UAV-MAPVQA: EVALUATING FOUNDATION MODELS ON INSTANCE-TO-ENTITY GROUNDING QUESTION ANSWERING ACROSS UAV IMAGERY AND VECTOR MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Current aerial perception benchmarks largely evaluate class-level recognition within the image plane, such as detection, counting, and scene understanding, but do not test whether models can ground visual instances to their corresponding real-world entities in a vector map. This capability is critical for UAV applications that rely on vector maps as compact, structured world knowledge for linking observations to semantics, attributes, and spatial context. We introduce UAV-MapVQA, the first benchmark for cross-view instance-level grounding and geospatial reasoning between UAV images and vector maps. The benchmark contains 336K multiple-choice QA pairs generated from 14,000 real-world UAV images collected across 3 countries (USA, China, and Germany) and more than 1,000 km of flights, using an automatic registration and formally verified QA generation pipeline with human validation. It spans eight tasks across four levels, from cross-view instance–entity grounding and change detection to viewpoint-aware localization and object-centric geospatial reasoning. Each question is provided with two alternative representations of the same vector map—a rendered raster map and a structured entity table—enabling controlled study of how map representation affects cross-view grounding and geospatial reasoning. Evaluating 13 foundation models (FMs), spanning proprietary and open-weight general-purpose models such as GPT-5.5 and Claude Opus 4.7 as well as domain-specific geospatial FM GeoChat, we find that no zero-shot model exceeds 50% overall accuracy, against 94.8% for humans. Supervised fine-tuning on UAV-MapVQA raises Qwen2.5-VL-32B from 23.0% to 57.2%, surpassing all zero-shot models, which confirms that the dataset provides an effective training signal; yet a large gap to humans remains, and aligning the map to the UAV heading shows that orientation errors propagate into downstream reasoning but explain only part of this gap. These findings expose major gaps in cross-view instance grounding and map-conditioned geospatial reasoning, establishing UAV-MapVQA as a new evaluation and training resource for map-aware multimodal intelligence at the intersection of aerial robotics, multimodal learning, and geospatial AI. An interactive demo of the benchmark is available at an anonymous URL: `https://storage.googleapis.com/uav_map_vqa/index.html`.

1 INTRODUCTION

Unmanned aerial vehicles (UAVs) are increasingly supporting tasks such as search-and-rescue, emergency response, infrastructure inspection, urban analytics, traffic monitoring, and law enforcement (Dhaouadi et al., 2025). In these applications, recognizing objects purely within the image plane is often insufficient (Xia et al., 2018): a UAV must often link what it sees to real-world entities represented in a map, such as a specific building, road, or landmark of interest (Zheng et al., 2020; Wang et al., 2025). This linkage enables access to map-defined semantics, attributes, and spatial context, such as building function, speed limit, address, topology, or nearby infrastructure (Lobry et al., 2020). Such *instance-to-entity grounding* is a prerequisite for many forms of map-aware geospatial reasoning and decision-making in aerial robotics (Kuckreja et al., 2024).

Among available world representations, vector maps are especially important for UAV systems because they provide a compact, structured, and queryable representation of the environment (Bai

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.

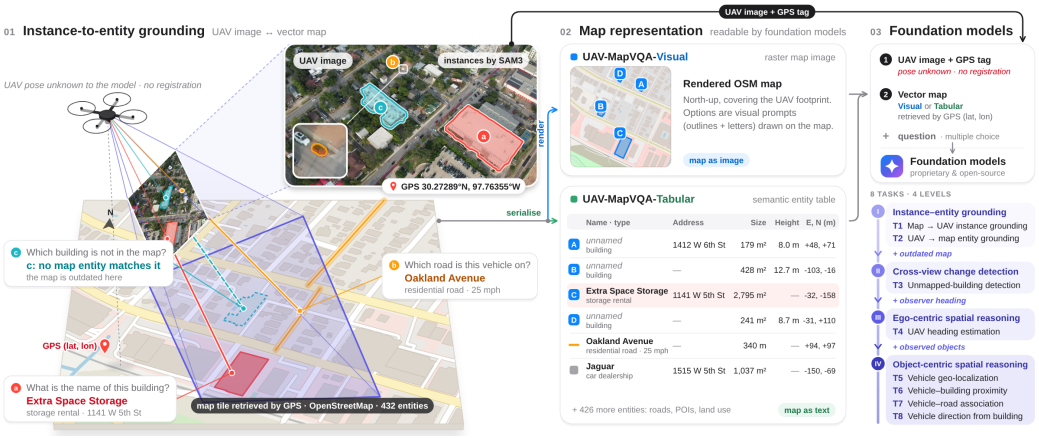


Figure 1: **UAV-MapVQA asks: can a foundation model ground a drone-observed visual instance to its exact real-world entity in a vector map, and reason using map knowledge?** We instantiate this question with 336K automatically generated QA pairs over 14K registered real-world UAV images from 3 countries and more than 1,000 km of flights, pairing UAV imagery with vector maps for instance-to-entity grounding and map-conditioned geospatial reasoning.

et al., 2025a). Unlike raster satellite imagery, vector maps encode explicit geometry, topology, and attributes for entities such as roads, buildings, intersections, and land-use regions. This is particularly useful for edge or onboard deployment, where storage and retrieval efficiency matter. For instance, while a full OpenStreetMap vector map of North America is around 13.3 GB (Geofabrik GmbH, 2026), the corresponding high-resolution satellite imagery can easily exceed hundreds of terabytes. Beyond compactness, vector maps expose the symbolic and relational structure needed for reasoning about visibility, adjacency, containment, proximity, and legal or functional attributes.

Despite rapid progress in foundation models (FMs) and multimodal spatial reasoning (Xing et al., 2025; Dihan et al., 2025; Danish et al., 2025; Liu et al., 2026; Yu et al., 2025; Bao et al., 2026), current benchmarks for aerial vision largely focus on class-level image understanding tasks such as object detection, segmentation, object counting, scene classification, captioning, visual grounding, or visual question answering over a single image (Liu et al., 2025; Kuckreja et al., 2024; Danish et al., 2025; Lobry et al., 2020). These tasks do not evaluate whether a model can determine which *specific* entity in a map corresponds to a visual instance in a UAV image, especially when the image and map are not explicitly aligned. Likewise, existing instance-recognition benchmarks, such as landmark recognition (Weyand et al., 2020), usually target highly distinctive categories and often rely on instance-specific visual memorization, which does not generalize to common buildings, roads, or urban structures that are visually repetitive and only identifiable through spatial layout and surrounding context. As a result, the ability of FMs to perform cross-view, cross-modal instance grounding between UAV imagery and vector maps remains largely unmeasured.

This problem is fundamentally challenging because it requires resolving both the modality gap and spatial misalignment (Tang et al., 2025; Lu et al., 2026): UAV images are often captured from oblique viewpoints with arbitrary camera pose, scale, illumination, and partial visibility (Ye et al., 2024), whereas maps are typically orthographic, north-up, and vector-encoded (Wang et al., 2025), so a model must bridge differences in appearance, projection, orientation, abstraction level, and field of view while distinguishing among many visually similar candidates. Solving the task typically requires combining local appearance cues with global spatial layout, object-to-context relations (e.g., a vehicle relative to a road or building), and map attributes not directly visible in the image—reasoning humans perform naturally but current FMs may not (Kuckreja et al., 2024; Zhou et al., 2026).

