MIND THE SPATIAL GAP: A BENCHMARK FOR GEOSPATIAL REASONING AND CONSTRAINED SPATIAL EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and manipulating spatial relations is fundamental to AI systems that reason about physical environments and construct grounded world models. However, existing evaluations of multimodal large language models (MLLMs) for spatial reasoning rely predominantly on static question-answering (QA) over small-scale or synthetic scenes, rarely probing whether models can modify spatial configurations while satisfying explicit geometric constraints. We introduce GeoSpatialBench, a benchmark designed to evaluate an escalating progression of spatial cognition, ranging from spatial relation understanding and geometric reasoning to constrained spatial editing. Grounded in real-world urban data across six geographically diverse cities, GeoSpatialBench evaluates a comprehensive taxonomy of 28 topological, directional, metric, and ordering relations across structured GeoJSON and map image modalities, with ground-truth references and constraints verified via GIS operations. Our empirical evaluation across frontier MLLMs uncovers a stark spatial gap: structured GeoJSON consistently outperforms map images, with multimodal combinations offering no reliable gain over GeoJSON alone, and while models show reasonable competence on qualitative topological and ordering, performance drops on quantitative metric relations. Crucially, editing performance on ordering relations declines relative to QA, and most failed edits violate at least one specified spatial constraint. These results highlight a gap between passively recognizing spatial relations and actively manipulating geometric configurations. GeoSpatialBench provides an open, rigorous diagnostic testbed to benchmark these spatial limitations, laying essential empirical groundwork for physically grounded world models and embodied spatial AI.

1 INTRODUCTION

The ability to understand spatial relations and use them to manipulate spatial configurations is fundamental to AI systems that reason about physical environments and construct world models (Ha & Schmidhuber, 2018; Assran et al., 2025). Consider planning a road across a region: it needs to avoid mountains and rivers, cross a densely populated district, remain connected, and meet a distance limit. Such reasoning underlies navigation, autonomous driving, spatial planning, embodied interaction, and world models (Zhou et al., 2025; Hong et al., 2026). Yet existing work on world models, embodied AI, and multimodal spatial reasoning is primarily evaluated in confined indoor or local scenes and over short trajectories (Cheng et al., 2024; Yang et al., 2025; Hong et al., 2026). Geospatial reasoning at city and regional scales introduces a different challenge: models must reason over heterogeneous geographic features and their topological, directional, metric, and ordering relations across larger spatial extents and multiple scales.

Recent multimodal large language models (MLLMs) have achieved impressive results on visual/geometric benchmarks. The proprietary GPT-6 Astra scores 95.9% with tools on BenchCAD, which evaluates programmatic CAD reasoning (Zhang et al., 2026; OpenAI, 2026). The open-weight Kimi K3 scores 81.6% without tools on MMMU-Pro, a benchmark of multidisciplinary visual understanding and reasoning (Yue et al., 2025; Kimi Team, 2026). These results demonstrate strong general spatial and visual reasoning capabilities but provide limited evidence about geospatial reasoning over real-world geographic configurations. Recent studies have begun to explore this

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

problem using different forms of geospatial representation, including maps (Dihan et al., 2025; Ung et al., 2025), imagery (Xu et al., 2026), vector geometry (Bao et al., 2026), and structured urban data (Ji et al., 2025). Across these settings, evaluations cover spatial relation families such as topological, directional, metric and ordering. However, the extent to which models can consistently recognize these relations and use them to modify spatial configurations remains unclear.

Most existing benchmarks use question-answering (QA) to assess whether models can understand and apply the underlying spatial relations. A correct answer, however, may not show whether that relation can guide a change to the scene (Vafa et al., 2024). More broadly, benchmark answers need not reflect coherent use of the underlying concepts (Mancoridis et al., 2025). Constrained spatial editing provides a complementary solution: object placements, geometric constructions, and urban layouts can be directly verified against explicit constraints (Abdelreheem et al., 2025; Huang et al., 2025b; Kim et al., 2026; Liu et al., 2026). Yet, existing studies usually evaluate editing separately from QA, rather than testing whether models can both recognize and apply the same spatial relations.

We introduce GeoSpatialBench, a benchmark that combines QA tasks on spatial relations with constrained spatial editing (Figure 1). By evaluating both tasks against the same set of spatial relations, GeoSpatialBench provides a stronger test of whether models can apply the relations they identify. It covers 28 formally defined relations in four families (topological, directional, metric, and ordering), drawing on established Geographic Information Science (GIS) and spatial-database concepts (Egenhofer & Herring, 1991; Open Geospatial Consortium, 2011). The benchmark contains 8,400 instances derived from real OpenStreetMap data across six cities. QA and editing use separate instance pools but share the same set of relations. Each scene is available as GeoJSON or as a map rendered from the same geometry. Our results show that GeoJSON generally supports stronger geospatial reasoning, whereas map images remain substantially weaker. This gap highlights an underexplored challenge in using maps as a visual medium for precise geospatial reasoning.

Our contributions are threefold. First, we introduce GeoSpatialBench, a comprehensive geospatial relation benchmark covering 28 spatial relations recognized in spatial database and GIS, with instances drawn from six cities worldwide. Second, we propose spatial editing with QA tasks to verify models' genuine spatial reasoning abilities. Third, we systematically compare leading models, including GPT-6 Astra, across input modalities using GIS-derived answers and deterministic checks of edited geometry. Together, these contributions enable more accurate measurement of spatial-relation understanding and reveal where current models still need improvement. Section 2 reviews related work, Section 3 describes the benchmark and evaluation protocol, Section 4 presents the results and analyses, and Section 5 concludes.

2 Related Work

2.1 Spatial relations and spatial reasoning of AI systems

Spatial relations describe how points, lines, and regions are arranged and relate to one another. GIS and qualitative spatial reasoning formalize a broad range of such relations and provide geometric predicates for evaluating them (Egenhofer & Franzosa, 1991; Frank, 1992; Cohn & Renz, 2008; Open Geospatial Consortium, 2011). Recent work has extended spatial reasoning to MLLMs. SpatialRGPT uses region and depth information to reason about relative direction and distance, SpatialReasoner incorporates explicit 3D representations, and SpatialScore evaluates a wider range of visual spatial abilities (Cheng et al., 2024; Ma et al., 2025; Wu et al., 2026). Related work on world models uses learned representations for prediction and planning, while VSI-Bench and ESI-Bench examine spatial recall and evidence gathering through action (Ha & Schmidhuber, 2018; Assran et al., 2025; Zhou et al., 2025; Yang et al., 2025; Hong et al., 2026). Together, these studies establish the importance of spatial knowledge across visual and embodied settings, but their evaluations largely center on local scenes with limited spatial extent. It remains less explored whether current MLLMs can reason over a broad set of formal spatial relations among georeferenced geographic features, derive geometric quantities from their configurations, and use those relations to modify spatial configurations.

