SPATIALGRAPH-EVAL: INTERVENTION-BASED EVALUATION OF SPATIAL-NETWORK REASONING IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal language models increasingly perform well on spatial and visual reasoning benchmarks, but aggregate accuracy does not show how model performance changes when the same problem is presented differently or when its underlying structure is modified. We introduce SPATIALGRAPH-EVAL, a diagnostic benchmark for spatial-network reasoning that separates graph structure from its presentation and enables targeted interventions on both. Graphs are procedurally generated with exact shortest-path solutions and controlled variation in path length and decision complexity, while graph size and mean branching are matched within path-length conditions. We evaluate three contemporary multimodal model systems on identical graph structures presented symbolically and visually, under topology-preserving rotations and reflections, and following minimal topology-changing counterfactuals. Representation sensitivity differed substantially across models, with optimal-route accuracy decreasing by 3.9 percentage points for Terra (68.1% to 64.2%), 42.8 for Claude (85.6% to 42.8%), and 17.2 for Gemini (100.0% to 82.8%) when comparing visual with symbolic presentation. Despite unchanged topology, optimal-route accuracy varied substantially across rotations and reflections, with significant heterogeneity across models (Model $\times$ Transform, $p = 1.83 \times 10^{-11}$). Among instances solved optimally before intervention, models found the new shortest route in only 2.0–7.1% of cases after a one-edge counterfactual rewiring changed the correct route. Failure analysis showed that degradation included graph-invalid transitions as well as valid but suboptimal routes. As an external validity check, we evaluated 50 subgraphs sampled from the San Francisco OpenStreetMap road network; across models, optimal-route accuracy decreased from 90–100% under symbolic presentation to 2–16% under visual presentation. These results reveal substantial model-dependent behavioral sensitivity to representation, topology-preserving presentation changes, and minimal structural interventions, motivating intervention-based evaluation of spatial-network reasoning beyond aggregate task accuracy.

1 INTRODUCTION

Spatial reasoning is central to navigation, planning, scene understanding, and interaction with structured environments. Controlled diagnostic benchmarks such as CLEVR have shown the value of isolating reasoning factors rather than relying only on aggregate visual-question-answering accuracy (Johnson et al., 2017). Recent multimodal benchmarks extend spatial evaluation from atomic and compositional capabilities to hierarchical spatial cognition and interactive tasks (Gong et al., 2026; Xu et al., 2025; Gao et al., 2026). In parallel, graph-reasoning benchmarks show that language-model performance can depend strongly on graph complexity, graph encoding, and graph structure (Wang et al., 2023; Fatemi et al., 2024). Yet a high aggregate benchmark score does not establish that a model behaves consistently when task-relevant structure is held fixed. A model may solve a graph when connectivity is written explicitly but fail when the identical graph is rendered visually; it may change its answer after a rotation that preserves topology; or it may fail to update after a small structural change that alters the correct route. These behaviors are difficult to isolate when perception, representation, topology, and task difficulty vary together.

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.

We introduce SpatialGraph-Eval, a controlled diagnostic benchmark designed around interventions on spatial-network problems. Each task is an undirected graph with a designated start node S and goal node G. The objective is to return a valid shortest route. Because the underlying graph is generated and stored explicitly, the benchmark provides exact algorithmic solutions and allows the same task to be rendered in multiple forms while preserving its structural identity.

The comparisons use matched graph instances. We change representation or image orientation while keeping connectivity fixed, then change one edge while retaining selected graph properties and requiring a different optimal route. Together, these interventions distinguish invariance when the graph problem is unchanged from adaptation when its structure changes. Aggregate task accuracy alone does not establish either property.

Our design separates four questions. First, how does performance change as shortest-path length and decision complexity vary under a size-matched procedural design? Second, how much does behavior change when the identical graph is presented symbolically rather than visually? Third, are model outputs stable under rotations and reflections that alter presentation but not topology? Fourth, can models adapt when a minimal one-edge rewiring changes the set of optimal routes while preserving other graph properties? We additionally evaluate 50 subgraphs sampled from the San Francisco OpenStreetMap road network as an external-validity check.