To address this gap, we introduce **UAV-MapVQA**, the first benchmark for instance-level grounding and map-based geospatial reasoning between UAV images and vector maps. Built from 14,000 real-world UAV images across 3 countries and over 1,000 km of flights, it uses an automatic UAV–map registration pipeline and a formal-method-inspired, human-verified question generation procedure to yield 336K multiple-choice QA pairs testing whether a model can identify the map entity for a visual instance and reason over map-defined attributes and relations. It spans eight tasks at four levels—bidirectional instance–entity grounding, change detection, ego-centric heading estimation,

and object-centric vehicle-to-map reasoning—each with two synchronized map representations from the same OSM data: **UAV-MapVQA-Visual** (rendered raster) and **UAV-MapVQA-Tabular** (structured entity table with tags, coordinates, and relations).

Evaluating 13 proprietary, open-weight, and domain-specific FMs, we find that all fall far short of human performance, and that a domain-specific remote-sensing FM does not transfer to cross-view map grounding. Fine-tuning narrows but does not close this gap; heading-aligned maps show orientation errors propagate into downstream reasoning without explaining most of the remaining gap. These results establish UAV-MapVQA as both a challenging benchmark and a training resource for map-aware multimodal intelligence at the intersection of aerial robotics, multimodal learning, and geospatial AI. Our contributions are summarized below.

- **Task formulation.** We formulate instance-to-entity grounding between UAV images and vector maps as a benchmark problem for aerial robotics and geospatial FMs.
- **Benchmark.** We present UAV-MapVQA: 336K human-verified multiple-choice QA pairs from 14,000 registered real-world UAV images collected across 3 countries and over 1,000 km of flights. It spans eight tasks at four levels, from instance–entity grounding and change detection to egocentric heading estimation and object-centric reasoning, with two synchronized map representations (raster Visual and structured Tabular) for every question.
- **Evaluation and training utility.** We evaluate 13 proprietary, open-weight, and remote-sensing FMs on both representations and show that all remain far below human performance. Fine-tuning an open-weight model on UAV-MapVQA surpasses all zero-shot models, demonstrating the dataset's effectiveness as a training signal, while heading-alignment experiments show that orientation errors are a measurable but partial bottleneck, leaving substantial room for improvement.

2 Related Work

UAV Imagery Benchmarks. Early aerial-image benchmarks (Ding et al., 2021; Waqas Zamir et al., 2019; Hellekes et al., 2024; Wu et al., 2024) targeted perception primitives such as detection, segmentation, and tracking, while recent UAV vision-language benchmarks push toward higher-level understanding: VQA-AID (Sarkar & Rahnemoonfar, 2021) studies post-disaster aerial VQA, AerialVG (Liu et al., 2025) and DVGBench (Zhou et al., 2026) evaluate aerial visual grounding, GeoText-1652 (Chu et al., 2024) aligns aerial images with language for navigation, and UAVBench (Zhan & Yuan, 2026) and UAVReason (Sun et al., 2026) scale evaluation across low-altitude tasks. All, however, treat drone imagery in isolation, leaving grounding to external maps unexplored.

Table 1: Comparison with representative datasets. MM: pairs imagery with a separate map modality; Inst.: instance-level; RW Ent.: real-world entity grounding; X-view: requires correspondence across viewpoints; ∼: partial support.

Benchmark	MM	QA	Inst.	RW Ent.	X-view
RSVQA (Lobry et al., 2020)	✓	✓	✗	✗	✗
GeoChat (Kuckreja et al., 2024)	✗	✓	∼	✗	✗
GEOBench-VLM (Danish et al., 2025)	✗	✓	∼	✗	✗
AerialVG (Liu et al., 2025)	✗	✗	✓	✗	✗
DVGBench (Zhou et al., 2026)	✗	✗	✓	✗	✗
Pix2Map (Wu et al., 2023)	✓	✗	✗	✗	✓
MapQA (Arnold et al., 2026)	✗	✓	✗	✗	✗
UAV-MapVQA (ours)	✓	✓	✓	✓	✓

Map-based and Remote-Sensing Benchmarks. Map-based benchmarks such as MapIQ (Srivastava et al., 2025), MapQA (Arnold et al., 2026), MapVerse (Bhat et al., 2026), MapEval (Dihan et al., 2025), and MSSG (Xing et al., 2025) evaluate MLLMs on map visualizations and geospatial reasoning, but typically treat the map as the sole visual input rather than requiring correspondence with imagery from another viewpoint. Remote-sensing VQA benchmarks—RSVQA (Lobry et al., 2020), Prompt-RSVQA (Chappuis et al., 2022), PAN-RSVQA (Chappuis et al., 2025)—incorporate geographic structure via OSM-derived supervision, prompting, pseudo-annotation, pretraining, or retrieval, but assume georeferenced overhead imagery aligned to global coordinates, unlike UAV observations captured from arbitrary 6-DoF viewpoints, leaving a gap in cross-modal reasoning between UAV images and maps.

Geospatial Reasoning for MLLMs. Evaluating and improving MLLMs for geospatial reasoning has attracted growing attention, spanning domain-specific fine-tuning and benchmarking (Kuckreja et al., 2024; Danish et al., 2025; Tsujimoto et al., 2026; Tong et al., 2024), instance-level recognition (Weyand et al., 2020), and retrieval- or agent-augmented reasoning over structured geospatial databases (Yu et al., 2025; Bao et al., 2026) (see Appendix A for an extended discussion). A



Figure 2: Dataset construction pipeline. UAV frames sampled by distance travelled (01) are registered to a satellite basemap (02). Projected OSM buildings and roads provide a map prior but retain residual offsets (03), so SAM3 proposals detected independently of OSM (04) are aligned and matched to the map prior, and only pairs supported by both sources are kept (05). The verified correspondence pool (06) supports QA generation for T1–T8.

closely related line of work is *cross-view geo-localization*, exemplified by University-1652/GeoText-1652 (Zheng et al., 2020; Chu et al., 2024) and Pix2Map (Wu et al., 2023): these match a UAV or ego-view image against satellite imagery or road-graph maps to retrieve the *camera's own* location or pose via image-image similarity. UAV-MapVQA targets a different problem—grounding individual visual *instances* (buildings, roads, vehicles) to persistent entities in a structured vector map and answering questions over their map-defined attributes and relations—combining arbitrary-view UAV imagery, structured vector maps, and instance-level cross-view reasoning in a single QA benchmark.

3 The UAV-MapVQA Dataset

UAV-MapVQA pairs real UAV images with the OpenStreetMap (OSM) vector map of the same area and asks multiple-choice questions whose answer requires grounding entities across the two views. The dataset has two synchronized representations: **UAV-MapVQA-Visual**, where the map is a rendered raster, and **UAV-MapVQA-Tabular**, where the same OSM window is serialized as entity tables. Both versions share the UAV images, questions, options, and answers.

