.`

We consider any image for which the model responds `Yes` as a candidate map.

#### B.1.1 NON-MAP EXAMPLES

We manually verify all map candidate images and remove those that we do not consider as maps. For example, although Figure 5 shows a silhouette of a city with subdivision, we consider it as a stylized graphic rather than a cartographic map. The image lacks essential map elements (i.e., map texts, legend, scale, and compass), which are needed to support cartographic reasoning. Without these components, we cannot reason about locations, distances, or spatial relationships; therefore, we exclude such images from our dataset and do not treat them as maps for FRIEDA.

### B.2 DEFINITION OF SPATIAL RELATION

Figure 6 visualizes the four topological spatial relations evaluated in FRIEDA: border, equal, intersect, and within.

### B.3 ANNOTATOR PROMPT

To standardize responses and minimize ambiguity, we supply annotators with a fixed instruction set (Figure 7). We introduce these guidelines during task onboarding and repeat them at the start of every question to promote a consistent answer format.

### B.4 LVLM SYSTEM PROMPT

To ensure consistency, we use the same instruction set provided to human annotators as the prompt for the LVLM system. As some LVLMs produce intermediate reasoning, we append a final line to standardize the output: `Give the final answer in 'Final answer: ⟨your answer⟩`. For the proprietary models, we additionally include the clause `Do not use online search` to prevent external browsing.[5]

[5] We add this clause as a precautionary measure; during the dataset construction phase, we verify that questions are not directly answerable through web search.


Figure 5: An example of a non-map image flagged by Idefics3-8B as a candidate map. The image is a graphic from the cover page of the document. We exclude it from the benchmark after manual verification, as we consider it a graphical image rather than a cartographic map.

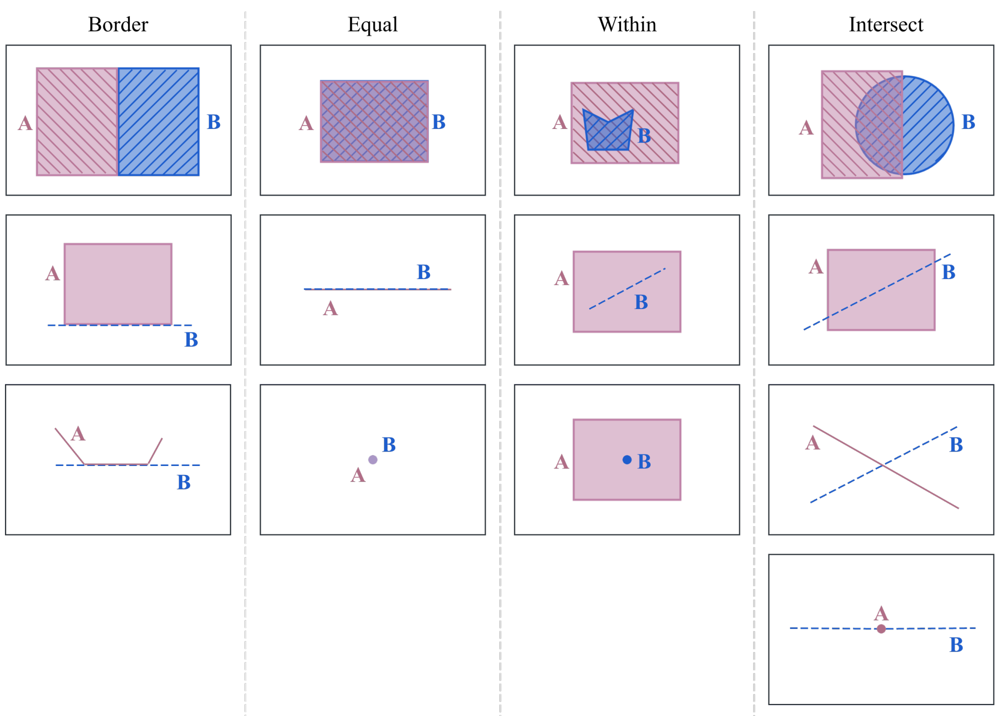



Figure 6: Illustrations of the spatial relations evaluated in the benchmark. Columns show *Border*, *Equal*, *Within*, and *Intersect*; rows provide representative cases across geometry types—areas, lines, and points.

For each one, please verify whether it can be answered (Q# Validation) using the provided map(s). If an image appears too small, click on the image. For question with multiple images, please mark whether all images were required to correctly answer the question (Q# M). You may use tools like a ruler or calculator, but do not use online search.

For each questions:
General:
- If question can be answered, write answer in short answer box
- If answer is a text from the map, copy it as it appears

Numerical Answers:
- Include units as indicated on the map *(Don't convert 1200m to 1.2km)*
- If both map frame and ruler scale is available, <u>use the ruler scale</u>
- If question asks for an area, use {unit}^2
- Use numerical values *(e.g., 4 instead of four)*

Directional Answers:
- Use 8 cardinal directions only: North, North East, East, South East, South, South West, West, North West
- Write 'North' or 'South' before 'East' or 'West'
- Notice that the north arrow compass do not always point upward

Multi-Part Answers:
- Separate with semicolon (;) *(e.g., Zone A; Zone B)*

Figure 7: Instruction provided to annotators at the beginning of every question.

```
Answer the questions based on the following criteria:
General:
* If question can be answered, write answer in short answer box
* If answer is a text from the map, copy it as it appears

Numerical Answers:
* Include units as indicated on the map (Don't convert 1200m to
1.2km)
* If both map frame and ruler scale is available, use the ruler
scale
* If question asks for an area, use {unit}^2
* Use numerical values (e.g., 4 instead of four)

Directional Answers:
* Use 8 cardinal directions only:  North, North East, East, South
East, South, South West, West, North West
* Write 'North' or 'South' before 'East' or 'West'
* Notice that the north arrow compass do not always point upward

Multi-Part Answers:
* Separate with semicolon (;) (e.g., Zone A; Zone B)
Give the final answer in 'Final answer:  ⟨your answer⟩'
{Do not use online search}
```

Figure 8: System prompt used for LVLM inference. For readability in the figure, newline characters (`\n`) are shown as actual line breaks.