On the same graphs, optimal-route accuracy changed from 68.1% to 64.2% for Terra, 85.6% to 42.8% for Claude, and 100.0% to 82.8% for Gemini when the input changed from symbolic to visual. Rotating or reflecting the images also changed performance, although the graph itself was unchanged. After one-edge rewiring, only 2.0–7.1% of previously optimal cases produced an optimal route. The 50 OpenStreetMap subgraphs also showed a symbolic-to-visual gap.

The benchmark combines exact graph-based scoring with paired changes to input format, image orientation, and connectivity. We also classify invalid responses and test whether the representation gap occurs on sampled road-network graphs.

2 Related Work

2.1 Diagnostic visual and spatial reasoning benchmarks

Diagnostic synthetic benchmarks have long been useful for separating reasoning capabilities from dataset biases. CLEVR introduced controlled synthetic scenes and functional annotations to diagnose compositional visual reasoning rather than relying only on aggregate visual-question-answering accuracy (Johnson et al., 2017). Recent spatial benchmarks broaden this agenda for multimodal models. SpaCE-10 evaluates spatial intelligence from atomic to compositional capabilities (Gong et al., 2026), while SpatialBench organizes spatial cognition into progressively complex levels spanning observation, topology, symbolic reasoning, causality, and planning (Xu et al., 2025). SpatialWorld moves toward ecological validity through interactive tasks (Gao et al., 2026), while SFI-Bench evaluates structured spatial and functional reasoning over egocentric indoor video (Zhang et al., 2026). Lost in Aggregation decomposes maze navigation across local, junction-level, and global scales and reports strong dependence on input format, including a large advantage for structured coordinate text over rendered images (Jiang et al., 2026).

In these prior studies, graph encoding and task difficulty can vary together. Our representation and orientation comparisons use the same underlying graph. A separate intervention changes one edge while preserving selected graph properties and forcing the optimal route to change.

2.2 Graph reasoning with language models

Graph reasoning benchmarks have shown that language-model performance depends on both task complexity and graph encoding. NLGraph contains natural-language graph problems ranging from connectivity and shortest path to maximum flow and other algorithmic tasks, and reports brittleness as graph problems become more complex (Wang et al., 2023). Talk Like a Graph and its GraphQA benchmark show that textual graph encoding and graph structure can substantially change LLM performance (Fatemi et al., 2024). GraCoRe broadens graph comprehension and reasoning evaluation across graph types and reports sensitivity to factors including node ordering (Yuan et al., 2025).

These studies motivate treating representation itself as an experimental factor rather than as a neutral interface.

These findings make graph encoding an experimental factor in our study. We pair symbolic and visual responses for the same graph, test transformations that preserve its edges, and use rewiring to test responses when the correct solution changes. Each response is scored against the graph.

3 SPATIALGRAPH-EVAL

3.1 TASK DEFINITION AND SCORING

Each instance is an undirected graph G=(V,E) with designated start S and goal G. A model returns an ordered route through visible node labels. The ground-truth shortest-path length L* is computed algorithmically using NetworkX. A predicted route is valid if it starts at S, ends at G, and every consecutive node pair is an edge in E. It is optimal if it is valid and its number of traversed edges equals L*. When multiple shortest routes exist, any valid minimum-length route is accepted.

Optimal-route accuracy is the primary outcome. Valid-route accuracy is secondary and distinguishes failures of connectivity from valid but inefficient routing. For diagnostic analysis, non-optimal outputs are further classified into valid non-optimal routes, parse failures, wrong-start or wrong-goal responses, and invalid-edge transitions.

3.2 PROCEDURAL BENCHMARK AND CONTROLLED COMPLEXITY

The frozen main benchmark contains 360 held-out graphs. Shortest-path length is varied over L $\in \{16,18,20,22\}$, and decision complexity over D $\in \{2,4,6\}$, with 30 instances in each L$\times$D cell. Each cell contains 15 aligned and 15 misleading geometry instances. Branching is fixed at B=2.

We operationalize decision complexity D as the number of real decision patterns among six fixed pattern slots. Real decision patterns alter viable choice structure, whereas filler patterns preserve the graph-size budget. Consequently, within each path-length condition, D varies while total node count, total edge count, and mean branching remain matched. Specifically, N=L+13 and E=L+18, yielding N=$\{29,31,33,35\}$ and E=$\{34,36,38,40\}$ for L=$\{16,18,20,22\}$. This permits a controlled comparison of decision complexity without mechanically increasing graph size. Appendix Table 2 summarizes the size-matched design.