2.2 BENCHMARKS FOR GEOSPATIAL REASONING IN MLLMS

Recent benchmarks have extended spatial reasoning to larger-scale geographic settings and real-world geospatial data. MapEval and CartoMapQA assess map reading and route reasoning, while FRIEDA evaluates multi-step reasoning over map images (Dihan et al., 2025; Ung et al., 2025; Pyo et al., 2026). EarthSpatialBench combines Earth imagery with explicit geographic coordinates to evaluate distance, direction, and topology, while UrbanGeoEval provides raw city-scale geometry for spatial inference (Xu et al., 2026; Bao et al., 2026). Ji et al. (2025) further study topological inference from WKT and relation-conditioned geometry generation. These studies demonstrate that geospatial reasoning can be evaluated using real geographic data and a range of representations, including map images, imagery, and structured or vector geometry. However, existing benchmarks differ substantially in their representations, relation sets, and task formulations, making it difficult to compare how models reason across spatial modalities and relation families. In particular, few systematically compare GeoJSON with a map image rendered from the same underlying geometry while evaluating a broad and formally defined set of spatial relations.

2.3 SPATIAL EDITING AND RELATION-GROUNDED EVALUATION

Most existing geospatial benchmarks evaluate models via QA tasks, testing their ability to recognize spatial relationships and compute geometric quantities. However, obtaining correct answers does not prove that a model can leverage these relationships to modify spatial configurations. Benchmark success can mask gaps in a model's underlying understanding (Vafa et al., 2024; Mancoridis et al., 2025). Spatial editing provides a complementary test because the resulting configuration can be checked against explicit spatial constraints. Recent work has explored language-guided object placement in 3D scenes, executable CAD editing, and iterative modification of urban GeoJSON (Abdelreheem et al., 2025; Zhang et al., 2026; Liu et al., 2026). These studies show the feasibility of evaluating whether models can translate spatial instructions into valid geometric changes. However, spatial editing has generally been studied separately from geospatial QA, rather than evaluating whether models can recognize and then apply the same spatial relations.

GeoSpatialBench brings the two together, scoring each task by relation family on separate instance pools and verifying answers and edits with GIS operations. This design provides a direct test of whether such spatial reasoning ability transfers to editing geographic geometry so that the models actually understand the underlying spatial relations.

3 GEOSPATIALBENCH

GeoSpatialBench evaluates spatial reasoning through spatial question answering and constrained geometric editing. Figure 1 summarizes the benchmark construction, model inputs, and evaluation procedures.

3.1 TASK DEFINITION

GeoSpatialBench covers 28 geospatial relations over points, lines, and polygons at the city scale. These relations are categorized into four groups: topology, direction, metric, and order. Instances of these relations include touching, relative direction, distance, and connectivity. Each relation is defined by a precise geometric or attribute-based rule. For instance, two roads are considered connected only if they share an endpoint, while vertical order is ascertained by road-layer attributes.

In QA, models are tasked with providing an answer based on a specific scene and an associated question. Depending on the types of relations involved, the answer may take the form of a category, a number, a feature ID, or an ordering. Each QA instance is presented in both open-ended and four-option multiple-choice formats.

Spatial editing requires the model to modify the scene instead of describing it. Given an initial scene, an instruction, and a target feature F_0, the model is expected to make edits exclusively to the specified feature F_0, while leaving other features unchanged. Editing accuracy is assessed by whether the scene resulting from the model's edit satisfies the requested spatial constraints. The aim of the editing task is to evaluate whether the models truly comprehend spatial relations.

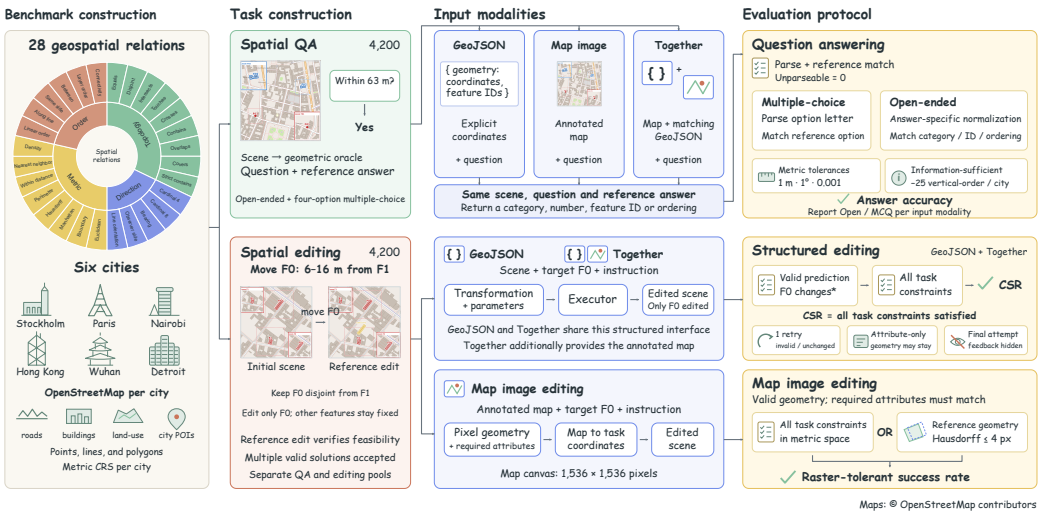

Figure 1: Overview of GeoSpatialBench. QA and spatial editing are constructed from urban vector data across six cities and 28 spatial relations. Both tasks support GeoJSON, image, and combined inputs. QA is evaluated by answer accuracy. GeoJSON and Together editing use the Constraint Satisfaction Rate (CSR), while map-image editing uses a raster-tolerant success rate.

3.2 Benchmark Construction

The benchmark covers six cities with diverse geographic contexts: Stockholm, Paris, Hong Kong, Wuhan, Detroit, and Nairobi. Each city contributes 700 QA and 700 editing instances, giving 8,400 canonical instances in total.