3.1 Problem Definition

Each sample is a tuple $(I, M, q, \mathcal{O}, a, p)$, where I is a UAV image, M is the map input for the registered area, q is a natural-language question, $\mathcal{O}$ is the set of answer options, $a \in \mathcal{O}$ (or $a \subseteq \mathcal{O}$ for multi-answer questions) is the ground-truth answer, and p is an optional visual prompt marking the queried instance or candidate options on I and/or M. The model predicts a from $(I, M, q, \mathcal{O}, p)$. Both dataset versions derive $M \in \{M_{\text{vis}}, M_{\text{tab}}\}$ from the same OSM data: M_{vis} is a north-up raster map that preserves geometry and layout, while M_{tab} lists the same entities, attributes, geometry, and relations as structured text, exposing semantics that the raster only implies.

3.2 Task Design

A UAV operating over a mapped area needs to answer operational questions that move from cross-view identification to map-aware action (Figure 3; construction rules and additional examples in Appendix C). **Level I, instance–entity grounding** tests whether the UAV can connect ordinary buildings across the two views: **T1** gives a dispatcher-specified mapped building that the model must find in the UAV image; **T2** reverses this, reporting which mapped entity an observed building is. **Level II, change detection** addresses map staleness: since OSM may miss newly built, damaged, or modified structures, **T3** asks the model to identify labelled buildings visible in the UAV image but absent from the map, as in post-event survey or routine updating. **Level III, ego-centric reasoning** tests whether the UAV can orient itself relative to the map: **T4** estimates the compass direction faced by the image's top edge, needed when onboard heading or GNSS is unreliable. **Level IV, object-centric reasoning** grounds transient objects visible only in the UAV image: a marked vehicle is localized

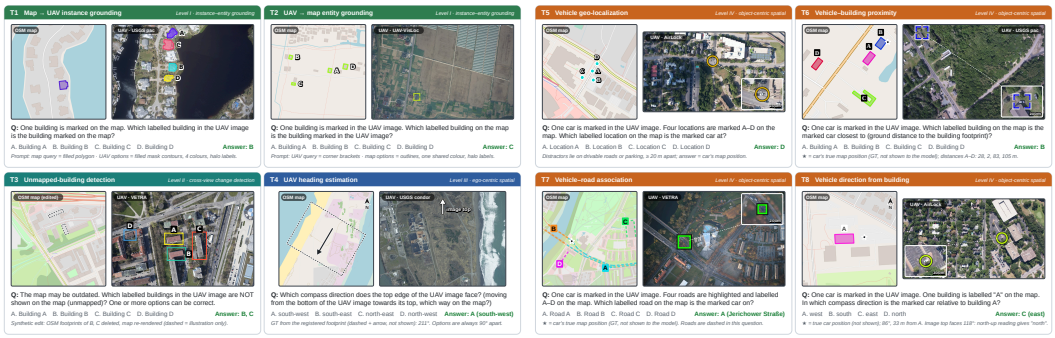


Figure 3: One real question per task (Visual version). Left: T1–T4; right: T5–T8. Each card shows the map and the UAV image with the question's randomized visual prompts, the question, the options and the answer. Dark dashed lines and ⋆ mark ground truth for the reader and are never shown to the model. The same questions in the Tabular version are shown in Appendix C.1 (Figures 7 and 8).

on the map (**T5**), linked to the nearest mapped building (**T6**), assigned to the road it occupies (**T7**), and described by compass direction from a reference building (**T8**). The levels deliberately add one dependency at a time, so per-task accuracy identifies whether a model fails at correspondence, map-update reasoning, orientation recovery, or downstream object-context reasoning.

3.3 Dataset Construction

Figure 2 summarizes the pipeline; each stage is described below.

UAV data collection. We aggregate UAV imagery from four public sources, VETRA (Hellekes et al., 2024), UAV-VisLoc (Xu et al., 2024), AirLock (Deng et al., 2026), and USGS-RSCC (Ritchie et al., 2023), covering urban, suburban, and mixed scenes in three countries (United States, China, Germany) and more than 1,000 km of flights (Appendix B.1). Frames carry approximate GPS but not 6-DoF poses. To avoid near-duplicate frames, we subsample each flight by *distance travelled*, keeping a frame only once the platform has moved far enough that its footprint barely overlaps the previous one, spreading samples along the trajectory and increasing scene diversity. Since cross-view grounding is informative only when an entity must be distinguished from similar neighbours—a scene with one or two buildings can often be solved from presence or coarse position alone—we run SAM3 (Carion et al., 2025) with the prompt "building" and keep only frames with at least four detected instances: the minimum for every option in a T1–T3 question to be a real, distinct building in the same image, and enough anchors to estimate the residual image–map offset.

UAV-to-map registration. We register each frame to Mapbox Satellite imagery (Mapbox, 2026): the GPS position selects the basemap region, RoMa (Edstedt et al., 2024) matches the UAV frame to it, and a homography is estimated from the correspondences, rejecting registrations with too few inliers or implausible projected footprints (non-convex shape, extreme angles, aspect-ratio or area distortion, or degenerate quadrilaterals) after inspecting sequences with playback videos and overlays. From the registered frames that pass the building-count threshold, we sample 3,500 images per source, yielding 14,000 image–map pairs with bidirectional pixel–geocoordinate mappings (Appendix Figure 5). To quantify residual error, six annotators manually georeferenced a stratified sample of 1,440 retained images (10.3%) in QGIS (QGIS Development Team, 2026) with ground control points: comparing the manual and automatic transformations on a 64×64 grid per image gives a median error of 1.9 m (P25 1.3 m, P75 2.6 m, P95 5.4 m), with 78%, 93%, and 99% of grid points within 3, 5, and 9 m (Appendix Table 7). This suffices for map-window placement and heading estimation, but not by itself for instance labels in dense neighborhoods, where a few metres of residual error, roof–footprint parallax, or relief displacement can move a projected footprint onto an adjacent building.

UAV–map pairing. From Geofabrik OSM extracts (Geofabrik GmbH, 2026), we clip for each image a square, north-up map window centred on the registered footprint, with side length $1.2\times$ the footprint's longest extent; the buffer adds context beyond the UAV field of view (e.g., roads and buildings continuing past the image border) and prevents the footprint from being trivially identified by the map boundary. We extract roads, buildings, land use, amenities, waterways, and railways with their geometry and normalized semantic tags: the Visual window is rendered in the OSM

Carto style (OpenStreetMap Contributors, 2026), where road classes differ in colour and width and buildings, land use, water, and railways have distinct fills or patterns, so the raster preserves semantic categories in addition to geometry; the Tabular version serializes the same entities as text with tags, geometry, and relations such as intersection and containment.

Reliable cross-view correspondences. To obtain instance-level correspondences despite residual registration error and roof–footprint misalignment, we avoid treating projected OSM polygons as ground truth. Instead, we combine independent map-based evidence (OSM building polygons projected into the UAV image, the *map prior*) with image-based evidence (SAM3 building masks detected without OSM input, the *visual evidence*) and keep only high-confidence agreements: we estimate the shared per-image offset by cross-correlating the two mask unions, refine it with a RANSAC similarity fit, and match corrected OSM polygons to SAM3 masks one-to-one via Hungarian assignment on IoU. A pair is accepted only if IoU ≥ 0.5, the SAM3 score is ≥ 0.5, the match is unique (margin ≥ 0.2), a second SAM3 mask prompted by the corrected OSM box agrees with the first (IoU ≥ 0.75), and the mask does not touch the image border; strongly parallaxed, truncated, or ambiguous buildings fail these tests and are dropped rather than mismatched. Images used for building-grounded QA require at least four such pairs, subsequently verified by human annotators (procedure and examples from all four sources in Appendix B.5, Figure 6).