3.3 HELD-OUT CONSTRUCTION

The 360-task evaluation set uses unique held-out seeds disjoint from development and calibration seeds. Evaluation instances were fixed before model evaluation and held constant across all compared systems. Exact seeds, manifests, and generation scripts are intended for release with the benchmark artifacts.

3.4 REPRESENTATION INTERVENTION

For every main-benchmark graph, we construct two matched inputs. The symbolic condition explicitly serializes graph connectivity, whereas the visual condition renders the same nodes, edges, start, and goal as a spatial network. The underlying graph, task identity, and shortest-path solution are unchanged. This paired design measures the behavioral effect of representation while holding graph structure fixed.

3.5 TOPOLOGY-PRESERVING TRANSFORMATIONS

The invariance experiment evaluates 72 base graphs under six visual transforms: original, 90° rotation, 180° rotation, 270° rotation, horizontal reflection, and vertical reflection. Graph connectivity and the shortest-route problem are unchanged. This yields 432 visual conditions per model and 1,296 evaluations across the three model systems.

3.6 Topology-changing counterfactuals

The counterfactual experiment contains 144 original/counterfactual graph pairs. A counterfactual is created by minimally rewiring one edge while preserving node identities, node positions, visible labels, graph size, shortest-path length, decision complexity, and mean branching. Original and counterfactual graphs have disjoint optimal-path sets, ensuring that the required optimal route changes solely because connectivity changes. Across two conditions and three models, this experiment contains 864 evaluations.

3.7 External validation on OpenStreetMap topology

To test whether the representation effect extends beyond procedurally generated topology, we sampled 50 drivable-road subgraphs from the San Francisco OpenStreetMap network using OSMnx (Boeing, 2017; OpenStreetMap contributors, 2026). Each subgraph was rendered from geographic coordinates and evaluated in matched symbolic and visual form. We use undirected topology and unweighted hop-count shortest paths to preserve the benchmark routing objective. Any valid route with length equal to the algorithmic shortest length is scored optimal. The experiment contains 300 evaluations (50 tasks $\times$ 2 representations $\times$ 3 models). It is intended as a small external-validity check, not as an evaluation of real-world navigation.

The OSM task set spans shortest-path lengths $\{8,10,12,14\}$; subgraphs contain 12–35 nodes.

4 Experimental Setup

4.1 Model systems

We evaluate three contemporary general-purpose multimodal model systems from distinct model families: Terra (gpt-5.6-terra), Claude (claude-sonnet-5), and Gemini (gemini-3.8-flash). All systems receive the same frozen benchmark instances and task instructions within each experimental condition. Because proprietary systems differ in architecture, inference procedures, and token accounting, comparisons are interpreted as behavioral comparisons under standardized benchmark conditions rather than controlled comparisons of model scale.

4.2 Prompting and response parsing

The evaluation uses a standardized JSON-only response protocol and visible node labels. The prompt and parser were frozen before held-out evaluation. Routes are parsed into a canonical sequence and scored against the frozen graph edges. Tolerant route parsing is used for the primary route score, while strict JSON compliance is reported separately. Malformed outputs are retained as parse failures rather than silently converted into successful routes.

4.3 Statistical analysis

For the main benchmark, the primary explanatory analysis models optimal-route success as Model $\times$ Representation $\times$ L plus Model $\times$ Representation $\times$ D, treating L and D categorically and using task-clustered standard errors. The targeted Model $\times$ Representation test evaluates heterogeneity in representation dependence across systems. Three-way interactions assess whether representation effects vary differently across models as L or D changes.

For invariance, a joint Model $\times$ Transform Wald test evaluates whether transformation sensitivity differs across systems. For counterfactuals, paired original/counterfactual comparisons are summarized with percentage-point changes, task-level bootstrap confidence intervals, exact paired tests, and a joint Model $\times$ Condition test. For OSM, symbolic and visual outcomes are paired by subgraph; we report paired bootstrap intervals and exact McNemar tests. Sensitivity analyses exclude confirmed output-limit cases where that status was recorded.

Table 1: Main benchmark performance on identical graph structures under symbolic and visual presentation.

