Look Sharp: Resolution Scaling for Cartographic Visual Question Answering

David N. Juboor
Independent

Abstract

Cartographic VQA demands multi-step spatial reasoning over complex map imagery, where frontier VLMs achieve only 38.2% versus 84.9% human accuracy on the FRIEDA benchmark. We systematically evaluate six inference-time scaling strategies—direct prompting, chain-of-thought, few-shot retrieval, uniform tiling, adaptive zoom, and zoom with SAM segmentation—across three model scales at n=500. All higher-resolution conditions significantly improve over the baseline (McNemar's $p<0.05$), with uniform 2×2 tiling achieving the highest accuracy at 44.0%. A critical ablation reveals that tiling—which provides no model-directed region selection—matches adaptive zoom (43.4%) and zoom+SAM (43.6%), all statistically indistinguishable. The bottleneck is resolution, not attention: simply providing higher-resolution crops in any form is sufficient. Cross-scale evaluation uncovers a capacity gradient: the same augmentation pipeline yields +4.8 pp on Opus, +1.6 pp on Sonnet, and −4.8 pp on a 7B model—all at n=500. Fine-tuning consistently fails, establishing inference-time resolution scaling as the preferred paradigm for capable models.

1. Introduction

Maps encode geographic information through symbols, colors, boundaries, and spatial layouts refined over centuries. Automated map understanding has direct applications in disaster response (interpreting FEMA flood maps), urban planning (parsing zoning documents), and natural resource management (reading geological survey maps). Yet computationally interpreting maps requires both reading fine-grained visual elements—legend entries, road labels, scale bars—and reasoning about spatial relationships between them. The FRIEDA benchmark [6] measures this across six spatial relation types on 500 questions spanning USGS, World Bank, FEMA, and urban planning maps. The gap is large: humans achieve 84.9% while the best VLMs score only 38.2% [6].

Domain-specific fine-tuning does not help. We find that SFT on cartographic data consistently degrades performance due to catastrophic forgetting of visual perception (Sec. 4.5). Instead, we investigate inference-time scaling: allocating additional compute at test time without modifying weights. We evaluate six conditions on FRIEDA across three model scales (7B, Sonnet 4, Opus 4) at n=500.

Key Findings

1. Resolution is the bottleneck. Uniform 2×2 tiling (44.0%) matches adaptive zoom (43.4%) and zoom+SAM (43.6%)—all $p<0.05$ vs. direct, all mutually indistinguishable. The gain comes from higher-resolution input, not model-directed selection. CoT and few-shot do not reach significance.
2. Per-relation specificity. Zoom improves perceptual tasks (Within +10.4 pp). SAM suggests structural gains (Border +5.6 pp) offset by perceptual regressions—non-significant individually but directionally consistent.
3. Capacity gradient. Augmentation yields +4.8 pp (Opus), +1.6 pp (Sonnet), and −4.8 pp (7B)—all at n=500. Benefit scales monotonically with model capacity.
4. Prompt engineering matters. Our direct baseline (39.2%) already matches the published FRIEDA best (38.2%), suggesting prompt design is an under-explored factor in cartographic VQA evaluation.

2. Related Work

Cartographic VQA. FRIEDA [6] benchmarks multi-step cartographic reasoning, establishing SOTA at 38.2%. CartoMapQA [11] evaluates symbol recognition; ReasonMap [3] targets transit maps via multi-stage RL; GTR-Bench [13] requires geo-temporal reasoning. We provide the first inference-time scaling

study on cartographic VQA.

Inference-Time Scaling for Vision. AwaRes [9] trains VLMs to zoom via GRPO tool-calling. Mini-Monkey [4] addresses sawtooth effects from fixed tiling. We extend this to cartographic imagery, where information density is spatially heterogeneous, and test whether model capacity modulates the benefit.

Spatial Reasoning in VLMs. PeBR-R1 [2] separates perception and reasoning stages. Geo-CoT [5] forces perceptually-grounded chains. TerraScope [10] combines segmentation masks with pixel-grounded reasoning. Our per-relation analysis provides empirical evidence that zoom and SAM address distinct failure modes, supporting the perception–reasoning decomposition.

