MAPSPATIAL: A BENCHMARK FOR REPRESENTATION-ROBUST SPATIAL REASONING OVER REAL MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable map use requires stable answers across views and orientations, and revised answers when roads or buildings change; single-image accuracy cannot measure both. We introduce **MapSpatial**, a geographic information system (GIS)-grounded benchmark of 77,800 instances spanning 14 sub-tasks in four families: qualitative direction, metric comparison, route-relative building identification and counting, and road-network path reasoning. Because every scene is stored as GIS geometry, each answer can be recomputed exactly when the same question is posed across three registered views (satellite imagery, a text-free structural map, and a conventional text-bearing map), under rotations and mirror flips, and, where applicable, after solver-verified structural edits paired with answer-preserving shams. We evaluate 25 multimodal models. Among six representative models, direct accuracy is 35.2–48.3%, but 33.8–60.7% of registered-view triplets yield different semantic predictions across views that share an answer but differ in visual cues. After a single declared geometric action, 21.1–34.3% of base-correct answers become wrong; the stricter base-plus-seven joint score is 11.2–22.8%. After an answer-changing edit, 9.6–26.2% of base-correct pairs update correctly, while 50.4–84.3% retain the pre-edit answer. Gemini-3-Flash correctly updates 50.4% of its base-correct edit pairs. Controlled visual aids show task-dependent effects. MapSpatial quantifies cross-view disagreement, transform sensitivity, and limited correct updating through paired, GIS-verifiable views and edits. Anonymous code and dataset access: `https://anonymous.4open.science/r/anonymous-code-8k4m2q7-AF2B`.

1 INTRODUCTION

Maps occupy an unusual position in multimodal reasoning. They are visual artifacts, but their meaning is not determined by appearance alone. Satellite, structural, and conventional cartographic encodings can depict the same geographic structure, and the displayed frame can be rotated or reflected by a known geometric action. Conversely, a visually small edit—closing one road or introducing one barrier—can change connectivity and the correct route. Spatial intelligence over maps therefore requires more than recognizing map pixels: a model must identify the spatial relationships relevant to the question and use them consistently across different map representations and controlled changes.

Current spatial benchmarks reveal substantial weaknesses in relative position, distance estimation, object recognition and counting in top-view imagery, and planning (Fu et al., 2024; Chen et al., 2024; Kamath et al., 2023). Top-view and Earth-imagery benchmarks extend spatial evaluation to overhead scenes, geometric relations, and object-level queries (Li et al., 2024; Xu et al., 2026). Cartographic benchmarks examine symbol recognition, information extraction, scale and legend interpretation, and multi-step map reasoning (Ung et al., 2025; Pyo et al., 2026; Srivastava et al., 2025). However, performance on a single map rendering cannot by itself reveal whether a model relies on appearance-specific cues rather than task-relevant spatial relations, or whether it would update its answer when the underlying map structure changes.

Maps also offer a controlled setting for studying models that both understand and generate images. Because task-relevant entities and relations are available from geographic information system (GIS)

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.

geometry, one geometrically faithful form of intermediate evidence can be rendered as a reference for model-generated evidence. This enables separate measurements of evidence selection, artifact construction, and downstream use, motivating RQ4 as a secondary diagnostic over the same map questions.

We introduce **MapSpatial** to study these questions under a common geospatial ground truth. Each case begins from a latent geographic structure containing coordinates, road links, buildings, route records, and task constraints. It is rendered into three registered views: satellite imagery, a controlled text-free structural map, and a conventional text-bearing map. The same source geometry determines the question, its answer under task-aware rotations and reflections, the reference visual evidence, and the counterfactual variants. This paired design links the same question across map views, geometric actions, and verified edits, making representation stability and correct structural updating measurable on linked cases.

Scope and research questions. MapSpatial is a benchmark rather than a training method. We organize the study around four research questions (RQs): **RQ1**, direct performance on the four task families; **RQ2**, consistency across registered views and task-aware geometric actions; **RQ3**, response to solver-verified changes in task-relevant structure; and **RQ4**, the utility and failure modes of controlled or model-generated visual evidence. All 25 models (17 understanding-only multimodal large language models (MLLMs) and 8 unified multimodal models) enter RQ1 and RQ2 per-view analyses; paired RQ2/RQ3 diagnostics use six fixed anchors and two scale references, with RQ3 restricted to valid T3/T4 edits. RQ4 is a secondary diagnostic for unified models whose interfaces support image editing.

Contributions. Our contributions are:

1. We formalize map reasoning over a latent geographic structure and render each scene as registered satellite, text-free structural, and conventional text-bearing views for the same questions.
2. We introduce 14 sub-tasks in four capability families, with task-aware geometric transforms on the two non-text-bearing views and solver-verified answer-changing edits with matched shams for eligible T3/T4 sub-tasks.
3. We add a secondary, information-stratified evidence protocol that distinguishes question-conditioned, oracle-conditioned, and model-generated visual evidence, and release recomputable labels, scene-grouped evaluation units, audits, and an executable evaluation harness.
4. We evaluate 25 models and find that direct accuracy alone misses paired behavior: among six anchors with 35.2–48.3% direct accuracy, 33.8–60.7% of registered-view triplets yield different semantic predictions, 65.7–78.9% of base-correct answers remain correct after one geometric action, and 9.6–26.2% of base-correct edit pairs update correctly. Visual evidence helps mainly when it resolves a task's critical selection step.

2 Related Work

General visual-spatial reasoning. BLINK, SpatialVLM, and What's Up expose weaknesses in relative position, metric estimation, and spatial relations (Fu et al., 2024; Chen et al., 2024; Kamath et al., 2023); OmniSpatial and SpatialBench broaden the diagnosis across spatial primitives and image domains (Jia et al., 2025; Xu et al., 2025). TopViewRS derives realistic and semantic top-view maps from shared simulated scenes (Li et al., 2024); it reports separate question collections for the two map types rather than a same-question paired-view metric. Adjacent benchmarks study embodied navigation or structured indoor layouts (Qiao et al., 2025; Rodionov et al., 2026). These benchmarks evaluate single renderings and do not pair the same question across views or controlled changes.

Consistency and counterfactual evaluation. Contrast sets and behavioral tests evaluate models on minimally changed inputs whose labels may change (Gardner et al., 2020; Ribeiro et al., 2020), and consistency-oriented VQA checks agreement across rephrased or entailed questions (Shah et al., 2019; Ray et al., 2019). MapSpatial applies this paired logic to maps, where GIS geometry and a solver generate and verify the variants: registered views that preserve the answer, task-aware rotations and reflections with recomputed answers, and answer-changing edits with matched shams.

Map, Earth-imagery, and route reasoning. Geospatial benchmarks cover Earth imagery (Danish et al., 2024; Xu et al., 2026), choropleth and general map reading (Chang et al., 2022; Ung et al., 2025; Srivastava et al., 2025; Pyo et al., 2026; Bhat et al., 2026; Arnold et al., 2026), and route reasoning over pixel, transit, or heterogeneous topological maps (Xing et al., 2025; Feng et al., 2026; Shang et al., 2026). Related systems answer directly over structured map databases or vector geometry (Li et al., 2025; Saeedan et al., 2026). MapIQ tests map-design changes (Srivastava et al., 2025); MapVerse reports input-corruption and resolution ablations (Bhat et al., 2026). These studies measure accuracy under design or image perturbations; they do not report same-instance joint correctness or task-aware answer-update metrics. MapBench provides a structured map-scene representation for indexing and evaluation (Xing et al., 2025), and MapTab pairs map images with structured route-attribute tables (Shang et al., 2026). These are useful partial analogues, but do not jointly provide question-aligned registered views, controlled variants, and solver-verified answer-changing world interventions from one latent geographic record. Table 1 summarizes the design affordances rather than incomparable raw scores.

Visual intermediates for reasoning. Prior work externalizes visual reasoning through prompting, explicit drawing interfaces, and native multimodal generation. Visualization-of-Thought prompts language models to visualize intermediate states in grid-world spatial tasks; Visual Sketchpad gives multimodal models a canvas, drawing tools, and optional specialist vision tools; and ViLaSR trains interwoven textual reasoning with elementary drawing operations (Wu et al., 2024; Hu et al., 2024; Wu et al., 2025). A complementary line develops unified models that natively interleave image and text generation, including BAGEL and ThinkMorph (Deng et al., 2025; Gu et al., 2026).

Generation-to-understanding (G2U) benchmarks and methods find strongly task- and protocol-dependent effects from self-generated visual intermediates (Shi et al., 2026; Zhou et al., 2026; Wen et al., 2026; Xie et al., 2025; Wu et al., 2026; Yang et al., 2026; Gao et al., 2026). End-to-end answer deltas can conflate choosing what to depict, rendering it faithfully, and using the resulting artifact. Most closely, VGAU-Diag stratifies six synthetic visual-planning task families by difficulty and replaces model-generated aids with externally constructed oracle aids, finding that current unified models often fail to use even faithful aids (Zhu et al., 2026). MapSpatial shares this diagnostic goal on real maps, where reference evidence is rendered from latent GIS geometry and stratified by information rather than fidelity: question-conditioned evidence marks queried entities and complete candidate sets without selecting an answer, whereas oracle-conditioned evidence selects it. RQ4 reports these conditions, generated-artifact fidelity, restart-based reading, and stateful continuation separately as a secondary diagnostic.

Table 1: **Design comparison with representative benchmarks.** ✓: documented; ✗: not reported. Paired Views: same question/answer across same-source visual maps. Paired Variants: same-question map-image changes (MapIQ design, MapVerse corruption/resolution, MapSpatial rotations/reflections). Answer Edits: world changes that alter gold answers. Structured Data: geometry, graphs, or tables used for inputs or labels. Visual G2U: generated images reused for answering.

Benchmark	**Setting**	**Paired design**			**Structured**	**Visual**
		Views	**Variants**	**Answer Edits**	**Data**	**G2U**
TopViewRS	Indoor top-view maps	✓	✗	✗	✓	✗
EarthSpatialBench	Earth imagery	✗	✗	✗	✓	✗
CartoMapQA	Cartographic maps	✗	✗	✗	✓	✗
FRIEDA	Document maps	✗	✗	✗	✗	✗
MapIQ	Thematic maps	✓	✓	✗	✓	✗
MapVerse	Real-world maps	✗	✓	✗	✗	✗
MapBench	Map-image navigation	✗	✗	✗	✓	✗
MapTab	Constrained routing	✗	✗	✗	✓	✗
ReasonMap	Transit maps	✗	✗	✗	✓	✗
UniG2U-Bench	Visual G2U tasks	✗	✗	✗	✗	✓
VGAU-Diag	Visual-planning puzzles	✗	✗	✗	✓	✓
MapSpatial	**Registered GIS views**	✓	✓	✓	✓	✓

3 BENCHMARK DESIGN AND FORMALIZATION

Each latent scene W contains entities, road relations, geometry, semantics, and task constraints. A question q of sub-task type t, one of 14 types grouped into families T1–T4, has semantic answer $y = f_t(W, q)$. Representation $r \in \{\mathrm{sat}, \mathrm{struct}, \mathrm{text}\}$ renders the scene as an image x_r; text denotes the conventional text-bearing map, not a text-only input. All questions are multiple choice with a single correct option, except detour waypoint traversability (T4), which is multi-select and scored by exact set match without partial credit; option counts and uniform-choice reference rates appear in Table 8.

Three paired changes probe different behavior on the same underlying case. A representation switch replaces x_r with the registered $x_{r'}$ while preserving q and its semantic answer y. A geometric action g is one of the seven non-identity elements of the dihedral group D_4: three rotations and four reflections. It maps the image to gx_r, the question and displayed candidates to $\tau_g(q)$, and the semantic answer to $\rho_g(y)$; transformed answers are recomputed from geometry rather than obtained by relabeling option letters. These actions apply to satellite and structural views, not the conventional view, whose text would rotate with the image. For T1 angular order, north remains display-up, so the clockwise answer is recomputed after each action. A world intervention changes task-relevant structure from W to W'; its paired question q' retains the query intent, referenced entities, and candidate set, while the solver recomputes $y' = f_t(W', q')$; $\hat{y}'$ denotes the prediction on the edited case. The question wording is condition-specific: T3 refers to the building layout shown after an occupancy or frontage edit, while T4 asks about traversability under the displayed road status. Thus q' need not be textually identical to q, even though the queried entities and candidate set are fixed. Retained edits change the semantic answer, $y' \neq y$, and are each paired with a visually matched sham edit that leaves the semantic answer unchanged. This pairing tests answer updates against baseline instability under an answer-preserving edit. Transform pairs define $P_{\mathrm{transform}}$, and eligible world-edit pairs define S_{cf}, as detailed in Section 5; task-specific actions and edit eligibility are detailed in Appendices A.4 and B.

Matched map representations. The satellite view retains natural visual texture; the controlled text-free structural map makes task-required roads, footprints, boundaries, and markers explicit; and the conventional map includes Chinese road and place labels, icons, and cartographic detail. Reading the Chinese labels is not required to answer the questions: referenced entities are identified by colored points or lines, and reference answers are computed from map geometry. All use the same crop, coordinate frame, resolution, question, and semantic answer. Satellite-to-structural differences probe the effect of making map geometry explicit; structural-to-conventional differences probe the combined effects of labels, icons, style, and visual density. Marker-only images are retained only as a T1/T2 ablation.

3.1 TASK FAMILIES

A capability progression from geometry to planning. The families progress conceptually from qualitative direction (T1) and metric distance (T2) to route-relative building identification and counting (T3) and road-network path reasoning (T4). This organization does not imply a model solves one family before the next or that accuracy decreases monotonically.

T1 comprises pairwise direction, dual-anchor direction, egocentric side, and angular order. T2 comprises nearest point, directional nearest point, composite Euclidean distance, and composite network distance. T3 comprises segment building count and route cumulative count; T4 comprises route validity, waypoint ordering, detour waypoint traversability, and detour waypoint shortest. Labels are computed from structured geometry, and task-specific ambiguity filters remove borderline cases. Figure 1 shows one example per family; Appendix E.3 gives the complete task matrix and Appendix A.4 gives the operational rules.