We use OpenStreetMap data for roads, buildings, land-use regions, and city-level points of interest (POIs). All features are projected to a city-specific metric coordinate system before distance and area calculations. We then filter them by geometry type and city-specific length or area thresholds. Road-layer attributes are retained for vertical-order questions. QA and editing use separate feature pools, so their instances are not paired.

For each geospatial relation, we construct a canonical scene from the filtered features. The scene records the required geometries, roles, source attributes, and construction provenance. It may use observed geometries or controlled constructions when the source data do not provide enough cases. A geometric oracle evaluates the scene and produces the reference answer. Source identities and construction records are kept for auditing but excluded from model inputs.

Each canonical scene is serialized as a GeoJSON FeatureCollection and also rendered as an annotated map. GeoJSON preserves the task-coordinate geometry and uses anonymous feature identifiers, with only task-required properties. For the map, the same scene is converted to WGS84 and rendered in a square Web Mercator view. Task features are highlighted and labeled, with a legend, detail insets, and attribute cards added when needed. All images use a $1{,}536 \times 1{,}536$ pixel canvas.

QA questions are generated from the oracle output. Open-ended questions ask the model to return the answer in the required form. Multiple-choice distractors depend on the relation. Among all, directional relations use nearby bearings, ordering relations use alternative sequences, and numerical relations use adjacent intervals. Topological relations and feature-selection tasks use plausible competing labels or feature IDs. Each item is checked before use. We verify the geometry and reference answer, then check that the options are distinct.

Editing instances are constructed separately from QA. Candidates are rejected if the spatial constraints specified by the instruction are already satisfied in the initial scene, and retained only when a reference edit confirms that a valid solution exists. The reference edit is used only to verify feasi-

bility; any valid edit that satisfies the instruction is accepted. Geometric edits must change the target feature F_0, while attribute edits may leave its geometry unchanged.

3.3 Input Modalities

We evaluate three input modalities: GeoJSON, map image, and their combination (Together). GeoJSON provides explicit coordinates and anonymous feature identifiers, while map image shows the same scene as an annotated map. Together pairs the annotated map with the same GeoJSON used in the GeoJSON condition. For QA, all three conditions use the same scene, question, and reference answer, with both open-ended and four-option multiple-choice forms. The answer format remains unchanged across conditions, allowing us to compare the effect of input representation on spatial reasoning.

For editing, GeoJSON uses a structured interface. Models receive the scene, target feature F_0, and instruction, and return a transformation and its parameters (e.g., translation, rotation, scaling, or attribute editing). An executor applies the transformation to F_0, and the evaluator checks the resulting scene against the required spatial constraints. Map image editing instead returns edited geometry in pixel coordinates with any required attributes. The evaluator maps the prediction back to the task coordinate system before checking the same constraints. All images use a $1{,}536 \times 1{,}536$ pixel canvas.

Together follows the same structured editing interface as GeoJSON while additionally providing the corresponding annotated map. This allows a direct comparison with GeoJSON to test whether visual information improves structured editing. By contrast, map image editing changes both the input representation and the output interface because the model must generate pixel-coordinate geometry. Differences between GeoJSON and map image editing should therefore not be interpreted as a pure input-only ablation.

3.4 Evaluation Protocol

Question answering. We use deterministic scoring rather than a learned judge. Multiple-choice outputs are parsed as option letters, while open-ended outputs use answer-specific normalization. Distance and perimeter answers allow a one-meter tolerance, bearing answers a one-degree circular tolerance, and density answers a tolerance of 0.001 in the stated units. Unparseable answers are counted as incorrect, and open-ended and multiple-choice accuracy are reported separately.

GeoJSON, map image, and Together use the same canonical scene, question, and reference answer. GeoJSON exposes the exact task-coordinate geometry, map image shows its annotated rendering, and Together provides both. For geometry-only conditions, we also report an information-sufficient score that removes the 25 vertical-order questions per city whose required layer attributes are not provided. Annotated maps may provide these attributes through attribute cards, but raster measurements still limit visual precision.

Structured editing. GeoJSON and Together share the same structured editing protocol. Given an initial scene, target feature F_0, and instruction, the model returns a transformation and its parameters. Let S_i and S_i' denote the initial and edited scenes, and let $\mathcal{C}_i$ be the task constraints. Let V_i indicate a parseable, nonempty, valid prediction. Let U_i indicate that a geometric edit changes F_0; for attribute-editing tasks, we set $U_i = 1$. We define

$$E_i = V_i U_i \prod_{c \in \mathcal{C}_i} \mathbf{1}[c(S_i, S_i')], \qquad \mathrm{CSR} = \frac{1}{N} \sum_{i=1}^{N} E_i. \tag{1}$$

Here, E_i is a binary indicator of whether edit i is valid and satisfies all task constraints, and Constraint Satisfaction Rate (CSR) is the proportion of successful edits over the evaluation set. With one corrective retry, an undecodable or unchanged edit triggers a second attempt requesting a valid, nonzero edit. We score only the final attempt; the feedback does not reveal whether the task constraints were satisfied.

Map image editing. The evaluator first maps the predicted pixel geometry back to the task coordinate system. It then accepts either metric-space constraint satisfaction or pixel geometry within

