CARTA: A TRACEABLE BENCHMARK FOR CROSS-REPRESENTATION MAP UNDERSTANDING IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same place can be shown as satellite imagery, a street map, a land-cover map, a contour sheet, or a weather chart. Reading across these visual languages is a basic requirement for models that navigate, monitor, or respond in the physical world, yet existing benchmarks rarely change the representation while holding the geographic referent fixed. We introduce **CARTA**, a benchmark that asks what breaks when the place stays the same but its representation changes. CARTA derives images, questions, answers, and distractors from shared geospatial state, so each of its 16,408 items is traceable from source data through rendering, audit, and frozen image bytes. The items form matched comparison families for representation profiling, continuous-field reading, and cross-representation correspondence, together with breadth and grounding tracks. Evaluating twelve MLLMs against a matched human reference reveals failures at three successive stages. Evidence: human accuracy on forest and cropland/bare targets falls to 15–45% on street maps, and an objective purity screen shrinks the human satellite–minimalist gap from 20.0 to 8.6 pp. Correspondence: models match categories more readily than places, as cross-class distractors inflate correspondence accuracy by up to 22 pp (Qwen3-VL-8B: 34.7% to 56.9%), and every model loses 3.5–14.4 pp when the two views are not aligned. Geometry: on point grounding, open-weight models do no better than copying the source mark's coordinates (17.6% versus 17.5%), and even GPT-5.5 reaches only 28.9%, against 88.3% for readers on a matched subset. On matched multiple-choice items GPT-5.5 scores 74.4% against the readers' 83.3%. Reasoning-enabled serving lifts two open-weight models from about 61% to 71% on MCQs but provides no reliable benefit for geometric grounding. CARTA thus provides a matched test bed for distinguishing representation-specific success from transferable geographic understanding. Data and code will be released after acceptance.

1 INTRODUCTION

Maps are among the most widely used visual interfaces to the physical world, and agents that plan routes, monitor land use, or interpret a weather briefing must read them in whichever form they arrive. A single place can be rendered as photographic texture, vector topology, symbolic colour, or quantitative contours; cartography treats these as distinct visual languages built from different visual variables (MacEachren, 1995; Roth, 2017). For a multimodal large language model (MLLM), reading across them means connecting markedly different pixels to the same geographic referent (Figure 1).

Existing benchmarks approach this capability only partially. Collected-map benchmarks evaluate navigation and document reasoning over real maps, but each map arrives in one representation (Xing et al., 2025; Bhat et al., 2026; Pyo et al., 2026; Chen et al., 2026b). Rendered thematic-map benchmarks vary visual design, but within a single map family such as choropleths (Chang et al., 2022; Mukhopadhyay et al., 2025; Srivastava et al., 2025; Ung et al., 2025). General perception suites probe visual encodings and spatial reasoning without a geographic referent (Fu et al., 2024; Yang et al., 2025; Mukherjee et al., 2026; Liu et al., 2025). In none of these settings can a fixed real-world place be re-rendered across photographic and symbolic idioms, so a change of representation cannot

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.

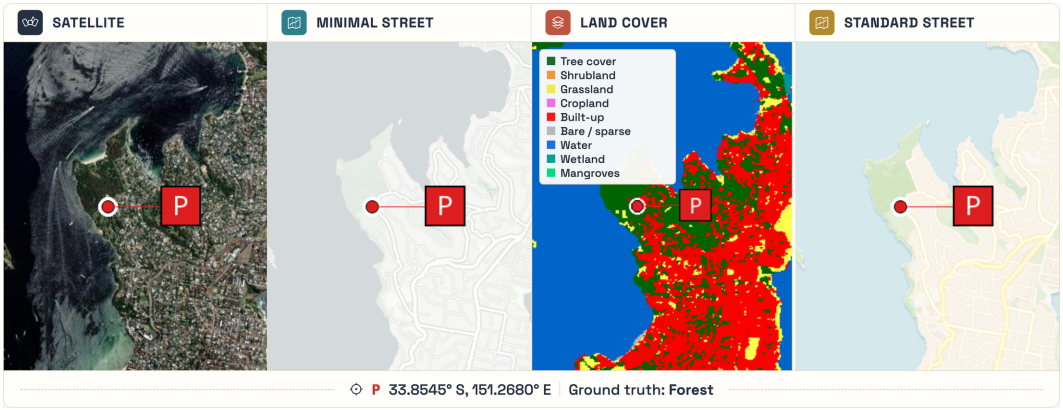

Figure 1: **One place, four visual languages.** C1 fixes the coordinate, question, and land-cover target; the representation and the class evidence it displays change together. Markers are enlarged for print clarity.

be separated from a change of place. **CARTA asks a single question: what breaks when the place is held fixed but its representation changes?**

We answer it by construction rather than collection. CARTA synthesises visual evidence, question text, ground truth, distractors, and matched relations from shared geospatial state (Johnson et al., 2017; Greff et al., 2022). Ground truth is therefore traceable to its source, and matched slices scale without per-view relabelling (Gebru et al., 2021; Pushkarna et al., 2022). Three comparison families structure the benchmark. **C1** profiles land-cover answers across four representations of the same coordinate; **C2** reads terrain and weather fields while crossing the physical field with its visual encoding; and **C3** matches a referent across representations under same- or cross-class distractors and aligned or unaligned framing. A breadth track covers direction, multilingual label-relative reasoning, and ranking, and a separate track evaluates point and region grounding. Because programmatic scale can amplify construction errors (Northcutt et al., 2021), candidates are accepted only after automated and human audits.

Evaluating ten open-weight and two proprietary MLLMs against a matched non-expert reference reveals failures at three successive stages. *Evidence: visible evidence is class-dependent.* Human accuracy on forest and cropland/bare targets falls to 15–45% on street maps, and an objective purity screen reduces the satellite–minimalist gap from 20.0 to 8.6 pp; the same screen raises all twelve models by 8–14 pp without producing a shared idiom preference. *Correspondence: models match categories more readily than places.* Cross-class distractors raise Qwen3-VL-8B's C3 accuracy from 34.7% to 56.9%, and every model loses 3.5–14.4 pp under unaligned framing. *Geometry: recognition does not guarantee precise transfer.* Copying the source mark's normalised coordinates without looking at the target reaches 17.5% point accuracy, matching the best open-weight model (17.6%); GPT-5.5 reaches 28.9%. On the copy-hard slice (a post-hoc diagnostic), open-weight models reach at most 4.1% and GPT-5.5 reaches 15.5%; readers reach 87.5% on the matched 20-item subset. Reasoning lifts two models' MCQ macro from about 61% to 71% but yields no reliable grounding gain.

Our contributions are:

1. **A cross-representation map benchmark.** CARTA links matched images, questions, answers, and provenance across natural and symbolic map representations: 14,008 multiple-choice items in five matched slices and 2,400 point and box grounding items.
2. **Validity by construction.** Ground truth is derived from source state, and a generate–audit–seal workflow keeps the benchmark valid at scale; a retrospective audit quantifies what its gates remove.
3. **Matched evaluation and diagnostics.** Twelve MLLMs, serving and input controls, and a matched human reference locate failures at three successive stages: evidence, correspondence, and geometry.

Benchmark	Source	Scope	Comparison unit
MapBench; MapVerse	Collected	Navigation; real-world maps	QA over fixed maps
FRIEDA; OmniMapBench	Collected	Map documents; reasoning	Reasoning over fixed documents
MAPWise; MapIQ; CartoMapQA	Rendered	Thematic and street maps	Within-family design variation
TopViewRS	Rendered/collected	Indoor top views	Task and map-type variation
MMVM; MC-Bench	Collected	Visual matching; grounding	Multi-image correspondence
CARTA (ours)	Generated	Natural + symbolic map idioms	**Matched fixed referents**

Table 1: **Map-understanding benchmarks by construction and comparison unit.** CARTA re-renders a fixed geographic referent across visual idioms so that representation can change while the place does not.

2 Related Work

Map and thematic-map benchmarks. Collected-map suites evaluate navigation, document parsing, and multi-step cartographic QA over real maps (Xing et al., 2025; Bhat et al., 2026; Chen et al., 2026a; Pyo et al., 2026; Chen et al., 2026b). Their strength is ecological realism; their limitation for our question is that each map is fixed, so representation and place change together. Rendered thematic and street-map benchmarks vary visual design within choropleth, cartogram, or symbol-map families (Chang et al., 2022; Mukhopadhyay et al., 2025; Srivastava et al., 2025; Ung et al., 2025), which isolates design choices but stays inside one map family. TopViewRS studies recognition, localisation, and spatial reasoning on realistic and semantic indoor top views (Li et al., 2024), the closest analogue to a fixed referent under two renderings. CARTA extends this idea to real geography and to five idioms (photographic, vector, thematic, terrain, and weather), with matched interventions inside each family (Table 1).

Geospatial and correspondence benchmarks. Remote-sensing benchmarks evaluate overhead imagery (Danish et al., 2025; Zhang & Wang, 2024; Xu et al., 2026), and cross-view geo-localisation retrieves places across camera viewpoints (Lu et al., 2025; Xia et al., 2024); both keep the photographic modality fixed. MMVM evaluates visual matching through multi-image multiple choice, and MC-Bench evaluates multi-context instance grounding (Zhou et al., 2025; Xu et al., 2025); both use collected natural images with manually annotated instances. CARTA's C3 and grounding tracks pose the same matching and transfer questions across symbolic idioms, with ground truth derived from geometry rather than annotation.

Generated benchmarks and diagnostic evaluation. Generating items rather than collecting them yields ground truth known by construction (Johnson et al., 2017; Greff et al., 2022) and makes matched interventions cheap. Work on visual encodings and spatial reasoning (Fu et al., 2024; Yang et al., 2025; Mukherjee et al., 2026; Liu et al., 2025; Do et al., 2026; Chen et al., 2025; Vaishnav & Tammet, 2026) shows that MLLMs can see an image without perceiving the quantity it encodes; CARTA brings that diagnostic style to maps. Because generated data can also propagate construction errors at scale (Northcutt et al., 2021), we pair generation with an audited acceptance protocol and report what it removes.