Question–answer generation. T1–T3 draw the query and all four options from the verified building pool, with T3 deleting the OSM footprints of $k \in \{1, 2, 3\}$ labelled options plus 0–2 unmarked decoy buildings before re-rendering or re-serializing the map. T4 uses the UAV heading inferred from the registered footprint (the bearing of the image's top edge), rejecting frames whose heading is unstable across the footprint. For T5–T8, SAM3-detected cars kept only if car-sized are lifted onto the map via the registration and per-image offset correction (error estimated as the median residual of verified building pairs), with answers computed by deterministic geometric predicates: map position, nearest building footprint, carriageway membership, and compass bearing. A question is kept only if every decision margin exceeds three times the per-image error, with closed-way building targets and candidates at least 100 m^2. Distractors are rule-generated to be plausible but verifiably wrong: locations at least $\max(20\,\mathrm{m}, 5 \times \text{error})$ away (T5), other verified buildings (T6), other drivable roads including one of the same class (T7), and the remaining compass directions (T8). Each image yields three QA variants per task on average by shuffling the selected buildings, roads, vehicles, and distractors, for 42K questions per task and 336K multiple-choice QA pairs in total (construction rules and distractor strategies in Appendix C).

3.4 Quality Control

Visual prompting supports grounding while avoiding shortcuts. Because current VLMs are unreliable at exact coordinate grounding, we mark the queried instance and each option directly on the images—via two randomized styles, colour masks and numeric tags (Yang et al., 2023; Cai et al., 2024)—with colour, fill, line width, shape, and option order randomized and frequency-balanced per question. Every mark passes a CIELAB contrast check, distractors are stratified by exploitable attributes (size, centrality, road class, compass direction), and UAV images are never rotated to north-up, so heuristics such as "pick the largest building" remain near chance (Appendix C).

Manual verification validates QA quality. The coauthors audited 1,200 sampled QA pairs in a web interface[1] against four criteria: clarity, unique grounding in the joint image–map context, prompt correctness, and answer correctness, rating 92.73% *accept*, 3.32% *minor fix*, 1.58% *major fix*, and 2.37% *reject*.

Dataset statistics. UAV-MapVQA totals 14,000 aligned UAV image–map pairs and 336K QA pairs across three countries, with 3.34M OSM features and realistic variation in footprint size and map density that we preserve rather than equalize; evaluation splits are stratified over sources and tasks. Full statistics are in Appendix B.2.

4 Experiments

We evaluate whether multimodal FMs can *ground a visual instance in a UAV image to its entity in a vector map* and *reason over the grounded entity using map structure and attributes*. Each question

[1] https://storage.googleapis.com/uav_map_vqa/verify.html

has two representations of the same OSM scene: a rendered raster map (**UAV-MapVQA-Visual**, M_{vis}) and a structured entity table (**UAV-MapVQA-Tabular**, M_{tab}). We address three questions:

- **Q1: Benchmark validity.** Does UAV-MapVQA provide a meaningful evaluation setting that is challenging for current models while remaining reliably solvable by humans?
- **Q2: Vector Map representation.** How do raster and structured maps affect grounding and reasoning, zero-shot and after fine-tuning?
- **Q3: Capability diagnosis.** Does training on UAV-MapVQA improve FMs, and does the orientation discrepancy between the views limit reasoning?

4.1 Experimental Protocol

Inputs. In Visual, the OSM Carto map and the UAV image (640 px) carry randomized visual prompts (Yang et al., 2023) to avoid visual bias. In Tabular, the map is replaced by tables of the same window (tags, normalized boxes, centroids, orientations, junctions), and map-side prompts become row IDs or coordinates (Appendix C) to avoid shortcuts.

Stratified test sets. Open-weight and API models are evaluated on **Test-3.2K** (400 questions per task); the more expensive proprietary models on **Test-0.8K** (100 per task), a nested subset on which we compare the two groups. Visual and Tabular share question IDs, so their comparison is paired. The 95% confidence half-width is at most $\pm$4.9 points per task and $\pm$1.7 overall on Test-3.2K, and $\pm$9.8 and $\pm$3.5 on Test-0.8K.

Prompting and metric. All models share one prompt template with zero-shot chain-of-thought (CoT) prompting (Kojima et al., 2022): "*Think step by step, then end with Answer: <letter>.*" (comma-separated letters for the multi-answer T3). The prediction is parsed from the last `Answer:` line; unparseable answers and over-length prompts count as wrong, so no question is dropped. Models without built-in reasoning decode greedily; GPT-5, GPT-5.5, and the Claude models reason at medium effort within 16K tokens (full prompting and evaluation protocol in Appendix D). We report exact-match accuracy (for T3, the exact set of unmapped buildings); overall is the mean over the eight tasks. Chance is 25% per single-answer task, 7.1% on T3 (1 of 14 answer sets), and 22.8% overall.

Models. We evaluate seven proprietary models (GPT-5.5 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), Claude Opus 4.7 (Anthropic, 2025b), Claude Sonnet 4.6 (Anthropic, 2025c), GPT-4o (OpenAI, 2024a), Claude Haiku 4.5 (Anthropic, 2025a), GPT-4o Mini (OpenAI, 2024b)), five open-weight or API models (Qwen VL Max (Alibaba Cloud, 2025a), Qwen VL Plus (Alibaba Cloud, 2025b), Minimax-01 (MiniMax, 2025), Qwen2.5-VL-7B/32B (Bai et al., 2025b)), and the remote-sensing FM GeoChat (Kuckreja et al., 2024) (single-image only, so map and UAV image are placed side by side). To test whether our data teaches the task, we fine-tune Qwen2.5-VL-7B/32B with 4-bit QLoRA (Dettmers et al., 2023) on a single A100 (40 GB) using a 12.8K-question, scene-disjoint training split (1,600 per task, one adapter per representation). As a human reference, six college students each answer 240 Visual questions (30 per task, same inputs, one minute per question); we skip a Tabular human reference since the serialized tables are not an intuitive interface for people, and most Tabular questions (Figures 7–8) cannot be solved by manually cross-referencing rows within a reasonable time.

4.2 Results and Analysis

UAV-MapVQA is challenging but solvable. No zero-shot model reaches 50% overall accuracy: the best are GPT-5.5 on Visual (44.0%, Table 2) and GPT-5 on Tabular (49.3%, Table 3), and four of the 13 models, including both open-weight Qwen2.5-VL models, stay within 3 points of chance on Visual. Humans reach 94.8% with one minute per question, so the questions are answerable from the given evidence. Relative to chance, T4 (heading) and T8 (compass direction), which depend most on the relative orientation of the views, are hardest: on Visual, zero-shot models exceed chance there by only 0.9 and 4.8 points on average, versus 9.2–14.3 elsewhere.