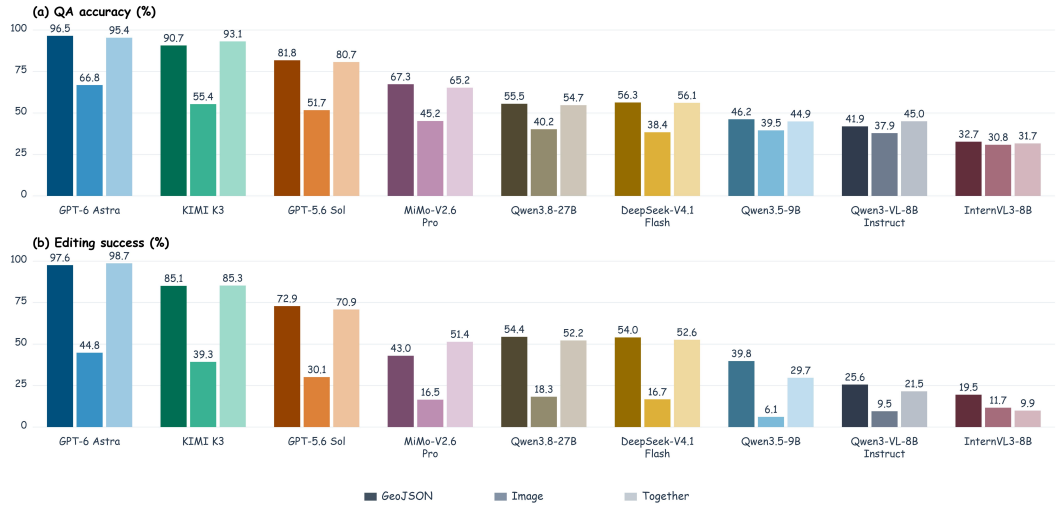

Figure 2: Average QA and spatial editing performance across models and input modalities. (a) QA accuracy averaged over open-ended and multiple-choice settings. (b) Spatial editing success. GeoJSON and Together editing use strict CSR, whereas Image editing uses raster-tolerant success.

four pixels of the reference under Hausdorff distance, provided that required attributes match. We report this raster-tolerant success rate separately from strict CSR. The reference geometry is used for feasibility verification and image-level scoring.

We aggregate scores by task, input modality, and QA form, and report results by city, relation family, and individual relation. Comparisons across GeoJSON, map image, and Together use matched scenes and instructions whenever the model supports the corresponding input modality.

4 Evaluation on GeoSpatialBench

4.1 Evaluation Setup

We evaluate nine models across the six cities, including four open-weight models (Qwen3.5-9B, Qwen3.8-27B, Qwen3-VL-8B-Instruct, and InternVL3-8B) and five API-accessed models (GPT-6 Astra, GPT-5.6 Sol, KIMI K3, MiMo-V2.6-Pro, and DeepSeek-V4.1-Flash).

For question answering, each model is evaluated in both open-ended and four-option multiple-choice settings. We consider three input settings: GeoJSON, map image, and Together, where Together combines the GeoJSON representation with the corresponding annotated map. Spatial editing is evaluated under the same three input settings using the task-specific output interfaces described in Section 3. We also report performance separately for open-ended QA, multiple-choice QA, and spatial editing. Additional model- and setting-specific results are provided in Appendix A.

4.2 Overall Performance

Figure 2 summarizes overall QA and editing performance across models and input modalities. QA accuracy is averaged over the open-ended and multiple-choice settings. Under GeoJSON, GPT-6 Astra achieves the strongest overall performance, with 96.5% QA accuracy and 97.6% editing success. KIMI K3 follows at 90.7% and 85.1%, respectively. Performance then declines across the remaining models, although the relative order is not identical between QA and editing.

QA and editing performance are broadly aligned under GeoJSON. Models with higher QA accuracy generally also achieve higher editing success, but several exceptions remain. MiMo-V2.6-Pro reaches 67.3% QA accuracy, above Qwen3.8-27B at 55.5%, whereas Qwen3.8-27B performs better

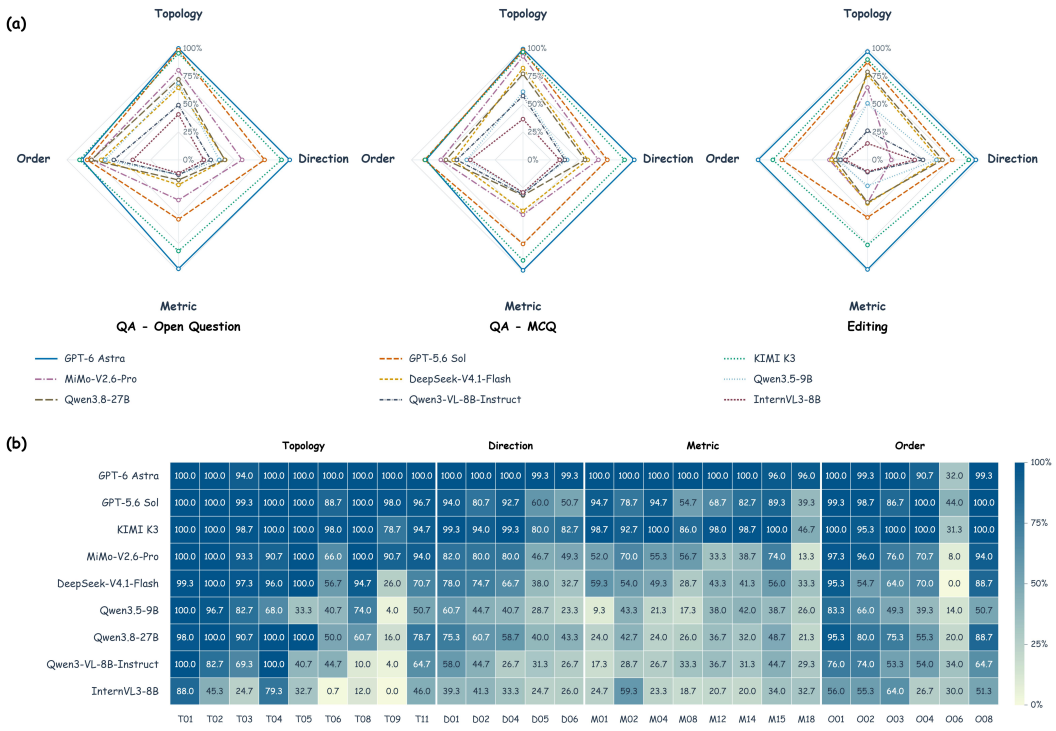

Figure 3: Performance across spatial relations under GeoJSON input. (a) Mean performance across the four relation families for open-ended QA, multiple-choice QA, and spatial editing. (b) Relation-level multiple-choice QA accuracy for each model across the 28 spatial relations. Metric relations are generally more challenging, while QA and spatial editing show different patterns across relation families.

on editing (54.4% vs. 43.0%). Thus, strong relation understanding does not always translate into equally strong constrained editing.

Performance also differs across input modalities. Map image scores are generally lower than GeoJSON scores for both QA and spatial editing. For MiMo-V2.6-Pro, mean QA accuracy decreases from 67.3% with GeoJSON to 45.2% with map image, while Together reaches 65.2%. In spatial editing, the same model achieves 43.0% CSR with GeoJSON and 51.4% with Together, whereas its raster-tolerant success rate with map image is 16.5%. Together therefore remains close to GeoJSON in QA, but its effect on spatial editing varies across models. Because map image editing uses a different output interface and evaluation protocol, its score is not directly comparable to strict CSR.