3. Experimental Setup

3.1. Benchmark

We evaluate on FRIEDA [6]: 500 questions across 210 map documents. Six spatial relation types: Within (116), Distance (91), Orientation (88), Intersect (80), Border (71), Equal (54). Scoring: exact set-match for textual (LLM-judge fallback for borderline cases), ± 1 cardinal adjacency for orientation, MAPE $\leq 20\%$ for distance.

3.2. Models

Primary: Claude Opus 4 (all six conditions at n=500). Cross-scale: Qwen2.5-VL-7B [1] (8.3B params, local) and Claude Sonnet 4 (API). All use greedy decoding.

3.3. Inference-Time Techniques

Direct prompting. Single-call VQA: image (max 2048 px) + question $\rightarrow$ answer.

Structured CoT. Step-by-step reasoning: identify legend, locate features, determine spatial relationship, answer.

Few-shot retrieval. CLIP ViT-B/32 [7] retrieves k=3 similar examples from FRIEDA's 9.5K training set as text-only demonstrations.

Uniform tiling. Single-call pipeline: the image is split into a fixed 2×2 grid (4 quadrants at 1500 px each), and the model receives the overview plus all 4 crops. No model-directed selection—this serves as an ablation to isolate whether resolution or adaptive selection drives the gain.

Table 1. FRIEDA results. †Published baselines [6] using the FRIEDA standardized evaluation harness.

Method	Overall	Orient.	Dist.	Border
Human†	84.9	91.8	78.3	89.0
Ours: Tiling (2×2)	44.0	71.6	36.3	29.6
Ours: Zoom+SAM	43.6	69.3	35.2	36.6
Ours: Zoom only	43.4	77.3	33.0	31.0
Ours: Direct prompt	39.2	73.9	30.8	29.6
Gemini-2.5-Pro†	38.2	71.6	25.3	32.4
GPT-5-Think†	37.2	69.3	27.5	25.4
Claude Sonnet 4†	31.6	56.8	23.1	33.8

Adaptive zoom. Two-call pipeline (Fig. 1). Call 1: model selects 2–4 regions from 11 named positions. Call 2: model answers using overview + up to 3 high-resolution crops (max 1500 px). 93.4% of questions receive 3 crops.

Zoom + SAM. Adaptive zoom augmented with SAM 2.1 Tiny [8] segmentation. SAM extracts up to 25 regions with per-region features (color, centroid, area, OCR text) and pairwise spatial relations (adjacency, containment, direction). Structured context ($\leq$3000 chars) is injected into the answer prompt.

4. Results

4.1. Comparison with Published Baselines

Table 1 situates our results against published FRIEDA baselines. Our best result, uniform 2×2 tiling, achieves 44.0%—statistically indistinguishable from adaptive zoom (43.4%) and zoom+SAM (43.6%).

A note on baselines. Our direct prompting baseline (39.2%) already matches the published best (38.2%) before any augmentation. This 1 pp gap likely reflects prompt engineering differences rather than model capability: our prompt includes structured output formatting and a "final answer" delimiter that improves answer extraction, while published baselines use FRIEDA's standardized harness. We therefore focus our analysis on within-system comparisons (ablation across conditions with identical scoring), which are not confounded by prompt differences. The cross-system numbers in Table 1 should be interpreted with this caveat.