3 Benchmark Construction

Figure 2 illustrates the task families with frozen items, and Figure 3 the construction pipeline. All comparisons are defined over matched groups selected before any model is evaluated.

3.1 Design Principles

Terminology. A *representation* is a cartographic or photographic visual idiom, such as satellite imagery, a vector street map, or a thematic land-cover rendering. *Correspondence* means identifying the same geographic referent across representations; *re-identification* names the C3 matching task. *Grounding* means transferring that referent to target-image coordinates or a bounding box. The tasks share geospatial state but use different outputs and are scored separately.

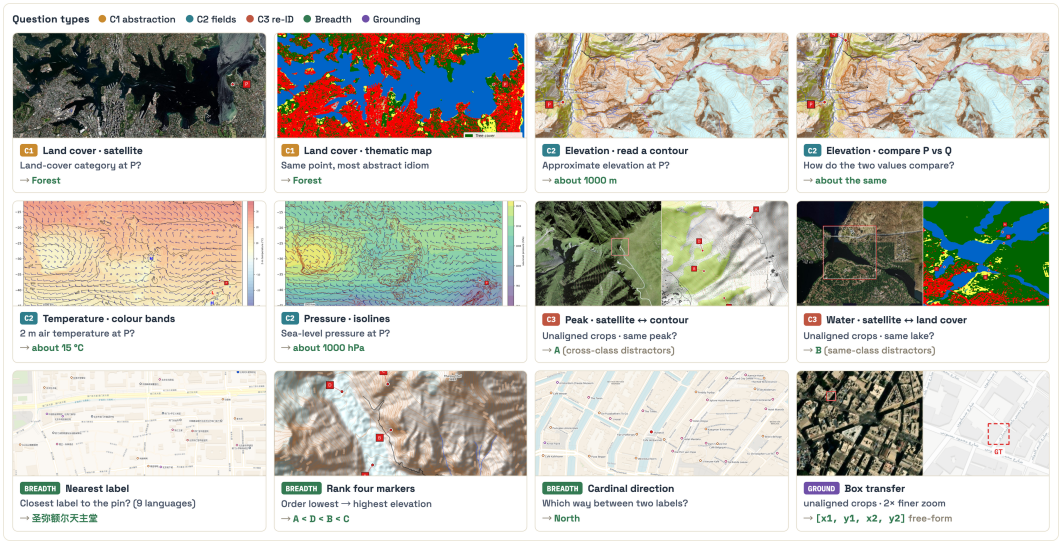

Figure 2: **Frozen CARTA examples, one per task template.** Rows show C1 and C2-contour (top), C2-weather and C3 (middle), and breadth and box grounding (bottom). Colour tags mark the task family; "C1 abstraction" and "C3 re-ID" are internal template names for C1 and C3. All panels are sealed artefacts; annotations are figure-only.

Matched units and balance. Each comparison family defines its matched unit before model evaluation. C1 preserves a complete four-idiom point group, C2-weather a four-cell field–encoding group, and C3 an aligned/unaligned twin. Selection uses only geographic metadata, structural coverage, and quality gates—never model performance. Rather than maximising volume, the benchmark balances task families and retains rare regions, classes, styles, scripts, and representations without allowing the largest generator family to dominate the aggregate score (Table 2).

Anti-shortcut validation. Automated audits limit statistical and visual shortcuts by checking option balance, text and length cues, dead options, image decoding, answer consistency, and matched-group integrity. Input ablations test whether the images are necessary. For any item that refers to a named or marked entity, an answerability gate verifies that the entity survives final rendering and decluttering.

Slice	Items (units)
C1 representation profile	3,200 (800×4)
C2-contour	1,200
C2-weather	3,200 (800×4)
C3 correspondence	4,000 (2,000 twins)
Breadth	2,408
Grounding (point + box)	2,400
Total	**16,408**

Table 2: **CARTA composition.** Grounding is scored outside the MCQ macro.

3.2 Task Families

C1: Representation profile. For a given coordinate, C1 queries a coarse land-cover category across satellite, minimalist street, standard street, and thematic land-cover views. The target is the dominant coarse ESA WorldCover class in a 100 m disk around the point, not the centre pixel alone (Zanaga et al., 2022). Coordinates, answers, and questions remain fixed within each `paired_group_id`, while texture, labels, symbology, and available class evidence co-vary. C1 therefore profiles how much of the answer each representation makes visible.

C2: Continuous-field reading. The contour arm queries elevation using SRTM (Farr et al., 2007). The weather arm uses GFS fields (NOAA National Centers for Environmental Prediction, 2026) and crosses {temperature, pressure} with {colour band, iso-line}. Standard frames encode temperature with colour and pressure with iso-lines; swapped frames reverse the mapping. Let $y_{f,e}$ denote accuracy for field f and encoding e. The conventional interaction is

$$I = (y_{T,I} - y_{T,C}) - (y_{P,I} - y_{P,C}) = 2(S_{\text{swap}} - S_{\text{std}}), \tag{1}$$

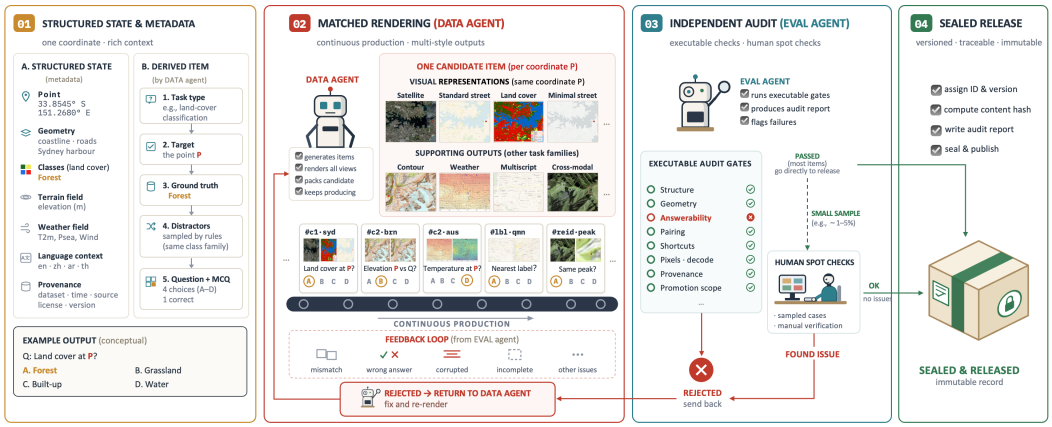

Figure 3: **Overview of the CARTA construction pipeline.** Structured geospatial state and task specifications feed deterministic rendering. The role-separated *generate–audit–seal* workflow separates candidate synthesis (**DATA**) from artefact verification (**EVAL**). Executable gates and human escalation determine acceptance; manifests and content hashes then bind the sealed benchmark to the accepted bytes.

where $S_{\text{std}} = (y_{T,C} + y_{P,I})/2$ and $S_{\text{swap}} = (y_{T,I} + y_{P,C})/2$. The interaction is therefore indistinguishable from frame difficulty in this design, so we report the frame-balanced marginal encoding contrast (iso-line minus colour band) as the primary quantity (Appendix D.1).

C3: Cross-representation correspondence. Given a highlighted feature in Image A, models identify the same referent among four candidates in Image B. Same-class candidates require within-class matching; cross-class candidates permit category elimination. Aligned twins preserve the frame, whereas unaligned twins jointly change translation, scale, crop, and visible context; a factorised control separates the first two (Appendix D.2).

Breadth, grounding, and auxiliary controls. The breadth track contains 2,408 items covering cardinal direction, multilingual label-relative reasoning, and spatial ranking. The grounding track evaluates point and region transfer separately from the MCQ macro. For 1,126 of the 1,200 box items, the target is a canonical source window centred on the same OSM coordinate; the remaining 74 project the same OSM polygon independently into both views. Grounding prompts are anonymous; a named-prompt counterpart serves as an OCR control.

3.3 Data Sources and Quality Assurance

Vector geometry and street networks come from OpenStreetMap (OpenStreetMap contributors, 2026; Olbricht, 2026), with CARTO basemap styling (CARTO, 2026). Satellite imagery comes from Esri World Imagery (Esri, 2026); land cover, terrain, and weather fields come from ESA WorldCover (Zanaga et al., 2022), SRTM (Farr et al., 2007), and NOAA GFS (NOAA National Centers for Environmental Prediction, 2026). The generator records region registries, source snapshots, and deterministic seeds alongside rendered images, item metadata, matched-group identifiers, provenance logs, and audit manifests.

Generate–audit–seal. To prevent programmatic scale from propagating rendering errors, construction is split between two role- and context-separated LLM-based agents: a **DATA** agent materialises candidates, and an **EVAL** agent verifies geometry, pairing, shortcut risks, and provenance. Executable gates reject formalised failure modes; edge cases go to human inspection, and recurring findings become new permanent rules. A retrospective audit on 600 archived candidates shows the gates reducing the visible-defect rate from 8.7% to 1.1% (Appendix B.1).