Model	Symbolic optimal	Visual optimal	Δ optimal	Symbolic valid	Visual valid
Terra	68.1%	64.2%	-3.9 pp	96.4%	82.5%
Claude	85.6%	42.8%	-42.8 pp	90.3%	58.3%
Gemini	100.0%	82.8%	-17.2 pp	100.0%	88.6%

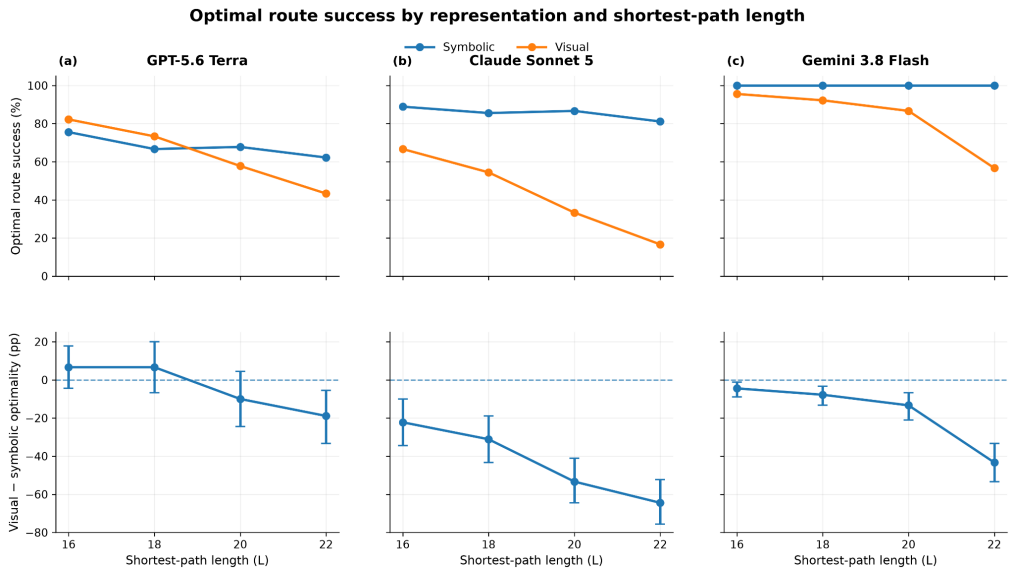


Figure 1: Optimal-route performance across path length for matched symbolic and visual graph representations on the frozen main benchmark.

5 Results

5.1 Representation sensitivity on identical graphs

Table 1 summarizes performance on identical graphs under symbolic and visual presentation. Across the 360 held-out graphs, representation sensitivity differed sharply by model. Terra achieved 68.1% optimal-route accuracy symbolically and 64.2% visually, a -3.9 percentage-point gap. Claude decreased from 85.6% to 42.8% (-42.8 points), while Gemini decreased from 100.0% to 82.8% (-17.2 points). Valid-route accuracy also declined from symbolic to visual presentation for all three systems.

The joint Model $\times$ Representation test gave $p < 0.001$, indicating that the magnitude of representation dependence differed across model systems. Under the Terra reference parameterization, the Representation $\times$ L term was significant ($p = 0.017$); the Representation $\times$ D term was not ($p = 0.916$). The Model $\times$ Representation $\times$ L interaction was not significant ($p = 0.175$), so we do not claim that the path-length dependence of the representation gap differs significantly across models. Likewise, Model $\times$ Representation $\times$ D was not significant ($p = 0.290$).

Figure 1 shows optimal-route accuracy across shortest-path lengths. Descriptively, visual performance worsened with increasing path length in all three systems. A task-bootstrap comparison of the visual-symbolic gap from L=16 to L=22 was -25.56 points for Terra (95% CI [-43.33,-7.78]), -42.22 for Claude ([-58.89,-25.56]), and -38.89 for Gemini ([-50.00,-27.78]). These model-specific descriptive patterns should be interpreted alongside the non-significant three-way interaction.