4.2. Ablation and Statistical Significance

Table 2 shows all six conditions with per-relation accuracy. McNemar's paired tests (Table 3) confirm that

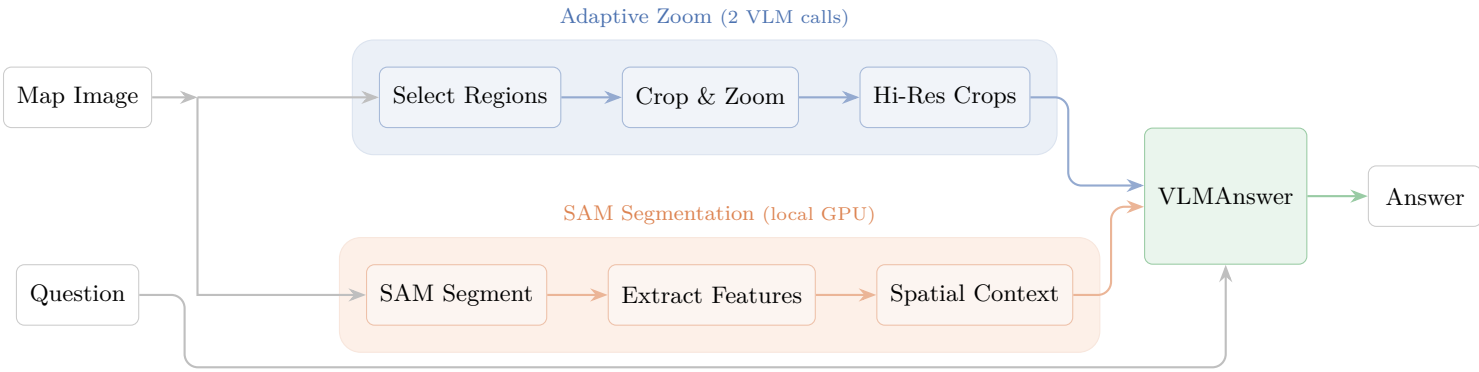


Figure 1. Pipeline. Map image forks to adaptive zoom (self-directed high-res cropping) and SAM segmentation (structured spatial context). Both merge with the question into a final VLM answer call.

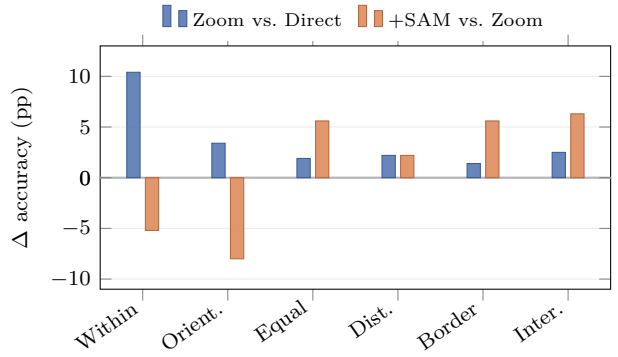


Figure 2. Per-relation marginal effects. Zoom helps perceptual tasks (left); SAM shows suggestive gains on structural tasks (right) but hurts perceptual ones.

only higher-resolution conditions significantly improve over the baseline.

All three higher-resolution conditions—tiling, zoom, and zoom+SAM—significantly beat the baseline ($p<0.05$) but are statistically indistinguishable from each other. Critically, uniform 2×2 tiling (44.0%) matches adaptive zoom (43.4%) with a single API call and no model-directed region selection ($\chi^2=0.06$, n.s.), indicating that the gain comes from resolution, not from the model's ability to select where to look. CoT and few-shot do not reach significance despite 1.6–2.0 pp gains.

4.3. Per-Relation Analysis

The aggregate equivalence of zoom and zoom+SAM masks opposing per-relation effects (Fig. 2). Per-relation McNemar's tests are all non-significant (see Supplementary Table 6), so we present these as descriptive patterns.

Zoom helps perception: Within (+10.4 pp) and Orientation (+3.4 pp) require reading fine labels and features—bottlenecked by resolution. High-res crops directly address this.

SAM suggests a structural pattern: Border (+5.6 pp over zoom) and Intersect (+6.3 pp) require determining shared boundaries—topology implicit in pixels but explicit in SAM's structured output. These gains are offset by Orientation (−8.0 pp) and Within (−5.2 pp) regressions, consistent with the 3000-char structured context competing for attention with visual information. The opposing effects cancel in aggregate ($\chi^2=0.00$). Per-relation McNemar's tests are all non-significant (Supplementary Table 6), so we present this as a descriptive hypothesis: the consistent directionality (all structural types gain, both perceptual types lose) is suggestive but unconfirmed at our sample size.