Frontier proprietary models lead, but no model family solves the task. On Test-0.8K (Table 8), the best open-weight/API model, Minimax-01, trails GPT-5.5 by 10.5 points on Visual and GPT-5 by 13.0 on Tabular, yet outperforms the smaller proprietary models, Claude Haiku 4.5 and GPT-4o Mini.

Table 2: Accuracy (%) on **UAV-MapVQA-Visual**. Blocks: proprietary (zero-shot CoT, Test-0.8K); open-weight/API and remote-sensing (GeoChat) (zero-shot CoT, Test-3.2K); fine-tuned (FT, 4-bit QLoRA on our 12.8K split, Test-3.2K). Levels: I grounding (T1–T2), II change detection (T3), III ego-centric (T4), IV object-centric (T5–T8). **Bold**: best; amber: best zero-shot; blue: fine-tuned. **Chance**: expected accuracy of uniform random guessing among the options (25% per single-answer task, 7.1% on the 14-way multi-answer T3, 22.8% overall). Reported accuracies are rounded to one decimal place.

Model	T1	T2	T3	T4	T5	T6	T7	T8	Overall
GPT-5.5	57.0	43.0	31.0	28.0	45.0	52.0	56.0	40.0	44.0
GPT-5	49.0	49.0	33.0	24.0	43.0	38.0	46.0	40.0	40.3
Claude Opus 4.7	53.0	34.0	33.0	30.0	49.0	46.0	50.0	28.0	40.4
Claude Sonnet 4.6	45.0	47.0	28.0	**34.0**	49.0	38.0	42.0	33.0	39.5
GPT-4o	47.0	39.0	23.0	26.0	39.0	38.0	51.0	38.0	37.6
Claude Haiku 4.5	35.0	32.0	19.0	33.0	22.0	30.0	37.0	27.0	29.4
GPT-4o Mini	32.0	32.0	10.0	26.0	25.0	26.0	31.0	29.0	26.4
Qwen VL Max	36.8	38.5	20.3	27.3	30.8	45.0	43.5	31.8	34.2
Minimax-01	42.8	37.0	28.3	23.5	36.0	32.3	45.0	30.0	34.3
Qwen VL Plus	27.0	30.0	14.0	22.5	28.5	31.0	25.8	24.3	25.4
Qwen2.5-VL-32B	27.8	23.0	8.0	21.3	22.8	26.3	28.3	27.0	23.0
Qwen2.5-VL-7B	30.5	26.3	12.0	20.3	26.0	26.5	28.5	20.8	23.8
GeoChat	21.0	21.0	7.5	20.8	28.3	22.8	26.8	19.0	20.9
Qwen2.5-VL-32B FT	**64.3**	**64.0**	**58.0**	33.3	**60.8**	**62.5**	**71.8**	**43.0**	**57.2**
Qwen2.5-VL-7B FT	56.0	51.5	30.5	30.8	51.8	62.0	57.5	40.0	47.5
Chance	25.0	25.0	7.1	25.0	25.0	25.0	25.0	25.0	22.8
Human	98.9	95.6	92.2	86.7	95.0	97.8	98.9	93.3	94.8

Table 3: Accuracy (%) on **UAV-MapVQA-Tabular**, same questions, settings, and blocks as Table 2. Δ: overall change vs. Visual (gain/loss). GeoChat is omitted (input exceeds its context); Human is omitted (Tabular is not an intuitive interface for people, and most questions cannot be solved manually within a reasonable time). Reported accuracies are rounded to one decimal place.

Model	T1	T2	T3	T4	T5	T6	T7	T8	Overall	Δ
GPT-5.5	59.0	54.0	39.0	**42.0**	47.0	54.0	57.0	37.0	48.6	+4.6
GPT-5	53.0	**64.0**	47.0	29.0	44.0	52.0	53.0	**52.0**	49.3	+9.0
Claude Opus 4.7	48.0	44.0	34.0	35.0	49.0	48.0	47.0	36.0	42.6	+2.2
Claude Sonnet 4.6	50.0	43.0	25.0	28.0	43.0	51.0	37.0	36.0	39.1	−0.4
GPT-4o	40.0	43.0	20.0	24.0	38.0	37.0	41.0	32.0	34.4	−3.2
Claude Haiku 4.5	31.0	32.0	22.0	23.0	27.0	39.0	35.0	23.0	29.0	−0.4
GPT-4o Mini	27.0	24.0	13.0	16.0	33.0	29.0	21.0	21.0	23.0	−3.4
Qwen VL Max	40.3	38.5	17.0	26.0	33.0	40.5	41.5	28.0	33.1	−1.1
Minimax-01	46.3	42.3	32.5	22.3	33.8	34.0	40.3	33.8	35.6	+1.3
Qwen VL Plus	26.8	29.0	10.3	24.8	28.5	28.0	28.0	23.8	24.9	−0.5
Qwen2.5-VL-32B	38.0	36.5	21.3	23.5	37.3	32.8	36.8	28.3	31.8	+8.8
Qwen2.5-VL-7B	30.8	26.5	12.3	21.5	24.0	26.8	26.5	25.0	24.2	+0.4
Qwen2.5-VL-32B FT	**66.0**	60.3	**50.0**	38.0	**56.8**	**59.3**	**60.8**	41.8	**54.1**	−3.1
Qwen2.5-VL-7B FT	59.0	50.0	36.3	28.3	47.0	53.0	56.5	35.8	45.7	−1.8

Table 4: FT models, sparse (n=1,708) vs. dense (n=1,492) scenes; last row: median map tokens (Qwen2.5-VL tokenizer).

	Visual		Tabular	
	Sp.	De.	Sp.	De.
7B FT	48.2	46.6	**51.2**	39.4
32B FT	58.1	56.1	**60.5**	46.8
Tokens	529	529	6.2K	16.6K

Table 5: Map frame, accuracy (%) over all tasks but T4: north-up (N-up), randomly rotated (Rand., control), and heading-aligned (Align.) map. Δ: Align. − N-up per level (Visual). GPT-5.5, Opus: 700 questions; Qwen: 2,800. Blue: aligned.

	Visual						Tabular		
Model	N-up	Rand.	Align.	Δ_{I}	Δ_{II}	Δ_{IV}	N-up	Rand.	Align.
GPT-5.5	46.3	45.6	**55.0**	+10.0	+10.0	+7.8	49.6	48.4	**52.6**
Claude Opus 4.7	41.9	39.1	49.9	+11.5	+16.0	+4.3	43.7	45.0	47.9
Qwen2.5-VL-32B	23.3	22.9	24.1	+2.4	+0.5	+0.2	33.0	31.9	35.2
Qwen2.5-VL-7B	24.4	24.2	26.6	+2.8	+6.3	+0.9	24.5	25.1	26.0

Remote-sensing pretraining does not help. GeoChat scores 20.9%, the lowest of all models and 1.9 points below chance, plausibly because its single-view satellite training covers neither vector maps nor cross-view correspondence; its single 504 px canvas also shrinks each view to about 250 px. It cannot take Tabular: about 2.7K tokens of its 4K context remain after the image, but the Tabular prompts of the questions in Figure 3 take 8.5K–71K tokens. These results answer **Q1**.