The frozen benchmark contains 20,566 unique PNGs (51.3 GiB) across 89 regions. At sealing, all 358 human-reviewed images matched the final manifest byte-for-byte; every image is bound by its

Model	C1	C2-cont.	C2-wx	C3-same	C3-cross	Breadth	Macro [95% CI]
Open-weight models							
Molmo2-8B	38.2	35.3	28.5	26.1	34.5	31.9	32.4 [31.4, 33.5]
Phi-4-Reasoning-Vision-15B	34.5	37.3	29.4	38.0	48.3	38.2	37.6 [36.5, 38.7]
InternVL3.5-8B	44.1	41.6	39.3	31.0	40.7	38.9	39.2 [38.2, 40.3]
Kimi-VL-A3B-Instruct	42.2	42.8	34.9	36.4	48.0	35.9	40.0 [39.0, 41.1]
Qwen3-VL-8B-Instruct	51.6	45.3	41.0	34.7	56.9	51.1	46.7 [45.6, 47.8]
GLM-4.6V-Flash	42.3	47.1	45.3	48.2	66.7	47.9	49.6 [48.4, 50.7]
Gemma-4-26B-A4B-it	43.4	52.7	**49.9**	43.6	58.1	53.2	50.2 [49.1, 51.2]
Qwen3.5-9B	46.5	48.8	**49.9**	52.5	69.9	56.8	54.1 [53.0, 55.1]
Qwen3.6-35B-A3B	44.9	**55.2**	48.4	66.4	**86.2**	62.7	60.6 [59.6, 61.7]
Qwen3.5-35B-A3B	**52.8**	51.8	45.8	**66.7**	84.8	**64.1**	**61.0 [59.9, 62.0]**
Open-weight models, reasoning-enabled serving control							
Qwen3.6-35B-A3B	48.7	60.3	69.1	73.7	87.0	82.3	70.2
Qwen3.5-35B-A3B	53.3	64.3	69.5	73.7	85.8	82.1	71.5
Proprietary models							
Gemini 3 Pro	47.5	63.8	65.1	68.9	81.2	74.7	66.8 [65.9, 67.8]
GPT-5.5	54.0	68.5	75.7	87.5	92.6	85.1	77.2 [76.4, 78.0]

Table 3: **Single-pass MCQ accuracy on the full benchmark (%).** The macro equally weights six slices. Each row is an end-to-end serving configuration. Bold marks the best open-weight result with reasoning disabled; grey rows re-run the same two checkpoints under reasoning-enabled serving (Appendix C). The matched human comparison is reported in Section 5.1.

SHA-256 hash, and every prediction record is bound to the dataset, code, model, and image identities it was produced from (Appendix B).

4 Evaluation Protocol

4.1 Models and Inference

Evaluated systems. We evaluate ten open-weight systems: Molmo2-8B, Phi-4-Reasoning-Vision-15B, InternVL3.5-8B, Kimi-VL-A3B-Instruct, Qwen3-VL-8B-Instruct, GLM-4.6V-Flash, Gemma-4-26B-A4B-it, Qwen3.5-9B, Qwen3.6-35B-A3B, and Qwen3.5-35B-A3B (Clark et al., 2026; Aneja et al., 2026; Wang et al., 2025; Kimi Team et al., 2025; Bai et al., 2025; Z.ai, 2026; Gemma Team et al., 2026; Qwen Team, 2026a;b), and two proprietary systems, Gemini 3 Pro and GPT-5.5 (Google DeepMind, 2025; OpenAI, 2026). Each prediction record stores the model or endpoint, preprocessing, inference settings, and parser version.

Inference execution. All models receive the same task instructions and original PNGs. Open-weight models run locally with their native processors, greedy decoding, and test-time reasoning disabled by default. Gemini 3 Pro uses a 1,024-token thinking budget; GPT-5.5 uses its native reasoning mode. We deterministically transcode 552 of the 18,008 unique MCQ image references that exceed API limits. A paired serving control re-runs both 35B Qwen variants on the MCQ and grounding tracks with native reasoning enabled while holding items, images, prompts, and greedy decoding fixed within each track (Appendix C).

4.2 Metrics and Controls

Primary multiple-choice metric. Multiple-choice items are evaluated in their frozen option order. Blank, malformed, and otherwise unparsed outputs count as incorrect. Let A_s denote accuracy on slice s. The primary macro is

$$A_{\text{macro}} = \tfrac{1}{6}\big(A_{\text{C1}} + A_{\text{C2c}} + A_{\text{C2w}} + A_{\text{C3same}} + A_{\text{C3cross}} + A_{\text{breadth}}\big), \tag{2}$$

where A_{breadth} is itself the unweighted mean of five templates: cardinal direction, label-relative reasoning, and three ranking variants. C2-contour and Breadth are single-representation slices; they

enter the macro so that reading a field and reasoning over labels are scored alongside the matched families, and every slice carries equal weight. Grounding does not enter this macro. Uniform guessing on the four-option MCQs yields 25.0%.

Answer-position and shortcut controls. Correct A/B/C/D positions account for 25.38/25.34/25.31/23.97% of the benchmark; position-balanced reweighting changes every system by less than 0.2 pp (Section 5.5). Because C3 option letters are rendered into the image, we keep them fixed to preserve spatial context. For grounding, we report Euclidean point error normalised by the target-image diagonal (thresholds 0.05 and 0.10) and end-to-end bounding-box IoU.

4.3 Uncertainty and Human Reference

All 95% intervals bootstrap the dependence-preserving unit: complete matched groups for C1 and C2-weather, aligned/unaligned twins for C3, individual items for unmatched MCQ slices, and regions for grounding. Six compensated non-experts completed 400 unique items (800 responses): 30 per primary MCQ slice, 40 complete C1 groups, and 30 point and 30 box targets. Human intervals resample items while retaining both ratings. Participant flags are kept as answerability evidence and do not alter the frozen benchmark (Appendix F.1). Human–model comparisons use the same 180 MCQs and 30 point targets.

5 Results

5.1 Overall Performance and Reasoning-Enabled Serving

GPT-5.5 obtains 77.2%, Gemini 3 Pro 66.8%, and the strongest open-weight system, Qwen3.5-35B-A3B, 61.0% (Table 3). The two 35B Qwen intervals overlap almost completely, and both sit well above the smaller open-weight systems.

The largest gains occur when visible evidence requires computation. Under the reasoning-enabled serving configuration, Qwen3.5-35B rises from 61.0% to 71.5% (95% CI of the gain: 9.5–11.4 pp) and Qwen3.6-35B from 60.6% to 70.2% (8.6–10.5 pp), placing both above Gemini 3 Pro (Table 3, grey rows). The gain is far from uniform. C2-weather rises from 45.8% to 69.5% and from 48.4% to 69.1%, and Breadth from 64.1% to 82.1% and from 62.7% to 82.3%: on these slices the answer is on the map and the remaining work is reading a legend, comparing values, or ordering markers. C3-same rises to 73.7% for both, whereas C3-cross (already near 85%) and C1 change by at most 3.8 pp. The contrast also includes a larger token budget, retry policy, and terminal parser.

Matched human comparison. On the same 180 items, non-expert readers score 83.3% and GPT-5.5 scores 74.4%, a paired gap of 8.9 pp (95% CI: 2.5–15.6). The sample is answerable to untrained readers yet remains challenging for current MLLMs.

5.2 Evidence: What the Map Makes Visible

C1 representation profiles. C1 separates models less than the macro: GPT-5.5 reaches 54.0% and Qwen3.5-35B 52.8%. Readers reach 78.8% on satellite and land-cover views, 58.8% on minimalist street maps, and 68.8% on standard street maps (Appendix F.2). The reduction on street maps is class-conditional: built-up and water targets stay at 90–100% in every idiom, whereas forest and cropland/bare targets fall to 15–45% on street maps, which draw little or no land-cover fill. Screening the human sample to locations with local class purity ≥ 0.80 shrinks the satellite-minus-minimalist gap from 20.0 [8.8, 32.5] to 8.6 [-1.7, 20.7] pp, and on the same 40 groups raises C1 accuracy for all twelve models by 8.0–13.8 pp. These groups are class-balanced (10 per class, including the hard forest and cropland/bare classes), so their accuracy is lower than on the full C1 slice (Table 21). Model satellite-minus-minimalist gaps remain mixed in sign (-5.0 to $+5.0$ pp unscreened; -13.8 to $+6.9$ screened). Evidence sufficiency thus limits readers and models alike, without inducing a shared preference among idioms.

Model	C2 iso − colour	C3 aligned − unaligned
Qwen3.5-35B-A3B	-6.5 [-9.6, -3.5]	$+12.7$ [10.6, 14.8]
Qwen3-VL-8B	-1.7 [-4.9, 1.6]	$+10.2$ [8.0, 12.3]
Gemini 3 Pro†	$+19.2$ [16.3, 22.1]	$+14.1$ [11.8, 16.4]
GPT-5.5†	$+4.5$ [1.8, 7.4]	$+14.4$ [12.8, 16.2]

Table 4: **Matched contrasts (pp).** Positive C2 values favour iso-lines; positive C3 values favour aligned framing. All twelve systems are listed in Table 13. †: proprietary model.

C2 encoding preferences are model-specific. No encoding is universally easiest. Across the twelve systems, the 95% intervals favour colour bands for six, favour iso-lines for three, and include zero for three (Table 4; full list in Table 13). Because the field-by-encoding interaction is aliased with frame difficulty (Section 3.2), we report the frame-balanced marginal contrast and read it as an encoding preference, not an encoding hierarchy.

5.3 Correspondence: Categories Before Places

Cross-class distractors inflate scores. In C3, replacing same-class with cross-class distractors raises Qwen3-VL-8B from 34.7% to 56.9%, and even GPT-5.5 from 87.5% to 92.6%: whenever category elimination is available, models use it. The source image still matters. Removing it or replacing it with an unrelated source drops the two tested Qwen systems from 77–80% to 37–42% (Appendix E.3), so correspondence is not solved from candidate semantics alone. Target-only performance is 58–59% on cross-class but 19–21% on same-class items (statistically indistinguishable from the 25% chance level), directly exposing category elimination as the residual shortcut.

Every model prefers aligned framing. Aligned twins beat unaligned twins for all twelve systems, by 3.5–14.4 pp, with every interval excluding zero. Unaligned twins jointly change translation, scale, crop, and visible context, so we add a 2×2 control on 200 matched C3 units that crosses translation with scale/context while holding representations, candidates, questions, and answers fixed (Appendix D.2). With only the representation changed, Qwen3.5-35B and Qwen3.6-35B reach 81.0% and 82.5%. Translation alone lowers them to 69.0% and 64.5%, scale/context alone to 69.5% and 68.0%, and both together to 69.5% and 68.0%. Either change alone produces at least the full-package loss (for Qwen3.6, translation alone is worse than both together), so the two penalties overlap rather than add. A controlled resolution sweep on the same two models finds no monotonic C3 benefit from higher input resolution (Appendix D.3), so the loss is not a resolution artefact.