4.4. Capacity Gradient

The same augmentation approach that significantly improves Opus (+4.8 pp, $p<0.05$) yields diminishing returns at smaller scales: a modest +1.6 pp on Sonnet and an outright −4.8 pp degradation on the 7B model (Table 4). Augmentation benefit scales monotonically with model capacity.

The pattern parallels findings in chain-of-thought scaling [12], where intermediate reasoning traces improve large language models but degrade smaller ones. The mechanism is analogous: high-resolution crops and structured spatial context provide scaffolding that capable models can selectively attend to, while smaller models lack the context processing capacity to distinguish signal from noise. The 7B model, processing the overview image plus additional zoom crops simultaneously, is flooded with information it cannot effectively filter, degrading performance below what it achieves from the image alone.

The practical implication is that augmentation

Table 2. Inference-time ablation (n=500, Claude Opus 4). Per-relation accuracy (%) with Δ vs. Direct. Best per-column in bold.

Condition	Overall (500)	Orient. (88)	Within (116)	Equal (54)	Dist. (91)	Border (71)	Inter. (80)
Direct prompt	39.2	73.9	33.6	37.0	30.8	29.6	28.8
+ CoT	40.8 (+1.6)	73.9	31.9 (-1.7)	40.7 (+3.7)	38.5 (+7.7)	32.4 (+2.8)	27.5
+ Few-shot	41.2 (+2.0)	72.7	38.8 (+5.2)	42.6 (+5.6)	31.9	26.8	32.5 (+3.8)
+ Tiling (2×2)	44.0 (+4.8)	71.6 (-2.3)	41.4 (+7.8)	46.3 (+9.3)	36.3 (+5.5)	29.6	37.5 (+8.8)
+ Zoom (adaptive)	43.4 (+4.2)	77.3 (+3.4)	44.0 (+10.4)	38.9	33.0	31.0	31.3
+ Zoom+SAM	43.6 (+4.4)	69.3 (-4.5)	38.8 (+5.2)	44.4 (+7.4)	35.2 (+4.4)	36.6 (+7.0)	37.5 (+8.8)

Table 3. McNemar's test (n=500). b/c: discordant pairs.

A → B	b	c	χ^2	Sig.
Direct → CoT	34	26	0.82	n.s.
Direct → Few-shot	34	24	1.40	n.s.
Direct → Tiling	56	32	6.01	p<.05
Direct → Zoom	51	30	4.94	p<.05
Direct → Zoom+SAM	53	31	5.25	p<.05
Tiling → Zoom	33	36	0.06	n.s.
Tiling → Zoom+SAM	37	39	0.01	n.s.
Zoom → Zoom+SAM	32	31	0.00	n.s.

Table 4. Cross-scale analysis (n=500). Best augmented condition per model. †7B uses fixed zoom crops (condition closest to tiling available at scale); pilot zoom+SAM (n=50) also degrades to 6.0%.

Model	Baseline	+ Augmented	Δ
Qwen-7B (8.3B)†	13.2	8.4	−4.8
Claude Sonnet 4	17.6	19.2	+1.6
Claude Opus 4	39.2	44.0	+4.8

should not be applied blindly. For smaller models, better prompting is a more productive investment than inference-time augmentation. For frontier models, adaptive zoom provides significant gains at modest cost (2× API calls). A quick baseline evaluation is essential before committing to a more expensive inference-time pipeline.

4.5. Error Analysis and Fine-Tuning

Qualitative examples illustrating per-condition behavior—SAM resolving boundary queries, zoom resolving small symbols, and SAM causing perceptual regressions—are provided in Supplementary C.

Error taxonomy. Of 434 errors from the 7B baseline: 72.6% factually incorrect, 12.0% correct answer in wrong format (∼10 pp recoverable), 9.4% verbose reasoning traces, 6.0% near-misses. Manual inspection reveals a roughly even perception–reasoning split, consistent with zoom addressing perception and SAM targeting reasoning.