Visual evidence targets. Question-conditioned evidence $e_t^{\mathrm{qry}}(W, q)$ is rendered using a fixed allowlist of marks. It shows reference frames or complete measurement geometry for T1/T2 and candidate-complete route or counting structure for T3/T4, without selecting a winner. Oracle-conditioned evidence $e_t^{\mathrm{oracle}}(W, q, y)$ additionally uses the answer and selects a correct counted set or route-relevant state for T3/T4; it is N/A for T1/T2 because highlighting a winning relation

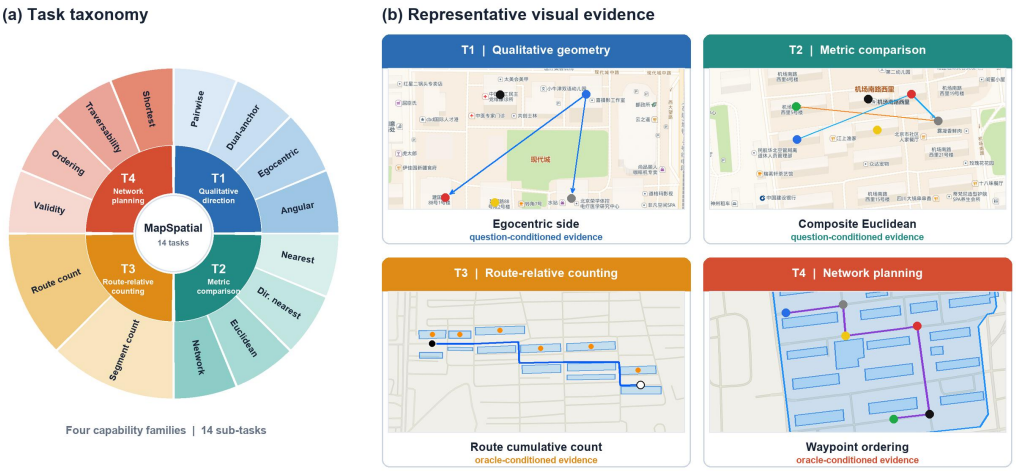


Figure 1: **MapSpatial tasks and representative evidence.** Four families cover 14 sub-tasks. Right: T1 asks where the red point lies when facing from the blue point toward the gray point; T2 compares gray–green and gray–red–blue routes; T3 counts first-row buildings cumulatively left of the blue route; T4 orders waypoints from blue to green along connected roads. T1/T2 show question-conditioned geometry; T3/T4 show oracle-conditioned selections.

or metric alternative would nearly encode the answer. Model-generated evidence $\hat{e}_t$ is produced from the displayed map and question. The deterministic evidence targets and generation prompts prohibit answer text, numeric answers, and option identifiers, because answer leakage through visual intermediates can inflate their apparent utility (Zhu et al., 2026); generated artifacts are audited for violations. Oracle-conditioned marks remain an answer-relevant readout reference rather than an information-matched baseline. Evidence primitives are stored in map and image coordinates; protocol interpretation is given in Section 5.

4 Data Construction

The 1,400 base questions span 585 geographic scenes in China. Source records include road links, building footprints, points of interest, routes, and georeferenced imagery. Each scene uses one crop and coordinate transform to register its canonical views. Questions and distractors are generated from latent geometry: T1 avoids near-boundary directions, T2 enforces distance margins, T3 excludes zero-count cases, and T4 uses network-infeasible route distractors. Eligible T3/T4 cases also receive solver-verified answer-changing edits and matched answer-preserving sham edits.

A deterministic verification pass recomputes retained labels from the source geometry and checks the declared transform and intervention actions. T3/T4 candidate images additionally undergo a manual visibility and rendering-quality screen, retaining approximately 12% and 34% of candidates, respectively. After this screen, both reviewers answered every retained T3/T4 image and matched its GIS label (100% post-screen agreement); this is an answerability check, not a human-performance ceiling. Screening may favor clearer scenes. The 12%/34% figures are candidate-image retention rates. All derived records from one geographic scene remain in one evaluation group, and confidence intervals resample scenes rather than rendered rows. Further construction and audit details appear in Sections B, E.2, and E.4 of the appendix; Figure 2 summarizes the three matched comparisons.

The 77,800 instances follow a condition-dependent expansion rather than a full Cartesian product. T1 and T2 each contribute $400(3 + 14 + 8) \times 2 = 20{,}000$ instances. T3 contributes $200(3 + 14 + 3 + 3) \times 3 = 13{,}800$. For T4, the 200 detour cases contribute $200(3 + 14 + 3 + 3) \times 3 = 13{,}800$, and the 200 non-detour cases contribute $200(3 + 14) \times 3 = 10{,}200$. Here 3 denotes the canonical views, 14 is seven transforms on two eligible views, and 8 is the marker-only base plus seven transforms for T1/T2. Each additional 3 is a sham or intervention rendering in the canonical views. The final factor counts Direct and question-conditioned evidence inputs (2), plus oracle-conditioned evidence

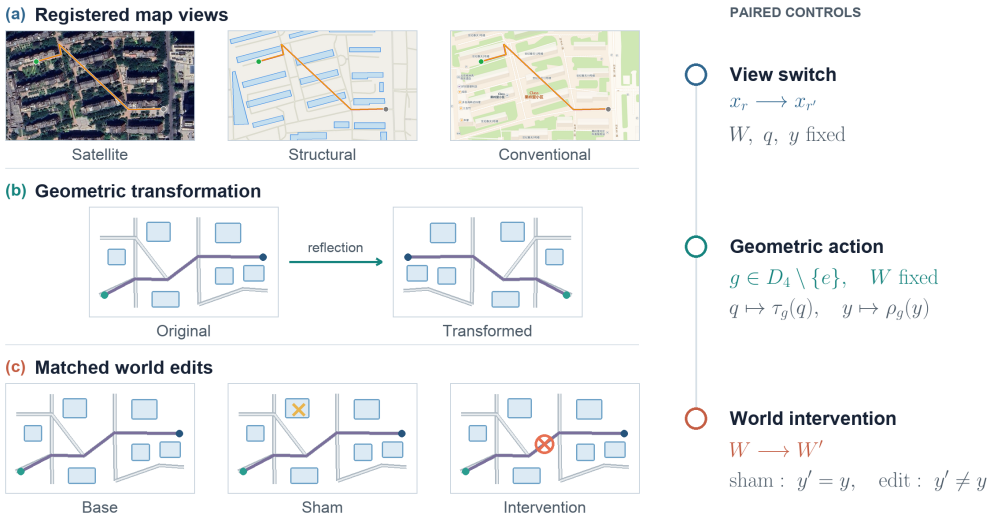


Figure 2: **Controlled evaluation design.** Top: the same geographic scene and query are rendered as co-registered satellite, structural, and conventional maps. Middle: a task-aware geometric transformation is applied consistently to the map, question, and label while preserving the underlying world. Bottom: each eligible T3 counting or T4 detour-planning case is paired with an answer-preserving sham edit and a solver-verified answer-changing intervention. In the right-hand notation, W, q, and y are the latent scene, question, and semantic answer; r indexes a map view and x_r is its rendering. The non-identity dihedral action $g \in D_4 \setminus \{e\}$ rotates or reflects the image, with τ_g and ρ_g mapping the question and answer. W' is an edited scene: the sham preserves the answer ($y' = y$), whereas the intervention changes it ($y' \neq y$).

Table 2: **Dataset statistics after filtering.** Released evaluation instances are correlated variants of 1,400 base questions and are not treated as independent statistical units.

Statistic	Count	Breakdown
Base questions	1,400	T1/T2/T3/T4: 400/400/200/400
Canonical base-view units	4,200	1,400 questions × 3 views
Question–transform groups	9,800	7 actions × 1,400
Canonical transform pairs	19,600	satellite + structural
Marker-only pairs	5,600	800 T1/T2 bases × 7 actions
World-edit base pairs	400	(200 T3 + 200 T4) × 3 views = 1,200 view-specific triples
Answer-preserving sham pairs	400	one per eligible base pair
Evidence targets	2,000	question/oracle: 1,400/600
Released instances	77,800	T1/T2/T3/T4: 20,000/20,000/13,800/24,000

for T3/T4 (3). Thus 77,800 counts condition-specific evaluation records, including evidence-bearing inputs, not independent questions or distinct image files.

5 Evaluation Protocol

Direct understanding. All models answer from (x_r, q) without task-specific visual evidence, using a shared prompt template and temperature 0. Truncated or unparseable initial outputs are retried with a larger output-token budget; each evaluation input contributes one final answer. This is the common leaderboard and the only track that compares understanding-only and unified models.

Table 3: **Primary results across RQ1–RQ3 for six category anchors and two scale references (%).** RQ1 reports direct accuracy for the four task families. RQ2 reports registered-view joint correctness (A_{tri}) and joint correctness on the base and all seven prescribed actions (Base+7). RQ3 reports accuracy on the eligible-pair base (Base$_E$), answer-preserving sham, and answer-changing edited maps, plus paired counterfactual correctness (S_{cf}). RQ3 includes only eligible T3/T4 cases with solver-verified answer-changing edits.

Model	RQ1 Direct capability				RQ2 Robustness		RQ3 World edit			
	T1	T2	T3	T4	A_{tri}	Base+7	Base$_E$	Sham	Edit	S_{cf}
Understanding										
GLM-4.6V-Flash	70.0	49.2	27.8	46.2	32.9	16.5	33.0	31.7	27.3	6.8
Qwen2.5-VL-7B-Instruct	57.8	42.8	34.8	41.0	28.4	13.4	37.4	34.9	27.8	6.3
MiMo-Embodied-7B	67.1	52.2	32.0	39.8	31.1	16.8	33.4	28.1	33.3	8.8
Spatial-MLLM	39.8	50.3	24.2	26.7	20.6	11.2	20.8	20.3	23.7	2.1
Understanding + generation										
Bagel-7B-MoT	70.2	43.5	25.8	35.7	35.5	22.8	27.7	26.2	31.6	3.3
SenseNova-U1-8B-MoT	48.1	45.3	37.8	35.0	20.3	13.3	33.9	32.6	28.3	3.3
Scale references										
Qwen3-VL-32B-Instruct	67.8	58.1	45.5	58.2	42.6	27.1	46.8	48.5	25.2	4.8
Gemini-3-Flash	99.6	87.1	78.3	80.4	72.8	47.2	74.4	54.2	50.0	37.5

Controlled and model-generated evidence. For RQ4, evidence is drawn onto the input map as an image edit. Question-conditioned evidence e^{qry} (T1–T4) tests whether non-selective marks help; oracle-conditioned evidence e^{oracle} (T3/T4) tests whether a correct, selective state can be exploited, and serves as an answer-relevant readout reference rather than a deployable baseline or an upper bound. Model-generated evidence requires map-conditioned image editing, which BLIP3o-8B and Janus-Pro-7B do not support. We evaluate two training-free re-consumption protocols: restart-based reading (C-R), in which the model answers from its own edited map in a fresh pass, and stateful continuation (C-F), which preserves generation state (Appendix F.2). We report paired changes from Direct by task, evidence level, and interface. The oracle-to-generated gap mixes information access with generation fidelity; artifact scores and C-R/C-F comparisons help assess construction versus consumption.

Metrics. RQ1 reports per-view accuracy and family macro-averages. For RQ2, tri-view correctness A_{tri} requires correct answers on all three registered views, and tri-view instability U_{tri} is the fraction of cases whose semantic predictions differ across views; it is not $1 - A_{\text{tri}}$, because views can agree on the same wrong answer. Paired transform correctness $P_{\text{transform}}$ requires correct answers on a base case and one transformed counterpart, and Base+7 requires correctness on the base and all seven transforms. For RQ3, S_{cf} requires correct answers before and after an answer-changing edit; among base-correct pairs, the update rate S_{up} is the fraction also correct after the edit, and the stale rate is the fraction whose edited answer keeps the pre-edit content. 95% percentile intervals use 10,000 scene-level bootstrap replicates; further metrics are in Appendices C and I.

6 Results and Analysis

We evaluate 17 understanding-only models (7 general, 7 spatial-specialized, and 3 scale references) and 8 UMMs, 6 of which support image editing; model settings and interfaces are in Appendices D and F.2. Table 3 includes a fixed six-model anchor set spanning two general understanding-only models (GLM-4.6V-Flash, Qwen2.5-VL-7B-Instruct), two spatial-specialized models (MiMo-Embodied-7B, Spatial-MLLM), and two unified multimodal models (Bagel-7B-MoT, SenseNova-U1-8B-MoT), fixed by category rather than performance and shared across RQ1–RQ3; their direct accuracy (35.2–48.3%) lies within the range of the other open models. Qwen3-VL-32B-Instruct and Gemini-3-Flash appear alongside them as scale references, and Qwen3.5-Plus is reported for direct understanding. All 25 direct-understanding results, with scene-bootstrap intervals, are in Table 13.

Table 4: **Anchor and scale-reference comparison (%).** Rows are the six anchors followed by two scale references. Direct is the T1–T4 macro-average on the three untransformed, unedited base views (Table 13). Base+7 requires a correct base and all seven transformed answers. S_{cf} requires correct answers before and after a solver-verified edit on the eligible world subset.

Model	Direct	Base+7	S_{cf}	Update | base	Stale | base
GLM-4.6V-Flash	48.3	16.5	6.8	20.7	57.6
Qwen2.5-VL-7B-Instruct	44.1	13.4	6.3	16.9	69.5
MiMo-Embodied-7B	47.8	16.8	8.8	26.2	50.4
Spatial-MLLM	35.2	11.2	2.1	10.0	76.4
Bagel-7B-MoT	43.8	22.8	3.3	11.8	78.6
SenseNova-U1-8B-MoT	41.6	13.3	3.3	9.6	84.3
Qwen3-VL-32B-Instruct	57.4	27.1	4.8	10.1	83.6
Gemini-3-Flash	86.3	47.2	37.5	50.4	35.1

6.1 RQ1: Direct Spatial Capability

Qualitative direction (T1) is generally the strongest family, whereas building counting along roads and routes (T3) is the weakest for five of the six anchors (24.2–37.8%). Scale does not remove this profile: Qwen3-VL-32B-Instruct, Qwen3.5-Plus, and Gemini-3-Flash reach 57.4%, 64.9%, and 86.3%, but T3 remains their weakest family (45.5%, 43.0%, and 78.3%). Spatial specialization also gives no consistent advantage: the best spatial-specialized model (MiMo-Embodied-7B, 47.8%) trails the best general 7–10B models (48.3% and 48.9%). A text-only control with Qwen2.5-VL-7B-Instruct scores 35.6% versus 45.7% on the same questions with satellite images (Table 9).