5.4 Geometry: Recognition Without Precise Transfer

Point localisation. On all 1,200 targets, coordinate copy scores 17.5% at 0.05, matching the best open-weight model (17.6%); GPT-5.5 reaches 28.9% and target-centre prediction 1.0%. On the 652 *copy-hard* targets (copy error > 0.10), open-weight point@.05 is at most 4.1% and GPT-5.5 reaches 15.5%, while readers reach 87.5% on the matched 20-item subset. On all 30 matched targets, readers reach 88.3% and GPT-5.5 20.0% (paired gap 68.3 pp; 95% CI: 51.7–83.3; Appendix E.4). Parse rates are at least 99.5%.

Extent is harder than centre. The best bounding-box IoU@0.5 is 3.1% (GPT-5.5), although GPT-5.5 box centres fall within 0.05 of the target diagonal on 29.8% of items (Table 19). The coordinate-copy baseline places 19.2% of box centres within 0.05, equal to the best open-weight model (19.2%), and reaches 0.8% IoU@0.5. Models can approximately locate a footprint but not reproduce its extent. Higher input resolution helps points slightly but not box IoU (Appendix D.3).

What does not help. For the two tested Qwen systems, a cartographic primer lowers aggregate accuracy by 3.1–6.1 pp, restored names change point@.05 by -2.2/-1.6 pp, and reasoning changes it by -0.9/-1.6 pp; none reliably improves grounding (Appendices E.3 and C.1).

Model	Full @ .05	Full @ .10	Matched-30 @ .05
Molmo2-8B	11.9	31.8	3.3
Phi-4-Reasoning-Vision-15B	1.3	5.0	0.0
InternVL3.5-8B	10.8	32.8	**13.3**
Kimi-VL-A3B	12.5	33.8	10.0
Qwen3-VL-8B	17.1	44.4	10.0
GLM-4.6V-Flash	15.2	43.3	10.0
Gemma-4-26B-A4B	12.8	36.1	6.7
Qwen3.5-9B	16.8	44.3	10.0
Qwen3.6-35B-A3B	**17.6**	**45.6**	10.0
Qwen3.5-35B-A3B	17.3	45.3	10.0
Gemini 3 Pro†	16.2	45.7	6.7
GPT-5.5†	28.9	58.8	20.0
Target centre	1.0	5.0	0.0
Coordinate copy‡	17.5	45.7	10.0
Human reference	–	–	88.3

Table 5: **Anonymous point-grounding accuracy (%).** Euclidean error is normalised by the target-image diagonal. Full uses 1,200 targets; Matched-30 is the human subset. Bold marks the best open-weight result. ‡: exact source-mark coordinates rescaled without reading target pixels.

5.5 Robustness and Sensitivity

Position reweighting changes every system by less than 0.2 pp; equal-region aggregation lowers accuracy by 1.0–7.9 pp, but no region explains the C3 result. Reader agreement is 75.0–86.7%, and removing four flagged MCQs changes no conclusion (Appendices G and F.1).

6 Discussion

Evidence, correspondence, geometry. Missing evidence limits readers and models; with evidence, category cues and aligned frames raise correspondence. Coordinate copy matches the best open-weight point score, performance falls further on copy-hard targets, and neither resolution, names, nor reasoning closes the gap. This ordering matters: aggregate accuracy can conceal whether failure lies in evidence, correspondence, or geometry.

Scope of claims. CARTA's contrasts capture packages, not isolated causes: C1 changes texture, labels, symbology, and evidence; unaligned C3 changes crop, scale, translation, and context; C2 is aliased with frame difficulty. MCQ and grounding are not linked one-to-one, and six readers establish answerability, not a ceiling. The effects are diagnostic within CARTA's rendering contract, not universal representation rankings.

Limitations. Matched construction limits geographic and script diversity and cannot rule out public-source contamination; the defect rate comes from archived candidates, not a fresh sealed sample. Readers were Chinese-speaking, and reasoning controls cover two 35B Qwen systems. Grounding mixes nearly registered and high-displacement targets (coordinate copy reaches 45.7% at point@.10); the copy-hard slice separates them but is defined post hoc and does not replace the full-track metric. Future versions should add a sealed, balanced copy-hard split.

7 Conclusion

CARTA holds the place fixed while changing its representation, with every item traceable to frozen bytes. Across twelve MLLMs, evidence gaps, correspondence shortcuts, and weak geometric transfer form successive failures that aggregate accuracy obscures; robust systems must recover evidence, match referents across visual languages, and transfer geometry precisely.

Reproducibility Statement

The frozen benchmark, image manifest, provenance records, generation code, evaluation scripts, and prediction records bind every reported result to its data, images, prompt and parser configuration, code, and model identity. The appendices document the serving contract, bootstrap units, audit ledger, controls, and human-study sensitivities. We will release the code and data after acceptance.

AI Use Statement

Generative AI tools were used in three ways. First, the DATA and EVAL roles of the generate–audit–seal workflow (Section 3) are LLM-based agents: the DATA agent generates and renders candidate items, and the EVAL agent audits them. Their outputs enter the benchmark only after passing executable gates, and edge cases are escalated to human review. Second, AI tools assisted with writing and debugging code for rendering, audit scripts, and analysis; all such code was reviewed, tested, and executed by the authors. Third, during manuscript preparation they suggested structural changes, edited English for readability and consistency, and assisted with LaTeX formatting. They were not used as a source of evidence for any empirical claim. The authors take responsibility for the final text, code, analyses, and claims.

Ethics Statement

Human validation was voluntary, informed, compensated, and pseudonymised. The study was designed to establish task answerability rather than to assess the participants, and no personally identifying participant information is included in the benchmark. CARTA evaluates visual understanding of geographic representations for academic research; it is not intended for surveillance, targeting, or high-stakes geospatial decisions. Rendered images retain the attribution required by each source (OpenStreetMap under ODbL; ESA WorldCover, SRTM, and NOAA GFS as open data; Esri World Imagery and CARTO basemaps under their providers' terms). Following common practice for benchmarks built on third-party map imagery, CARTA will be released for non-commercial research use after acceptance, together with the rendering scripts and source metadata needed to regenerate every image; users remain responsible for complying with the terms of the underlying providers.

References

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. Phi-4-reasoning-vision-15B technical report. *arXiv preprint arXiv:2603.03975*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

CARTO. CARTO basemaps (Positron, Voyager). https://carto.com/basemaps, 2026. Basemap tiles © CARTO, © OpenStreetMap contributors.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps. In *NeurIPS 2022 Workshop on Table Representation Learning*, 2022.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. ClimateIQA: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5322–5333, 2025. doi: 10.1145/3711896.3737406.

Minxin Chen, He Zhu, Junyou Su, Wen Wang, Yijie Deng, and Wenjia Zhang. PlanBench-V: A spatial planning map benchmark for vision-language models. *arXiv preprint arXiv:2606.05744*, 2026a.

Yang Chen, Yunwen Li, Yufan Shen, Minghao Liu, Tianyu Zheng, Bin Fu, Qunshu Lin, Zhi Yu, and Botian Shi. OmniMapBench: Benchmarking visual-centric reasoning on diverse map documents. *arXiv preprint arXiv:2607.09068*, 2026b.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7132–7142, 2025.

Patrick Phuoc Do, Chau M. Ta, and Chaoli Wang. Benchmarking multimodal large language models for scientific visualization literacy. *arXiv preprint arXiv:2607.15176*, 2026.

Esri. World Imagery basemap. https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9, 2026. Acknowledgments: Esri, Vantor, Earthstar Geographics, and the GIS User Community.

Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, et al. The Shuttle Radar Topography Mission. *Reviews of Geophysics*, 45(2), 2007. doi: 10.1029/2005RG000183.

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 *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Google DeepMind. Gemini 3 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, 2025.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti Liu, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

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

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

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 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807, 2024. doi: 10.18653/v1/2024.emnlp-main.106.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of VLMs for chart understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10829–10841, 2025. doi: 10.18653/v1/2025.findings-emnlp.573.

Xudong Lu, Zhi Zheng, Yi Wan, Yongxiang Yao, Annan Wang, Renrui Zhang, Panwang Xia, Qiong Wu, Qingyun Li, Weifeng Lin, Xiangyu Zhao, Peifeng Ma, Xue Yang, and Hongsheng Li. GLEAM: Learning to match and explain in cross-view geo-localization. *arXiv preprint arXiv:2509.07450*, 2025.

Alan M. MacEachren. *How Maps Work: Representation, Visualization, and Design*. Guilford Press, New York, 1995. ISBN 9780898625899.

Kushin Mukherjee, Donghao Ren, Dominik Moritz, and Yannick Assogba. EncQA: Benchmarking vision-language models on visual encodings for charts. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):648–658, 2026. doi: 10.1109/TVCG.2025.3634249.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. MAPWise: Evaluating vision-language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9348–9378, 2025. doi: 10.18653/v1/2025.naacl-long.473.

NOAA National Centers for Environmental Prediction. NOAA Global Forecast System (GFS). https://registry.opendata.aws/noaa-gfs-bdp-pds/, 2026. NOAA Big Data Program public archive (AWS `noaa-gfs-bdp-pds`).

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Roland M. Olbricht. Overpass API. https://overpass-api.de, 2026.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, 2026.

OpenStreetMap contributors. OpenStreetMap. https://www.openstreetmap.org/copyright, 2026. Open map data licensed under the Open Database License (ODbL).

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022. doi: 10.1145/3531146.3533231.

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 *International Conference on Learning Representations*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026a. Evaluated checkpoints: https://huggingface.co/Qwen/Qwen3.5-9B and https://huggingface.co/Qwen/Qwen3.5-35B-A3B.

Qwen Team. Qwen3.6-35B-A3B. https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026b. Official model card.

Robert E. Roth. Visual variables. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pp. 1–11. Wiley, 2017. doi: 10.1002/9781118786352.wbieg0761.

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*, 2025.

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

Mohit Vaishnav and Tanel Tammet. Symbolic grounding reveals representational bottlenecks in abstract visual reasoning. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 318–343, 2026. doi: 10.18653/v1/2026.conll-main.19.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, Yingying Zhang, Jiangwei Lao, Jingdong Chen, Ming Yang, and Yongjun Zhang. Cross-view geo-localization with street-view and vhr satellite imagery in decentrality settings. *arXiv preprint arXiv:2412.11529*, 2024.

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.