### B.5 Annotation Platform

We built a web interface (Figure 9) to collect annotator responses. We provide the annotator instruction (Figure 7) at the top of every question similar to how LVLMs receives the system instruction for each question. For each question, annotators see the question and its associated map image(s), then (1) enter a short free-text answer if it is considered answerable, (2) mark answerability by selecting either "Can be answered" or "Map doesn't contain information to answer the question" (the latter requires a brief justification), and (3) for multi-map questions, indicate whether all images are necessary to precisely answer the question without guessing.

### B.6 LLM to Generate Questions

We use GPT-4 and GPT-o3[6] with a tailored prompt (Figure 10) to draft candidate questions for FRIEDA. In addition to the prompt, we supply 10 randomly selected map images for question generation. Two of the authors then manually review each candidate question, editing or discarding questions as needed to ensure correctness, clarity, and coverage of targeted spatial relations before adding them to the benchmark.

## C Extended Benchmark Details

### C.1 Question Count per Spatial Relation

In Table 4, we report the number of questions in FRIEDA by spatial relation, including totals as well as the counts split into single-map and multi-map questions. The distribution is roughly balanced: *Within* is the largest class (23.0%), while *Equal* is the smallest (10.8%).

| **Spatial Relation** | **Total Q Count** | **Single-map Q Count** | **Multi-map Q Count** |
|---|---|---|---|
| Border | 71 (14.2%) | 41 (8.2%) | 30 (6.0%) |
| Distance | 91 (18.2%) | 42 (8.4%) | 49 (9.8%) |
| Equal | 54 (10.8%) | 0 (0.0%) | 54 (10.8%) |
| Intersect | 80 (16%) | 38 (7.6%) | 42 (8.4%) |
| Orientation | 89 (17.8%) | 32 (6.4%) | 57 (11.4%) |
| Within | 115 (23.0%) | 49 (9.8%) | 65 (13.0%) |

Table 4: Question statistics in FRIEDA across six spatial relations. The table reports the total number of questions per relation, along with their breakdown into multi-map and single-map settings.

### C.2 Example Question per Spatial Relation

In Table 5, we present one sample question for each spatial relation, split by map count (single-map vs. multi-map).

### C.3 Nation and Domain Coverages

**Nation Coverage** FRIEDA includes maps from government documents and multilateral reports from 32 countries across six continents (Figure 11; Table 6). We also report the ten most-represented countries by question count in Figure 12.

**Domain Coverage** We source maps from domains where spatial reasoning is essential: geologic reports (SEDAR+, 2025), national park management reports (National Park Service, 2025), investment and infrastructure reports (AIIB, 2025), disaster and hazard assessments (Federal Emergency Management Agency, 2025), city and regional planning documents (Seattle Planning and Community Development, 2024; City of Cape Town, 2025; Department of Community Development, 2025; Urban Redevelopment Authority, 2025), and environmental reviews (U.S. Environmental Protection

[6]Questions are generated before the release of GPT-5.

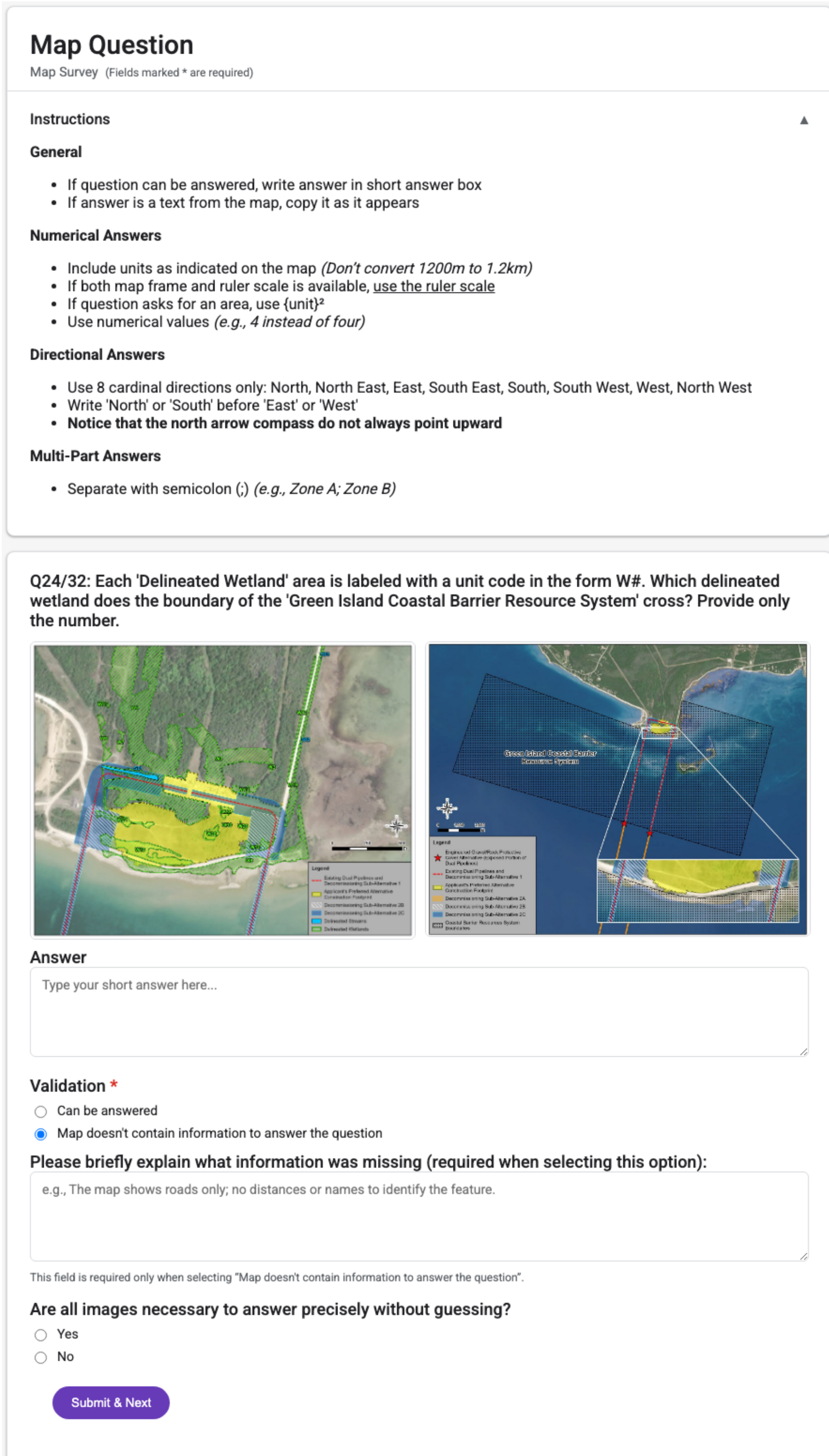



Figure 9: Annotation interface for validating questions of FRIEDA.