4.3 Performance across Relations and Cities

As shown in Figure 3(a), performance varies substantially across the four relation families under GeoJSON. In open-ended QA, topology and ordering achieve 74.4% and 71.9% accuracy, followed by direction at 55.3% and metric at 38.6%. Multiple-choice QA shows a similar pattern, with 77.9% for topology, 69.4% for ordering, 62.0% for direction, and 53.4% for metric. Spatial editing shows a different distribution. Topology and direction achieve the highest CSR at 65.1% and 63.4%, whereas ordering falls to 47.3% and metric remains the lowest at 42.9%.

Figure 3(b) further shows substantial variation among individual relations in multiple-choice QA. Relation codes used in the figure are listed in Appendix B, Table 4. Averaged across models, Equality (T01) achieves 98.4% accuracy and Touching (T04) reaches 92.7%, whereas Boundary Coverage (T09) drops to 46.4%. Metric relations are generally more difficult: POI Density (M18) achieves 37.6%, and Hausdorff Distance (M08) reaches 46.8%. Ordering relations also vary widely, from 89.2% for Along-Line Order (O01) to 23.7% for Layer Hierarchy (O06). Layer Hierarchy is

attribute-dependent, and the required road-layer attribute is omitted from the geometry-only GeoJSON QA input. Its raw accuracy therefore reflects information availability in addition to geometric reasoning difficulty.

Performance varies relatively little across the six evaluated cities. For open-ended GeoJSON QA, the within-model range across cities is 1.00–5.71 percentage points. By comparison, pooled accuracy differs by 67.03 percentage points between the highest- and lowest-performing models. The observed variation across models is therefore substantially larger than the within-model variation across these six cities. These results do not establish generalization to unseen cities.

4.4 Error Analysis and Discussion

We further examine whether spatial-relation recognition transfers to spatial intervention, which relations remain difficult, and how representation affects spatial reasoning. In GeoSpatialBench, this capability is operationalized through spatial editing. Figures 4 and 5 show representative QA and editing cases, while complete relation-level statistics are provided in Appendix B.

The representative cases reveal distinct sources of error across tasks and representations. In QA, some relations remain stable across inputs, while precise metric estimates are more fragile. In the Touching case, DeepSeek-V4.1-Flash is correct with GeoJSON but predicts *intersects* from the map image. For Boundary Distance, KIMI K3 gives an accurate open-ended value with GeoJSON but fails under the map image, while still selecting the correct multiple-choice interval. Spatial editing introduces a different class of errors. In the Bearing case, KIMI K3 succeeds under both inputs, whereas InternVL3-8B moves the target in the wrong direction with GeoJSON. In the Euclidean Distance case, GPT-6 Astra succeeds under both inputs, while KIMI K3 succeeds with GeoJSON but produces a 42.8 m distance from the map image, outside the required 8–22 m range.

Relation recognition can approach saturation before spatial intervention. Several basic topological relations are already recognized reliably in GeoJSON MCQ-4: Equality, Touching, and Disjointness have error rates of only 1.6%, 7.3%, and 8.4%. However, this does not consistently transfer to editing. Touching and Disjointness increase to 71.2% and 47.3% editing error, whereas Equality remains easy at 3.3%. Thus, recognizing a relation and constructing a geometry that satisfies it are distinct capabilities. Figure 5 also illustrates this gap: models may choose a plausible operation but still produce an incorrect direction, distance, or final geometry. Editing therefore remains discriminative even when some recognition tasks approach saturation. The gap is also relation-dependent. For direction relations, GeoJSON editing reaches 63.4% CSR, close to the 62.0% MCQ-4 accuracy.

Metric reasoning remains the most persistent bottleneck. Metric relations remain the most difficult family under both GeoJSON and map-image inputs. Even with explicit GeoJSON geometry, open-ended metric error reaches 61.4%. Several numerical relations are particularly difficult: Minimum Distance, Hausdorff Distance, Total Perimeter, and POI Density exceed 70% error in open-ended QA. Figure 4 shows a related pattern. Models can sometimes identify the correct distance interval while failing to produce an accurate numerical value. These results point to a persistent weakness in quantitative geometric reasoning. Access to explicit coordinates reduces representational ambiguity, but does not by itself ensure reliable computation of continuous spatial quantities.

Explicit geometry better supports precise spatial reasoning. For QA, GeoJSON and map-image conditions use the same scene, question, and reference answer, enabling a direct comparison of representations. GeoJSON has lower family-level error across all four relation families, with the largest gap in metric reasoning. Figure 4 shows the same tendency: coarse directional or categorical judgments may remain correct from maps, while precise boundary and distance estimates are less reliable. Together inputs also do not consistently improve over GeoJSON alone, suggesting that additional visual context does not automatically strengthen an explicit geometric representation. For editing, the GeoJSON–image gap should be interpreted cautiously because image editing also uses a different output interface.

Toward explicit spatial representations. These patterns suggest that precise spatial reasoning may require representations that preserve geometric and relational structure more explicitly. Metric reasoning depends on continuous quantities, while rasterization requires exact geometry to be re-

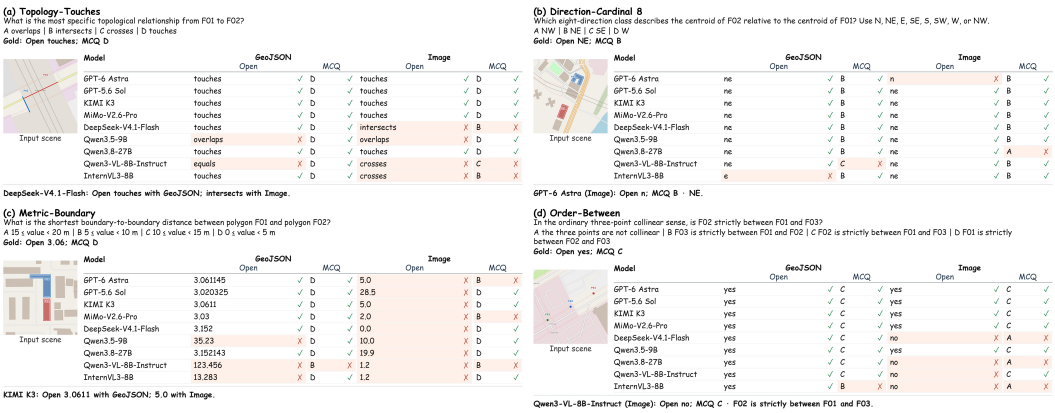

Figure 4: Representative QA cases across the four spatial relation families. Map images often preserve coarse spatial judgments but are less reliable for precise boundary and metric reasoning.