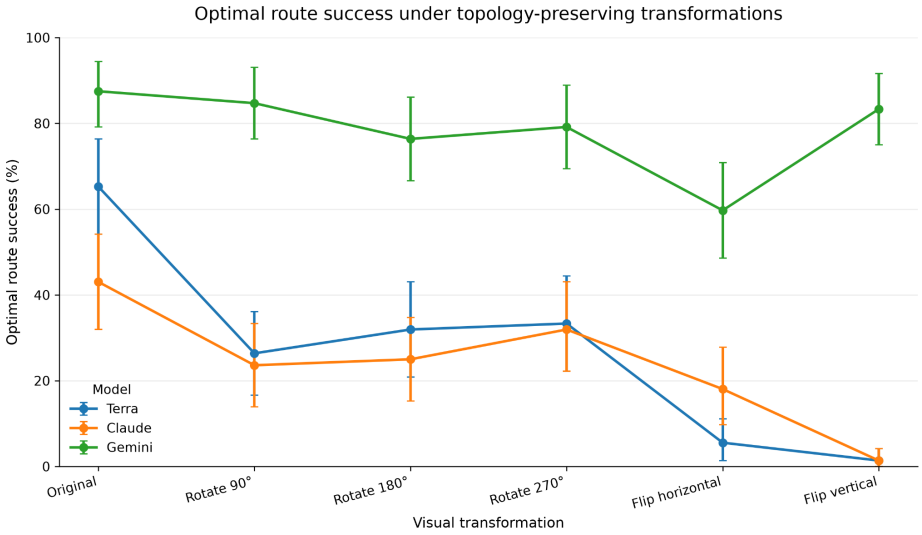


Figure 2: Transformation-specific optimal-route accuracy in the frozen invariance experiment.

5.2 Sensitivity to topology-preserving transformations

Figure 2 shows that optimal-route accuracy varied substantially after rotations and reflections even though graph topology and the correct shortest-route problem were unchanged. Transformation-specific values are reported in Appendix Table 3. The joint Model $\times$ Transform interaction was significant ($F(10, 71) = 11.53$, $p < 0.001$), showing that transformation sensitivity differed across systems.

Terra and Claude showed particularly large degradation under reflections. Gemini was comparatively stable to rotations and vertical reflection but remained sensitive to horizontal reflection. These results measure behavioral invariance, not internal reasoning processes: a topology-preserving change that should leave the graph problem equivalent can nonetheless change the probability of an optimal response.

The paired outcomes show what aggregate accuracy misses. Of the 47 graphs Terra solved optimally in the original orientation, only 4 remained optimal after horizontal reflection and 1 after vertical reflection. For Gemini under 90° rotation, 7 originally optimal graphs became non-optimal while 5 originally non-optimal graphs became optimal. Thus, the small aggregate change from 63 to 61 successes across the 72 graphs includes 12 instances whose optimality status changed. Appendix Table 7 reports conditional optimality for each model and transformation.

5.3 Adaptation to minimal topology-changing counterfactuals

Figure 3 shows that counterfactual rewiring caused large declines in optimal-route performance. Terra decreased from 27.78% on original graphs to 1.39% after rewiring (-26.39 points, 95% bootstrap CI [-34.03,-18.75], exact paired $p < 0.001$). Claude decreased from 9.72% to 0.69% (-9.03 points, [-13.89,-4.86], $p < 0.001$), and Gemini from 35.42% to 0.69% (-34.72 points, [-42.36,-27.08], $p < 0.001$). The joint Model $\times$ Condition test was significant (Wald $\chi^2(2)$=45.95, $p < 0.001$).

Appendix Table 4 reports the paired results and conditional retention. Because baseline performance differs across models, the conditional retention analysis is particularly informative. Among instances solved optimally before intervention, only one instance per model produced an optimal route after rewiring: 2.5% for Terra, 7.1% for Claude, and 2.0% for Gemini. Thus, among instances solved optimally before intervention, 92.9–98.0% were no longer solved optimally after a single edge changed the required route. Comparing parsed routes within these originally optimal pairs, the counterfactual response exactly repeated the original route in 11 of 40 Terra cases (27.5%), 4 of 14 Claude cases (28.6%), and 17 of 51 Gemini cases (33.3%). Other responses changed the route