> I'm trying to create a benchmark dataset to test out generative AI's ability on complex cartographical reasoning on maps. The hard questions we should provide in this benchmark should leverage information from one or a few of the given maps above, and should involve some reasoning. Also, the questions should follow these criteria:
>
> - Answer should be self-contained, non-binary, and not-multiple choice questions.
> - Question should not be solved by searching online - We assume that the image to refer to is not known when answering the question.
> - We assume that the image to refer to is not known when answering the question.
>
> Give a set of questions, the maps to refer to, and the answer to the question. Target spatial relation is {Spatial Relation}.

Figure 10: Question-generation prompt used to prompt for candidate questions to either GPT-4 or GPT-o3.

| **Spatial Relation** | **Map Count** | **Question Example** |
|---|---|---|
| Border | Single | Which DMMUs along the Inner Harbor Navigation Canal share a boundary with 'DMMU 4'? Answer in the form DMMU #. |
| | Multi | Identify the 'National Road [map1]' that crosses the 'Ou Ta Paong' area. Which two 'Irrigation Schemes [map2]' does this road serve as a border for? Provide the names without the word 'Area'. |
| Distance | Single | What is the approximate straight-line distance between the SLC-6 Launch Site and the 2 psf contour of the Falcon Heavy Launch line? |
| | Multi | In Tinian, each 'Heritiera longipetiolata' species observation area is associated with a name. Which 'observation area[map1]' is located closest to the Noise Sensitive Receptor labeled 'T15[map2]'? |
| Equal | Multi | Which 'feature[map1]' on the infrastructure map corresponds to 'Existing Component 60[map2]' of the Santander project? |
| Intersect | Single | How many Asanko tenement blocks does the Haul Road intersect? |
| | Multi | Which 'claim block(s)[map1]' of the UEX Christie Lake Project are crossed by the 'power line[map2]'? |
| Orientation | Single | What is the name of the northernmost 'National Air Monitoring Site' as recorded by Ordnance Survey Ireland? |
| | Multi | In the Lumberton Loop Project Area, what is the orientation of the 'Crosswalk Stripping [map1]' in relation to the 'Walnut Street Component[map2]'? |
| Within | Single | Along Pine Street and Pike Street, how many 'Future Redevelopment & Renovation Project' areas overlap with the 'West Focus Area'? |
| | Multi | Identify the area of Nighthawk Gold Property located North of the 'Winter Road[map1]'. How many 'Gold Deposits[map2]' are located within this area? |

Table 5: Example questions by spatial relation and map count. For multi-map questions, entities are annotated with [map1]/[map2] only for illustration, indicating different source maps; these tags are not part of the actual questions. We underline the word/phrase that denotes the target spatial relation.

Agency, 2025; Ministry of Agriculture Climate Change and Environment, 2025; Environmental Protection Agency, 2025). Several of these are umbrella categories that can be further subdivided. For example, environmental assessments may target facilities, hydrology, land use/land cover, or habitat. For consistency, we retain the top-level labels used by the source repositories. Across these domains, maps employ varied symbol conventions (legends, scale bars, north arrows) and heterogeneous geometry types (areas, lines, points), encouraging generalization beyond any single map style. Figure 13 summarizes the domain coverage.

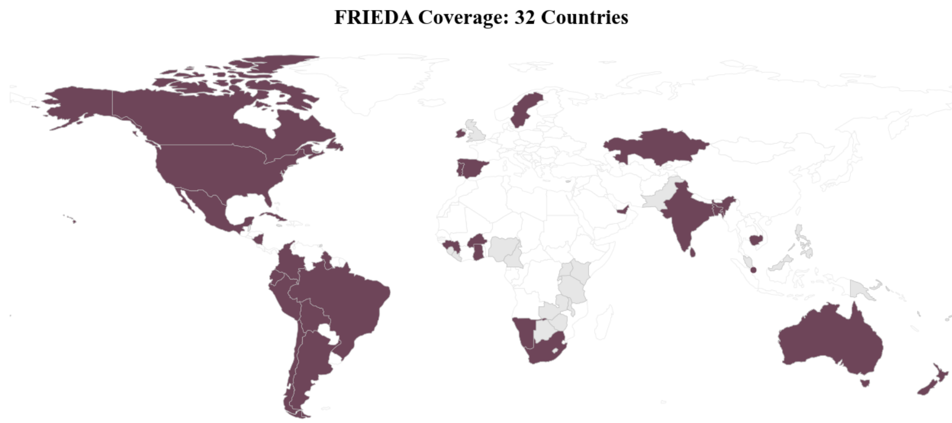



Figure 11: Global country coverage of FRIEDA. Countries included in the dataset are shown in purple; countries where English is a primary or official working language but not covered by FRIEDA are shaded light gray. Coverage spans six continents (32 countries).

| Country | Count | Country | Count |
|---|---|---|---|
| United States | 251 | Mexico | 18 |
| Canada | 82 | Portugal | 2 |
| South Africa | 32 | New Zealand | 1 |
| Peru | 9 | Chile | 4 |
| Burkina Faso | 1 | Brazil | 2 |
| Guyana | 2 | Guinea | 3 |
| Ireland | 24 | Colombia | 2 |
| Seychelles | 14 | Ecuador | 1 |
| Singapore | 9 | Cuba | 1 |
| Kazakhstan | 6 | Argentina | 3 |
| Cambodia | 5 | Bolivia | 2 |
| India | 7 | Spain | 1 |
| Bangladesh | 6 | Sweden | 1 |
| Sri Lanka | 3 | Australia | 1 |
| United Arab Emirates | 3 | Namibia | 2 |
| Ghana | 1 | Nicaragua | 1 |

Table 6: Country coverage in FRIEDA. Count reflects the number of questions whose maps originate from each country.