Fine-tuning fails. SFT on Qwen2.5-VL-3B with LoRA (rank 16, 2000 FRIEDA examples): zero-shot 7.4%, +SFT 7.4%, +CoT traces 7.0%, +multi-crop 8.4%. The zero-shot 7B (13.2%) outperforms every fine-tuned 3B variant. Scaling the model dominates fine-tuning; scaling inference-time compute then adds further gains.

5. Analysis

Resolution is the bottleneck, not attention. Cartographic maps are information-dense documents often produced at resolutions of 5000–10000 px, but VLM APIs typically accept images downscaled to 2048 px or fewer tokens. At this resolution, fine text (street names, contour labels, legend entries) and small symbols (point features, boundary markers) become illegible. Providing higher-resolution crops directly resolves this bottleneck.

The tiling result (Table 2) isolates the mechanism: uniform 2×2 tiling—which provides no model-directed selection—achieves 44.0%, matching adaptive zoom (43.4%, χ^2=0.06, n.s.) at half the API cost (1 call vs. 2). This means the gain comes from showing the model higher-resolution pixels, not from the model's ability to choose which pixels to examine. The Within and Equal improvements in tiling (+7.8 pp, +9.3 pp) are comparable to adaptive zoom (+10.4 pp, +1.9 pp), confirming that resolution-limited perception errors are the primary failure mode. Adaptive zoom's advantage in Orientation (+3.4 pp vs. −2.3 pp for tiling) suggests that selective cropping may help when the relevant information is spatially concentrated, but this is offset by tiling's broader coverage of the full image.

Table 5. Cost and reproducibility. Per-question costs at Opus 4 API pricing ($15/$75 per 1M input/output tokens). SAM runs locally on a single GPU.

Condition	API calls	Images	$/question	Total
Direct	1	1	∼0.05	∼$25
CoT	1	1	∼0.06	∼$30
Few-shot	1	1	∼0.08	∼$40
Tiling (2×2)	1	1+4	∼0.11	∼$55
Zoom	2	1+3	∼0.11	∼$55
Zoom+SAM	2	1+3	∼0.13	∼$65
Full evaluation (6 cond. × 500q)				∼$270

SAM: possible structural scaffolding with attentional cost. SAM's structured context makes spatial topology explicit—region boundaries, pairwise adjacency, containment relationships—information that is implicit in pixel data but requires integrating visual evidence across large spatial extents. For structural relations like Border and Intersect, where a model must determine whether two regions share a boundary, this explicit representation may provide scaffolding that complements the visual input. However, the 3000 characters of structured text compete for the model's finite attention budget. For perceptual tasks already well-served by zoom (Orientation, Within), this additional context appears counterproductive—the model's attention may be diverted from visual evidence that already contains the answer. This attention competition hypothesis is consistent with zoom and zoom+SAM being statistically equivalent in aggregate (χ^2=0.00), but we emphasize that the per-relation effects are non-significant individually (Supplementary Table 6). An adaptive pipeline that selects augmentation strategy based on the question type could in principle capture the structural gains without the perceptual cost, but validating this requires larger per-relation evaluation sets.

Computational cost. Zoom doubles the number of API calls (2 vs. 1 for direct), with the answer call processing up to 4 images (overview + 3 crops). Table 5 reports per-condition costs. SAM inference runs locally on GPU at negligible marginal cost, but since zoom and zoom+SAM are statistically equivalent, the simpler zoom-only pipeline is preferred on cost-efficiency grounds.

Per-source variation. Augmentation benefit varies substantially by map type (Supplementary Table 7). The largest gains come from structured geographic maps with clean visual layouts: national park maps (+18.2 pp, n=22) and Seychelles development maps (+14.3 pp, n=14), where zoom-enhanced legend reading has the highest payoff. Mining technical reports (NI43-101, n=166) and FEMA flood maps (n=83) show moderate gains (+6–7 pp). Dense urban zoning maps show regressions (Cape Town −6.7 pp, n=30), suggesting that complex overlapping polygon structures may confuse the pipeline. These patterns, while based on small subgroup samples, could inform selective deployment strategies.