6.2 RQ2: Representation and Geometric Robustness

Across the six anchors, tri-view joint correctness is 20.3–35.5% (Table 3), while 33.8–60.7% of matched triplets yield non-identical semantic predictions. Retained views pass task-required visibility checks but differ in visual cues; disagreement does not isolate its cause. Per-view results for all models (Appendix H) show no uniform gain from conventional maps; tri-view intervals are in Table 15. Paired base–transform correctness ranges from 25.2% to 36.5%; among initially correct predictions, 65.7–78.9% remain correct after one declared action. Even when the transformed reference answer is unchanged, only 54.1–75.1% of predictions stay the same. Across complete base-plus-seven groups, 29.5–49.3% of base-correct cases remain correct on all seven actions; the corresponding joint Base+7 rate is 11.2–22.8% (Tables 19 and 20).

6.3 RQ3: Structural Intervention Sensitivity

For the six anchors in the RQ3 columns of Table 3, edited-map accuracy is 23.7–33.3%, and sham accuracy is 20.3–34.9% (base: 20.8–37.4%); paired counterfactual correctness is only 2.1–8.8% and conditional updating among base-correct pairs is 9.6–26.2% (Table 21). Among the six anchors, the most common observed outcome on base-correct pairs is retaining the pre-edit answer after the edit (50.4–84.3%). Among the scale references, Qwen3-VL-32B-Instruct keeps the pre-edit answer on 83.6% of base-correct pairs, whereas Gemini-3-Flash correctly updates 50.4% of base-correct pairs (Table 4).

6.4 RQ4: Controlled and Model-Generated Evidence

Evidence effects vary by task and condition. Across the four common models (Bagel, ThinkMorph, SenseNova-U1, and LatentUM; Table 5), question-conditioned T1 point estimates rise for all four; T2 varies, T3 changes little, and T4 falls for all four. Oracle-conditioned T3 gains range from 0.0 to 4.2 points; T4 improves for three of the four models. Even with answer-selecting oracle evidence, T3/T4 accuracy stays at or below 45.4% (Table 24), but this readout does not isolate evidence-consumption failure: selective marks may also obscure or distract from map content. These four-model contrasts are descriptive.

Table 5: **Evidence effects vs. Direct (percentage points; calculated before rounding).** Q/O: question-/oracle-conditioned (O: T3/T4); C-R/C-F: generated-evidence macro changes. Direct is evidence-matched (%) and may differ from full Track-U Direct; absolute scores: Table 24.

		T1	T2	T3		T4		Generated	
Model	**Direct**	Q	Q	Q	O	Q	O	C-R	C-F
Bagel	44.2	$+1.8$	$+1.6$	$+1.2$	$+4.2$	-4.5	$+8.3$	-0.9	-4.5
ThinkMorph	42.6	$+3.9$	$+3.8$	$+0.7$	$+1.8$	-2.4	$+9.6$	-0.9	$+1.2$
SenseNova-U1	42.1	$+3.1$	-1.6	0.0	$+0.3$	-7.5	-0.7	$+2.2$	-10.3
LatentUM	38.2	$+5.6$	-0.4	$+0.3$	0.0	-1.1	$+3.9$	-2.1	-3.6

C-R changes macro accuracy by -2.1 to $+2.2$ points (six models), and C-F by -10.3 to $+1.2$ (four); generated artifacts often omit evidence or map context, or hallucinate text (Appendix L).

7 Discussion and Limitations

Interpretation. MapSpatial separates image-level accuracy from behavior across registered views, declared geometric actions, and edited maps. Prediction disagreement and conditional correctness capture behavior that direct accuracy alone misses. View and transform gaps indicate sensitivity to rendering and display frame; they are not a pure measure of internal geometric reasoning because visibility and image quality also change.

These paired tests reveal two behaviors hidden by marginal accuracy: inconsistent answers across representations of the same scene and retention of pre-edit answers after verified world edits. Disagreement and conditional correctness measure these behaviors; joint correctness sets a stricter success criterion. Because shams also flip predictions, neither edited accuracy nor a prediction flip alone establishes correct updating. The benchmark does not identify the internal representation that produces consistency; it measures the observable behavior expected of a system that tracks task-relevant map structure.

Limitations. The geographic sample and map providers are Chinese; transfer elsewhere is untested. Rotations exclude text-bearing maps, and schematic T3/T4 edits and shams cannot match every visual cue. Scene bootstrap handles within-scene dependence, not source bias.

The setting is narrower than real navigation. Questions are static and option-based, with one multi-select sub-task; they do not test sequential interaction, uncertainty-aware route execution, or recovery from localization error. We verify labels exhaustively from GIS structure rather than estimate a human-performance ceiling. Each evaluation input contributes one final answer at temperature 0, so intervals reflect scene sampling rather than run-to-run variance, and Gemini-3-Flash and Qwen3.5-Plus are the only proprietary models evaluated. Oracle-conditioned evidence is not information-matched to model generation; C-R/C-F deltas describe complete generation-and-reading pathways rather than isolating evidence construction from subsequent use.

8 Conclusion

MapSpatial tests GIS-grounded spatial questions across registered views, geometric transforms, and solver-verified edits with matched shams. **First, single-map accuracy misses paired behavior:** across six anchors, 33.8–60.7% of matched view triplets disagree, single-action retention is 65.7–78.9%, and correct edit updating among base-correct pairs is 9.6–26.2%. **Second, visual-evidence effects depend on the task and evidence condition.** Question-conditioned marks raise T1 accuracy but lower the T4 point estimate for all four evaluated unified models; oracle-conditioned selections mainly help T4, while generated artifacts can omit evidence, lose map context, or hallucinate content. Together, the paired protocols and GIS-grounded labels make representation consistency, correct updating, and evidence use measurable targets beyond direct accuracy. The released paired groups support evaluating, and potentially training, models on linked equivalent and counterfactual inputs rather than independent map images.

AI USE STATEMENT

During this project, we used generative AI and coding assistants for literature exploration, formalization, implementation and debugging of data-generation and evaluation scripts, figure and LaTeX editing, and drafting or revising prose. The authors reviewed and validated these outputs and remain responsible for the research questions, experimental design, implementation, labels, analyses, claims, and final manuscript. Benchmark images and annotations were produced by the deterministic pipeline, and generative AI was not used as the score-producing evaluator.

ETHICS AND BROADER IMPACT

The authors have read and will adhere to the ICLR Code of Ethics. The work does not study human behavior; manual review includes visual-quality screening and two-reviewer image-based answer checks. Released scenes use anonymized identifiers and desensitized coordinates where appropriate, with vector and imagery licenses audited separately. Map benchmarks may expose infrastructure or movement-relevant structure and may inherit geographic or cartographic bias. Outputs must not be used as ground-truth guidance for safety-critical navigation, surveillance, or other high-stakes decisions.

REPRODUCIBILITY STATEMENT

The anonymous repository provides the generation and evaluation code and links to the released benchmark dataset. Together, these resources provide evaluation instances and images, registered transforms, labels, evidence primitives, the evaluation-group manifest, prompts, parsers, and scene-grouped metric outputs. The main paper specifies the benchmark claims and primary protocol; the appendices provide task rules, construction details, interface manifests, validity checks, complete tables, and extended analyses needed to reproduce the reported scores.

REFERENCES

Christian M. A. Arnold, Andrew Alini, Jonathan Wang, Pieter M. Feenstra, Conner Arnold, Jan DeWitt, Natalie C. Ritsema, Jung Hyun Yae, Boris Katz, Andrei Barbu, and Brian Cheung. MapQA: A map-question-answering benchmark for visual language model reasoning. In *ICLR Workshop on Multimodal Intelligence*, 2026.

Sharat Bhat, Harshita Khandelwal, Tushar Kataria, and Vivek Gupta. MapVerse: A benchmark for geospatial question answering on diverse real-world maps. *arXiv preprint arXiv:2602.10518*, 2026. URL https://arxiv.org/abs/2602.10518.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. URL https://arxiv.org/abs/2211.08545.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian 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)*, 2024. URL https://arxiv.org/abs/2401.12168. arXiv:2401.12168.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a. URL https://arxiv.org/abs/2505.09568.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b. URL https://arxiv.org/abs/2501.17811.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024. URL https://arxiv.org/abs/2411.19325.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. ReasonMap: Towards fine-grained visual reasoning from transit maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41077–41088, 2026. URL https://arxiv.org/abs/2505.18675. arXiv:2505.18675.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2404.12390. arXiv:2404.12390.

Wenxi Gao, Guanxi Lu, Didi Zhu, Hao Mark Chen, Quan Deng, Zhican Wang, Jiankang Deng, and Hongxiang Fan. Model guides you how to draw: Adaptive visual gating for unified multimodal reasoning, 2026. URL https://arxiv.org/abs/2607.10004.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mB3vxfrQZM.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4423.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025. URL https://arxiv.org/abs/2506.03135.

Jiachun Jin, Zetong Zhou, Xiao Yang, Hao Zhang, Pengfei Liu, Jun Zhu, and Zhijie Deng. LatentUM: Unleashing the potential of interleaved cross-modal reasoning via a latent-space unified model, 2026. URL https://arxiv.org/abs/2604.02097.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9161–9175, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.15826. arXiv:2311.15826.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1786–1807, 2024. doi: 10.18653/v1/2024.emnlp-main.106. URL https://arxiv.org/abs/2406.02537. arXiv:2406.02537.

Zekun Li, Malcolm Grossman, Eric Qasemi, Mihir Kulkarni, Muhao Chen, and Yao-Yi Chiang. MapQA: Open-domain geospatial question answering on map data. *arXiv preprint arXiv:2503.07871*, 2025. URL https://arxiv.org/abs/2503.07871.

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. arXiv:2512.08016.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. NavBench: Probing multimodal large language models for embodied navigation. *arXiv preprint arXiv:2506.01031*, 2025. URL https://arxiv.org/abs/2506.01031.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5860–5865, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1596.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2507.07644. arXiv:2507.07644.

Majid Saeedan, Muhammad Shihab Rashid, Ahmed Eldawy, and Vagelis Hristidis. GS-QA: A benchmark for geospatial question answering. *arXiv preprint arXiv:2605.22811*, 2026. URL https://arxiv.org/abs/2605.22811.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6649–6658, June 2019. doi: 10.1109/CVPR.2019.00681.

Ziqiao Shang, Ling-Yue Ge, Zian Xu, Zi-Jian Cheng, Shi-Yu Tian, Zhenyu Huang, Wenbo Fu, Weiming Wu, Yang Chen, Xiangwen Zhang, Yulan Hu, Bin Liu, and Lan-Zhe Guo. MapTab: A diagnostic benchmark for long-horizon multi-criteria multimodal reasoning on heterogeneous topological graphs. *arXiv preprint arXiv:2602.18600*, 2026. URL https://arxiv.org/abs/2602.18600.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. RealUnify: Do unified models truly benefit from unification? a comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22488–22497, 2026. URL https://arxiv.org/abs/2509.24897. arXiv:2509.24897.

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. MapIQ: Evaluating multimodal large language models for map question answering. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=VSwRuGtB5n. arXiv:2507.11625.

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. *arXiv preprint arXiv:2512.03558*, 2025. URL https://arxiv.org/abs/2512.03558.

Zimo Wen, Boxiu Li, Wanbo Zhang, Junxiang Lei, Xiaoyu Chen, Yijia Fan, Qi Zhang, Yujiang Wang, Lili Qiu, Bo Li, Ziwei Liu, Caihua Shan, Yifan Yang, and Yifei Shen. UniG2U-Bench: Do unified models advance multimodal understanding? *arXiv preprint arXiv:2603.03241*, 2026. URL https://arxiv.org/abs/2603.03241.

Jialong Wu, Xiaoying Zhang, Hongyi Yuan, Xiangcheng Zhang, Tianhao Huang, Changjing He, Chaoyi Deng, Renrui Zhang, Youbin Wu, and Mingsheng Long. Visual generation unlocks human-like reasoning through multimodal world models, 2026. URL https://arxiv.org/abs/2601.19834.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025. URL https://arxiv.org/abs/2506.09965.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.03622.

Wulin Xie, Yi-Fan Zhang, Chaoyou Fu, Yang Shi, Bingyan Nie, Hongkai Chen, Zhang Zhang, Liang Wang, and Tieniu Tan. MME-Unify: A comprehensive benchmark for unified multimodal understanding and generation models. *arXiv preprint arXiv:2504.03641*, 2025. URL https://arxiv.org/abs/2504.03641.

Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human? *arXiv preprint arXiv:2503.14607*, 2025. URL https://arxiv.org/abs/2503.14607.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025. URL https://arxiv.org/abs/2511.21471.

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. *arXiv preprint arXiv:2602.15918*, 2026. URL https://arxiv.org/abs/2602.15918.

Qian Yang, Ankur Sikarwar, Huy Le, Le Zhang, Zhuan Shi, Perouz Taslakian, and Aishwarya Agrawal. How and what to imagine? visual thinking in unified multimodal models for cross-view spatial reasoning, 2026. URL https://arxiv.org/abs/2605.27310.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Chaorui Deng, Shen Yan, Haoqi Fan, Yejin Choi, James Zou, Cihang Xie, Huaxiu Yao, and Qinghao Ye. When visualizing is the first step to reasoning: MIRA, a benchmark for visual chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26154–26164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_When_Visualizing_is_the_First_Step_to_Reasoning_MIRA_a_CVPR_2026_paper.html.