# D EXAMPLES OF FRIEDA

We store each example as a JSON record containing the question, the gold answer, the required map image(s), any contextual image(s), and metadata such as the number of maps, target spatial relation, and answer type. Figure 14 illustrates an example of a single-map question, and Figure 15 shows an example of a multi-map question.

# E DETAILED BENCHMARK RESULT AND ANALYSIS

## E.1 LLM-AS-JUDGE PROMPT

To evaluate free-form textual answers, we employ LLM-as-Judge (Gu et al., 2025) using Mistral-Small-3.1 (Mistral AI, 2024). Since not all models follow our requested output format ("Final answer: ⟨your answer⟩") and minor wording differences may occur (e.g., '15.00%' vs. '15'), we first require the LLM to extract the answer span based on the question and then compare the extracted portion to the gold answer with tolerance for minor variants (Figure 16).

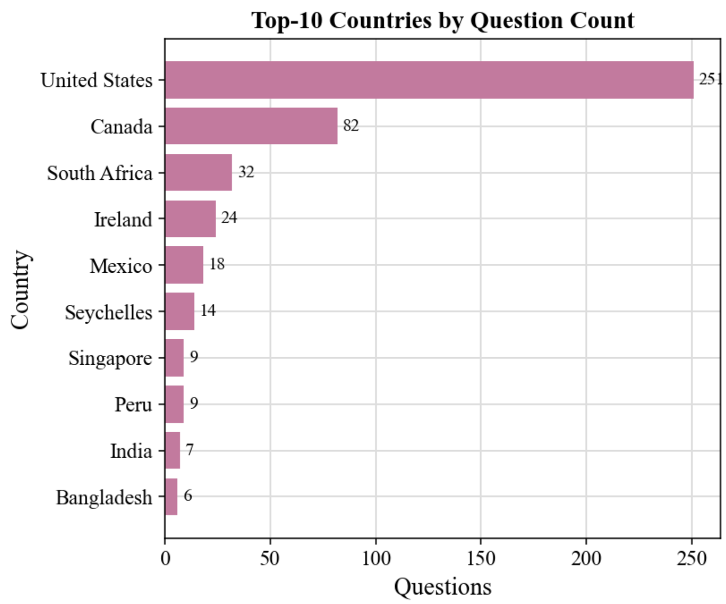



Figure 12: Top 10 countries by question Count

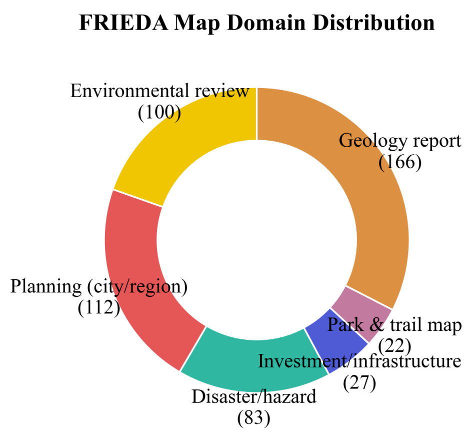



Figure 13: Domain distribution in FRIEDA by document category (e.g., geology, planning). Slices indicate categories, and parentheses denote question counts.

### E.2 Performance on FRIEDA-contextual

Table 7 reports overall and per-spatial relation performance for FRIEDA-contextual. As noted in Section 5, models show little difference between the FRIEDA-direct and FRIEDA-contextual settings. Figure 17 summarizes overall accuracy across models on FRIEDA-contextual.

### E.3 Per Map Count & Answer Type Result Breakdown

We also report performance by map count and answer type for FRIEDA-direct (Table 8) and FRIEDA-contextual (Table 9). In the FRIEDA-direct setting, GPT-5-Think leads on multi-map questions, outperforming the next-best model (Gemini-2.5-Pro) by roughly 5%. Claude-Sonnet-4 performs best on *Distance* answers but underperforms on directional (i.e., *Orientation*) questions.

```
"question_ref":"q_1093",
"question_text":"What is the orientation of 'Bryan Palmer & Barry
Maust' in relation to 'Gary Blocher' within the Meyersdale Study
Area?",
"expected_answer":"South",
"image_urls":[
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image21_m1.png"
],
"map_count":"Single",
"spatial_relationship":"Orientation",
"answer_type":"cardinal",
"contextual_urls":[
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image21_m1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image21_m0.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image20_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image22_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image19_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image26_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image15_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image11_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image10_1.png",
  "EIS/Vol-3-FEISAppendixA-M-May-2025/image12_1.png"
]
```


Figure 14: Example question single map


```
"question_ref":"q_0150",
"question_text":"The Aberdeen-Hoquiam North Shore Levee is
classified into three categories.  In which category is the
'Hoquiam Police Station' located?",
"expected_answer":"North Shore Levee (West)",
"image_urls":[
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image116_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image101_1.png"
],
"map_count":"Multi",
"spatial_relationship":"Intersect",
"answer_type":"textual",
"contextual_urls":[
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image116_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image118_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image136_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image138_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image101_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image139_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image140_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image141_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image142_1.png",
  "FEMA/BRIC-EMS-2020-BR-102-0002_WA-NorthShoreLeveeWest-DEA-
    20241126/image137_1.png"
]
```


Figure 15: Example question multi map

> You will be given a triple consisting of a question, an expected answer, and a given response. Your task is to output either ‘yes’ or ‘no’. Given the question and response, extract only the exact portion of the text that serves as the answer from the given response. Then output ‘yes’ if the user response conveys the same meaning as the expected answer in relation to the question. Output ‘no’ if it does not. For questions with multiple correct answers, the expected answers are separated by semicolons. The user response is correct if it matches all required answers, regardless of order. When the user provides more items than required, the response is treated as incorrect. If the user lists fewer items than expected, mark the response as incorrect. Differences in plurality, extra details such as acronyms or counts, minor typographical errors, and differences in wording style do not affect correctness. Focus only on whether the meaning matches.
>
> Question: {Question}
> Expected answer: {Expected Answer}
> Given response: {User Response}
>
> Does the response correctly answer the question based on the expected answer?
> Answer strictly ‘yes’ or ‘no’

Figure 16: The input prompt to generate questions.