The two map representations have complementary strengths. Zero-shot, Tabular helps only some models (Table 3): it raises GPT-5 and GPT-5.5 by 9.0 and 4.6 points, as tags, coordinates, and junctions are explicit rather than read off pixels, and lifts Qwen2.5-VL-32B from chance (23.0%) to 31.8%, while the other nine change by at most 3.4. After fine-tuning, Visual leads by 1.8 (7B) and 3.1 (32B) points overall, but only because of dense scenes (Table 4). We call a scene dense or sparse by the number of OSM features in its map window, split at a single dataset-wide threshold

chosen with Otsu's method (Otsu, 1979): the cut on the feature-count histogram that maximizes the between-class variance of the two groups, so the split follows the data rather than a hand-picked value. The serialized map grows with every entity and needs 2.7× more tokens in dense scenes, where fine-tuned Tabular models lose 11.8 (7B) and 13.7 (32B) points relative to sparse scenes; the fixed-size raster loses at most 2.0, while Tabular stays ahead in sparse scenes. These results answer **Q2**: structured maps help strong reasoners and sparse scenes, whereas the raster is robust to density and gains more from fine-tuning (32B: +34.2 vs. +22.3 points). Neither solves the task, so the difficulty is not an artifact of one map encoding.

4.3 Improving Cross-view Instance–Entity Grounding and Reasoning

Supervised fine-tuning verifies the training signal. Fine-tuning brings large gains for both sizes and representations: Qwen2.5-VL-32B improves from 23.0% to 57.2% on Visual and from 31.8% to 54.1% on Tabular (Tables 2–3), becoming the strongest model. On Test-0.8K (Table 8) it exceeds the best zero-shot result, GPT-5 on Tabular (49.3%), by 7.8 points on Visual and 4.7 on Tabular. Training and test scenes are disjoint, and the fine-tuned model stays at chance when either view is withheld (Table 9), so the gains reflect cross-view grounding learned from UAV-MapVQA rather than scene memorization or single-view shortcuts. Gaps remain: it is 37.6 points below humans and, relative to chance, again weakest on T4 and T8 (+8.3 and +18.0 points vs. $\geq$+35.8 elsewhere).

Orientation discrepancy is a measurable but partial bottleneck. The weakness on T4 and T8 in every model group suggests a perceptual cause: UAV images have arbitrary headings while maps are north-up, and misperceiving the rotation between the views corrupts grounding and every later step. To test this, we rotate the map to the registered UAV heading: the Visual map is re-rendered from rotated vector data (identical prompts, north arrow), and Tabular coordinates are expressed in the rotated frame with the north bearing stated. Alignment leaves the question, answer, and reasoning steps unchanged and only removes the rotation to be perceived. T4 is excluded (its answer becomes readable), and a random rotation controls for re-rendering.

Alignment raises GPT-5.5 and Claude Opus 4.7 by 8.7 and 8.0 points on Visual and by 3.0 and 4.2 on Tabular, whereas random rotation stays within 2.8 points of north-up (Table 5); the Qwen models, near chance, gain at most 2.2. For GPT-5.5 and Opus, Visual gains are largest on Levels I–II (+10.0 to +16.0) but also appear on object-centric Level IV (+7.8, +4.3), whose reasoning steps alignment leaves untouched, so orientation errors propagate into downstream reasoning. On north-up maps, their accuracy falls from 54.9% for images facing within 45° of north to 35.8% for south-facing ones; alignment shrinks this 19-point gap to 7 points. As the heading comes from the registration, these gains upper-bound what heading estimation can recover, yet aligned GPT-5.5 still reaches only 55.0% on these seven tasks, versus 96.0% for humans.

These results answer **Q3**: UAV-MapVQA provides effective supervision, and orientation errors reach downstream reasoning, but most of the gap to humans persists under aligned views.

5 Conclusion, Limitations, and Future Work

We introduced **instance-to-entity grounding** between UAV images and vector maps and presented **UAV-MapVQA**: 336K multiple-choice questions from 14,000 real UAV images in 3 countries, spanning eight tasks at four levels with raster and tabular maps. No zero-shot FM, general-purpose or remote-sensing, reaches 50% overall accuracy, versus 94.8% for humans; fine-tuning narrows but does not close this gap, and heading alignment shows orientation errors reach reasoning but explain only part of what remains.

Limitations and future work. UAV-MapVQA inherits OSM incompleteness and residual registration error (median 1.9 m, P95 5.4 m); we bound ambiguous supervision by excluding buildings under 100 m^2 and requiring each decision margin to exceed 3× the per-image error, leaving small objects and fine-grained spatial distinctions out of scope. T3 captures only new construction absent from the map, not demolition or structural change, since synthesizing such cases by editing UAV imagery risks artifact-driven shortcuts. Our multiple-choice format enables exact, reproducible evaluation without lexical-matching or LLM-judge confounds; the released entity IDs, geometries, OSM attributes, and reasoning traces nonetheless support future free-form tasks—agentic OSM querying, cross-view/onboard heading estimation, and map-aware pretraining.

AI USE STATEMENT

We used generative AI tools to help polish the writing after the authors drafted the manuscript. For figures, AI tools assisted in refining the visual design based on the authors' initial layouts; all visual content, including UAV images, maps, and task examples, is taken from the UAV-MapVQA dataset. QA pairs are generated from templates over OSM geometry and verified correspondences (Section 3.3), without the use of generative AI. As this work studies foundation models, these models also appear in our experiments as the subjects of evaluation (Section 4.1). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

UAV-MapVQA is built to be released openly. An interactive preview of the benchmark is already available online at `https://storage.googleapis.com/uav_map_vqa/index.html`. Section 3.3 and Appendix B describe the full construction pipeline, including UAV sources, sampling, registration, map pairing, correspondence verification, and QA generation, and Appendix B.4 reports registration accuracy. Section 4.1 and Appendix D specify models, providers, prompts, decoding, scoring, retry rules, and the fixed-seed stratified test splits. The fine-tuning setup (4-bit QLoRA, 12.8K scene-disjoint split) is given in Section 4.1. We will release the dataset, evaluation code, prompt configurations, and fine-tuned model checkpoints.

REFERENCES

Alibaba Cloud. Qwen VL Max. `https://www.alibabacloud.com/`, 2025a.

Alibaba Cloud. Qwen VL Plus. `https://www.alibabacloud.com/`, 2025b.

Anthropic. Claude Haiku 4.5. `https://www.anthropic.com/claude`, 2025a.

Anthropic. Claude Opus 4.7. `https://www.anthropic.com/claude`, 2025b.

Anthropic. Claude Sonnet 4.6. `https://www.anthropic.com/claude`, 2025c.

Christian Michael Arnold, Andrew Alini, Jonathan Wang, Pieter M Feenstra, Conner Arnold, Jan DeWitt, Natalie C Ritsema, Jung Hyun Yae, Boris Katz, Andrei Barbu, et al. MapQA: A map-question-answering benchmark for visual language model reasoning. 2026.