Limitations. Adaptive zoom uses a fixed vocabulary of 11 named regions with predefined pixel fractions, which may miss relevant information in non-standard map layouts; a learned bounding-box approach (as in AwaRes [9]) could be more flexible but requires training. All results are on a single benchmark (FRIEDA); generalization to CartoMapQA [11] or GTR-Bench [13] is untested. Our scoring includes an LLM-judge fallback for borderline textual matches—all conditions use identical scoring, so relative comparisons are valid, but absolute numbers may differ from strictly exact-match evaluations. The cross-system comparison in Table 1 is confounded by prompt engineering differences (Sec. 4); our within-system ablation is the more reliable analysis.

6. Conclusion

Higher-resolution input is the only statistically validated inference-time technique for cartographic VQA, improving Claude Opus 4 from 39.2% to 44.0% on FRIEDA (p<0.05). A critical ablation shows that uniform 2×2 tiling matches adaptive zoom and zoom+SAM—the bottleneck is resolution, not attention. SAM segmentation shows suggestive per-relation patterns that cancel in aggregate. Augmentation benefit scales monotonically with model capacity: +4.8 pp (Opus), +1.6 pp (Sonnet), −4.8 pp (7B). Fine-tuning consistently fails. The practical recommendation is simple: scale the model, then provide higher-resolution crops—a single-call tiling pipeline suffices. A 41 pp gap to human performance remains.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025. 2

[2] Yan Chen, Long Li, Teng Xi, Long Zeng, and Jingdong Wang. Perception before reasoning: Two-stage reinforcement learning for visual reasoning. arXiv preprint arXiv:2509.13031, 2025. ICLR 2026. 2

[3] Sicheng Feng, Kaiwen Tuo, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. ReasonMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. arXiv preprint arXiv:2510.02240, 2025. ICLR 2026. 1

[4] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. arXiv preprint arXiv:2408.02034, 2024. ICLR 2025. 2

[5] Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Perceptually-grounded GeoSpatial chain-of-thought for VLMs. arXiv preprint arXiv:2509.22221, 2025. 2

[6] 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. arXiv preprint arXiv:2512.08016, 2025. ICLR 2026. 1, 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandeep Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021. 2

[8] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 2

[9] Nimrod Shabtay, Moshe Kimhi, Artem Spector, Sivan Haray, Ehud Rivlin, Chaim Baskin, Raja Giryes, and Eli Schwartz. Look where it matters: High-resolution crops retrieval for efficient VLMs. arXiv preprint arXiv:2603.16932, 2026. 2, 5

[10] Yan Shu, Bin Ren, et al. TerraScope: Pixel-grounded visual reasoning for earth observation. arXiv preprint arXiv:2603.19039, 2026. CVPR 2026. 2

[11] 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: Cartographic map question answering. arXiv preprint arXiv:2512.03558, 2025. SIGSPATIAL 2025. 1, 5

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS, 2022. 3

[13] Qinghongbing Xie, Zhaoyuan Xia, Feng Zhu, Lijun Gong, Ziyue Li, Rui Zhao, and Long Zeng. GTR-Bench: Geo-temporal reasoning benchmark. arXiv preprint arXiv:2510.07791, 2025. 1, 5

Supplementary Material

Look Sharp: Resolution Scaling for Cartographic VQA

A. Per-Relation McNemar's Test

Table 6 shows McNemar's test for zoom vs. zoom+SAM broken out by spatial relation type. No individual relation reaches significance, though the direction is consistent: all structural types (Border, Intersect, Equal) favor SAM while both perceptual types (Orientation, Within) favor zoom alone.

Table 6. McNemar's: Zoom vs. Zoom+SAM by relation.