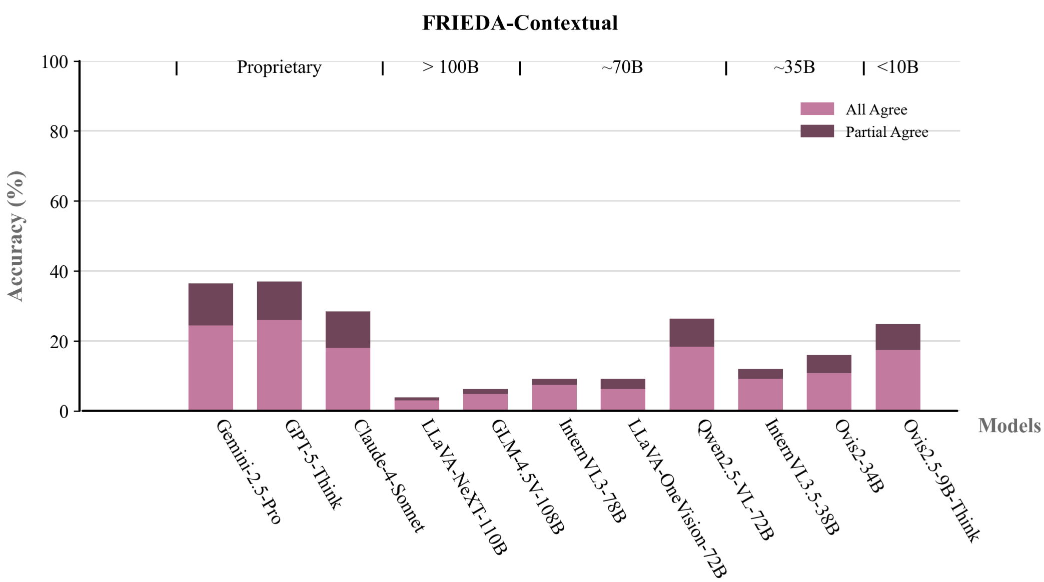



Figure 17: Overall accuracy across models in the FRIEDA-contextual setting.

| Model Size | Accuracy (%) |
|---|---|
| 1B | 9.40 |
| 2B | 12.80 |
| 4B | 20.00 |
| 8B | 23.20 |
| 14B | 23.00 |
| **30BA3B** | **24.20** |
| 38B | 14.20 |
| 241BA28B | 11.40 |

Table 10: InternVL3.5 performance by size

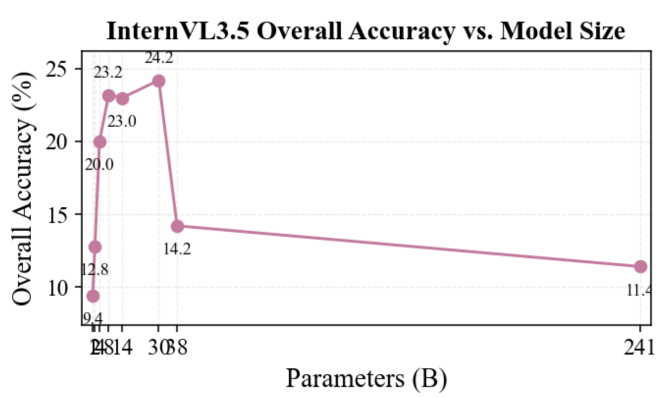



Figure 18: Performance of InternVL3.5 by model parameter size

| | Overall (500) | Border (71) | Distance (91) | Equal (54) | Intersect (80) | Orientation (89) | Within (115) |
|---|---|---|---|---|---|---|---|
| ***Proprietary LVLMs*** | | | | | | | |
| Gemini 2.5 Pro | <u>36.60</u> | **28.17** | **29.67** | **50.00** | 21.25 | **64.77** | **30.17** |
| GPT-5-Think | **37.00** | **28.17** | <u>27.47</u> | <u>40.74</u> | **36.25** | <u>61.36</u> | **30.17** |
| Claude Sonnet 4 | 28.40 | 19.72 | 19.78 | 27.78 | 23.75 | 55.68 | 23.28 |
| ***Open Source LVLMs*** | | | | | | | |
| LLaVA-NeXT-110B | 3.80 | 2.86 | 3.30 | 3.64 | 7.50 | 0.00 | 5.17 |
| GLM-4.5V-108B | 7.40 | 9.46 | 0.00 | 6.00 | 13.41 | 1.12 | 12.82 |
| InternVL3-78B | 9.20 | 2.82 | 5.49 | 5.56 | 3.75 | 30.68 | 5.17 |
| LLaVA-OneVision-72B | 9.20 | 7.04 | 5.49 | 3.7 | 7.5 | 17.05 | 11.21 |
| Qwen2.5-VL-72B | 26.40 | 12.68 | 16.48 | 29.63 | 16.25 | 55.68 | 25.86 |
| InternVL3-78B | 9.20 | 2.82 | 5.49 | 5.56 | 3.75 | 30.68 | 5.17 |
| InternVL3.5-38B | 12.00 | 8.45 | 4.40 | 7.41 | 7.50 | 34.09 | 8.62 |
| Ovis2-34B | 16.00 | 21.13 | 14.29 | 18.52 | 18.75 | 2.27 | 21.55 |
| Ovis2.5-9B-Think | 24.80 | 18.31 | 9.89 | 27.78 | 21.25 | 56.82 | 17.24 |

Table 7: Performance of humans and 11 LVLMs across 6 spatial relationships on FRIEDA-contextual setting. Values represent performance scores (in %) for each spatial relationship and the overall accuracy.