Lubian Bai, Xiuyuan Zhang, Siqi Zhang, Zepeng Zhang, Haoyu Wang, Wei Qin, and Shihong Du. Geolink: Empowering remote sensing foundation model with openstreetmap data. *arXiv preprint arXiv:2509.26016*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Riyang Bao, Cheng Yang, Dazhou Yu, Zhexiang Tang, Gengchen Mai, and Liang Zhao. Spatial-agent: Agentic geo-spatial reasoning with scientific core concepts. *arXiv preprint arXiv:2601.16965*, 2026.

Sharat Bhat, Harshita Khandelwal, Tushar Kataria, and Vivek Gupta. Mapverse: A benchmark for geospatial question answering on diverse real-world maps. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8168–8178, 2026.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12914–12923, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou,

Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-rsvqa: Prompting visual context to a language model for remote sensing visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1372–1381, 2022.

Christel Chappuis, Gencer Sümbül, Syrielle Montariol, Sylvain Lobry, and Devis Tuia. Pan-rsvqa: Vision foundation models as pseudo-annotators for remote sensing visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3005–3016, 2025.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pp. 213–231. Springer, 2024.

Muhammad 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.

Zhiyun Deng, Austin Case, and Luis Sentis. AirLock+: Scaling UAV-to-satellite image registration for target geolocalization and geospatial augmented reality. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3340–3349, 2026.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Oussema Dhaouadi, Riccardo Marin, Johannes Michael Meier, Jacques Kaiser, and Daniel Cremers. OrthoLoC: UAV 6-DoF localization and calibration using orthographic geodata. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. MapEval: A map-based evaluation of geo-spatial reasoning in foundation models. *arXiv preprint arXiv:2501.00316*, 2025.

Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44 (11):7778–7796, 2021.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19790–19800, 2024.

Geofabrik GmbH. Openstreetmap data extracts. `https://download.geofabrik.de/`, 2026. Extracts of OpenStreetMap data. Accessed: 2026-05-04.

Jiajin Guan, Haibo Mei, Bonan Zhang, Dan Liu, Yuanshuang Fu, and Yue Zhang. UAV-VL-R1: Generalizing vision-language models via supervised fine-tuning and multi-stage GRPO for UAV visual reasoning. *arXiv preprint arXiv:2508.11196*, 2025.

Jens Hellekes, Manuel Mühlhaus, Reza Bahmanyar, Seyed Majid Azimi, and Franz Kurz. Vetra: A dataset for vehicle tracking in aerial imagery–new challenges for multi-object tracking. In *European Conference on Computer Vision*, pp. 52–70. Springer, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27831–27840, 2024.

Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5177–5187, 2025.

Zeping Liu, Lao Ni, Zhangyu Wang, Junfeng Jiao, and Gengchen Mai. GAIR: Location-aware self-supervised contrastive pre-training with geo-aligned implicit representations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:166–182, 2026.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Jun Lu, Zehao Sang, Haoqi Wei, Xiangyun Liu, Kun Zhu, Haitao Guo, Zhihui Gong, and Lei Ding. Vfm-loc: Zero-shot cross-view geo-localization via aligning discriminative visual hierarchies. *arXiv preprint arXiv:2603.13855*, 2026.

Mapbox. Mapbox satellite. https://docs.mapbox.com/data/tilesets/reference/mapbox-satellite/, 2026.

MiniMax. MiniMax-01. https://www.minimaxi.com/, 2025.

OpenAI. GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024a.

OpenAI. GPT-4o Mini. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024b.

OpenAI. GPT-5.5. https://openai.com/, 2025a.

OpenAI. GPT-5. https://openai.com/, 2025b.

OpenStreetMap Contributors. Openstreetmap carto. https://github.com/openstreetmap-carto/openstreetmap-carto, 2026.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

QGIS Development Team. QGIS geographic information system, 2026. URL https://qgis.org. Version 4.0.2.

Andrew C. Ritchie, Peter J. Triezenberg, Jonathan A. Warrick, Gerald A. Hatcher, and Daniel D. Buscombe. Remote sensing coastal change simple data distribution service. https://doi.org/10.5066/P9M3NYWI, 2023.

Argho Sarkar and Maryam Rahnemoonfar. Vqa-aid: Visual question answering for post-disaster damage assessment and analysis. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 8660–8663. IEEE, 2021.

Liang Shi, Wei Li, Kevin M Beussman, Lin Chen, and Yun Fu. Iir-vlm: In-context instance-level recognition for large vision-language models. *arXiv preprint arXiv:2601.14188*, 2026.

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. MapIQ: Evaluating multimodal large language models for map question answering, 2025. URL https://arxiv.org/abs/2507.11625.

Jintao Sun, Hu Zhang, Donglin Di, Gangyi Ding, and Zhedong Zheng. UAVReason: A unified, large-scale benchmark for multimodal aerial scene reasoning and generation. *arXiv preprint arXiv:2604.05377*, 2026.

Yong Tang, Jingyi Zhang, Jianhua Gong, Yi Li, and Banghui Yang. City-level aerial geo-localization based on map matching network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229: 65–77, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Mai Tsujimoto, Junjue Wang, Weihao Xuan, and Naoto Yokoya. Geo3DVQA: Evaluating vision-language models for 3D geospatial reasoning from aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4994–5003, 2026.

Zhen Wang, Dianxi Shi, Chunping Qiu, Songchang Jin, Tongyue Li, Ziteng Qiao, and Yang Chen. VecMapLocNet: Vision-based UAV localization using vector maps in GNSS-denied environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225:362–381, 2025.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. iSAID: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 28–37, 2019.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Yuxiang Liu, Maojun Zhang, and Shen Yan. UAVD4L: A large-scale dataset for UAV 6-DoF localization. In *2024 International Conference on 3D Vision (3DV)*, pp. 1574–1583. IEEE, 2024.

Xindi Wu, KwunFung Lau, Francesco Ferroni, Aljoša Ošep, and Deva Ramanan. Pix2Map: Cross-modal retrieval for inferring street maps from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17514–17523, 2023.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. Grounding language models for visual entity recognition. In *European Conference on Computer Vision*, pp. 393–411. Springer, 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.

Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. UAV-VisLoc: A large-scale dataset for UAV visual localization. *arXiv preprint arXiv:2405.11936*, 2024.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

Qin Ye, Junqi Luo, and Yi Lin. A coarse-to-fine visual geo-localization method for GNSS-denied UAV with oblique-view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212: 306–322, 2024.

Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai, and Liang Zhao. Spatial-rag: Spatial retrieval augmented generation for real-world geospatial reasoning questions. *arXiv preprint arXiv:2502.18470*, 2025.

Yang Zhan and Yuan Yuan. UAVBench and UAVIT-1M: Benchmarking and enhancing MLLMs for low-altitude UAV vision-language understanding. *arXiv preprint arXiv:2603.14336*, 2026.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pp. 1395–1403, 2020.

Yue Zhou, Jue Chen, Zilun Zhang, Penghui Huang, Ran Ding, Zhentao Zou, PengFei Gao, Yuchen Wei, Ke Li, Xue Yang, et al. DVGBench: Implicit-to-explicit visual grounding benchmark in UAV imagery with large vision–language models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232:831–847, 2026.