Figure 5: Representative spatial editing cases across four relation families, illustrating failures in geometric execution and constraint satisfaction.

covered indirectly from a finite-resolution image. Our experiments do not isolate a specific architectural cause, but they suggest that additional multimodal visual information alone may be insufficient. Spatial tokenization methods such as SPOK (Huang et al., 2025a) provide one possible direction by representing geographic space and spatial relations explicitly. Such representations may help bridge spatial recognition, geometric reasoning, and spatial intervention.

5 Conclusion

We introduced GeoSpatialBench, a benchmark for evaluating spatial relation understanding, geometric computation, and constrained spatial editing in urban geographic scenes using structured GeoJSON and rendered map images. Across nine models, GeoJSON generally yields stronger performance than map images, while combining the two inputs does not consistently improve performance. Precise metric computation and boundary relations remain challenging, and constraint violations account for most editing failures. By evaluating spatial reasoning and editing within a common, automatically verifiable framework, GeoSpatialBench shows that recognizing a spatial relation does not necessarily mean a model can construct geometry that satisfies it. A promising next step is to test whether dedicated, relation-aware representations can help bridge this gap. Difficulty with basic 2D geometric constraints also highlights the challenge of more complex spatial generation and provides a baseline for future work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the six OpenStreetMap source cities, instance counts, GeoJSON and map-image construction, task formats, and GIS-based procedures for generating answers and checking edited geometry. It also defines the parsing rules, numerical tolerances, corrective retry, and separate raster-tolerant criterion for image editing. The evaluation code and supporting benchmark materials will be provided by the authors upon request.

AI USE STATEMENT

AI-based tools were used to assist with language editing and LaTex formatting. The authors used OpenAI's ChatGPT (5.6 Sol) to assist with code development, data processing, and language editing of the manuscript. The authors reviewed and tested the AI-assisted code, checked the data-processing outputs, and reviewed and revised the AI-edited text. The authors take full responsibility for the accuracy and integrity of the methods, results, and final manuscript.

REFERENCES

Ahmed Abdelreheem, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Abdelrahman Eldesokey, Peter Wonka, Gabriel Brostow, Sara Vicente, and Guillermo Garcia-Hernando. PlaceIt3D: Language-guided object placement in real 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6645–6655, 2025. doi: 10.1109/ICCV51701.2025.00626. URL https://openaccess.thecvf.com/content/ICCV2025/html/Abdelreheem_PlaceIt3D_Language-Guided_Object_Placement_in_Real_3D_Scenes_ICCV_2025_paper.html.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985. Preprint.

Mutian Bao, Qiuyi Qi, Tian Liang, Jinjian Zhang, Wei Zhou, Ming Kong, Linjian Mo, and Qiang Zhu. UrbanGeoEval: A city-scale benchmark for evaluating large language models in geospatial reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40183–40223, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1867. URL https://aclanthology.org/2026.acl-long.1867/.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024. doi: 10.52202/079017-4293. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Anthony G. Cohn and Jochen Renz. Qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter (eds.), *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pp. 551–596. Elsevier, 2008. doi: 10.1016/S1574-6526(07)03013-1. URL https://doi.org/10.1016/S1574-6526(07)03013-1.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13774–13813. PMLR, 2025. URL https://proceedings.mlr.press/v267/dihan25a.html.

Max J. Egenhofer and Robert D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991. doi: 10.1080/02693799108927841. URL https://doi.org/10.1080/02693799108927841.

Max J. Egenhofer and John R. Herring. Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical report, Department of Surveying Engineering, University of Maine, Orono, Maine, 1991. URL https://escholarship.org/uc/item/5nj6647c. Reprinted in NCGIA Technical Report 94-1.

Andrew U. Frank. Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages & Computing*, 3(4):343–371, 1992. doi: 10.1016/1045-926X(92)90007-9. URL https://doi.org/10.1016/1045-926X(92)90007-9.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Fei-Fei Li, Jiajun Wu, and Yejin Choi. ESI-Bench: Towards embodied spatial intelligence that closes the perception-action loop, 2026. URL https://arxiv.org/abs/2605.18746.

Fei Huang, Jian-Rong Lv, Guo-Long Li, and Yang Yue. Spok: Tokenizing geographic space for enhanced spatial reasoning in geoai. *International Journal of Geographical Information Science*, 39(12):2768–2808, 2025a. doi: 10.1080/13658816.2025.2497810.

Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. Fireplace: Geometric refinements of llm common sense reasoning for 3d object placement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13466–13476, 2025b.

Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. Foundation models for geospatial reasoning: Assessing the capabilities of large language models in understanding geometries and topological spatial relations. *International Journal of Geographical Information Science*, 39(9):1866–1903, 2025. doi: 10.1080/13658816.2025.2511227. URL https://doi.org/10.1080/13658816.2025.2511227.

Jinwoong Kim, Rui Yang, and Huishuai Zhang. GeoBuildBench: A benchmark for interactive and executable geometry construction from natural language, 2026. URL https://arxiv.org/abs/2605.13167.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Rui Liu, Steven Jige Quan, Zhong-Ren Peng, Zijun Yao, Han Wang, Pengyang Wang, Kunpeng Liu, Yanjie Fu, and Dongjie Wang. City editing: Hierarchical agentic execution for dependency-aware urban geospatial modification, 2026. URL https://arxiv.org/abs/2602.19326. Preprint; accepted by ACM SIGSPATIAL 2026.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140751–140774, 2025. doi: 10.52202/085713-4703. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cdd63c6c813a6ec826051b6a49f11b5e-Abstract-Conference.html.

Marina Mancoridis, Bec Weeks, Keyon Vafa, and Sendhil Mullainathan. Potemkin understanding in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42857–42881. PMLR, 2025. URL https://proceedings.mlr.press/v267/mancoridis25a.html.