| | | Map Count | | Answer Types | | |
|---|---|---|---|---|---|---|
| | Overall (500) | Single (202) | Multi (298) | Textual (372) | Distance (45) | Direction (83) |
| Human Average | **84.87** | **84.91** | **88.08** | **87.93** | **67.18** | **92.15** |
| ***Proprietary LVLMs*** | | | | | | |
| Gemini 2.5 Pro | **38.20** | **32.67** | <u>41.95</u> | **33.06** | 15.56 | **73.49** |
| GPT-5-Think | <u>37.20</u> | 23.76 | **46.31** | <u>30.65</u> | <u>26.67</u> | <u>72.29</u> |
| Claude Sonnet 4 | 31.60 | <u>24.26</u> | 36.58 | 25.81 | **28.89** | 59.04 |
| ***Open Source LVLMs*** | | | | | | |
| LLaVA-NeXT-110B | 8.60 | 7.43 | 9.40 | 10.48 | 8.89 | 0.00 |
| GLM-4.5V-108B | 6.40 | 4.81 | 7.53 | 8.33 | 0.00 | 1.23 |
| InternVL3-78B | 11.00 | 6.93 | 13.76 | 6.18 | 4.44 | 36.14 |
| LLaVA-OneVision-72B | 13.00 | 15.35 | 11.41 | 9.41 | 11.11 | 30.12 |
| Qwen2.5-VL-72B | 25.60 | 21.78 | 28.19 | 21.24 | 8.89 | 54.22 |
| InternVL3.5-38B | 14.20 | 11.88 | 15.77 | 9.68 | 6.67 | 38.55 |
| Ovis2-34B | 17.80 | 17.33 | 18.12 | 22.58 | 11.11 | 0.00 |
| Ovis2.5-9B-Think | 25.80 | 22.28 | 28.19 | 20.43 | 20.00 | 53.01 |

Table 8: Performance of humans and 11 LVLMs across the two map count types and three answer types on FRIEDA-direct.

| | Overall (500) | Map Count Single (202) | Map Count Multi (298) | Answer Types Textual (372) | Answer Types Distance (45) | Answer Types Direction (83) |
|---|---|---|---|---|---|---|
| ***Proprietary LVLMs*** | | | | | | |
| Gemini 2.5 Pro | <u>36.60</u> | <u>25.25</u> | <u>44.30</u> | **31.99** | <u>17.78</u> | **67.47** |
| GPT-5-Think | **37.00** | **26.24** | **44.30** | **31.72** | **28.89** | <u>65.06</u> |
| Claude Sonnet 4 | 28.40 | 20.30 | 33.89 | 23.12 | <u>17.78</u> | 57.83 |
| ***Open Source LVLMs*** | | | | | | |
| LLaVA-NeXT-110B | 3.80 | 1.99 | 5.03 | 4.85 | 2.22 | 0.00 |
| GLM-4.5V-108B | 7.40 | 6.19 | 8.28 | 9.95 | 0.00 | 0.00 |
| InternVL3-78B | 9.20 | 6.93 | 10.74 | 4.57 | 4.44 | 32.53 |
| LLaVA-OneVision-72B | 9.20 | 7.43 | 10.40 | 7.53 | 6.67 | 18.07 |
| Qwen2.5-VL-72B | 26.40 | 18.32 | 31.88 | 21.24 | 11.11 | 57.83 |
| InternVL3.5-38B | 12.00 | 8.42 | 14.43 | 7.53 | 4.44 | 36.14 |
| Ovis2-34B | 16.00 | 14.36 | 17.11 | 19.62 | 15.56 | 0.00 |
| Ovis2.5-9B-Think | 24.80 | 20.79 | 27.52 | 19.62 | 6.67 | 57.83 |

Table 9: Performance of humans and 11 LVLMs across the two map count types and three answer types on FRIEDA-contextual.

# F Extended Analyses

## F.1 Examples of Each Error Category

We illustrate the three most frequent error categories for Gemini-2.5-Pro and show each example alongside answers and reasoning from Gemini-2.5-Pro, GPT-5-Think, and Claude-Sonnet-4.

**Misinterpretation of legend** Listing 1 presents a case where the model fails to map a legend symbol or color to its intended semantic class, leading to the selection of the wrong feature despite the correct evidence being present.

**Cross-map interpretation failure** Listing 2 shows a failure that arises when reasoning requires aligning information across multiple maps or overlays; the model identifies the wrong subject when the maps must be cross-referenced.

**Spatial-relation semantics error** Listing 3 illustrates a case where the model misinterprets the key spatial relation, yielding an incorrect answer.

## F.2 Analyses on Model Size

In the main evaluation (Figure 3), the results deviate from the usual scaling law (Kaplan et al., 2020), which states that the performance of the model improves with size. Among open-source models, LLaVA-NeXT, despite having the most parameters, ranks near the bottom, whereas Ovis-2.5-9B, the smallest model, ranks near the top. We, therefore, hypothesize that cartographic reasoning is not an emergent ability (i.e., a capability absent in smaller models but present in larger ones). To test this, we evaluate the InternVL3.5 family (Wang et al., 2025b) on FRIEDA: 1B, 2B, 4B, 8B, 14B, 30BA3B, 38B, 241BA28B where 'A' denotes parameters active at inference. The trend (Figure 18, Table 10) shows modest gains up to roughly 30B parameters, followed by degradation thereafter.

# G Extended Related Works

We compare FRIEDA with prior MapVQA benchmarks (Table 11), reporting (1) the spatial abilities evaluated, (2) diversity of map elements (via country and domain counts), (3) whether multi-map reasoning is tested, and (4) whether a contextual setting is included. Orange checkmarks (✓) denote partial coverage. For example, in the topological spatial relation category, we treat questions such as "how many points lie along the route to location A?" as an acceptable form of *intersect*. However, these models do not examine the spatial relation in the depth or rigor as of FRIEDA.

| | Spatial Relation | | | Map Elements | | | |
|---|---|---|---|---|---|---|---|
| | Topological | Metric | Directional | Country | Domain | Multi-Map | Contextual |
| MapQA (Chang et al., 2022) | ✗ | ✗ | ✗ | 1 | 1 | ✗ | ✗ |
| MapWise (Mukhopadhyay et al., 2025) | ✓ | ✗ | ✓ | 3 | 3 | ✗ | ✗ |
| MapIQ (Srivastava et al., 2025) | ✓ | ✗ | ✗ | 1 | 6 | ✗ | ✗ |
| MapEval (Dihan et al., 2025) | ✓ | ✓ | ✓ | 54 | 1 | ✗ | ✗ |
| ReMi (Kazemi et al., 2025) | ✗ | ✗ | ✓ | 100? | 1 | ✓ | ✗ |
| PEACE (Huang et al., 2025) | ✓ | ✓ | ✓ | 2 | 1 | ✗ | ✗ |
| ReasonMap (Feng et al., 2025) | ✗ | ✗ | ✗ | 13 | 1 | ✗ | ✗ |
| **FRIEDA** | ✓ | ✓ | ✓ | 32 | 6 | ✓ | ✓ |

Note: ReMi (Kazemi et al., 2025) reports counts by city, not by country; consequently, the corresponding country total is less than 100.