Yubo Zhu, Zhehan Kan, Jingyi Yang, Miaolin Chen, Jinbo Xing, Kai Zhu, Zijian Wang, Sheng Zhong, and Wei Tong. When does visual generation help visual understanding in unified multimodal models?, 2026. URL https://arxiv.org/abs/2608.22174.

APPENDIX: SUPPLEMENTARY MATERIAL

The following appendices provide detailed task definitions, data-construction procedures, evaluation interfaces, validity controls, complete model tables, and extended analyses referenced by the main paper.

A THE MAPSPATIAL BENCHMARK

A.1 DESIGN GOALS

MapSpatial follows five design requirements. **Shared latent structure:** every answer, rendering, and evidence target must be independently recomputable from W. **Matched representations:** satellite, controlled text-free structural-map, and conventional text-bearing map images from one scene form a registered group rather than independent samples. **Map-native dependence:** core tasks require metric, topological, or route-relative information; marker-only geometry is retained as a diagnostic rather than presented as map understanding. **Controlled intervention:** appearance changes and structural changes are generated separately. **Statistical validity:** all variants of one scene retain one group identifier, and confidence intervals are grouped by scene.

A.2 VISUAL INPUTS AND LABELS

Each case is grounded in one latent map scene stored as a structured group with world geometry, matched representations, evaluation instances, interventions, question-conditioned visual evidence, and oracle-conditioned visual evidence (Appendix E.1). All representations share a map-coordinate frame, and each renderer stores the map-to-image transform.

A.3 MATCHED MAP REPRESENTATIONS

Each question is evaluated on a registered representation group chosen to control representation confounds, rather than on unrelated map subsets. The **satellite view** has high visual realism but leaves roads and buildings implicit in texture. The **controlled text-free structural map** uses a fixed road-map rendering with deterministic overlays of task-relevant building footprints, boundaries, and benchmark markers. It removes satellite texture, points-of-interest labels, and uncontrolled answer text while retaining the minimum task-sufficient roads and footprints. The **conventional text-bearing map** retains ordinary Chinese road names, points-of-interest labels, icons, and cartographic density. These labels provide realistic map context but are not required for solving the benchmark: whenever a task uses marked candidates, answer-critical entities and relations are encoded with color-coded points or lines, and the question, candidate set, and reference label are generated from map coordinates rather than optical character recognition. The text-free structural view provides the corresponding optical-character-recognition-free representation, while the conventional view tests robustness to the additional cartographic content. All three views are co-registered to the same latent geographic record, crop, coordinate frame, question, and answer.

The controlled structural map is distinct from a marker-only blank canvas. A blank canvas that contains only colored points is answer-equivalent for some T1/T2 geometry questions, but it is not a valid representation for tasks whose semantics depend on buildings or roads. We therefore use marker-only blank images only as a geometry ablation, while the controlled structural and conventional maps are main representations for T1–T4. This is a deliberate control choice: the registered group preserves enough map structure for every task while varying the visual encoding of that structure. All views use matched crops, aspect ratios, output resolution, marker sizes, and question text.

The three views expose different renderings of the same scene. The satellite-to-structural gap probes the burden of recovering explicit geometry from natural appearance. The structural-to-conventional gap probes the combined effects of realistic labels, icons, style, and visual density; because these layers are not perfectly factorized in the source tiles, we do not interpret it as a pure optical-character-recognition or text gain. Disagreement across the three views measures output variation across renderings without identifying its cause.

A.4 TASK FAMILIES

A capability progression from geometry to planning. The four task families form a capability-stratified progression. T1 tests qualitative spatial relations, and T2 adds metric comparison. T3 introduces semantic map structure by requiring models to identify and aggregate building obstacles relative to a directed route, including visibility, occlusion, and duplicate-counting constraints. T4 then tests whether a model can reason over road connectivity and plan routes while avoiding non-traversable structures or localized obstructions. This progression is conceptual rather than computationally cascaded: T4 does not consume a model-produced T3 output, but requires a richer combination of geometric, semantic, and topological map understanding. It is a capability ordering rather than a claim that empirical accuracy must decrease monotonically from T1 to T4; the observed difficulty can depend on the model and sub-task.

T1: Qualitative Direction Reasoning. T1 is a foundational geometry diagnostic: it tests whether a model can infer direction and transform between global and route-relative reference frames before map topology is required. Each scene contains 4–6 colored building markers and four sub-tasks:

1. *Pairwise direction*: determine the diagonal direction of one point relative to another.
2. *Dual-anchor direction*: find the point that simultaneously satisfies two directional constraints from two anchors.
3. *Egocentric side*: facing from one point toward another, classify a third point as *left-front*, *left-back*, *right-front*, or *right-back*.
4. *Angular order*: order same-quadrant candidates by clockwise bearing from north.

Pairs near cardinal-axis boundaries are filtered to avoid ambiguous labels. Question-conditioned visual evidence marks relevant entities and reference axes without highlighting the correct option.

T2: Metric Distance Reasoning. T2 evaluates metric comparison and the distinction between image-space, geodesic, and network-constrained distance. It contains four sub-tasks:

1. *Nearest point*: select the closest candidate to a reference point.
2. *Directional nearest point*: select the closest candidate among points that lie in a specified diagonal direction from the reference.
3. *Composite Euclidean distance*: compare a direct segment A→B with a two-segment route A→C→D using straight-line distances.
4. *Composite network distance*: compare the same alternatives using road-network link distances.

Question-derived prompts draw measurement rays to all candidates or compared route segments; numerical distances and final answers are not rendered.

T3: Route-Relative Building Identification and Counting. T3 evaluates route-relative aggregation under spatial constraints. It contains two sub-tasks:

1. *Segment building count*: count buildings on a specified side of one directed road segment.
2. *Route cumulative count*: accumulate counts across a multi-segment route while updating the left/right side after each turn and avoiding duplicate counting at segment boundaries.

For both sub-tasks, the counted set follows a frontage rule: buildings are assigned to the closest relevant segment and must be visible from the route side rather than hidden behind another building. The question-conditioned rendering draws the complete route with one arrow per segment, the queried side, and eligible candidate regions without marking the final counted set. A separate oracle-conditioned visualization marks counted buildings and is used only as a selective diagnostic reference and a target for model-generated-evidence quality.

T4: Road-Network Path Reasoning. T4 evaluates road-network topology and constraint-sensitive planning through four sub-tasks:

1. *Route validity*: select one of four separately rendered candidate-route images. The correct candidate must connect the start and end points along connected road geometry without shortcutting through buildings or omitting required route structure.
2. *Waypoint ordering*: select the waypoint order associated with the feasible road-network route.
3. *Detour waypoint traversability*: given two natural routes with the same origin and destination, select all waypoint sequences that remain traversable under a localized non-traversable obstruction.
4. *Detour waypoint shortest route*: under the same two-route setup, select the shortest traversable waypoint sequence. Cases are retained only when the route-length difference exceeds the ambiguity margin.

All route evidence is rendered from complete road-link coordinate strings rather than by connecting only sparse turn points. The detour-planning sub-tasks separate the road-status image from the waypoint-only image, and the obstruction is represented visually by a red cross indicating a location that must be avoided. T4 uses satellite, controlled structural-map, and conventional text-bearing map views; the marker-only blank control is omitted because road topology and building layout are essential to the task. Question-conditioned visual evidence is candidate-complete: route validity retains the full road-following polyline but removes non-traversed points; waypoint ordering overlays all four candidates on one map with matched line styles and distinct colors; detour traversability reuses the existing two-route rendering; and detour shortest shows both alternatives with the oracle-conditioned line style. Oracle-conditioned visual evidence instead selects the correct route.

Appendix E.3 summarizes the full task matrix, including sub-tasks, question forms, evidence targets, and answer types.

A.5 Visual Evidence Targets

Every evidence-bearing instance uses a vocabulary of `point`, `line`, `arrow`, `region`, `polyline`, `barrier`, and `scale_bar`. Each primitive stores its semantic role, map coordinates, and per-view image coordinates. Their union defines task-specific controlled evidence at two information levels.

Question-conditioned visual evidence. e^{qry} is answer-agnostic: it externalizes the operations needed to solve a question without selecting the answer. Examples include reference axes, rays to every distance candidate, a directed route and side buffer, or all candidate paths together with obstacles. It is defined for all four tasks.

Oracle-conditioned visual evidence. e^{oracle} is answer-conditioned: it identifies the counted set or the feasible route or route set for T3/T4. Its gain is interpreted under the boundary defined in Section 5. It is N/A for T1/T2 because selecting only the winning relation or metric alternative would nearly encode the answer. We report the two evidence conditions separately.

Automated leakage checks reject answer text, numeric measurements, correct option identifiers, and other marks that directly serialize the final choice in either target. A deterministic renderer provides a calibrated reference.

B Data Construction

Source data. Scenes are sampled from geospatial source records in China, including road-network links, building footprints, points-of-interest locations, route records, and georeferenced imagery. We report both the underlying question cases and the benchmark evaluation instances. The latter include correlated views, transformations, world variants, and evidence conditions and are therefore not treated as independent statistical units; uncertainty estimates instead resample geographic scenes. Final post-filter statistics are reported in Table 2. Benchmark identifiers are anonymized, and imagery licensing is audited separately from vector-data licensing.

Representation generation. For each scene, one crop and coordinate frame define all matched views. Satellite, controlled text-free structural-map, and conventional text-bearing map inputs

are registered to this frame. The structural-map renderer receives deterministic overlays for task-sufficient footprints, boundaries, and markers. All three views are derived from the same latent geographic record and use the same crop and map-coordinate transform; the satellite pixels themselves are not assumed to be rendered from road records. Resolution, aspect ratio, marker size, and question text are held fixed across the canonical representation triplet. We do not claim a systematic zoom, visible-layer, or within-provider style intervention because those factors are not independently varied in the current benchmark. Instead, each eligible satellite and structural-map image receives seven deterministic map transformations: three rotations, one horizontal reflection, and three reflected rotations. The displayed image frame defines up/north, down/south, left/west, and right/east. The question, answer, and pixel-space evidence therefore follow the task-specific action ρ_g defined in the main paper (Table 6); no compass or post-transform north cue is added. The conventional text-bearing map participates in canonical north-up representation comparisons and world-intervention evaluation. It is excluded only from raw raster rotation and reflection: these operations would rotate or mirror text and confound spatial equivariance with text-orientation robustness. Consequently, the benchmark is not a full Cartesian product of views and transforms. The transform subset uses satellite and text-free structural views, whereas canonical representation and eligible world-intervention subsets use all three main views.

Question generation. Questions are generated from structured geometry rather than written manually. T1 uses bearings between colored points, with angular margins to avoid boundary cases. T2 uses GPS distances and rendered pixel-distance consistency checks. T3 uses segment-side assignment and visibility/frontage filters. T4 uses full road-link coordinate strings and waypoint structures from road-network paths. Network-distance questions are retained only when Euclidean and network rankings provide a controlled contrast. All map-derived answers are computed in an appropriate projected coordinate reference system; geodesic checks are used where the crop extent makes planar approximation non-negligible. The current base design contains 1,400 question cases before representation, transform, intervention, and evidence expansion: T1 contains four sub-tasks with 100 cases each (400), T2 contains four with 100 each (400), T3 contains two with 100 each (200), and T4 contains four with 100 each (400). Derived renderings and evidence conditions are correlated observations of these base cases and are never presented as additional independent cases.

Structure-changing interventions. The current benchmark evaluates world interventions only on eligible T3/T4 sub-tasks. T3 edits route-relative building occupancy or frontage, and T4 adds a localized non-traversable obstruction to the two detour-planning sub-tasks when a solver-verified alternative exists. The obstruction is represented by a red cross and is interpreted as a task-level region that a valid route must avoid; it is not presented as a photorealistic counterfactual satellite scene. The current construction does not provide paired world interventions satisfying the required controls for T1/T2, route validity, or plain waypoint ordering, so these cells are N/A rather than evidence that such counterfactuals are impossible in principle. The query intent, referenced entities, and candidate set are held fixed. An intervention is retained only when an independent solver verifies the intended answer change; a visually matched sham must preserve it.

Structured distractors and auditing. T1/T2 candidates obey the angular and distance margins in Appendix E.2; T3/T4 use the frontage and route checks there.

Labels are computed from the latent GIS record and retained only when an independent deterministic pass reproduces the answer and the declared transform or intervention action. Human quality control has a different role: T3/T4 candidate renderings are screened for task-relevant visibility, readability, and obvious imagery defects. Two reviewers also answer the retained images directly and match the GIS-derived labels; their visual check does not replace the GIS-derived reference label. This stringent image screen retains approximately 12% of T3 candidates and 34% of T4 candidates; these are candidate-to-retained image rates, not estimates of label accuracy or a human performance ceiling.

Splits and statistical unit. The current benchmark is evaluation-only and therefore does not define a train/test or region-held-out generalization split. All questions, renderings, evidence conditions, and interventions derived from one geographic scene share one group identifier. Confidence intervals resample these scene groups rather than treating correlated QA rows as independent. Exact scene, question, view, and intervention counts are reported after all filters.

Table 6: **Declared task actions for the seven map transformations.** In every row, g acts on the image and all pixel-space evidence. The table makes explicit how the question action τ_g and answer action ρ_g are produced; transformed labels are recomputed and audited from geometry, not edited as strings after inference.