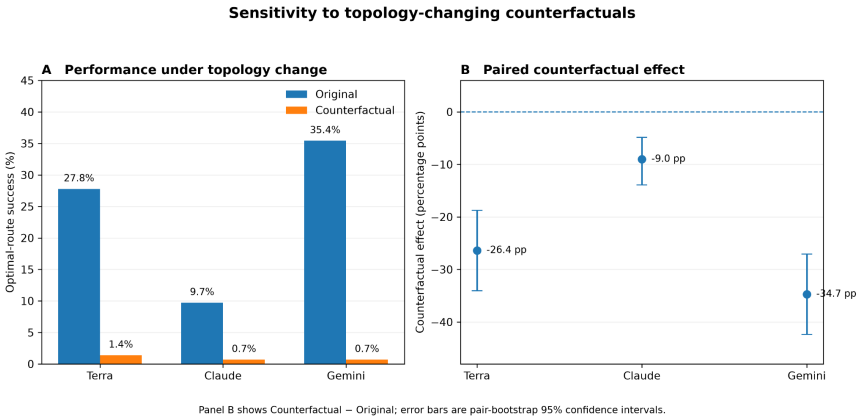


Figure 3: Optimal-route performance before and after minimal topology-changing counterfactual rewiring.

or could not be parsed. Low counterfactual accuracy therefore does not imply that models simply repeated their previous answers.

5.4 Failure modes

Appendix Table 5 shows that performance degradation was not limited to valid but suboptimal routing. Visual presentation increased graph-invalid transitions, particularly invalid-edge moves, and the failure distribution differed across model systems. For example, Claude's invalid-edge rate increased from 0.56% symbolically to 23.33% visually, Gemini's from 0% to 10.00%, and Terra's from 3.33% to 17.22%. Claude additionally showed a substantial visual parse-failure rate (18.33%).

Validity was independently reconstructed from saved routes and graph edges and matched the stored validity field for all checked rows. We do not attribute these failures to a specific internal mechanism because the present analyses do not identify one.

5.5 External validity on San Francisco road-network topology

Appendix Table 6 summarizes the external-validation results. A large symbolic-to-visual decrease also occurred on the sampled road-network topology. Terra decreased from 90.0% optimal symbolically to 16.0% visually (-74.0 points, 95% paired bootstrap CI [-86,-62], exact McNemar $p < 0.001$). Claude decreased from 96.0% to 2.0% (-94.0 points, [-100,-86], $p < 0.001$), and Gemini from 100.0% to 8.0% (-92.0 points, [-98,-84], $p < 0.001$).

For optimality, every discordant matched pair favored symbolic presentation: 37–0 for Terra, 47–0 for Claude, and 46–0 for Gemini (symbolic success/visual failure versus symbolic failure/visual success). Validity also collapsed under visual presentation. The same direction of difference appears in this small sample of road-network subgraphs; these results do not establish real-world navigation ability.

6 Discussion

The same graph often produced different results in symbolic and visual form. Results also changed when visual graphs were rotated or reflected, even though connectivity stayed fixed. Gemini's 90° rotation illustrates how losses and gains on different graphs can partly cancel in aggregate accuracy. Finally, changing one edge substantially reduced accuracy on the counterfactual tasks.

The profiles also differed across systems. The small average symbolic-to-visual gap for Terra on the main benchmark coexisted with strong transformation sensitivity; Claude showed a larger representation gap with visual invalid-edge and parse failures; and Gemini's high symbolic accuracy coexisted with visual, reflection, and counterfactual degradation. The OSM experiment showed a

large symbolic-to-visual decrease on sampled road-network topology. Because these graphs differ from the procedural graphs in geometry, topology, and path-length range, the size of the gap need not match the main-benchmark result. The small OSM sample does not establish performance on practical navigation.

7 Limitations

The benchmark uses abstract spatial networks to isolate the effects of representation and controlled graph changes. Its findings do not establish performance in photorealistic or embodied navigation. Symbolic and visual inputs encode the same topology, while the different interfaces make visual parsing part of the measured representation effect. The generator matches selected graph properties; other structural features may also influence difficulty. The interventions identify changes in behavior, but cannot attribute them to model architecture or training because the evaluated systems are proprietary. The road-network check covers 50 simplified, undirected, unweighted subgraphs from one city. Results may also change as deployed model APIs are updated.

8 Conclusion

Across three multimodal systems, paired interventions revealed changes in behavior that aggregate accuracy could obscure. Rotations and reflections tested invariance while preserving the graph problem, and one-edge rewiring tested adaptation when the required optimal route changed. We also observed a large symbolic-to-visual gap on 50 OpenStreetMap subgraphs. Spatial-reasoning evaluations should assess invariance and adaptation alongside overall accuracy.