Table 11: A comparison of FRIEDA with prior map VQA benchmarks. FRIEDA covers a broader set of map-reading abilities and exhibits greater geographic and thematic diversity.

## H THE USE OF LARGE LANGUAGE MODELS

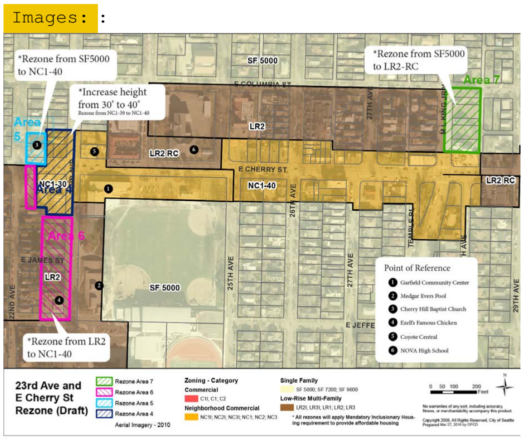
We acknowledge the use of large language models (LLMs) for benchmark question curation, revision, and polishing of this paper. The details of usage, the exact prompt used, and all related information are provided in the main paper or appendices. All questions created with the assistance of a large language model have been verified and modified by the authors. The paper's main contribution remains with the authors.

Listing 1: Legend misinterpretation example of Gemini-2.5-Pro on FRIEDA. Other models are shown for reference. Orange : Task Instruction. Green : Correct Answer. Red : Incorrect Answer.

System: :

```
Answer the questions based on the following criteria:
General:
    * If question can be answered, write answer in short answer box
    * If answer is a text from the map, copy it as it appears

Numerical Answers:
    * Include units as indicated on the map (Don't convert 1200m to 1.2km)
    * If both map frame and ruler scale is available, use the ruler scale
    * If question asks for an area, use {unit}^2
    * Use numerical values (e.g., 4 instead of four)

Directional Answers:
    * Use 8 cardinal directions only: North, North East, East, South East, South, South West,
    West, North West
    * Write 'North' or 'South' before 'East' or 'West'
    * Notice that the north arrow compass do not always point upward

Multi-Part Answers:
    * Separate with semicolon (;) (e.g., Zone A; Zone B)

    Give the final answer in 'Final answer: <your answer>'
    Do not use online search.
```

Images: :




Question: The 23rd Ave and E Union St Rezone (Draft) classifies each area into four distinct Rezone Areas. With which Rezone Area does 'Rezone Area 4' share its westernmost boundary. Give the name of the area in form Rezone Area #.

Gold Answer: Rezone Area 6

---

Gemini-2.5-Pro:
I've determined the user's objective: to find the "Rezone Area" west of "Rezone Area 4". I've analyzed the map and deconstructed the task. "Rezone Area 4" is the blue-hatched area. Its western border is along 22nd Ave. I've identified the adjacent pink area. Now, according to the legend, the pink/magenta area corresponds to "Rezone Area 5". Final answer: Rezone Area 5

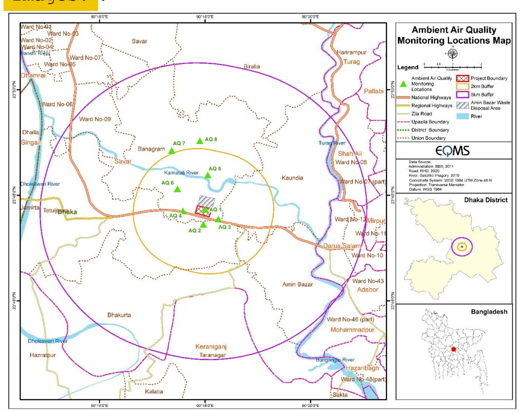
GPT-5-Think: The image shows a zoning map with areas labeled 4-7, and the title mentions "23 rd Ave and E Cherry St Rezone (Draft)." The question asks about Rezone Area 4's westernmost boundary. I'll need to find Area 4 on the map and check which area lies directly to the west of it, likely along 23rd Ave. The map shows colored boxes for Areas 4, 5, and 6. Area 4, labeled in light blue, is near the left margin, adjacent to "NC1-30" and "NC1-40." It seems that Area 4's westernmost boundary touches Area 5, which is directly to the left of Area 4. So , I suspect Area 5 shares this boundary. Final answer: Rezone Area 5

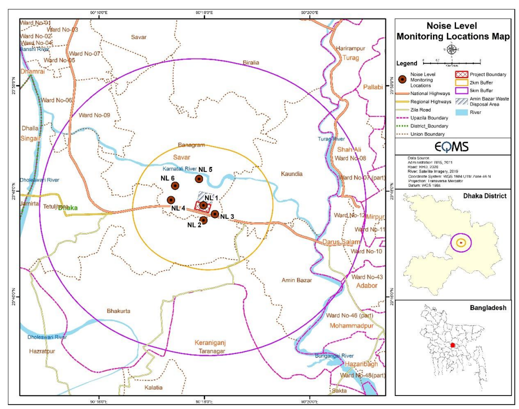
Claude-4-Sonnet: Looking at the map, I can identify Rezone Area 4 by its blue diagonal line pattern. This area is located in the western portion of the map.
Examining the westernmost boundary of Rezone Area 4, I can see that it directly borders the magenta/pink colored area, which according to the legend corresponds to Rezone Area 6. Final answer: Rezone Area 6

Listing 2: Cross-map interpretation failure example of Gemini-2.5-Pro on FRIEDA. Other models are shown for reference. Orange: Task Instruction. Green: Correct Answer. Red: Incorrect Answer.