Sub-task	$\tau_g(q)$: **question/candidate action**	$\rho_g(y)$: **answer action**
T1 pairwise direction	Map display-frame direction words and all point coordinates by g.	Map the direction class by the same action.
T1 dual-anchor	Map both directional constraints and all point coordinates by g.	Preserve target identity; remap its displayed option position.
T1 egocentric side	Transform observer, heading target, and query point; swap handedness under reflections.	Recompute front/back and left/right class from transformed vectors.
T1 angular order	Transform all candidate coordinates; keep north fixed to the vertical display axis in every transformed view.	Recompute clockwise order from display-frame north after the geometric action.
T2 nearest point	Transform reference and candidates; keep the metric and candidate identities fixed.	Preserve nearest-target identity.
T2 directional nearest	Transform points and map the directional eligibility constraint.	Preserve the selected target identity after geometric recomputation.
T2 composite Euclidean	Transform all endpoints; keep route identities fixed.	Preserve the shorter-route identity because the prescribed map transformations preserve length.
T2 composite network	Transform the rendered network and route endpoints jointly.	Preserve the shorter-network-route identity.
T3 segment count	Transform the directed segment and side query; swap left/right under reflections.	Recompute the eligible set; preserve its cardinality after a valid action.
T3 route cumulative	Transform every directed route segment in route order and swap the left/right query under reflections.	Recompute segment assignment and deduplication; preserve the final count.
T4 route validity	Transform each complete candidate route and map; preserve candidate identity.	Preserve the valid candidate after solver verification.
T4 waypoint ordering	Transform the map, endpoints, waypoints, and all candidate routes jointly.	Preserve the feasible ordering identity.
T4 detour traversability	Transform both routes, waypoints, and obstruction region jointly.	Preserve the solver-verified feasible route set.
T4 detour shortest	Transform both routes, waypoints, and obstruction region jointly.	Preserve the shortest feasible route identity.

After final filtering, the benchmark is distributed across ten JSONL files. T1/T2 provide two evaluation conditions, whereas T3/T4 provide three. World edits are retained for 400 base questions (200 T3 and 200 T4 detour questions), each with a solver-verified answer-changing edit and a matched answer-preserving sham. Across three canonical views, these yield the 1,200 view-specific triples used by RQ3. Canonical satellite/structural transform pairs exclude marker-only blank views.

C Evaluation Protocol

C.1 Protocol for Understanding-Only and Unified Models

The main-text definitions of the evaluation protocols are authoritative; see Section 5. This appendix does not repeat their conceptual descriptions. Implementation details for track assignment and protocol execution are provided in Sections D, F.1, and F.2.

C.2 Evidence Conditions

The main-text evidence-condition definitions and metric contrasts are authoritative; see Section 5. This appendix focuses on reproducibility details: the locked reader prompt and answer parser are shared across conditions wherever the interface permits, and model-specific track support is listed in Table 11. The reported end-to-end contrasts do not localize a failure to artifact construction or evidence consumption.

C.3 Metrics

Representation robustness. Besides per-view accuracy, we use the main-text tri-view correctness A_{tri}, the fraction of cases answered correctly in all three canonical representations. Tri-view instability U_{tri} is the fraction for which the three predictions disagree. Pairwise correctness, wrong consistency, and disagreement matrices are reported to localize which representation transition causes an error. We also report the structural-minus-satellite and conventional-minus-structural accuracy differences. The latter measures the net effect of conventional labels, icons, style, and visual density, not a pure OCR gain.

Transform consistency. For every declared transform g, we report paired transform correctness and complementary prediction-level diagnostics. Let A_{base} and A_{trans} denote marginal correctness on the paired base and transformed members. Joint correctness uses the main-text $P_{\text{transform}}$, where $y_g = \rho_g(y)$ is the transformed reference answer generated by the task-aware label builder:

$$P_{\text{transform}} = \Pr\left[\hat{y}(x_r, q) = y \ \wedge \ \hat{y}(gx_r, \tau_g(q)) = \rho_g(y)\right]. \tag{1}$$

For pairs whose reference answer is unchanged, we also report raw prediction agreement, $C_{\text{same-answer}} = \Pr[\hat{y}(gx_r, \tau_g(q)) = \hat{y}(x_r, q) \mid y_g = y]$. Equivariance retention among initially correct predictions is $R_{\text{eq}} = P_{\text{transform}}/A_{\text{base}}$, with a scene-grouped interval computed on paired rows. The two quantities are kept separate: raw option equality is not a valid equivariance test when a transform changes the correct option. This replaces an ambiguous single "render invariance" score with paired correctness and a conditional stability diagnostic.

Counterfactual sensitivity. For interventions selected to change an answer from y to y', paired counterfactual accuracy uses the main-text S_{cf}:

$$S_{\text{cf}} = \Pr\left[\hat{y} = y \ \wedge \ \hat{y}' = y' \mid y \neq y'\right]. \tag{2}$$

On the same eligible population $E = \{i : y_i \neq y_i'\}$, the update rate conditioned on a correct base answer is

$$S_{\text{up}} = \frac{\sum_{i \in E} w_i \mathbb{1}[\hat{y}_i = y_i] \mathbb{1}[\hat{y}_i' = y_i']}{\sum_{i \in E} w_i \mathbb{1}[\hat{y}_i = y_i]}. \tag{3}$$

It is N/A when the denominator is zero. The weights, parser rules, and eligible pair set must be identical in the numerator and denominator; in particular, S_{up} must not be obtained by dividing S_{cf} by a base accuracy computed on all T3/T4 cases. We additionally report sham-edit stability and the rate at which a model changes its answer in the correct versus incorrect direction.

Model-generated evidence is scored at three coupled levels: **primitive quality** against reference map-coordinate geometry, **target-type fidelity** to the declared question-conditioned or oracle-conditioned information level, and **downstream utility** after the artifact is consumed. Final-answer serialization is prohibited in every generated artifact. Spatial understanding uses multiple-choice accuracy and deterministic parsers, with task, representation, and scene-bootstrap breakdowns. Primitive definitions, transform consistency, and extended gains are in Appendix I.

Primary accuracy uses all eligible assigned instances as the denominator. Initial truncated or unparseable outputs are retried with a larger output-token budget before scoring. All final scored responses in this release are parseable; each eligible instance contributes once. Unsupported model–track combinations remain N/A.

Unless otherwise stated, the tables report point estimates. We compute two-sided 95% percentile intervals with a scene-level bootstrap: all canonical views and derived variants from the same geographic scene are resampled together, rather than treating rendered rows as independent observations. The complete intervals, effective scene counts, and data-integrity audit are provided in Appendix J. These intervals quantify sampling uncertainty over the evaluated scene set and do not account for geographic coverage, provider, or model-interface bias.

The benchmark therefore separates spatial-structure inference, representation robustness, counterfactual sensitivity, evidence construction, and evidence use. No aggregate score combines model interfaces that participate in different tracks. Appendix E.4 gives the complete validity matrix.

D Experimental Setup

D.1 Models and Setup

The completed direct-understanding suite contains 25 models: 17 understanding-only MLLMs and 8 unified multimodal models (UMMs). All enter the direct-understanding protocol on the canonical question set. The 600 T3/T4 questions (200 T3 and 400 T4) form the diagnostic subset receiving every representation, intervention, and evidence condition. Among the 8 UMMs, 6 support the map-conditioned generation and re-consumption interfaces used in the evidence experiments. BLIP3o-8B and Janus-Pro-7B are evaluated only under direct understanding in this release: their available interfaces are text-to-image rather than map-conditioned image editing, so external-draw and downstream closed-loop results are N/A rather than missing measurements. Show-o2-7B uses separate autoregressive text and flow-based image-generation heads, while JoyAI-Image couples an MLLM to a diffusion transformer. Their generation paths do not form a single stateful image-to-answer trajectory in the evaluated architectures. Both models support C-R, which reads the generated image in a fresh pass, but C-F is not applicable to them. Protocol eligibility follows interface capability: a UMM with separate generation and understanding calls can enter the restart-based reading protocol; only a backend that preserves or reconstructs the generation state for continuation can enter the stateful continuation protocol; Incompatible interfaces are never merged into one score. The unified suite includes both base UMMs and models post-trained for interleaved reasoning (Deng et al., 2025; Gu et al., 2026; Chen et al., 2025a;b; Jin et al., 2026). The spatial-specialized group includes the remote-sensing model GeoChat-7B (Kuckreja et al., 2024).

The results tables identify the 25 evaluated models. The release manifest records the reproducibility details for each model, including its exact checkpoint or API snapshot, inference code, image and decoding settings, parser version, and supported interfaces. All completed models are included in the reported model count and task-accuracy tables under the same direct-evaluation protocol.

Within each model and backend, restart-based reading and stateful continuation use the same scene, question, view, generation instruction, seed, applicable guidance settings, resolution, generation steps, image count, decoding budget, and answer parser wherever the backend permits. Settings are matched within a model rather than forced to be numerically identical across heterogeneous architectures. Generation is image-first: the model produces exactly one task-evidence image before free-form answer text. Generation instructions request only the declared task-evidence primitive schema and prohibit answer text, option identifiers, numerical answers, and unsupported background edits. The release manifest records checkpoints, resolution, decoding, generation conditioning, re-consumption mode, and post-image prompt mode. The two re-consumption protocols are compared under matched generation settings wherever the backend permits.

We do not report a human-performance ceiling: the two reviewers' 100% agreement with GIS labels applies to retained T3/T4 images after quality screening and is an answerability check, not a timed or representative human study. Label correctness is verified exhaustively by construction: every retained reference answer is computed from the latent GIS record and released only when an inde-

pendent deterministic pass reproduces it, so verification covers all released labels rather than relying on a sampled human label audit.

Scientific-validity audit. The construction pipeline recomputes retained base labels from coordinates and road-network geometry, checks canonical-view registration, and validates each declared map-transformation action. Eligible world variants are retained only when the solver changes the intervention answer and preserves the matched-sham answer. Duplicate identifiers and scene-group assignments are checked before scoring. Because the benchmark is evaluation-only, conventional train–test leakage is not applicable; scene grouping instead prevents correlated derived rows from being treated as independent observations. These deterministic checks assess whether records realize the intended constructs and remain separate from both manual image-quality screening and the model leaderboard.

Evidence provenance and calibration. The deterministic renderer feeds either question-conditioned primitives or, for T3/T4, oracle-conditioned primitives to the same reader. The gap between oracle-conditioned and closed-loop performance measures the end-to-end oracle-realization gap, while the question-conditioned–Direct contrast isolates the effect of answer-agnostic externalization. Attributing either gap to construction, re-encoding, or stateful consumption requires generated-artifact scores and the comparison of restart-based reading with stateful continuation. Primitive provenance determines whether an evidence condition is question-conditioned, oracle-conditioned, or model-generated. Same-model restart-based reading is a generation-to-understanding baseline. Stateful continuation is reported only for models with the required architecture. Model-wise closed-loop accuracy and generated-artifact diagnostics are in Appendix L.

E Dataset Specification, Validity, and Release

E.1 Released Group Schema

Each benchmark group stores:

- **world**: entities, coordinates, road links, buildings, routes, semantics, and constraints;
- **representations**: registered satellite, controlled text-free structural-map, and conventional text-bearing map images plus protocol-eligibility masks;
- **transforms**: a declared map transformation, its task-specific label action ρ_g, and transformed pixel evidence;
- **evaluation_instances**: question, choices, answer key, images, and task metadata;
- **interventions**: solver-verified answer-changing world variants and visually matched answer-preserving sham edits for eligible sub-tasks;
- **supportive_evidence**: answer-agnostic primitives;
- **solution_visualization**: answer-conditioned primitives stored separately.

E.2 Additional Dataset Details

This section records the principal ambiguity filters and label checks; task forms and evidence targets appear in Section A.4. Distractors are hard but unambiguous: T1 avoids boundary directions, T2 enforces distance margins, T3 filters zero-count cases, and T4 uses network-infeasible route distractors. Labels are verified by deterministic answer recomputation and primitive checks. Separately, T3/T4 candidate images undergo manual visibility and rendering-quality screening, with approximate candidate retention rates of 12% and 34%, respectively. Two reviewers also answer retained images from the rendered view and match the GIS-derived labels; the GIS-derived label remains the scoring reference.

All source geometry is stored in WGS84 longitude/latitude. Metric decisions use haversine distances or a documented local small-area metre approximation, while map rendering uses the Web-Mercator tile transform. Images are rendered at zoom 18 and cropped to at least 720 pixels on each undersized dimension. Because ground sampling distance varies with latitude and crop, we do not report a

single nominal GSD; every image record stores its WGS84 boundary, crop box, canvas/output sizes, and independently computed horizontal and vertical metres per pixel.

T1 angular filters. Pairwise-direction points must be at least 20 m apart, and bearings within 15 degrees of a cardinal axis are excluded. Dual-anchor direction choices require 20 degrees of cardinal-axis clearance; egocentric side questions require 20 degrees of clearance from their local front/back and left/right axes. Angular-order candidates require at least 10 degrees between adjacent bearings.

T2 distance margins. Nearest-point questions require a first-to-second distance gap of at least 6 m and 5–25% of the nearest distance. Directional nearest-point questions require a gap of at least 20 m and 10% to the next same-direction candidate. Euclidean and road-network route comparisons require a length gap of at least 20 m and 5–25% of the shorter route; image-space nearest-point rankings are checked against the geographic ranking.

T3 deterministic counting rule. The solver keeps footprints within 50 m of a directed route, assigns their side by the local directed cross product, and uses overlap between longitudinal footprint projections to retain at least 10 m of exposed frontage. This is an analytic interval test rather than a sampled raycast. Internal endpoints use a 5 m exclusion margin; endpoint matching and spatial deduplication use 8 m and 1.5 m, respectively. Route-cumulative questions apply the same rule per segment and union stable building identifiers to avoid double counting at turns.

T4 deterministic route rule. Candidate routes are reconstructed from complete ordered road-link geometry. Turns require at least 45 degrees, waypoint tasks retain two to four turns, and the main ambiguity margins are 20 m for waypoint spacing and shortest-route length difference. For detour interventions, the solver closes a road link that belongs to the base route but not its natural alternative; the red cross marks that link visually but is not itself a metric collision region. Alternatives longer than 1.5 times the base route or with more than 12 m reverse overlap are discarded, and sham marks remain at least 30 m from both candidate routes.

E.3 Task Matrix

Table 7: MapSpatial task matrix. Controlled evidence comprises question-conditioned visual evidence and, for T3/T4 only, oracle-conditioned visual evidence; the two information levels are reported separately.