9 Reproducibility Statement

Evaluation sets were fixed before model evaluation and held constant across systems. The intended release includes the procedural generator, exact evaluation manifests and seeds, rendering code, prompts, parser and scoring code, quality-control tests, statistical-analysis scripts, and result tables, subject to model-provider terms.

10 AI Use Statement

Large language models (LLMs) assisted with literature discovery, coding, debugging, code generation and refinement for benchmark construction and analysis, synthetic-data workflow development, and drafting and editing portions of the manuscript. The author determined the research questions and experimental decisions, executed the experiments, inspected and validated benchmark instances and outputs, verified the reported results, and takes responsibility for the paper.

References

Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017. doi: 10.1016/j.compenvurbsys.2017.05.004.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*, 2024.

Hongcheng Gao, Hailong Qu, Jingyi Tang, Jiahao Wang, Zihao Huang, Hengkang Qiao, Shihong Huang, Junming Yang, Yi Li, Hongyixuan Yuan, Wenjie Li, Bohan Zeng, Wenbo Li, Bo Wang, Jianhui Liu, Olive Huang, Haoyang Huang, Wentao Zhang, Guoqing Huang, Nan Duan, and Yinpeng Dong. SpatialWorld: Benchmarking interactive spatial reasoning of multimodal agents in real-world tasks. *arXiv preprint arXiv:2606.09669*, 2026.

Ziyang Gong, Wenhao Li, Oliver Ma, Songyuan Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models

in compositional spatial intelligence. In *International Conference on Learning Representations*, 2026.

Yuhan Jiang, Peng Luo, and Liqiu Meng. Lost in aggregation: A multi-scale diagnostic benchmark for LLM spatial navigation. *arXiv preprint arXiv:2606.22219*, 2026.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

OpenStreetMap contributors. OpenStreetMap. `https://www.openstreetmap.org/`, 2026.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

Zike Yuan, Ming Liu, Hui Wang, and Bing Qin. GraCoRe: Benchmarking graph comprehension and complex reasoning in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7925–7948. Association for Computational Linguistics, 2025.

Le Zhang, Jihan Yang, Soundarya Krishnan, Jimit Majmudar, Xiou Ge, Prasoon Puri, Prathamesh Saraf, Shruti Bhargava, Dhivya Piraviperumal, Yinan Ling, Cindy Pan, Hong Yu, Aishwarya Agrawal, and Bo-Hsiang Tseng. From where things are to what they are for: Benchmarking spatial-functional intelligence in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12052–12063, 2026.

A Additional Experimental and Analysis Details

A.1 Benchmark construction and provenance

The main evaluation set contains 360 unique held-out graphs with seeds disjoint from development and calibration. Evaluation sets for the main, transformation, counterfactual, and OSM experiments were fixed before model evaluation. Exact manifests, seeds, hashes, and generation scripts are intended for release as reproducibility artifacts rather than reported in the main manuscript.

The invariance stimuli were made from the frozen main-benchmark images. The original image was copied unchanged. Rotations by 90, 180, and 270 degrees and horizontal and vertical reflections used Pillow's `Image.Transpose` operations on the image pixels; labels were transformed with the rest of the image. No coordinate-based rerendering or interpolation was applied during these transformations.

A.2 Size-matched complexity design

Table 2: Size-matched complexity design.

Path length L	Nodes N	Edges E	Decision complexity D
16	29	34	2, 4, 6
18	31	36	2, 4, 6
20	33	38	2, 4, 6
22	35	40	2, 4, 6

A.3 Detailed intervention results

Table 3: Optimal-route accuracy under topology-preserving visual transformations.

Model	Original	Rot. 90	Rot. 180	Rot. 270	Flip H	Flip V
Terra	65.3%	26.4%	31.9%	33.3%	5.6%	1.4%
Claude	43.1%	23.6%	25.0%	31.9%	18.1%	1.4%
Gemini	87.5%	84.7%	76.4%	79.2%	59.7%	83.3%

Table 4: Counterfactual performance and success conditional on optimality before intervention.