Yunqiu Xu, Linchao Zhu, and Yi Yang. MC-Bench: A benchmark for multi-context visual grounding in the era of MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17675–17687, 2025.

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.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

Z.ai. GLM-4.6V-Flash. https://huggingface.co/zai-org/GLM-4.6V-Flash, 2026. Checkpoint `zai-org/GLM-4.6V-Flash`.

Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. ESA WorldCover 10 m 2021 v200. Zenodo dataset, 2022. © ESA WorldCover project 2021 / Contains modified Copernicus Sentinel data.

Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking GPT-4V on Earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7839–7849, 2024.

Yikang Zhou, Tao Zhang, Shilin Xu, Shihao Chen, Qianyu Zhou, Yunhai Tong, Shunping Ji, Jiangning Zhang, Lu Qi, and Xiangtai Li. Are they the same? Exploring visual correspondence shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17663–17674, 2025.

A Observable Serving Contract

Systems	Visual input contract	Generation	Unobserved serving factors
Ten open-weight primary configurations	Original frozen PNG; one or two images jointly packed; native processor; 4.19M-pixel configured ceiling per image	Greedy; reasoning off; 64 or 128 output tokens	Architecture-specific resizing, tiling, and visual-token allocation are recorded by processor/config where exposed
Two Qwen 35B MCQ reasoning controls	Original frozen PNG; the same one- or two-image item input	Greedy; native reasoning; adaptive 4,096–32,768 output-token caps	Architecture-specific multimodal packing and visual-token allocation
Two Qwen 35B grounding reasoning controls	Original frozen PNG; the same two-image grounding input	Greedy; native reasoning; 4,096-token cap; 128-token coordinate finalisation when needed	Architecture-specific multimodal packing and visual-token allocation
Two proprietary systems	The same frozen item images, supplied through a fixed evaluation wrapper	Fixed evaluation settings	Effective image resolution, visual-token allocation, and other service-side preprocessing are unobservable

Table 6: Observable serving contract. Exact open-weight model revisions, processor settings, prompt hashes, and decoding fields are stored in prediction metadata records. Differences in the last column are part of the evaluated systems and limit architecture-only interpretation.

The frozen images range from 388×385 to 6446×6257 pixels. Table 6 reports only the input contract visible to the evaluator. Service-side effective resolution and image-token allocation are unavailable and are not inferred from aggregate token counts.

B Frozen Benchmark and Reproduction Contract

The 16,408-item primary benchmark contains 14,008 MCQs and 2,400 grounding items; the 2,400 named-grounding records remain an auxiliary control. It references 20,566 unique PNGs (51.3 GiB) across 89 region identifiers. The seven frozen JSON files, selection and provenance records, image manifest, and audit outputs are bound by the benchmark digest `b8f0a6810edb3a9c2b9efe1bcad28cfe3ab8ae754c15b49c507eecbf1a6b389e`. The bound manifest stores every image SHA-256. Before selection, we fixed a coordinate-based exclusion list covering regions that straddle or lie within approximately 12 km of an international border; any item whose source or target footprint intersected a listed area was ineligible. Full decode/hash audit found zero failures, coordinate audit found zero such intersections, and all 358 human-reviewed images were byte-identical after sealing.

Benchmark-to-results checklist. The contract fixes seven steps: (1) verify the digest; (2) verify the image manifest and materialised hashes; (3) select a declared frozen slice; (4) initialise the evaluation record and run the standard prompt; (5) parse and validate outputs; (6) apply the declared matched-unit or region bootstrap; and (7) verify the final prediction–dataset–model binding. The prepared artefact package includes schemas for dataset, manifest, audit, and prediction metadata records, together with evaluator entry points, a smoke-test command, expected checksums, and output names. The package will be released after acceptance. Evaluation records are initialised before inference to reject incompatible resume and checked again before formal scoring.

B.1 Observed audit triggers and remediation

Tables 7–8 report observed failures for which the archived construction record preserves a defensible count. It is a remediation ledger, not an additive funnel: candidate pools changed across iterations, early generators did not uniformly log skipped proposals, and one item could violate more than one rule. We therefore do not reconstruct unobserved stage-wise denominators or assign overlapping failures to an arbitrary first-failure stage.

To quantify the action of the gates without reconstructing the full evolving generator history, we also ran a prespecified retrospective shadow audit on 600 archived candidates, stratified as 100

Family	Observed pre-release trigger and detection	Remediation and permanent rule	Final
C1	No family-specific historical reject count was retained; automated benchmark audits were preserved.	Require complete four-idiom groups, source/answer consistency, image integrity, and shortcut checks.	3,200
C2-contour	Coordinate-policy audit excluded 198 candidates from a then-current selection because they were border-proximate.	Apply the centralised coordinate exclusion before selection and sealing.	1,200
C2-weather	Independent design review found that a 6,772-item predecessor pool confounded field with encoding; this was a design replacement, not a stage-wise reject.	Replace the pool; require complete field×encoding four-cell groups.	3,200

Table 7: Observed audit triggers, remediation, and final primary-benchmark counts (part 1 of 2). Counts describe archived events from evolving predecessor pools and must not be subtracted from the final column as a single funnel.

Family	Observed pre-release trigger and detection	Remediation and permanent rule	Final
C3	Artefact review and structural audits found 110 border-proximate candidates, 268 items in an unanswerable vector–contour cell, 144 over-large extents, and 7 orphaned twins. The stale-letter count was not retained; 3,767 predecessor images were rerendered.	Remove the invalid cell; cap extent at 60% of frame width; require twin completeness and answer–marker consistency; regenerate rendering from metadata.	4,000
Breadth	Human review and an automated census found that 714 of 1,000 predecessor direction items referenced an undrawn anchor; coordinate audit separately excluded 24 border-proximate candidates.	Question anchors must survive final decluttering; otherwise visible letter markers are rendered. Questions request the nearest cardinal direction, which is geometry-audited.	2,408
Grounding	Human/artefact review found 811 predecessor records containing a blocked basemap response rather than a valid target map.	Remove the provider, replace affected records, and check source/target assets and footprints before sealing.	2,400

Table 8: Observed audit triggers, remediation, and final primary-benchmark counts (part 2 of 2). Grounding excludes the separate 2,400-item named control.

per family. Table 9 assigns an item to the first cumulative gate that rejects it. Thus its columns are mutually exclusive. This diagnostic candidate pool deliberately includes obsolete designs—in particular, incomplete C2-weather field×encoding groups—and is not a resampling of the sealed benchmark.

For an independent visible-defect check, a reviewer blindly scanned a uniformly sampled 300-item subset without solving the questions. The reviewer marked only pass/defect/uncertain; all 300 received determinate labels. Table 10 evaluates the same fixed labels on the candidates retained after each cumulative gate. Structural checks target group-level validity, which is often invisible on a single rendered card, so their purpose is not to monotonically reduce this item-visible metric. Artefact/render gates reduced the observed visible-defect rate to 2.7%, and rules derived from prior human review reduced it further to 1.1%. Wilson 95% intervals for the first and final rates are [6.0, 12.4]% and [0.3, 3.8]%, respectively.

Across families, the audit taxonomy covers missing or invalid source assets, projection and coordinate consistency, complete matched groups, marker containment, post-decluttering label visibility, nearest-direction uniqueness, distractor and footprint collisions, contour/legend readability, and duplicate or dead options. Mechanically decidable cases are executable gates; perceptual readability is escalated for human inspection. Historical pre-gate counters were not retained for every check; absence from Tables 7–8 therefore means "not quantified," not "never tested." Human inspection was decisive when metadata remained internally consistent but the rendered evidence was wrong or unanswerable: stale C3 marker lettering, the blank vector–contour cell, blocked grounding tiles, and undrawn spatial-direction anchors. These findings caused removal or rerendering and were converted into the executable rules summarised above. The 358 replacement spatial images were subsequently reviewed and verified byte-for-byte against the sealed benchmark.

Family	Pool	First rejection gate			Final
		Struct.	Render	Human	
C1	100	0	0	0	100
C2-contour	100	16	3	0	81
C2-weather	100	100	0	0	0
C3	100	2	0	12	86
Breadth	100	2	0	46	52
Grounding	100	0	38	0	62
All	600	120	41	58	381

Table 9: Retrospective shadow-audit funnel. Gate columns give mutually exclusive first-rejection counts under cumulative application.

Stage	Reviewed retained	Defects	Rate
No acceptance gates	300	26	8.7%
+ structural	240	26	10.8%
+ artefact/render	219	6	2.7%
+ human-derived	188	2	1.1%

Table 10: Blind visible-defect audit on a fixed 300-candidate subset.

The blind shadow review observed, without treating reviewer-selected categories as causal diagnoses, source-panel defects, missing referenced marks, coordinate inconsistencies, incomplete matched groups, and other visible construction problems. We report the set of observed issue types but not their category frequencies: the review was designed to establish visible defect status, whereas precise root-cause attribution requires a separate engineering trace. Two visible defects survived all replayed gates, providing concrete targets for the next permanent rules: one C3 land-cover panel defect and one grounding item whose referenced source mark was not visible. These two defects were observed in the archived shadow pool, not in a sample drawn from the sealed benchmark.

Each prediction metadata record binds dataset, image, digest, model, prompt, decoding, parser, and code identities. Any mismatch rejects resume, and earlier-digest results are not migrated.

Determinism and statistical replication. Recorded seeds fix selection and inference; each configuration uses one greedy pass, and controls retain separate identities. Confidence intervals use 10,000 unit-clustered bootstrap replicates. Guidance additionally uses exact paired McNemar tests; untested differences remain descriptive.

C Reasoning-Mode Serving Control

The primary open-weight comparison keeps reasoning disabled. To measure the importance of that serving choice, we evaluate the same Qwen3.5-35B-A3B and Qwen3.6-35B-A3B weights under reasoning-enabled serving. Items, images, prompts, option order, and greedy decoding remain fixed within each track. MCQ outputs that end before a final answer are retried at 8,192, 16,384, and 32,768 tokens; the auditable merge retains the first valid response in this declared order. Three Qwen3.5 and four Qwen3.6 responses remain empty and count as incorrect.