Task	Sub-task	Question form	Evidence	Answer
T1	Direction; dual-anchor; egocentric; angular order	Qualitative direction / order	Reference axes / N/A	Direction / point / order
T2	Nearest; directional; Euclidean; network	Closest target or route	Measurement geometry / N/A	Point / route
T3	Segment count; route cumulative count	Route-relative building count	Route, side buffer, candidates / counted set	Count
T4	Route validity / waypoint ordering / detour traversability / detour shortest	Validity, feasibility, or shortest path	Routes, barriers, constraints / feasible route set	Route / ordering

Every row of Table 7 is evaluated across registered representations and, where the view is eligible, with the seven prescribed map transformations. A task-native W-variant is generated only when the

sub-task admits a meaningful answer-changing intervention. In the current T4 design this applies to detour waypoint planning, not route validity or plain waypoint ordering.

Table 8: **Uniform-choice reference by sub-task (%).** The reference uses each original satellite question's option count K: $1/K$ for single-answer questions and $1/(2^K - 1)$ for nonempty-subset guessing in the multi-select traversability task. This analytic reference differs from the measured text-only control in Table 9 and does not assume the correct subset size is known.

Task	Sub-task	K	Reference
T1	angular_order	2/4	46.0
T1	direction	4	25.0
T1	dual_anchor_direction	4	25.0
T1	egocentric_side	4	25.0
T2	composite_euclidean_distance	2	50.0
T2	composite_network_distance	2	50.0
T2	directional_nearest_point	3/4/5	21.0
T2	nearest_point	4/5	20.8
T3	route_cumulative_count	4	25.0
T3	segment_building_count	4	25.0
T4	detour_waypoint_shortest	4	25.0
T4	detour_waypoint_traversability	4	6.7
T4	route_validity	4	25.0
T4	waypoint_ordering	2/4	34.5

Table 9: **Text-only control on canonical base questions (%).** Qwen2.5-VL-7B-Instruct receives the same question and answer options without map images. Satellite scores use the same question IDs and gold answers. The average weights T1–T4 equally; this single-model control probes question and option priors rather than chance performance.

Task	N	Text-only	Satellite
T1	400	44.5	59.8
T2	400	41.0	44.5
T3	200	25.5	38.0
T4	400	31.3	40.5
Macro-average	1,400	35.6	45.7

Benchmark views. Every T1–T4 question is evaluated on the registered satellite, controlled text-free structural-map, and conventional text-bearing map views. These expose implicit imagery, explicit geometry and topology, and conventional labels and icons, respectively. The marker-only blank view is restricted to an answer-equivalent T1/T2 geometry ablation; raw rotation and reflection exclude text-bearing maps. Eligible sub-tasks additionally receive structural world edits, while evidence overlays are reported by information condition.

E.4 Validity Controls

Table 10 lists the main validity threats, implemented controls, and interpretation boundaries.

Before scoring, retained instances are required to pass label recomputation, matched-view registration, declared geometric-action, world-edit, sham, and duplicate-identifier checks applicable to their condition. The benchmark is evaluation-only, so train–test split leakage is not applicable; all derived records remain grouped with their source case for evaluation.

Table 10: Validity threats, implemented checks, and interpretation boundaries in MapSpatial.

Threat	Control in MapSpatial
Evidence leakage and selection	Construct question-conditioned evidence without access to y; forbid answer text, values, option IDs, and answer-selecting marks; use candidate-complete overlays.
"Matched" views expose different required information	Verify task-required visibility per view, exclude invalid view–question pairs, and report structural/conventional gaps separately.
Models react to any visual edit	Pair answer-changing structural interventions with answer-preserving sham edits.
Scene-derived rows inflate significance	Keep all variants from one scene under one group identifier and bootstrap by geographic scene.
Generator or audit artifacts	Independently audit labels, cross-view consistency, transforms, interventions, registration, and group integrity.
Interface or evidence-use effects	Report per-task and per-interface evidence results; interpret C-R/C-F as end-to-end pathways without assigning failures to construction or consumption alone.

Released artifacts and benchmark card. The anonymous repository provides generation and evaluation code and links to the released benchmark dataset. Together, these resources include benchmark images, scene-grouped questions and labels, separate question-conditioned and oracle-conditioned evidence targets, the inference harness, metrics, and documentation. They support registration checks and reproducible model scoring.

F Prompts and Model Interfaces

F.1 Prompt Templates

All prompts append the task question and its fixed candidate list and require a single parseable answer. The frozen wording is:

> **Direct / evidence reader:** "Use the provided map image(s) to answer the question. Return only the requested option or value. Do not assume that an overlay identifies the correct candidate."
>
> **Generated scaffold:** "Before answering, create exactly one visual workspace image in the original map frame. Mark only the entities, reference frames, complete candidate set, and task constraints requested by the question. Do not write answer text, option identifiers, numeric answers, or selectively highlight a winning candidate. Then answer using the resulting image."

Question-conditioned and oracle-conditioned readers receive identical reader wording; only the supplied image differs. The benchmark uses fixed task-level reader templates: the question and scene-specific values are filled in for each example. The released materials provide these templates and renderer instructions verbatim so that wording changes cannot be conflated with evidence changes.

F.2 Model Interface and Track Eligibility

Table 11 maps interface classes to the conditions they can enter, and Table 12 lists the information available to the reader under each condition.

Table 11: Interface capability by condition. U is direct understanding; Q and O denote deterministic question-conditioned and oracle-conditioned evidence; G is model generation scored without re-consumption. C-R requires same-model generation followed by a fresh reader pass, and C-F requires stateful visual re-consumption.

Interface	**U**	**Q**	**O**	**G**	**C-R**	**C-F**
Understanding-only MLLM	✓	✓	✓	✗	✗	✗
UMM with separate gen/read paths	✓	✓	✓	✓	✓	✗
UMM with stateful continuation	✓	✓	✓	✓	✓	✓

Table 12: **Information-access boundary.** "Oracle primitives" means answer-selective geometry derived from y.

Condition	**Image**	**Question**	**Choices**	**Oracle primitives**	**Generated artifact**
U / Direct	✓	✓	✓	✗	✗
Question-conditioned	✓	✓	✓	✗	Deterministic non-selective
Oracle-conditioned	✓	✓	✓	✓	Deterministic selective
G	✓	✓	✓	✗	Model output, scored only
C-R / C-F	✓	✓	✓	✗	Model output, then consumed

G DIRECT-UNDERSTANDING RESULTS

Table 13: Track-U map-understanding accuracy on MapSpatial for the 25 evaluated models, including 17 understanding-only MLLMs and 8 unified multimodal models evaluated in direct mode without task-specific evidence. Track-U accuracy uses only the three canonical base views and a frozen input-instance identity; marker-only blank is reported separately. All models use the same canonical direct-evaluation definition. Avg is the macro-average of the four task accuracies; brackets give its scene-bootstrap 95% interval.

Model	T1	T2	T3	T4	Avg [95% CI]
Open-weight general					
Qwen3-VL-8B-Instruct	59.8	53.0	32.5	50.3	48.9 [46.40, 51.60]
Qwen2.5-VL-7B-Instruct	57.8	42.8	34.8	41.0	44.1 [41.90, 46.30]
Qwen2-VL-7B-Instruct	42.8	32.8	32.0	30.2	34.4 [32.20, 36.80]
InternVL3-8B-Instruct	56.8	55.4	29.5	31.9	43.4 [41.10, 45.90]
InternVL3.5-8B	43.5	54.6	30.8	28.7	39.4 [37.50, 41.30]
GLM-4.6V-Flash	70.0	49.2	27.8	46.2	48.3 [46.40, 50.20]
Step3-VL-10B	69.0	43.9	35.8	33.2	45.5 [43.50, 47.50]
Open-weight spatial-specialized					
Spatial-MLLM	39.8	50.3	24.2	26.7	35.2 [33.20, 37.30]
Cambrian-S-7B-LFP	27.2	29.8	20.8	24.3	25.5 [23.30, 27.80]
ViLaSR	51.2	52.5	39.5	30.7	43.5 [41.30, 45.70]
SenseNova-SI-1.3-Qwen3-VL-8B	62.0	47.1	28.0	49.4	46.6 [44.40, 49.00]
SenseNova-SI-1.5-InternVL3-8B	63.3	48.8	32.8	30.1	43.8 [41.50, 46.10]
MiMo-Embodied-7B	67.1	52.2	32.0	39.8	47.8 [45.60, 50.00]
GeoChat-7B	40.2	38.4	25.2	19.8	30.9 [28.80, 33.10]
Scale references					
Qwen3-VL-32B-Instruct	67.8	58.1	45.5	58.2	57.4 [55.20, 59.70]
Qwen3.5-Plus	74.7	69.8	43.0	72.0	64.9 [62.30, 67.30]
Gemini-3-Flash	99.6	87.1	78.3	80.4	86.3 [83.50, 89.00]
Unified multimodal					
ThinkMorph-7B	65.4	41.8	28.3	34.0	42.4 [40.10, 44.60]
Bagel-7B-MoT	70.2	43.5	25.8	35.7	43.8 [41.60, 46.20]
JoyAI-Image	48.6	52.1	38.5	43.7	45.7 [43.20, 48.20]
SenseNova-U1-8B-MoT	48.1	45.3	37.8	35.0	41.6 [39.20, 44.00]
LatentUM-Base	52.2	34.5	35.2	30.9	38.2 [35.80, 40.60]
Show-o2-7B	56.6	40.4	42.2	24.2	40.9 [38.70, 43.00]
BLIP3o-8B	45.4	40.2	26.5	33.8	36.5 [34.20, 38.80]
Janus-Pro-7B	31.7	35.0	24.8	10.9	25.6 [23.80, 27.50]

G.1 Marker-only Geometry Ablation

The marker-only blank view is a restricted geometry ablation rather than a main benchmark representation. It retains the point markers used in T1 and T2 but removes the roads, building footprints, and other map structure needed by T3 and T4. We therefore report it only for the two geometry-focused task families (Table 14) and do not include it in the primary RQ1 score.

Table 14: **Marker-only blank ablation.** Direct accuracy on T1 and T2 for all 25 evaluated models, using blank views that contain point markers but no full map structure. Avg is the macro-average of T1 and T2. These results are diagnostic and are not pooled with the three canonical views in Table 13.

Model	T1	T2	Avg
Open-weight general			
Qwen3-VL-8B-Instruct	58.8	52.2	55.5
Qwen2.5-VL-7B-Instruct	60.8	48.5	54.6
Qwen2-VL-7B-Instruct	49.5	45.0	47.2
InternVL3-8B-Instruct	60.0	59.0	59.5
InternVL3.5-8B	27.8	56.7	42.2
GLM-4.6V-Flash	73.0	60.2	66.6
Step3-VL-10B	81.5	52.5	67.0
Open-weight spatial-specialized			
Spatial-MLLM	49.8	53.2	51.5
Cambrian-S-7B-LFP	28.2	28.8	28.5
ViLaSR	56.8	56.2	56.5
SenseNova-SI-1.3-Qwen3-VL-8B	65.2	41.8	53.5
SenseNova-SI-1.5-InternVL3-8B	64.8	52.8	58.8
MiMo-Embodied-7B	70.5	58.5	64.5
GeoChat-7B	40.1	38.7	39.4
Scale references			
Qwen3-VL-32B-Instruct	69.8	56.2	63.0
Qwen3.5-Plus	72.2	65.2	68.8
Gemini-3-Flash	100.0	76.2	88.1
Unified multimodal			
ThinkMorph-7B	66.0	49.0	57.5
Bagel-7B-MoT	73.5	48.5	61.0
JoyAI-Image	57.2	49.0	53.1
SenseNova-U1-8B-MoT	61.2	52.5	56.9
LatentUM-Base	53.0	38.8	45.9
Show-o2-7B	58.8	41.8	50.3
BLIP3o-8B	45.0	33.2	39.1
Janus-Pro-7B	34.5	38.5	36.5

H Registered-View Accuracy and Contrasts

Tables 17 and 18 report the canonical Direct results for all 25 models. Each view score is the equal mean of T1–T4 accuracies; each contrast pairs the same question instances across views before averaging over tasks. Brackets are 95% scene-bootstrap intervals. Table 16 decomposes the six anchors' joint correctness and prediction disagreement by task family.

Table 15: **Tri-view robustness for the six category anchors shown in Table 3 (%).** A_{tri} is joint correctness across all three registered views; U_{tri} is prediction disagreement. Brackets give 95% percentile intervals from 10,000 scene-level bootstrap replicates over 1,400 matched triplets and 585 scenes per model.

Model	A_{tri} [95% CI]	U_{tri} [95% CI]
GLM-4.6V-Flash	32.9 [30.20, 35.40]	52.6 [50.00, 55.30]
Qwen2.5-VL-7B-Instruct	28.4 [26.00, 30.80]	46.5 [43.70, 49.20]
MiMo-Embodied-7B	31.1 [28.50, 33.60]	51.2 [48.80, 53.70]
Spatial-MLLM	20.6 [18.50, 22.60]	52.1 [49.40, 54.70]
Bagel-7B-MoT	35.5 [32.90, 38.20]	33.8 [31.40, 36.10]
SenseNova-U1-8B-MoT	20.3 [18.10, 22.50]	60.7 [57.90, 63.40]

Table 16: **Tri-view robustness by task family for the six anchors (%).** Each row uses complete matched canonical triplets: 400 for T1, 400 for T2, 200 for T3, and 400 for T4. Brackets give 95% percentile intervals from 10,000 scene-level bootstrap replicates within each family.