Open Geospatial Consortium. Opengis implementation standard for geographic information – simple feature access – part 1: Common architecture. OpenGIS Implementation Standard OGC 06-103r4, Open Geospatial Consortium, 2011. URL https://www.ogc.org/standards/sfa/. Version 1.2.1.

OpenAI. GPT-6 Astra: A new generation of intelligence. OpenAI research release, 2026. URL https://openai.com/index/gpt-6-astra/.

Jiyoon Pyo, Yuankun Jiao, Dongwon Jung, Zekun Li, Leeje Jang, Sofia Kirsanova, Jina Kim, Yijun Lin, Qin Liu, Junyi Xie, Hadi Askari, Nan Xu, Muhao Chen, and Yao-Yi Chiang. FRIEDA: Benchmarking multi-step cartographic reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QQCadccQqU.

Huy Quang Ung, Guillaume Habault, Yasutaka Nishimura, Hao Niu, Roberto Legaspi, Tomoki Oya, Ryoichi Kojima, Masato Taya, Chihiro Ono, Atsunori Minamikawa, and Yan Liu. CartoMapQA: A fundamental benchmark dataset evaluating vision-language models on cartographic map understanding. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 440–453. ACM, 2025. doi: 10.1145/3748636.3762755. URL https://doi.org/10.1145/3748636.3762755.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26941–26975, 2024. doi: 10.52202/079017-0846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31029–31041, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_SpatialScore_Towards_Comprehensive_Evaluation_for_Spatial_Intelligence_CVPR_2026_paper.html.

Zelin Xu, Yupu Zhang, Saugat Adhikari, Saiful Islam, Tingsong Xiao, Zibo Liu, Shigang Chen, Da Yan, and Zhe Jiang. EarthSpatialBench: Benchmarking spatial reasoning capabilities of multimodal LLMs on earth imagery. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mR9YFoMHBK.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Haozhe Zhang, Kaichen Liu, Miaomiao Chen, Lei Li, Shaojie Yang, Cheng Peng, and Hanjie Chen. BenchCAD: A comprehensive, industry-standard benchmark for programmatic CAD, 2026. URL https://arxiv.org/abs/2605.10865. Preprint; supplied version v2.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

A Additional Results

Table 1 jointly reports GeoJSON, Image, and Together performance for all nine models, ordered by six-city GeoJSON open-ended QA. Table 2 shows the six-city variation in GeoJSON open-ended QA. The differences between cities range from 1.00% to 5.71%.

Table 1: Overall performance (%; higher is better), ordered by GeoJSON open-ended QA. Each model uses 4,200 instances across six cities. Image editing reports raster-tolerant success; GeoJSON and Together editing report strict CSR. Bold marks the highest reported score in each column.

	GeoJSON			Map image			Together		
Model	Open	MCQ-4	Edit	Open	MCQ-4	Edit	Open	MCQ-4	Edit
GPT-6 Astra	**96.36**	**96.64**	**97.60**	**60.57**	**73.12**	**44.81**	**94.86**	**95.86**	**98.71**
KIMI K3	89.52	91.88	85.10	51.43	59.29	39.29	92.86	93.43	85.29
GPT-5.6 Sol	78.12	85.43	72.90	47.40	56.02	30.12	77.14	84.29	70.86
MiMo-V2.6-Pro	62.98	71.71	42.98	39.19	51.14	16.50	59.55	70.93	51.38
Qwen3.8-27B	52.45	58.64	54.38	38.43	42.00	18.29	53.52	55.93	52.19
DeepSeek-V4.1-Flash	49.50	63.17	54.02	33.69	43.10	16.74	52.79	59.48	52.60
Qwen3.5-9B	46.50	45.95	39.79	35.52	43.55	6.14	42.62	47.21	29.71
Qwen3-VL-8B-Instruct	37.19	46.69	25.62	34.86	40.90	9.55	42.48	47.57	21.55
InternVL3-8B	29.33	36.07	19.50	27.55	34.14	11.67	28.98	34.36	9.93

Table 2: Open-ended GeoJSON QA across six cities (%; higher is better), ordered by pooled accuracy. Each city has 700 questions per model. Overall pools 4,200 questions; Range is the maximum minus minimum city score in percentage points.

Model	Stockholm	Paris	Nairobi	Hong Kong	Wuhan	Detroit	Overall	Range
GPT-6 Astra	95.57	96.14	96.14	96.71	96.86	96.71	96.36	1.29
KIMI K3	89.29	90.00	89.86	89.43	89.00	89.57	89.52	1.00
GPT-5.6 Sol	77.86	79.57	79.57	77.57	76.57	77.57	78.12	3.00
MiMo-V2.6-Pro	62.57	64.14	62.14	64.43	61.43	63.14	62.98	3.00
Qwen3.8-27B	51.14	53.14	53.00	52.43	51.71	53.29	52.45	2.14
DeepSeek-V4.1-Flash	48.43	49.00	49.14	48.71	49.86	51.86	49.50	3.43
Qwen3.5-9B	44.14	48.00	46.57	47.00	46.86	46.43	46.50	3.86
Qwen3-VL-8B-Instruct	37.00	36.71	38.43	37.00	38.43	35.57	37.19	2.86
InternVL3-8B	26.14	31.86	31.00	29.00	30.00	28.00	29.33	5.71

B Additional Error Analysis

B.1 Distribution by Relation

Figure 6 and Tables 3 and 4 summarize relation-level error rates for the nine reported models.