Reasoning responses may contain option letters in their rationale before a bare final-line answer. We therefore use a distinct *reasoning-terminal* parser that prioritises this terminal commitment. The legacy parser and all primary main-table scores remain unchanged. Raw outputs, parser source, source-run identities, accepted repair contributions, and remaining failures are recorded in the merged prediction metadata records.

The MCQ gains are strongly task-dependent. Qwen3.5 rises from 45.8% to 69.5% on C2-weather and from 64.1% to 82.1% on Breadth, but only from 84.8% to 85.8% on C3-cross; Qwen3.6 rises from 48.4% to 69.1%, from 62.7% to 82.3%, and from 86.2% to 87.0%, respectively. This pattern shows that additional inference-time computation does not uniformly improve all forms of map reasoning.

Model	Serving	C1	C2-cont.	C2-wx	C3-same	C3-cross	Breadth	Macro
Qwen3.5-35B-A3B	reasoning off	52.8	51.8	45.8	66.7	84.8	64.1	61.0
	reasoning on	53.3	64.3	69.5	73.7	85.8	82.1	71.5
Qwen3.6-35B-A3B	reasoning off	44.9	55.2	48.4	66.4	86.2	62.7	60.6
	reasoning on	48.7	60.3	69.1	73.7	87.0	82.3	70.2

Table 11: Paired reasoning-mode serving control (%). Reasoning-on minus reasoning-off changes in the equal-slice macro are $+10.5$ pp [9.5, 11.4] for Qwen3.5 and $+9.6$ [8.6, 10.5] for Qwen3.6 (95% matched experimental-unit cluster-bootstrap intervals). The condition changes inference-time computation and serving, so the contrast is not an architecture-only or latent-mechanism effect.

It does not by itself identify the cause of each gain, and, because the control covers two 35B Qwen systems, its magnitude is not extrapolated to other model families.

C.1 Grounding under reasoning-enabled serving

We apply the same paired control to all 1,200 point and 1,200 box items. The reasoning-enabled arm uses a 4,096-token cap. If a response contains reasoning but no terminal coordinates, one greedy 128-token coordinate-only finalisation pass is applied; this occurs for 21 Qwen3.5 and 62 Qwen3.6 items, and all 2,400 outputs per model are parsed. Items, images, prompts, and scoring are otherwise identical to the reasoning-off arm.

Model	Serving	Pt@.05	Pt@.10	Mean IoU	IoU@.5
Qwen3.5-35B-A3B	reasoning off	17.3	45.3	1.51	0.67
	reasoning on	16.4	44.4	1.46	0.83
Qwen3.6-35B-A3B	reasoning off	17.6	45.6	1.51	0.75
	reasoning on	16.0	45.3	1.52	0.83

Table 12: Full-track grounding under the paired serving control (%). Mean IoU is shown as $100\times$IoU. All 2,400 outputs per configuration are parsed.

Reasoning-enabled minus reasoning-off changes for Qwen3.5 are -0.9 pp [-2.3, 0.4] for point@.05, -0.8 [-2.5, 0.9] for point@.10, -0.05 [-0.34, 0.23] for mean IoU, and $+0.17$ [-0.26, 0.66] for IoU@.5. The corresponding Qwen3.6 changes are -1.6 pp [-3.1, 0.0], -0.3 [-2.0, 1.4], $+0.02$ [-0.28, 0.31], and $+0.08$ [-0.43, 0.54]. Intervals use the same paired region-cluster bootstrap as the primary grounding analysis and all include zero. Thus the large MCQ gains do not transfer reliably to point or extent localisation.

D Full Matched-Effect Results

Table 13 reports every per-model contrast summarised in the main paper. Confidence intervals use a group-cluster bootstrap. C2's first column is the frame-balanced marginal encoding effect. C3 uses the two members of each aligned/unaligned twin.

D.1 Why the C2 diagonal is not separately identified

Let $y_{T,C}$, $y_{T,I}$, $y_{P,C}$, and $y_{P,I}$ denote accuracy for temperature/pressure crossed with colour/iso-line. The conventional interaction is

$$I = (y_{T,I} - y_{T,C}) - (y_{P,I} - y_{P,C}). \tag{3}$$

The standard composite frame assigns temperature to colour and pressure to iso-lines, whereas the swapped frame reverses that assignment:

$$S_{\text{standard}} = (y_{T,C} + y_{P,I})/2, \tag{4}$$

$$S_{\text{swapped}} = (y_{T,I} + y_{P,C})/2. \tag{5}$$

Model	C2 iso-line − colour-band	C2 aliased diagonal†	C3 aligned − unaligned
Molmo2-8B	+2.7 [−0.6, 5.9]	−0.1 [−6.0, 5.9]	+6.9 [4.7, 9.0]
Phi-4-Reasoning-Vision-15B	−3.5 [−6.5, −0.4]	−0.8 [−6.5, 5.1]	+9.9 [7.6, 12.2]
InternVL3.5-8B	−4.8 [−7.9, −1.9]	−2.8 [−8.6, 3.4]	+3.5 [1.1, 6.1]
Kimi-VL-A3B-Instruct	+2.7 [0.1, 5.6]	+1.1 [−4.4, 6.6]	+10.0 [7.6, 12.3]
Qwen3-VL-8B-Instruct	−1.7 [−4.9, 1.6]	−11.6 [−18.0, −5.6]	+10.2 [8.0, 12.3]
GLM-4.6V-Flash	−5.8 [−9.2, −2.4]	+1.4 [−5.0, 7.9]	+8.2 [6.2, 10.3]
Gemma-4-26B-A4B-it	−4.5 [−8.0, −1.2]	−12.5 [−18.9, −6.0]	+9.9 [7.6, 12.2]
Qwen3.5-9B	−6.8 [−10.2, −3.4]	−8.5 [−14.9, −1.8]	+12.2 [10.2, 14.3]
Qwen3.6-35B-A3B	−0.5 [−3.8, 3.0]	−14.0 [−20.1, −7.5]	+10.6 [8.8, 12.6]
Qwen3.5-35B-A3B	−6.5 [−9.6, −3.5]	−3.0 [−9.2, 3.8]	+12.7 [10.6, 14.8]
Gemini 3 Pro	+19.2 [16.3, 22.1]	+5.0 [−0.9, 11.1]	+14.1 [11.8, 16.4]
GPT-5.5	+4.5 [1.8, 7.4]	+14.5 [9.1, 19.8]	+14.4 [12.8, 16.2]

Table 13: Matched effects in percentage points (pp) with 95% group-cluster bootstrap intervals. †This diagonal is not an identified pure field-by-encoding interaction; it is exactly aliased with the composite-frame contrast derived below.

	Accuracy (%)				Matched effect (pp)			
Model	Rep.	Trans.	Scale/ctx.	Full	Trans. main	Scale main	Interaction	Rep. − full
Qwen3.5-35B-A3B	81.0	69.0	69.5	69.5	−6.0	−5.5	+12.0	+11.5
Qwen3.6-35B-A3B	82.5	64.5	68.0	68.0	−9.0	−5.5	+18.0	+14.5

Table 14: C3 factorised framing control over 200 matched units. Main effects average over the other factor; interaction is $y_{11} - y_{10} - y_{01} + y_{00}$.

Therefore

$$I = 2(S_{\text{swapped}} - S_{\text{standard}}). \tag{6}$$

No post-hoc re-analysis can distinguish a pure field-by-encoding interaction from this frame contrast in the present design. The evaluation script checks this equality numerically and labels the quantity as an alias diagnostic.

D.2 Factorising the C3 framing package

The primary C3 aligned–unaligned contrast intentionally changes a framing package rather than identifying a single mechanism. We therefore add a post-freeze auxiliary 2×2 control that crosses translation (0 versus 0.2 source-frame widths) with geographic scale/context (1× versus 2× extent), while holding the source–target representation change, source image, candidates, option order, question, and answer fixed. The four arms are representation-only, translation-only, scale/context-only, and the full unaligned package. The deterministic subset contains 200 matched units (50 in each satellite–contour/vector × same-/cross-class cell) and 800 items. Target-LULC cells are excluded because the shadow review identified a legend–marker occlusion mode; this control consequently estimates framing effects only on the four retained cells. Existing representation-only and full-package predictions are reused by exact frozen item mapping; only the 400 new arms are rendered and inferred. All new images pass full decode, SHA-256, viewport, answer–marker, and four-arm completeness audits.

Unit-clustered 95% bootstrap intervals exclude zero for both main effects: translation is −6.0 [−10.8,−1.3] and −9.0 [−14.3,−3.8] pp, while scale/context is −5.5 [−10.0,−1.0] and −5.5 [−10.8,−0.3] for Qwen3.5 and Qwen3.6, respectively. However, the positive interactions, +12.0 [4.0,20.0] and +18.0 [8.5,27.5], show that the penalties are strongly non-additive. Either translation or scale/context alone nearly reproduces the full-package loss; combining them adds little further damage. Thus the observed 11.5–14.5 pp package penalty cannot be assigned to one isolated transformation. Translation is somewhat larger for Qwen3.6, but the dominant conclusion is overlapping rather than additive failure modes. This auxiliary two-model result sharpens, but does not replace, the primary 12-model package contrast.

Model	Input cap	C3 same	C3 cross	Pt@.05	Pt@.10	Box ctr@.05	Box ctr@.10	Mean IoU
Qwen3-VL-8B	1 MP	30	52	16	40	25	53	2.10
	2 MP	32	56	17	43	26	55	2.25
	4 MP	24	50	16	43	26	55	2.10
	original	36	52	18	44	26	54	1.59
Qwen3.5-35B	1 MP	64	80	11	36	26	52	2.86
	2 MP	68	86	14	42	20	49	2.00
	4 MP	62	84	14	43	18	47	1.26
	original	66	80	18	43	26	45	2.45

Table 15: Controlled resolution sweep. All entries are percentages except that mean IoU is shown as $100\times$IoU. Box IoU@.5 remains between 1–3% in every arm and is omitted for space.