Model	**Task**	N	A_{tri} [95% CI]	U_{tri} [95% CI]
GLM-4.6V-Flash	T1	400	59.0 [53.9, 63.9]	26.8 [22.4, 31.4]
GLM-4.6V-Flash	T2	400	23.0 [19.0, 27.1]	68.2 [63.5, 72.9]
GLM-4.6V-Flash	T3	200	8.5 [4.9, 12.8]	66.0 [58.7, 73.1]
GLM-4.6V-Flash	T4	400	28.8 [24.5, 33.2]	56.2 [51.4, 61.1]
Qwen2.5-VL-7B-Instruct	T1	400	41.0 [36.3, 45.7]	43.5 [38.9, 48.2]
Qwen2.5-VL-7B-Instruct	T2	400	20.2 [16.2, 24.5]	62.5 [57.6, 67.2]
Qwen2.5-VL-7B-Instruct	T3	200	22.0 [16.4, 28.2]	34.5 [26.6, 42.1]
Qwen2.5-VL-7B-Instruct	T4	400	27.0 [22.9, 31.3]	39.5 [34.5, 44.4]
MiMo-Embodied-7B	T1	400	47.5 [42.4, 52.6]	44.8 [40.2, 49.4]
MiMo-Embodied-7B	T2	400	35.5 [31.0, 40.1]	43.5 [39.4, 47.5]
MiMo-Embodied-7B	T3	200	8.0 [4.4, 12.1]	66.0 [59.5, 72.9]
MiMo-Embodied-7B	T4	400	21.8 [17.8, 25.9]	58.0 [53.0, 62.9]
Spatial-MLLM	T1	400	18.5 [14.9, 22.3]	58.0 [53.6, 62.6]
Spatial-MLLM	T2	400	34.0 [29.4, 38.8]	47.8 [43.2, 52.1]
Spatial-MLLM	T3	200	13.5 [9.3, 18.1]	42.0 [34.9, 49.5]
Spatial-MLLM	T4	400	12.8 [9.5, 16.3]	55.5 [50.4, 60.8]
Bagel-7B-MoT	T1	400	58.5 [53.9, 63.1]	31.0 [27.0, 35.1]
Bagel-7B-MoT	T2	400	32.8 [28.1, 37.5]	27.2 [23.0, 31.3]
Bagel-7B-MoT	T3	200	11.5 [6.9, 16.8]	50.5 [42.2, 58.5]
Bagel-7B-MoT	T4	400	27.2 [22.9, 31.7]	34.8 [29.9, 39.6]
SenseNova-U1-8B-MoT	T1	400	20.5 [16.7, 24.4]	65.0 [60.4, 69.5]
SenseNova-U1-8B-MoT	T2	400	20.2 [16.3, 24.6]	64.8 [59.8, 69.6]
SenseNova-U1-8B-MoT	T3	200	27.5 [20.8, 34.2]	32.0 [25.3, 38.7]
SenseNova-U1-8B-MoT	T4	400	16.5 [12.9, 20.3]	66.8 [61.9, 71.3]

Table 17: **Accuracy by registered view (%).** Canonical satellite (Sat), text-free structural (Struct), and conventional text-bearing (Conv) views. Brackets give scene-bootstrap 95% intervals.

Model	Sat [95% CI]	Struct [95% CI]	Conv [95% CI]
Qwen3-VL-8B-Instruct	48.9 [46.10, 51.90]	48.7 [45.90, 51.70]	49.1 [46.20, 52.00]
Qwen2.5-VL-7B-Instruct	45.7 [43.10, 48.30]	41.9 [39.40, 44.70]	44.7 [42.10, 47.30]
Qwen2-VL-7B-Instruct	33.1 [30.50, 35.70]	35.8 [33.20, 38.50]	34.4 [31.90, 37.10]
InternVL3-8B-Instruct	42.4 [39.80, 45.10]	44.1 [41.50, 47.00]	43.8 [41.00, 46.70]
InternVL3.5-8B	40.1 [37.30, 42.90]	39.1 [36.40, 41.90]	39.0 [36.30, 41.90]
GLM-4.6V-Flash	47.7 [45.10, 50.40]	46.9 [44.50, 49.30]	50.3 [47.90, 52.70]
Step3-VL-10B	49.4 [46.50, 52.20]	48.2 [45.60, 50.80]	38.8 [36.00, 41.70]
Spatial-MLLM	36.0 [33.50, 38.60]	35.8 [33.10, 38.40]	33.9 [31.60, 36.40]
Cambrian-S-7B-LFP	25.8 [23.30, 28.20]	24.4 [22.10, 26.90]	26.4 [23.80, 29.20]
ViLaSR	44.2 [41.60, 46.80]	43.9 [41.20, 46.50]	42.4 [39.60, 45.20]
SenseNova-SI-1.3-Qwen3-VL-8B	48.2 [45.70, 50.90]	46.6 [43.90, 49.40]	45.0 [42.50, 47.60]
SenseNova-SI-1.5-InternVL3-8B	40.9 [38.10, 43.80]	46.4 [43.50, 49.10]	44.0 [41.20, 46.90]
MiMo-Embodied-7B	48.3 [45.60, 51.10]	49.1 [46.20, 52.00]	45.9 [43.40, 48.70]
GeoChat-7B	30.9 [28.70, 33.10]	31.1 [28.70, 33.60]	30.6 [28.30, 33.00]
Qwen3-VL-32B-Instruct	57.8 [55.10, 60.60]	58.7 [56.00, 61.50]	55.7 [53.10, 58.30]
Qwen3.5-Plus	62.8 [59.90, 65.60]	69.4 [66.10, 72.50]	62.3 [59.50, 65.20]
Gemini-3-Flash	87.2 [83.30, 90.70]	91.2 [87.70, 94.50]	80.6 [76.60, 84.30]
ThinkMorph-7B	42.8 [40.00, 45.60]	39.8 [37.40, 42.20]	44.6 [41.70, 47.40]
Bagel-7B-MoT	42.9 [40.10, 45.80]	44.1 [41.60, 46.60]	44.5 [41.80, 47.20]
JoyAI-Image	46.1 [43.30, 49.10]	46.2 [43.30, 49.20]	44.8 [41.90, 47.80]
SenseNova-U1-8B-MoT	43.4 [40.50, 46.30]	42.3 [39.50, 45.20]	39.0 [36.20, 41.90]
LatentUM-Base	37.3 [34.50, 40.10]	38.8 [36.10, 41.60]	38.4 [35.50, 41.40]
Show-o2-7B	38.8 [36.30, 41.40]	43.2 [40.80, 45.80]	40.5 [38.20, 42.90]
BLIP3o-8B	37.2 [34.60, 39.90]	36.9 [34.30, 39.60]	35.2 [32.50, 37.90]
Janus-Pro-7B	26.1 [23.80, 28.40]	25.4 [23.00, 28.00]	25.3 [23.00, 27.70]

Table 18: **Paired registered-view contrasts (percentage points).** Differences use matched question instances; brackets give scene-bootstrap 95% intervals. These contrasts combine changes in visual encoding and do not isolate OCR.

Model	Struct–Sat [95% CI]	Conv–Struct [95% CI]
Qwen3-VL-8B-Instruct	-0.3 [-2.37, 1.86]	0.4 [-1.87, 2.66]
Qwen2.5-VL-7B-Instruct	-3.8 [-6.14, -1.39]	2.8 [0.12, 5.32]
Qwen2-VL-7B-Instruct	2.8 [0.60, 4.94]	-1.4 [-3.89, 1.05]
InternVL3-8B-Instruct	1.8 [-0.44, 4.02]	-0.4 [-2.73, 2.02]
InternVL3.5-8B	-0.9 [-4.56, 2.58]	-0.1 [-3.61, 3.63]
GLM-4.6V-Flash	-0.8 [-3.49, 1.71]	3.4 [0.70, 6.25]
Step3-VL-10B	-1.1 [-4.46, 2.16]	-9.4 [-12.76, -6.05]
Spatial-MLLM	-0.3 [-2.59, 2.11]	-1.8 [-4.52, 0.81]
Cambrian-S-7B-LFP	-1.3 [-3.29, 0.72]	2.0 [0.02, 3.99]
ViLaSR	-0.3 [-2.77, 2.15]	-1.5 [-4.44, 1.45]
SenseNova-SI-1.3-Qwen3-VL-8B	-1.6 [-4.11, 0.82]	-1.6 [-3.92, 0.72]
SenseNova-SI-1.5-InternVL3-8B	5.5 [2.62, 8.32]	-2.4 [-5.06, 0.46]
MiMo-Embodied-7B	0.8 [-2.41, 3.93]	-3.1 [-6.00, -0.34]
GeoChat-7B	0.3 [-1.50, 2.03]	-0.5 [-2.20, 1.13]
Qwen3-VL-32B-Instruct	0.9 [-1.79, 3.44]	-3.0 [-5.43, -0.67]
Qwen3.5-Plus	6.6 [3.70, 9.41]	-7.1 [-10.03, -3.99]
Gemini-3-Flash	4.1 [0.59, 7.71]	-10.6 [-15.14, -6.18]
ThinkMorph-7B	-3.1 [-5.75, -0.28]	4.8 [2.14, 7.39]
Bagel-7B-MoT	1.2 [-1.10, 3.39]	0.4 [-1.82, 2.72]
JoyAI-Image	0.1 [-2.59, 2.81]	-1.5 [-3.94, 1.05]
SenseNova-U1-8B-MoT	-1.1 [-3.71, 1.66]	-3.3 [-6.13, -0.53]
LatentUM-Base	1.5 [-0.92, 3.89]	-0.4 [-3.04, 2.31]
Show-o2-7B	4.4 [2.07, 6.75]	-2.8 [-4.76, -0.65]
BLIP3o-8B	-0.3 [-2.82, 2.16]	-1.8 [-4.18, 0.68]
Janus-Pro-7B	-0.6 [-3.58, 2.09]	-0.1 [-2.79, 2.60]

I Extended Metrics and Diagnostic Analyses

We report paired gains for restart-based reading and, where supported, stateful continuation against Direct. These compare complete generation-and-reading pathways, not the isolated effect of consuming a generated artifact. For geometric transforms, answer-level evaluation uses the paired $P_{\text{transform}}$ and conditional $C_{\text{same-answer}}$ diagnostics of Appendix C. Generated primitives are compared over matched declared pixel-space targets after applying the same geometric action. Semantic outputs are likewise compared after the task-specific label action ρ_g; no separate answer-invariant raster-rotation protocol is reported.

Reported diagnostics include tri-view representation comparisons, Direct-versus-evidence contrasts, and the original/sham/answer-changing structure intervention.

I.1 Protocol Interpretation and Limitations

The three registered views are not information-identical: imagery can expose details absent from vector maps, the controlled structural map makes selected geometry explicit, and the conventional map adds labels, icons, and provider-specific styling. Counterfactual map edits do not simulate all real navigation dynamics. Geographic coverage, imagery licensing, map-source bias, and urban-form imbalance limit global claims, while map-coordinate tolerances and rendering choices affect primitive metrics and are recorded for sensitivity analysis.

Visual generation architectures also differ substantially. Some models edit an input image, some generate a new canvas, and some cannot preserve generation state when re-consuming their output. We therefore report same-model restart and stateful continuation separately, with eligibility determined by each model's generation and reading paths. Image-first stateful continuation may be less natural for models trained to verbalize before drawing. Target-type audits verify the declared information stratum and forbid final-answer serialization, but cannot eliminate all implicit answer leakage. The deterministic GIS solver and renderer are calibration tools, not competing foundation models.

I.2 Paired Transform Diagnostics

Table 19 reports the paired transform metrics behind the RQ2 columns of Table 3.

Table 19: **Paired transform diagnostics (%).** Rows are the six anchors of Table 3. Marginal base and transformed accuracy are reported beside joint correctness, equivariance retention, unchanged-reference-answer agreement, and joint correctness on the base plus all seven transformed variants. Base+7|base conditions the latter on a correct base answer within complete eight-way groups. The final values pool two transform-eligible canonical views and seven actions (19,600 view-specific pairs per complete model), excluding the 5,600 marker-only ablation pairs.

Model	A_{base}	A_{trans}	$P_{\text{transform}}$	R_{eq}	$C_{\text{same-answer}}$	**Base+7**	**Base+7|base**
GLM-4.6V-Flash	50.2	54.5	35.4	70.5	54.1	16.5	32.9
Qwen2.5-VL-7B-Instruct	45.4	45.6	29.8	65.7	59.5	13.4	29.5
MiMo-Embodied-7B	50.9	50.8	36.0	70.8	59.1	16.8	33.0
Spatial-MLLM	37.5	38.1	25.2	67.2	60.7	11.2	29.8
Bagel-7B-MoT	46.2	44.4	36.5	78.9	75.1	22.8	49.3
SenseNova-U1-8B-MoT	43.2	40.7	28.7	66.3	61.6	13.3	30.7

Table 20: **Equivariance retention by action type (%).** R_{eq} is the proportion of initially correct base predictions that remain correct after the paired action. Rotations pool three actions (8,400 pairs per model); reflections pool the horizontal mirror and its three rotated forms (11,200 pairs per model). Both groups cover T1–T4 on the satellite and structural views. Brackets are 95% intervals from 10,000 scene-grouped bootstrap samples; Δ is rotation minus reflection, resampled on the same scenes.

Model	Rotations	Reflections	Δ
GLM-4.6V-Flash	70.3 [68.0, 72.6]	70.6 [68.4, 72.8]	-0.3 [-2.0, 1.3]
Qwen2.5-VL-7B-Instruct	65.9 [63.0, 68.8]	65.5 [63.0, 68.1]	0.4 [-1.8, 2.5]
MiMo-Embodied-7B	72.3 [70.2, 74.2]	69.7 [67.5, 71.8]	2.6 [1.2, 4.1]
Spatial-MLLM	67.6 [65.4, 69.9]	66.9 [64.6, 69.3]	0.7 [-1.1, 2.5]
Bagel-7B-MoT	81.9 [79.9, 83.9]	76.7 [74.2, 79.2]	5.2 [3.2, 7.3]
SenseNova-U1-8B-MoT	70.7 [68.3, 73.1]	63.0 [60.5, 65.7]	7.6 [5.2, 10.0]

J Confidence Intervals and Statistical Audit

The main tables retain their point estimates for readability. This appendix reports the corresponding scene-level uncertainty estimates without replacing the primary results. We use a percentile bootstrap with 10,000 replicates and a fixed seed for reproducibility. The resampling unit is the geographic scene: when a scene is sampled, its canonical views, task instances, and applicable transform or intervention variants are sampled together. This preserves the dependence induced by shared latent geometry and questions. Direct-understanding task scores and averages use the same task weighting as the main tables; paired transform and intervention metrics resample their matched pairs jointly.