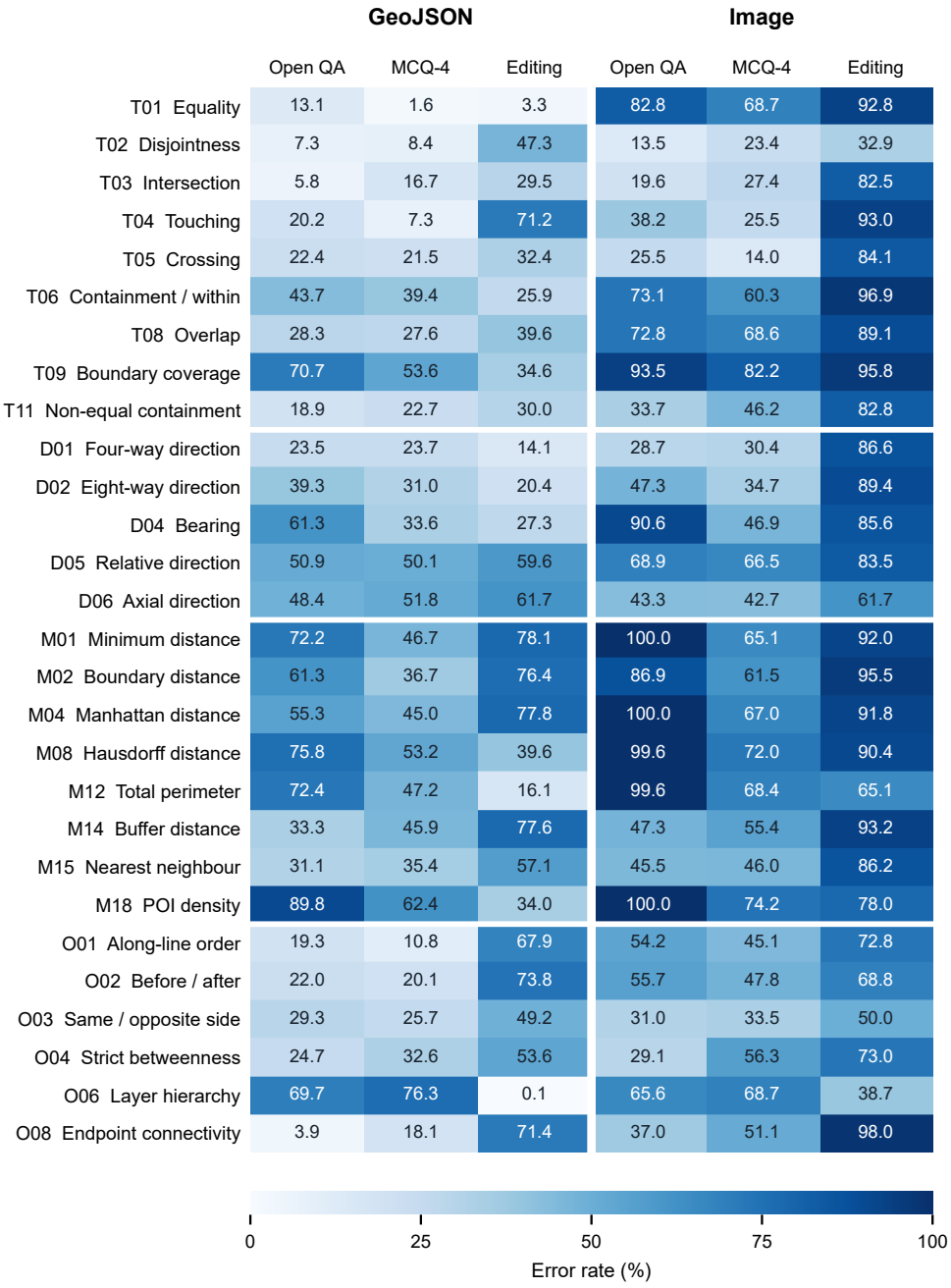

Figure 6: Error rates by spatial relation, input, and task (%; lower is better).

Table 3: Error rates by relation family (%; lower is better), pooled over the nine reported models. GeoJSON and Map Image use six cities for every model.

	GeoJSON			Map image		
Family	Open QA	MCQ-4	Editing	Open QA	MCQ-4	Editing
Topology	25.61	22.09	34.86	50.30	46.27	83.32
Direction	44.68	38.04	36.64	55.76	44.21	81.37
Metric	61.40	46.57	57.09	84.87	63.69	86.53
Ordering	28.14	30.59	52.67	45.43	50.44	66.87

Table 4: Error rates for the 28 spatial relations (%; lower is better).

			GeoJSON			Map image		
ID	Relation	Family	Open QA	MCQ-4	Editing	Open QA	MCQ-4	Editing
T01	Equality	Topology	13.11	1.63	3.26	82.78	68.73	92.82
T02	Disjointness	Topology	7.33	8.37	47.26	13.47	23.43	32.90
T03	Intersection	Topology	5.78	16.67	29.48	19.59	27.43	82.53
T04	Touching	Topology	20.22	7.33	71.19	38.20	25.47	92.98
T05	Crossing	Topology	22.44	21.48	32.44	25.47	14.04	84.08
T06	Containment / within	Topology	43.70	39.41	25.93	73.14	60.33	96.90
T08	Overlap	Topology	28.30	27.63	39.63	72.82	68.57	89.06
T09	Boundary coverage	Topology	70.74	53.63	34.59	93.55	82.20	95.84
T11	Non-equal containment	Topology	18.89	22.67	30.00	33.71	46.20	82.78
D01	Four-way direction	Direction	23.48	23.70	14.07	28.73	30.37	86.61
D02	Eight-way direction	Direction	39.26	31.04	20.44	47.27	34.69	89.39
D04	Bearing	Direction	61.33	33.56	27.33	90.61	46.86	85.63
D05	Relative direction	Direction	50.89	50.15	59.63	68.90	66.45	83.51
D06	Axial direction	Direction	48.44	51.78	61.70	43.27	42.69	61.71
M01	Minimum distance	Metric	72.15	46.67	78.07	100.00	65.06	92.00
M02	Boundary distance	Metric	61.33	36.74	76.37	86.94	61.47	95.51
M04	Manhattan distance	Metric	55.26	45.04	77.85	100.00	67.02	91.84
M08	Hausdorff distance	Metric	75.78	53.19	39.63	99.59	72.00	90.37
M12	Total perimeter	Metric	72.44	47.19	16.07	99.59	68.41	65.06
M14	Buffer distance	Metric	33.26	45.93	77.63	47.27	55.43	93.22
M15	Nearest neighbour	Metric	31.11	35.41	57.11	45.55	45.96	86.20
M18	POI density	Metric	89.85	62.44	34.00	100.00	74.20	78.04
O01	Along-line order	Ordering	19.33	10.81	67.93	54.20	45.06	72.82
O02	Before / after	Ordering	22.00	20.07	73.85	55.67	47.84	68.82
O03	Same / opposite side	Ordering	29.26	25.70	49.19	31.02	33.55	49.96
O04	Strict betweenness	Ordering	24.67	32.59	53.56	29.06	56.33	72.98
O06	Layer hierarchy	Ordering	69.70	76.30	0.07	65.63	68.73	38.69
O08	Endpoint connectivity	Ordering	3.85	18.07	71.41	36.98	51.10	97.96