D.3 Controlled input-resolution sweep

Because service-side resizing is unobservable for proprietary systems, we run a separate local control on Qwen3-VL-8B and Qwen3.5-35B-A3B. The matched subset contains 50 C3 same-class items, 50 C3 cross-class items, 100 point-grounding items, and 100 box-grounding items. Every selected item has both input images above 1 MP and at least one above 4 MP. Before model serving, each image is deterministically capped at 1, 2, or 4 million pixels, preserving aspect ratio and aligning dimensions to the models' 32-pixel patch–merge grid; images below a cap are never upsampled. A fourth arm supplies the original sealed PNG. The selected originals range up to 11.1 MP, below both checkpoints' configured 16.8 MP processor maximum. Ground-truth coordinates are transformed by the same realised resize factors. Prompts, weights, decoding, selected IDs, and coordinate parsing remain fixed. All 1,511 derived assets pass dimension, decode, and SHA-256 audits. C3 intervals use item bootstrap; grounding intervals resample regions.

There is no monotonic resolution benefit for C3 or box localisation. All C3 matched differences relative to 1 MP have bootstrap intervals spanning zero. For Qwen3-VL-8B, point@.10 rises from 40% at 1 MP to 44% at original resolution (95% CI of the gain: 1.0–7.9 pp), while point@.05 moves only from 16% to 18% (0.0–5.0 pp). For Qwen3.5-35B, point@.05 rises from 11% to 18% at original resolution (3.0–11.1 pp), and point@.10 from 36% to 43% at 4 MP (2.1–11.9 pp). These gains show that visual compression explains part of point-localisation error. They do not extend to boxes: centre accuracy and IoU do not improve consistently, and the 4 MP Qwen3.5 arm falls below 1 MP on centre@.05 (from 26% to 18%) and on mean IoU (from 2.86 to 1.26, $\times 100$). Thus increased input resolution alone does not remove the precise-localisation bottleneck, especially the estimation of object extent. This conclusion is restricted to the two local Qwen models and the deliberately high-resolution subset; it does not reveal the effective resolution of proprietary systems. Additional visual resolution can recover some point-localisation accuracy for the two tested local models, but it does not remove the broader localisation deficit. The separate full-track reasoning control likewise finds no reliable localisation gain (Section C.1); this sweep does not test an interaction between input resolution and reasoning.

E Predefined Matched Controls

E.1 Cartographic guidance

To separate missing convention knowledge from failure to apply visible evidence, we predefine a prompt-only matched control over 800 frozen MCQ items. The guided arm prepends the same concise primer to every item: it explains isoline and colour-legend reading, common but non-universal land-cover colour cues, the distinction among satellite, vector, land-cover, and topographic representations, and geometry-based cross-style matching. It neither refers to an item nor identifies an answer. Images, questions, options, answers, and option order remain byte-identical to the primary no-guidance condition.

The sample comprises 200 items from each primary MCQ family: 50 complete C1 four-idiom groups; 100 value and 100 comparison contour items; 50 complete weather field-by-encoding 2×2 groups; and 100 complete C3 aligned–unaligned twins, equally divided between same- and cross-class

Model	Primer − base	C3 full − target	C3 full − shuffled	Named − anonymous@.05
Qwen3.5-35B-A3B	-6.1 [-9.4, -2.8]	$+40.0$ [30.5, 49.0]	$+37.5$ [29.2, 45.8]	-2.2 [-4.0, -0.3]
Qwen3.6-35B-A3B	-3.1 [-6.2, -0.2]	$+38.0$ [28.5, 47.5]	$+40.1$ [31.8, 47.9]	-1.6 [-3.5, 0.2]

Table 16: Matched auxiliary controls in percentage points with 95% cluster-bootstrap intervals. Primer uses 800 items; target-only/shuffled use 200/192 C3 items; named grounding uses 1,200 point targets.

distractors. This yields 400 independent units. Selection is deterministic, binds the frozen benchmark and primer hashes, and is shared across models. We reuse each model's no-guidance predictions and run only the guided arm. Differences are reported as guided-minus-baseline accuracy with experimental-unit cluster-bootstrap intervals and exact paired McNemar tests. This auxiliary control does not replace or modify the primary MCQ score.

E.2 Source use and named grounding

The C3 input-necessity control preserves the question, target image, options, answer, and item ID mapping while comparing three conditions: the complete source–target input, target-only input, and a deterministically shuffled source drawn from another item. Target-only performance measures information available from candidate semantics and layout alone; shuffled-source performance tests whether an unrelated source image is ignored or spuriously used. Analysis is paired by the original C3 experimental unit and stratified by distractor type and alignment.

The named-grounding control reuses all 2,400 anonymous grounding targets and images but restores the source and target entity names in the prompt. Its matched difference estimates the benefit of visible-label/OCR access under the same geometric transfer target. It does not redefine anonymous grounding or enter the primary benchmark score. Control results are added only from provenance-valid outputs that bind the frozen image manifest and checkpoint identity.

E.3 Control results

The general primer does not rescue either system overall. Its only positive primary-family change is Qwen3.6 on C1 ($+7.0$ [2.5, 12.0]), which does not generalise to C2 or C3. The C3 ablations give the clearest necessity result: Qwen3.5 scores 79.5% on the complete-input target-only subset versus 39.5% without the source, and 79.7% on the shuffled subset versus 42.2% with an unrelated source. Qwen3.6 shows the same pattern (77.0/39.0% and 77.1/37.0%). Correct pairing, rather than candidate semantics alone, supplies a large fraction of C3 performance. The residual target-only accuracy is concentrated in cross-class items: 58.0% versus 21.0% on same-class items for Qwen3.5, and 59.0% versus 19.0% for Qwen3.6 (chance is 25%; with 100 items per stratum, the same-class values are statistically indistinguishable from chance). With shuffled sources, the corresponding cross-/same-class results are 58.2/25.5% and 52.0/21.3%. This separation directly identifies category elimination as the main target-only shortcut.

The negative aggregate primer effect should be read narrowly: the supplied generic guidance did not improve either evaluated system under this fixed prompt. The experiment does not distinguish among prompt interference, irrelevant guidance for already-known conventions, or a bottleneck in applying the visible evidence, and it does not imply that task-specific demonstrations or other forms of guidance cannot help.

Restoring names does not repair geometric transfer. At point@.10, named-minus-anonymous changes are -4.6 pp [-7.2, -2.0] for Qwen3.5 and -3.0 [-5.0, -1.1] for Qwen3.6; box IoU@.5 changes by less than 1 pp for both. Thus visible names are not a substitute for locating and transferring the source footprint.

E.4 Coordinate-only grounding baselines

We evaluate two deterministic controls on all 2,400 frozen grounding items. The *target-centre* control predicts the centre of the target image; for a box, it centres a rectangle there while preserving the source box's normalised width and height. The *coordinate-copy* baseline takes the source mark's annotation, normalises it by the source dimensions, and rescales it to the target dimensions. Neither control reads target-image pixels. Because the source mark is drawn in the source image, coordinate copy is a naive no-transfer baseline: it assumes the two views are registered and measures how far that assumption alone goes, not an upper bound.

Control	Pt@.05	Pt@.10	Box ctr.@.05	IoU@.5
Target centre	1.0	5.0	1.4	0.1
Coordinate copy	17.5	45.7	19.2	0.8

Table 17: Coordinate-only grounding controls (%). Point and box-centre errors are normalised by the target-image diagonal; IoU@.5 is the strict box metric.

Region-cluster bootstrap 95% intervals for point@.05 are [0.4, 1.6] for the target-centre control and [14.7, 20.6] for coordinate copy; for box IoU@.5 they are [0.0, 0.3] and [0.2, 1.6]. Coordinate copy nearly equals the strongest open-weight point result (17.6%) and box-centre result (19.2%), but remains far below GPT-5.5 (28.9% and 29.8%) and does not solve strict extent transfer. Thus much of the open-weight point score is consistent with coarse coordinate transfer, while precise localisation remains unresolved.

On the 30 point targets used for the human comparison, target-centre prediction scores 0.0% at both thresholds and coordinate copy scores 10.0% at 0.05 and 33.3% at 0.10. The readers' 88.3% at 0.05 therefore cannot be explained by the coordinate prior.

We further define a descriptive *copy-hard* slice using frozen geometry alone: its coordinate-copy error exceeds 0.10 of the target diagonal. This retains 652 of 1,200 point targets (54.3%). The slice is independent of model predictions but was defined after inspecting the coordinate-copy control, so it is a diagnostic rather than a replacement for the prespecified full-track metric.

System	Copy-hard @.05	Copy-hard @.10
Molmo2-8B	0.8	4.0
Phi-4-Reasoning-Vision-15B	0.0	0.8
InternVL3.5-8B	3.1	10.1
Kimi-VL-A3B	0.2	1.8
Qwen3-VL-8B	0.5	0.8
GLM-4.6V-Flash	**4.1**	**19.0**
Gemma-4-26B-A4B	0.8	6.3
Qwen3.5-9B	1.1	3.1
Qwen3.6-35B-A3B	0.5	2.1
Qwen3.5-35B-A3B	0.5	2.8
Gemini 3 Pro†	3.2	10.9
GPT-5.5†	15.5	25.3
Human reference (matched hard-20)	87.5	87.5

Table 18: Point accuracy on the copy-hard slice (%). Model rows use all 652 items. The human row uses the 20 copy-hard items in the matched sample and 40 responses; it is shown as an answerability reference, not a population estimate. Bold marks the best open-weight result.

Most open-weight systems collapse at the strict threshold once direct coordinate reuse is excluded. GLM-4.6V-Flash is the strongest open-weight system at 4.1% point@.05, GPT-5.5 reaches 15.5%, and the matched readers remain at 87.5%. At point@.10 the corresponding values are 19.0%, 25.3%, and 87.5%. The comparison isolates high-displacement cases and strengthens the conclusion that

Model	Centre@.05	IoU@.1	IoU@.5
Molmo2-8B	11.7	2.4	0.7
Phi-4-Reasoning-Vision-15B	1.2	0.1	0.1
InternVL3.5-8B	11.6	3.1	0.5
Kimi-VL-A3B	10.4	2.9	0.7
Qwen3-VL-8B	18.2	4.5	0.6
GLM-4.6V-Flash	19.2	6.1	0.8
Gemma-4-26B-A4B	7.6	1.9	0.4
Qwen3.5-9B	17.2	4.4	**1.0**
Qwen3.6-35B-A3B	**18.8**	4.5	0.8
Qwen3.5-35B-A3B	18.3	4.6	0.7
Gemini 3 Pro	14.7	3.8	0.7
GPT-5.5	29.8	9.2	3.1

Table 19: Bounding-box decomposition (%, 1,200 items). Centre uses normalised target-diagonal distance; IoU@.5 is the strict primary metric. Bold marks the best open-weight result. Additional extent and generalised-IoU diagnostics are stored in the analysis JSON.

precise geometric transfer, rather than output parsing or the mere presence of a source mark, remains the bottleneck.