Model	Original opt.	CF opt.	Δ	Originally opt.	CF success	Cond. CF success
Terra	27.78%	1.39%	-26.39 pp	40	1	2.5%
Claude	9.72%	0.69%	-9.03 pp	14	1	7.1%
Gemini	35.42%	0.69%	-34.72 pp	51	1	2.0%

Table 5: Focused failure taxonomy on the main benchmark.

Model / repr.	Optimal	Valid non-opt.	Parse	Wrong start	Wrong goal	Invalid edge
Claude sym.	85.56%	4.72%	9.17%	0%	0%	0.56%
Claude vis.	42.78%	15.56%	18.33%	0%	0%	23.33%
Gemini sym.	100.00%	0%	0%	0%	0%	0%
Gemini vis.	82.78%	5.83%	1.11%	0%	0.28%	10.00%
Terra sym.	68.06%	28.33%	0%	0.28%	0%	3.33%
Terra vis.	64.17%	18.33%	0%	0.28%	0%	17.22%

Table 6: External-validation performance on 50 San Francisco OpenStreetMap subgraphs.

Model	Symbolic optimal	Visual optimal	Δ optimal	Symbolic valid	Visual valid
Terra	90.0%	16.0%	-74.0 pp	96.0%	20.0%
Claude	96.0%	2.0%	-94.0 pp	100.0%	2.0%
Gemini	100.0%	8.0%	-92.0 pp	100.0%	10.0%

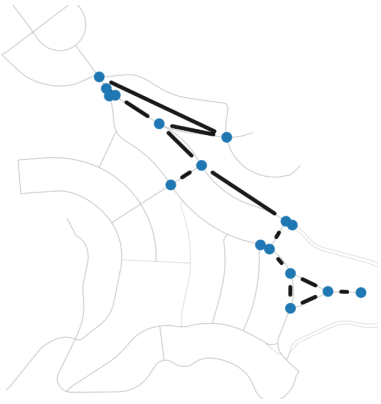
Table 7: Optimal-route success after each visual transformation, conditional on success for the same graph in the original orientation. Each cell gives successes divided by originally optimal graphs, with the conditional percentage.

Transform	Terra ($n = 47$)	Claude ($n = 31$)	Gemini ($n = 63$)
Rotate 90°	12/47 (25.5%)	10/31 (32.3%)	56/63 (88.9%)
Rotate 180°	18/47 (38.3%)	10/31 (32.3%)	51/63 (81.0%)
Rotate 270°	16/47 (34.0%)	12/31 (38.7%)	50/63 (79.4%)
Flip horizontal	4/47 (8.5%)	6/31 (19.4%)	41/63 (65.1%)
Flip vertical	1/47 (2.1%)	1/31 (3.2%)	54/63 (85.7%)

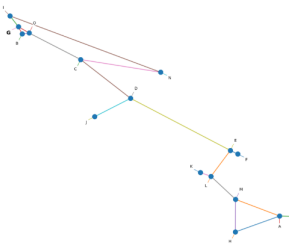
A.4 OSM QUALITATIVE EXAMPLES

Figure 4 shows illustrative OSM tasks; the aggregate results in Appendix Table 6 use all 50 subgraphs.

A. OSM road context (osm_sf_011)

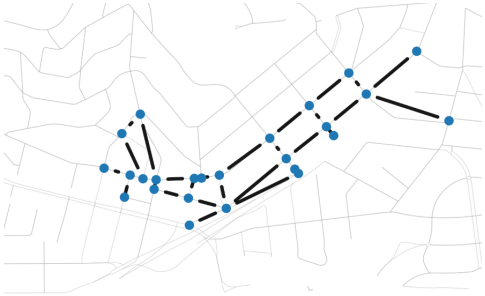


Evaluated graph stimulus

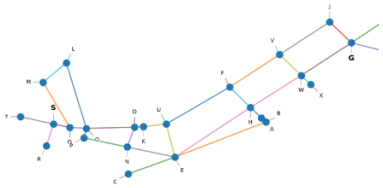


Optimal routes: symbolic—all three models; visual—Claude and Terra.

B. OSM road context (osm_sf_001)



Evaluated graph stimulus



Optimal routes: symbolic—all three models; visual—none.

Figure 4: Illustrative OSM external-validation examples. Geographic context is shown for visualization only; the evaluated graph stimulus is the rendered subgraph. Aggregate results use all 50 tasks.