---

```
System: :
Answer the questions based on the following criteria:
General:
    * If question can be answered, write answer in short answer box
    * If answer is a text from the map, copy it as it appears

Numerical Answers:
    * Include units as indicated on the map (Don't convert 1200m to 1.2km)
    * If both map frame and ruler scale is available, use the ruler scale
    * If question asks for an area, use {unit}^2
    * Use numerical values (e.g., 4 instead of four)

Directional Answers:
    * Use 8 cardinal directions only: North, North East, East, South East, South, South West,
    West, North West
    * Write 'North' or 'South' before 'East' or 'West'
    * Notice that the north arrow compass do not always point upward

Multi-Part Answers:
    * Separate with semicolon (;) (e.g., Zone A; Zone B)

    Give the final answer in 'Final answer: <your answer>'
    Do not use online search.

Images: :
```




```
Question: Identify the Air Quality Monitoring Location closest to the Noise Level Monitoring
Location 'NL 5'. What is the orientation of this point in relation to 'NL 5'?

Gold Answer: North East
```

---

Gemini-2.5-Pro: I've determined that AQ 6 is the closest point to NL 5. Now, I am analyzing their relative orientation. It appears that from NL 5, AQ 6 lies in a west-northwest direction . Final answer: AQ 6; <u>West</u>

GPT-5-Think: Looking at the maps, AQ5 seems nearest to NL5, based on their locations along the river. AQ5 seems to be slightly northeast of NL5, just a bit to the east of it. To answer the question, it makes sense to say AQ5 is closest to NL5. As for the orientation, I think AQ5 lies slightly southeast of NL5. The AQ5 triangle is located near the river, south of the river bend, while NL5 is to the north. I'll go with: AQ5 lies southeast of NL5. This seems like the most accurate conclusion. Final answer: AQ 5; <u>southeast</u>.

Claude-4-Sonnet: Looking at both maps, I need to locate NL 5 on the Noise Level Monitoring Locations Map and find the closest AQ location on the Ambient Air Quality Monitoring Locations Map. NL 5 is located near the Karnatali River area. Examining the AQ locations on the first map:
- AQ 5 is positioned very close to the same Karnatali River area
- AQ 6 is slightly further west
- AQ 2, AQ 3, AQ 4 are further south
- AQ 7 and AQ 8 are further north

AQ 5 is clearly the closest Ambient Air Quality Monitoring Location to NL 5.
Comparing the relative positions of AQ 5 and NL 5, AQ 5 appears to be positioned to the east of NL 5. Final answer: AQ 5; <u>East</u>"}

---

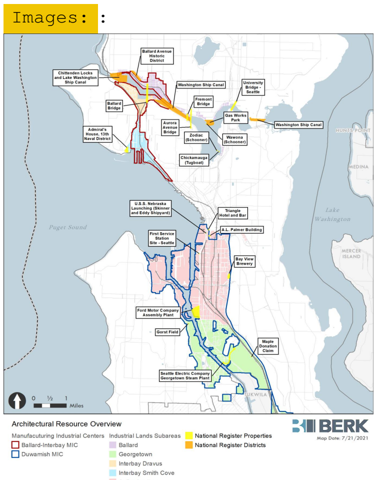
Listing 3: Spatial-relation semantics error example of Gemini-2.5-Pro on FRIEDA. Other models are shown for reference. Orange : Task Instruction. Green : Correct Answer. Red : Incorrect Answer.

---

System: :

```
Answer the questions based on the following criteria:
General:
    * If question can be answered, write answer in short answer box
    * If answer is a text from the map, copy it as it appears

Numerical Answers:
    * Include units as indicated on the map (Don't convert 1200m to 1.2km)
    * If both map frame and ruler scale is available, use the ruler scale
    * If question asks for an area, use {unit}^2
    * Use numerical values (e.g., 4 instead of four)

Directional Answers:
    * Use 8 cardinal directions only: North, North East, East, South East, South, South West,
    West, North West
    * Write 'North' or 'South' before 'East' or 'West'
    * Notice that the north arrow compass do not always point upward

Multi-Part Answers:
    * Separate with semicolon (;) (e.g., Zone A; Zone B)

    Give the final answer in 'Final answer: <your answer>'
    Do not use online search.
```

Images: :




Question: Which 'National Register Districts' overlap with both the 'Ballard' and 'Interbay Dravus' Industrial Lands Subareas within the Ballard-Interbay MIC?
Gold Answer: Washington Ship Canal

---

Gemini-2.5-Pro: I'm currently focused on dissecting the user's complex query. My goal is to determine the exact regions where the yellow "National Register Districts" overlap with both the purple "Ballard" and orange "Interbay Dravus" subareas, all contained within the "Ballard-Interbay MIC." I'm visualizing the overlay on the map. I've broken down the request into map elements and used the legend to visualize the features. I've concluded that the only district meeting all criteria is the "Chittenden Locks and Lake Washington Ship Canal." Final answer: Chittenden Locks and Lake Washington Ship Canal

GPT-5-Think: I need to figure out which National Register Districts overlap with both the 'Ballard' and 'Interbay Dravus' subareas within the Ballard-Interbay MIC. I've got the map, and I think the overlapping yellow polygons (representing districts) fall within both the Ballard (pink?) and Interbay Dravus (green?) subareas. I'll check for yellow dots (properties) and orange polygons (districts), focusing on labels in the north area, like the Ballard Avenue Historic District. The "Washington Ship Canal" National Register District seems to be the most likely answer, as it follows the canal and overlaps both subareas. Final answer: Washington Ship Canal

Claude-4-Sonnet: Looking at the map, I need to identify National Register Districts (shown in orange) that overlap with both the Ballard Industrial Lands Subarea (shown in purple) and the Interbay Dravus Industrial Lands Subarea (shown in orange/peach). After careful examination of the map, I do not see any National Register Districts that overlap with both the Ballard and Interbay Dravus Industrial Lands Subareas within the Ballard-Interbay MIC.Final answer: None"

---