Relation	n	b	c	χ^2	Sig.
Border	71	7	3	0.90	n.s.
Intersect	80	8	3	1.45	n.s.
Equal	54	4	1	0.80	n.s.
Distance	91	7	5	0.08	n.s.
Within	116	4	10	1.79	n.s.
Orientation	88	2	9	3.27	n.s.

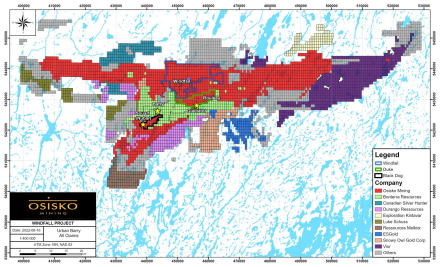
B. Per-Source Accuracy

Table 7. Accuracy by document source.

Source	n	Direct	Zoom+SAM	Δ
NI43-101 (mining)	166	31.9	38.6	+6.7
FEMA (flood)	83	50.6	56.6	+6.0
EIS (environment)	62	38.7	37.1	−1.6
Seattle planning	60	38.3	43.3	+5.0
Cape Town	30	46.7	40.0	−6.7
AIIB	27	25.9	29.6	+3.7
Ireland	24	45.8	45.8	0.0
National Parks	22	50.0	68.2	+18.2
Seychelles	14	35.7	50.0	+14.3

C. Qualitative Examples

q_7021 | Border | SAM helps

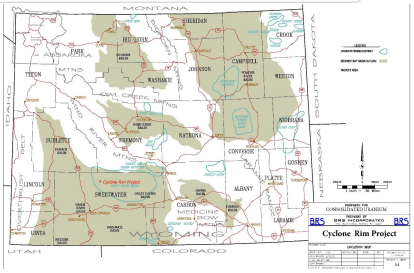


Q: "NE border of Black Dog shared with?"
Expected: Bonterra Resources

Direct:	Snowy Owl Gold	✗
Zoom:	Osisko Mining	✗
Zoom+SAM:	Bonterra Resources	✓

SAM boundary context resolves adjacent mining claims that both direct and zoom miss.

q_5021 | Within | Zoom+SAM helps

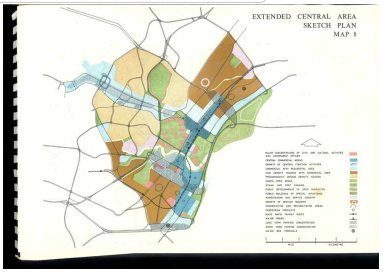


Q: "Which district is Cyclone Rim in?"
Expected: Great Divide Basin

Direct:	Crooks Gap	✗
Zoom:	Crooks Gap	✗
Zoom+SAM:	Great Divide Basin	✓

Near-boundary project marker corrected by combined pipeline.

q_0103 | Within | SAM hurts

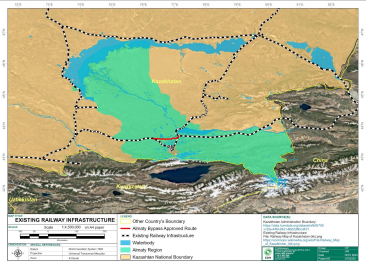


Q: “Bus terminals in High Density Housing?”
Expected: 1

Direct:	2	✗
Zoom:	1	✓
Zoom+SAM:	2	✗

SAM context causes count regression—structured text distracts from visual evidence.

q_0088 | Intersect | SAM enables precise crossing count

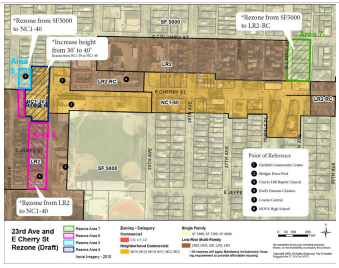


Q: “How many times does ‘Existing Railway Infrastructure’ cross the ‘Other Country’s Boundary’?”
Expected: 3