F Grounding and Human Diagnostics

F.1 Human-study protocol

Six compensated, Chinese-speaking non-experts without professional cartography experience gave informed, voluntary participation. They completed six self-contained HTML booklets independently, without Internet access or AI tools; geographic proper nouns retained their original script. The interface recorded multiple-choice selections and allowed point clicks and mouse-drawn boxes directly on the target image. Participants could flag an item as unanswerable and provide a confidence rating. Reported analyses use pseudonymous reviewer IDs; names are not reported.

The deterministic assignment contains 400 unique items and 800 responses, exactly two per item. It comprises 180 general MCQs (30 from each primary slice), 160 C1 items preserving 40 complete four-idiom groups, 30 point targets, and 30 box targets. Booklets contain 133 or 134 items. Model–reader comparisons restrict model predictions to the identical 180 MCQs or 30 point targets. Intervals resample items, preserving the two responses per item, and therefore condition on this six-participant pool rather than estimating a professional or population ceiling.

The first box instruction could be read as asking for a semantic-object boundary. Those annotations were discarded; all six participants repeated the box items after the instruction specified transfer of the exact geographic footprint highlighted in the source. Only the clarified annotations are reported. Participant flags are retained as answerability evidence rather than used to change the frozen benchmark; excluding the four flagged general-MCQ items is reported as a sensitivity check.

The six-participant reference reaches 83.3% [78.6, 87.8] on the 180 general MCQs, 71.3% [65.0, 77.5] on the 160 complete-group C1 items, and 88.3% [76.7, 96.7] on the 30 point targets. Corresponding exact agreement is 80.6%, 75.0%, and 86.7%; general-MCQ and C1 Fleiss' κ are 0.743 and 0.665. The clarified box sample has 48.3% accuracy at IoU 0.5 and 33.3% pairwise agreement. Because the task transfers an exact source footprint rather than a semantic-object boundary, both centre displacement and extent mismatch reduce IoU sharply; the main paper therefore treats this track as secondary. All intervals cluster the two responses per item and condition on this participant pool.

F.2 C1 evidence-sufficiency analysis

We further analyse the existing 40 complete C1 four-idiom groups (10 per land-cover class; two independent responses per item), without collecting new responses or changing the frozen benchmark. For each frozen point, we reopen the generation-time ESA WorldCover cache. C1 generation assigns

(a) By representation

Representation	Acc. [95% CI]	Agree.	Flag	Conf.
Satellite	78.8 [67.5, 88.8]	72.5	0.0	4.58
Minimal street	58.8 [45.0, 72.5]	72.5	0.0	4.66
Standard street	68.8 [55.0, 81.2]	82.5	0.0	4.74
Land cover	78.8 [68.8, 87.5]	72.5	2.5	4.74

(b) By target class

Class	Sat.	Min.	Std.	LC
Built-up	95	100	100	100
Cropland/bare	75	30	30	55
Forest	55	15	45	70
Water	90	90	100	90

Table 20: **C1 human evidence diagnostics** (%; confidence on a 1–5 scale). (a) 40 items and 80 ratings per representation; intervals resample the 40 matched groups. (b) 10 items and 20 ratings per cell; Sat./Min./Std./LC denote satellite, minimal street, standard street, and land-cover views. Cells are small and descriptive.

the dominant coarse class within a 100 m disk, rather than the centre pixel alone. Boundary distance is the distance to the nearest pixel outside the target *coarse* class, and local purity is the fraction of valid pixels in a 100 m disk that map to that class. These metadata strata are observational and class-composition dependent.

Table 20 shows that the street-map reduction is class-conditional rather than uniform: built-up and water remain at 90–100%, whereas forest and cropland/bare are much less recoverable from street-map layers. Confidence remains high and neither street condition is flagged unanswerable, so self-reported flags alone do not detect this form of evidence insufficiency.

The metadata stratification points in the same direction. Human accuracy across the four representations is 56.3% for the six groups within 20 m of a coarse boundary, 55.7% for the 11 groups at 20–50 m, and 82.6% for the 23 groups at least 50 m away. Accuracy is 45.5%, 78.8%, and 82.8% for local purity below 0.80, from 0.80 to 0.95, and at least 0.95, respectively. These bins are not randomised: the far/high-purity strata contain more built-up and water targets, so they establish sensitivity rather than a boundary-only causal effect.

Filtering the two participant-flagged groups leaves 38 groups and changes the minimal/standard street results only from 58.8/68.8% to 61.8/69.7%; the 12-system ranking on this human subset is essentially unchanged (Spearman $\rho = 0.998$). Requiring mean confidence at least 4 retains 39 groups with the same rank correlation. A stricter screen retaining only the 19 groups on which both participants agree in all four idioms raises satellite/minimal/standard/land-cover accuracy to 100/89.5/89.5/100%, but retains 9 built-up and 8 water groups versus only one forest and one cropland/bare group. It is therefore not a class-balanced replacement leaderboard. An objective purity ≥ 0.80 screen retains 29 groups and yields 82.8/74.1/82.8/84.5%; the matched satellite-minus-minimal gap falls from 20.0 [8.8, 32.5] to 8.6 [-1.7, 20.7] pp. Thus insufficient class-specific visual evidence accounts for part, but not demonstrably all, of the original human C1 representation gap. On the same groups, overall accuracy before/after screening is 40.0/51.7% for GPT-5.5, 41.9/54.3% for Gemini 3 Pro, 40.6/50.0% for Qwen3.5-35B, and 48.8/60.3% for Qwen3-VL-8B. Their matched satellite-minus-minimalist gaps are respectively 0.0/0.0, 0.0/0.0, $-2.5/-3.4$, and $+5.0/+6.9$ pp (full/screened). The screen therefore improves recoverability for both readers and models without revealing a common model preference between these idioms. This analysis refines, rather than overturns, the main-text interpretation of C1: the profile includes class-dependent differences in visible evidence across map idioms. We retain the full prespecified C1 slice and interpret it as a representation profile, not a pure abstraction mechanism.

System	Matched-40	Purity$\geq.80$	Δ	Sat.−min. Full	Sat.−min. Screened
Molmo2-8B	38.8	52.6	+13.8	−2.5	0.0
Phi-4-Reasoning-Vision-15B	30.0	39.7	+9.7	+5.0	+3.4
InternVL3.5-8B	45.0	56.9	+11.9	0.0	+3.4
Kimi-VL-A3B-Instruct	40.0	50.0	+10.0	−2.5	0.0
Qwen3-VL-8B-Instruct	48.8	60.3	+11.6	+5.0	+6.9
GLM-4.6V-Flash	41.2	50.9	+9.6	0.0	−6.9
Gemma-4-26B-A4B-it	36.9	44.8	+8.0	−5.0	−13.8
Qwen3.5-9B	43.8	55.2	+11.4	−2.5	+3.4
Qwen3.6-35B-A3B	41.9	52.6	+10.7	−5.0	−6.9
Qwen3.5-35B-A3B	40.6	50.0	+9.4	−2.5	−3.4
Gemini 3 Pro†	41.9	54.3	+12.4	0.0	0.0
GPT-5.5†	40.0	51.7	+11.7	0.0	0.0
Human reference	71.3	81.0	+9.8	+20.0	+8.6

Table 21: **C1 evidence-sufficiency sensitivity across all systems (%).** Matched-40 is the 40 complete four-idiom groups in the human study; the objective purity screen retains 29 groups. Δ accuracy is screened minus Matched-40, and the last two columns are paired satellite-minus-minimalist gaps. These are descriptive sensitivity estimates on the fixed sample, not replacement leaderboard scores. For both gap columns, 10,000-replicate group-cluster 95% intervals include zero for every system; point estimates support the absence of a common direction, not equivalence. †: proprietary system.

G Sensitivity Analyses

Operational taxonomy. C1 exposes sensitivity to available texture, layers, labels, and symbology as a package. C2 probes field-value extraction and legend or iso-line decoding. C3 distinguishes category elimination from within-class correspondence and measures sensitivity to the unaligned framing package. Grounding separates parsing, centre localisation, and extent transfer. These categories organise observable errors rather than assert mutually exclusive internal mechanisms; the matched controls test whether source images, displayed labels, or supplied cartographic conventions are actually used.

Answer position. Correct A/B/C/D positions comprise 25.38/25.34/25.31/23.97% of the frozen MCQ benchmark. Reweighting correct positions equally within each task-by-option-count stratum changes every system by less than 0.2 pp (range -0.04 to $+0.19$), although the largest predicted-position share varies from 25.8% to 32.1% across systems. Because C3 letters are rendered into the image, arbitrary text-only rotation would change the evidential contract; this diagnostic addresses the observed imbalance but does not claim permutation invariance.

Geography and script. The four largest C3 regions contribute 41.2% of items. Equal-region aggregation lowers system accuracy by 1.0–7.9 pp, the four largest regions outperform the remainder by 2.3–18.6 pp, and leave-one-region-out estimates span at most 4.8 pp within a system. Thus no single region explains the aligned-framing result, but geographic concentration remains a limitation. Label-relative data contain 724 Latin and 156 CJK items (72.4% and 15.6%); smaller scripts are exploratory. Gemini 3 Pro scores 60.9/60.3%, GPT-5.5 77.6/75.0%, and Qwen3.6-35B-A3B 61.0/64.1% on Latin/CJK, so neither script has a uniform advantage.

Human-reference sensitivity. Agreement, Fleiss' κ, and the participant-flag exclusion are reported in Appendix F.1; removing the four flagged MCQs leaves the conclusions and frozen benchmark unchanged.