World-edit eligibility compares the semantic gold-answer content in matched evaluation-input records across base, sham, and intervention conditions; shuffled option letters alone are not comparable across conditions.

The combined Direct table, Table 13, reports the four task scores and the scene-bootstrap interval for each macro-average. Per-representation intervals and contrasts appear in Tables 17 and 18.

Table 21: **Matched world-intervention results (%).** Each anchor model uses the same 1,200 answer-changing T3/T4 base–sham–intervention triples from 248 scenes (600 T3 and 600 T4 pairs). Eligibility is checked by matching each result to its evaluation-input record and comparing the semantic content of gold options across conditions, not shuffled multiple-choice letters or the coverage of a local benchmark copy. Base, sham, edited accuracy, S_{cf}, and S_{up} all use this eligible set; N_b is the number of base-correct pairs. Brackets give paired scene-bootstrap 95% intervals.

Model	N	N_b	Base$_E$	Sham$_E$	Edited$_E$
GLM-4.6V-Flash	1200	396	33.0	31.7	27.3
Qwen2.5-VL-7B-Instruct	1200	449	37.4	34.9	27.8
MiMo-Embodied-7B	1200	401	33.4	28.1	33.3
Spatial-MLLM	1200	250	20.8	20.3	23.7
Bagel-7B-MoT	1200	332	27.7	26.2	31.6
SenseNova-U1-8B-MoT	1200	407	33.9	32.6	28.3

Model	S_{cf} [95% CI]	S_{up} [95% CI]
GLM-4.6V-Flash	6.8 [5.40, 8.40]	20.7 [16.40, 25.20]
Qwen2.5-VL-7B-Instruct	6.3 [4.30, 8.50]	16.9 [11.70, 22.30]
MiMo-Embodied-7B	8.8 [6.80, 10.80]	26.2 [21.10, 31.30]
Spatial-MLLM	2.1 [1.20, 3.10]	10.0 [6.00, 14.50]
Bagel-7B-MoT	3.3 [2.00, 4.70]	11.8 [7.30, 16.70]
SenseNova-U1-8B-MoT	3.3 [2.00, 4.50]	9.6 [6.10, 13.20]

K STRONG-MODEL COMPARISON AND WORLD-EDIT AUDIT

Table 4 in Section 6 compares the anchors, Qwen3-VL-32B-Instruct, and Gemini-3-Flash on direct, transform, and world-edit results; its world columns use each model's full eligible set, as in Table 21. Table 22 decomposes how the same models change their predictions after world edits.

Table 22: **World-edit prediction changes (%).** Rows are the six anchors followed by two scale references. The first two columns compare observed paired correctness with the descriptive product of marginal base and edited accuracy on each model's own eligible set. Flip rates compare parsed semantic answers; the final columns partition base-correct pairs. A stale answer is the same option content as before the edit, not merely the same shuffled choice letter; "Other wrong" comprises changed-but-incorrect edited answers.

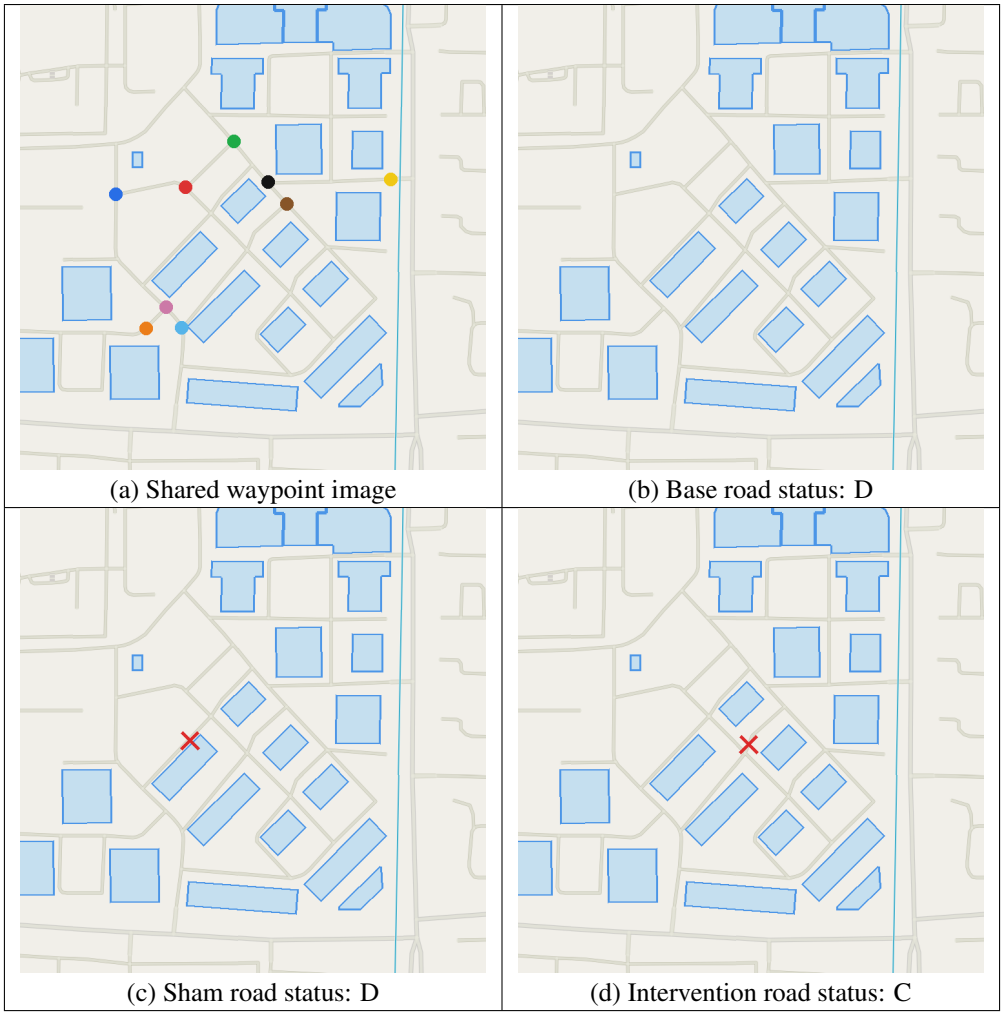
Model	S_{cf}	Indep. product	Edit flip	Sham flip
GLM-4.6V-Flash	6.8	9.0	48.7	47.5
Qwen2.5-VL-7B-Instruct	6.3	10.4	26.6	20.0
MiMo-Embodied-7B	8.8	11.1	50.4	48.6
Spatial-MLLM	2.1	4.9	22.8	18.6
Bagel-7B-MoT	3.3	8.7	23.2	18.3
SenseNova-U1-8B-MoT	3.3	9.6	22.8	20.4
Qwen3-VL-32B-Instruct	4.8	11.8	22.8	17.9
Gemini-3-Flash	37.5	37.2	60.6	39.7

Model	Correct update	Stale	Other wrong
GLM-4.6V-Flash	20.7	57.6	21.7
Qwen2.5-VL-7B-Instruct	16.9	69.5	13.6
MiMo-Embodied-7B	26.2	50.4	23.4
Spatial-MLLM	10.0	76.4	13.6
Bagel-7B-MoT	11.8	78.6	9.6
SenseNova-U1-8B-MoT	9.6	84.3	6.1
Qwen3-VL-32B-Instruct	10.1	83.6	6.2
Gemini-3-Flash	50.4	35.1	14.6

The independence product is a descriptive reference, not a null hypothesis: shared question difficulty can correlate success on both sides of a pair. Incorrect edited answers are frequently semantically unchanged from the correct base answer. Sham flips also occur without a solver-answer change, so flips alone do not establish correct updating.

K.1 Illustrative T4 World-Edit Case

One released detour-waypoint-shortest case shares one question and one option set across its base, sham, and intervention conditions: "Two registered map images describe the same area. The road-status image shows whether any road is closed, and the waypoint image shows colored checkpoints without drawing complete routes. Starting at the yellow point and ending at the orange point, which ordered color sequence describes the shortest traversable route under the current road state? Include the start point as the first item, but do not include the end point." The options are (A) yellow → sky_blue → pink → black → brown; (B) yellow → blue → green → black → red → pink; (C) yellow → black → green → red → blue → pink; (D) yellow → black → brown → sky_blue → pink. The solver-verified reference answer is D in the base and sham, and C after the closure. Qwen3-VL-32B-Instruct answers D in all three conditions on the satellite, structural, and conventional views alike: correct before the closure and on the sham, but stale after the answer changes—an illustrative failure, not a frequency estimate.



(a) Shared waypoint image (b) Base road status: D

(c) Sham road status: D (d) Intervention road status: C

Table 23: Structural-map rendering of the illustrative detour-waypoint-shortest case. (a) The shared waypoint image is unchanged across conditions; (b)–(d) the road-status image in the base, sham, and intervention conditions, with the solver-verified reference answer shown after each label. The red cross moves from a task-irrelevant location in the sham to a link of the base route in the intervention.

These are direct inputs without oracle route highlights.

L Model-Generated Evidence and Closed-Loop Results

L.1 Generated-Artifact Quality and Failure Modes

We audited the generated visual artifacts for the four models with both restart-based and stateful-continuation results: Bagel-7B-MoT, LatentUM-Base, SenseNova-U1-8B-MoT, and ThinkMorph-7B. The audit covers 57,745 generated artifacts, all of which were present on disk and decodable. It combines automated measurements of background preservation, marker completeness, near-blank outputs, blue-evidence pixels, and geometric alignment with a 56-image manual audit (40 stratified and 16 targeted examples). The manual audit was performed by one reviewer and is used for qualitative diagnosis rather than a population-level estimate.

The models exhibit distinct artifact patterns. ThinkMorph-7B nearly copies the input map (median normalized cross-correlation 0.995), but explicit evidence is often absent or misplaced; only half of the stratified examples were judged to contain adequate evidence. Bagel-7B-MoT frequently produces schematic redraws in restart-based reading, while its stateful outputs more often preserve the background and add colored marks. SenseNova-U1-8B-MoT produces usable evidence in many cases but often uses orange, white, or yellow instead of the requested blue, so blue-only automatic measures underestimate its evidence. LatentUM-Base frequently produces hallucinated UI panels, text grids, or near-blank images; the targeted manual sample found hallucinated content in 60% of examples and text-like leakage in 50%.

On the matched three-protocol cases (Table 25), restart-based reading is not uniformly better than Direct: the paired macro differences range from -2.1 to $+2.2$ points across models, while stateful continuation is lower than restart-based reading for three of the four models (by 1.5–12.5 points) and higher for ThinkMorph-7B by 2.1 points. A separate artifact audit suggests that artifact quality alone does not explain these answer failures: among cases where Direct is correct but restart-based reading is wrong, 10.1% were flagged as artifact-bad, compared with 11.2% among cases where both were correct. This pattern suggests that evidence consumption remains an important bottleneck, although it does not identify a causal mechanism.

L.2 Evidence-Condition Accuracy

Table 24: **Parallel RQ4 evidence conditions (%).** Each row uses the same question–view keys across the available conditions: Direct, question-conditioned, oracle-conditioned, C-R, and C-F. Question-conditioned evidence is answer-agnostic; oracle-conditioned evidence is answer-selective and deliberately N/A for T1/T2. C-R and C-F use model-generated evidence with different re-consumption protocols.

Model	Task	Direct	Question-conditioned	Oracle-conditioned	C-R	C-F
Bagel	T1	70.3	72.0	N/A	65.6	64.3
Bagel	T2	43.5	45.1	N/A	45.8	46.3
Bagel	T3	25.8	27.0	30.0	27.0	22.7
Bagel	T4	37.1	32.6	45.4	34.5	25.4
ThinkMorph	T1	65.7	69.6	N/A	56.5	66.7
ThinkMorph	T2	41.8	45.5	N/A	47.7	48.8
ThinkMorph	T3	28.3	29.0	30.2	25.3	30.5
ThinkMorph	T4	34.7	32.3	44.4	37.5	29.4
SenseNova-U1	T1	48.3	51.4	N/A	50.9	40.4
SenseNova-U1	T2	45.3	43.8	N/A	44.9	33.1
SenseNova-U1	T3	37.1	37.1	37.3	46.2	27.9
SenseNova-U1	T4	37.8	30.3	37.1	35.2	25.9
LatentUM	T1	52.2	57.8	N/A	43.2	46.3
LatentUM	T2	34.5	34.1	N/A	39.1	43.6
LatentUM	T3	35.2	35.5	35.2	32.5	28.2
LatentUM	T4	30.9	29.8	34.8	29.5	20.4

Table 25: Model-generated-evidence evaluation for the 8-model unified cohort. C-R is available for 6 models; C-F is available for the four models whose architectures support stateful visual re-consumption. Show-o2-7B and JoyAI-Image support C-R but not C-F. Direct uses no task-specific evidence. C-R uses a same-model restart, while C-F preserves the stateful generation path. For each model, reported differences use the same question–view keys as their corresponding Direct score and equal task weighting. Unsupported protocols are N/A; their Direct scores use the full Track-U set. Deltas use unrounded scores. Evidence geometry and leakage require separate artifact checks.

Model	**Direct**	**C-R**	$\boldsymbol{\Delta}_{\mathrm{R}}$	**C-F**	$\boldsymbol{\Delta}_{\mathrm{F}}$
Bagel-7B-MoT	44.2	43.2	-0.9	39.7	-4.5
ThinkMorph-7B	42.6	41.8	-0.9	43.8	+1.2
BLIP3o-8B	36.5	N/A	N/A	N/A	N/A
SenseNova-U1-8B-MoT	42.1	44.3	+2.2	31.8	-10.3
LatentUM-Base	38.2	36.1	-2.1	34.6	-3.6
Janus-Pro-7B	25.6	N/A	N/A	N/A	N/A
Show-o2-7B	40.9	38.9	-2.0	N/A	N/A
JoyAI-Image	45.7	45.5	-0.3	N/A	N/A