Direct:	(hallucinated trace)	✗
Zoom:	4	✗
Zoom+SAM:	3	✓

Baseline hallucinates. Zoom over-counts. SAM’s boundary overlay enables precise intersection counting.

q_0002 | Border | SAM enables correct boundary identification



Q: “With which Rezone Area does ‘Rezone Area 4’ share its westernmost boundary?”
Expected: Rezone Area 6

Direct:	Rezone Area 5	✗
Zoom:	Rezone Area 5	✗
Zoom+SAM:	Rezone Area 6	✓

Both baseline and zoom fail identically. SAM’s boundary segmentation correctly identifies the adjacent hatched zone.

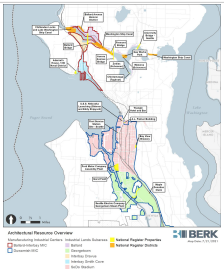
q_0024 | Within | Zoom alone resolves small symbols

Q: “How many ‘Terminal & Rail Yard Gate’ points are outside both ‘Intermodal Facilities’?”
Expected: 7

Direct:	(unconstrained narrative)	✗
Zoom:	7	✓
Zoom+SAM:	7	✓

Zoom resolves the small point symbols. SAM adds nothing—this is a pure perception task.

q_0026 | Within | SAM causes district hallucination



Q: “Which National Register Districts overlap with both Ballard and Interbay Dravus subareas?”
Expected: Washington Ship Canal

Direct:	Chittenden Locks…	✗
Zoom:	Washington Ship Canal	✓
Zoom+SAM:	Ballard Ave Historic Dist.	✗

Zoom alone succeeds. Adding SAM context redirects attention to the wrong district entirely—a concrete example of structured context causing regression.

D. Prompt Templates

Tiling Prompt (Sec. 3.3).

```
{instruction}
## Question
{question}
You are seeing the map at multiple
zoom levels:
- Image 1: Full map overview
- Image 2: Top-left quadrant (NW)
- Image 3: Top-right quadrant (NE)
- Image 4: Bottom-left quadrant (SW)
- Image 5: Bottom-right quadrant (SE)
Use the zoomed-in quadrant views to read
text labels, legend entries, and fine
details accurately.
1. Read legend from the zoomed views
2. Identify features in the question
3. Determine spatial relationship
Final answer: [your answer]
```

CoT Prompt (Sec. 3.3).

```
Analyze this map step-by-step.
{instruction}
Follow these steps carefully:
1. Legend: What do colors, symbols,
   patterns represent?
2. Compass: Which direction is north?
3. Scale: What distance info is shown?
4. Features: What map features are
   relevant? Name them specifically.
5. Spatial reasoning: How do features
   relate spatially?

## Question
{question}
Work through each step, then:
Final answer: [your answer]
```

Zoom Scan Prompt.

```
{instruction}
## Task: Identify Key Regions
Look at this map overview. I need to
answer: "{question}"
Tell me which regions to zoom into.
Output JSON of 2-4 regions:
{"zoom_regions": [
  {"region": "legend",
   "reason": "contains the legend"},
  {"region": "center",
   "reason": "contains target features"}
]}
Valid: top-left, top-right, bottom-left,
bottom-right, top, bottom, left, right,
center, legend, top-center
```

Zoom Answer Prompt.

```
{instruction}
## Question
{question}
You see the map at multiple zoom levels:
- Image 1: Full overview
{crop_descriptions}
Use zoomed views for text/legend detail.
Cross-reference overview for context.
1. Read legend from zoomed view
2. Identify features in the question
3. Determine spatial relationship
Final answer: [your answer]
```

Zoom+SAM Answer Prompt.

```
{instruction}
## Question
{question}
## CV Analysis (SAM 2.1 + OCR)
{spatial_context}
## Multi-Resolution Views
- Image 1: Full overview
{crop_descriptions}
Use ALL three sources:
1. Zoomed views for text/legend detail
2. SAM analysis for region boundaries
   and spatial relationships
3. Full overview for spatial context
Final answer: [your answer]
```