A Detailed Related Work

A.1 Aerial Image VQA: From Scene-Level to Image-Space Grounding

Many geospatial VQA research focused primarily on satellite and aerial remote sensing (RS) imagery. Prompt-RSVQA (Chappuis et al., 2022) pioneered prompting language models with visual context extracted from RS images, establishing the basic vision–language interface for the domain. Subsequent work revealed that general-domain VLMs fail on RS-specific queries due to the unique challenges of high-resolution imagery, dense small objects, and diverse scale variation. GeoChat (Kuckreja et al., 2024) addressed this by fine-tuning a domain-specific RS VLM capable of image-level and region-level dialogue with spatial coordinate grounding. GEOBench-VLM (Danish et al., 2025) and Geo3DVQA (Tsujimoto et al., 2026) further established comprehensive benchmarks evaluating scene understanding, object counting, localization, and height-aware 3D reasoning, finding that even state-of-the-art proprietary models achieve well below human performance.

A key enabler for RS VQA is that nadir-view satellite imagery is naturally georeferenced: pixel coordinates correspond directly to geographic coordinates, making map fusion straightforward. RSVQA (Lobry et al., 2020) exploited this by deriving question–answer pairs from OpenStreetMap (OSM) annotations aligned to RS imagery. PAN-RSVQA (Chappuis et al., 2025) and GeoLink (Bai et al., 2025a) further demonstrated that incorporating structured OSM data during training and inference consistently improves RS understanding. Critically, however, grounding in these works remains at the *semantic or class level*: map entities serve as category-level supervision rather than as individual geographic instances to be identified and localized.

UAV imagery introduces an important shift: oblique viewing angles, free camera rotation, and variable altitude break the direct pixel-to-coordinate correspondence that makes RS georeference effortless. Existing UAV VQA benchmarks (Sarkar & Rahnemoonfar, 2021; Guan et al., 2025; Zhan & Yuan, 2026) cover image-level tasks such as counting, scene classification, and spatial scene inference, but do not reach instance-level geographic reasoning. The most advanced UAV visual grounding benchmarks—AerialVG (Liu et al., 2025) and DVGBench (Zhou et al., 2026)—do achieve instance-level localization and require spatial relation reasoning (e.g., distinguishing the target vehicle by its position relative to neighboring objects). Yet grounding in these works is defined entirely within *image space*: descriptions use observer-perspective directions (left, right, above, below) and bounding box outputs are pixel coordinates. No geographic knowledge is required or tested; the same task could be solved from an image with all geographic context removed.

In summary, while UAV visual grounding has progressed toward instance-level image-space localization, *no existing benchmark requires associating UAV visual instances with persistent geographic entities in a structured map*—the central challenge our work addresses.

A.2 Geographic Grounding across Views and Modalities

Geographic grounding (i.e., establish correspondences between visual observations and real-world geographic entities) has been explored most thoroughly in the autonomous driving domain. Pix2Map (Wu et al., 2023) formulates the problem of inferring urban road map topology from ego-view street images as cross-modal retrieval, learning a joint embedding space for images and map graphs. This work demonstrates that cross-view geographic grounding between ground-level imagery and structured maps is feasible, serving as a proof of concept for the broader problem. However, Pix2Map targets street-level perspective images under near-frontal viewing geometry, operates at the *road-segment level* rather than at the level of individual geographic instances, and involves no question-answering or reasoning component.

On the UAV side, geo-localization methods such as GeoText-1652 (Chu et al., 2024) address the problem of estimating a UAV image's geo-location by aligning image with spatial-aware text descriptions. These works confirm that cross-view alignment between UAV imagery and geographic references is a meaningful and challenging problem, but they target the localization of the *camera* rather than of individual objects within the scene, and they produce pose estimates rather than natural-language answers.

The gap is therefore two-fold. First, no prior work achieves *instance-level* geographic grounding from UAV imagery: given a specific object detected in the air, identifying the corresponding entity

in a structured map. Second, no prior work frames geographic grounding as a *question-answering* task, precluding the evaluation of multi-step reasoning that combines visual observation with map semantics. Our benchmark directly targets both dimensions.

A.3 Spatial Reasoning and Map Understanding in MLLMs

Complementary to the grounding literature, a growing body of work evaluates the spatial reasoning and map understanding capabilities of multimodal large language models (MLLMs). (Tong et al., 2024) systematically identified nine categories of map understanding failure in current MLLMs, including orientation and direction, positional and relational context, and viewpoint and perspective, which are directly relevant to UAV-map reasoning. MapQA (Arnold et al., 2026) and MapIQ (Srivastava et al., 2025) introduced large-scale benchmarks for map-based VQA, finding that SOTA proprietary models answer fewer than half of questions correctly, with open-source models typically below 30%. Spatial-RAG (Yu et al., 2025) further showed that LLMs lack intrinsic spatial computing capabilities and benefit substantially from retrieval-augmented access to structured geospatial databases. Spatial-Agent (Bao et al., 2026) extends this line by incorporating GIScience domain knowledge into orchestration for multi-step geographic reasoning.

On the instance recognition side, fine-grained benchmarks such as GLDv2 (Weyand et al., 2020) and methods such as AutoVER (Xiao et al., 2024) have demonstrated that VLMs can be equipped with instance-level recognition capabilities, distinguishing specific real-world entities within a large label space. IIR-VLM (Shi et al., 2026) further improve performance on instances in-context in a zero-shot manner.

These capabilities, such as map understanding, spatial reasoning, and instance-level recognition, have been evaluated largely in isolation and always within a *single modality*. However, no existing benchmark simultaneously requires an MLLM to *ground a visual instance from a UAV image to a geographic entity in a map*, reason about spatial relationships in the map coordinate system, and answer questions conditioned on map-attributed semantics (e.g., speed limits, building types, road names).

B Dataset Construction Details

B.1 UAV Image Statistics

Image pool and sampling. After UAV-to-map registration and the building-count threshold, the candidate pool contains 25,583 frames: 3,508 from VETRA (Hellekes et al., 2024), 7,011 from AirLock (Deng et al., 2026), 5,202 from UAV-VisLoc (Xu et al., 2024), and 9,862 from USGS-RSCC (Ritchie et al., 2023). The USGS-RSCC frames come from two camera platforms: 9,173 Precision Airborne Camera (PAC) frames along the coasts of North Carolina, Florida, and Massachusetts, selected so that image centres are at least 200 m apart, and 689 Condor frames from the northern California coast, selected at a 500 m spacing. Six UAV-VisLoc frames whose altitude above ground could not be recovered reliably (below 50 m; see below) are excluded. To balance sources, we sample 3,500 frames from each, 14,000 aligned image–map pairs in total. Sampling is stratified over $\sim$1 km grid cells so that each cell receives the same quota, which spreads the sample across flight regions instead of concentrating it on densely captured trajectories. Within a cell, we prefer frames with more verified building anchors, which avoids scenes too sparse to support grounding questions.

Flight altitude. We report altitude above ground level (AGL), because the sources record altitude against different references. For VETRA, AGL is the recorded flight height minus the terrain height in the image metadata; for AirLock, it is the take-off-relative altitude from the DJI flight log; for UAV-VisLoc and Condor, it is the recorded absolute height minus terrain elevation from a global elevation model (AWS Terrain Tiles); for PAC, it is the GPS altitude over coastal terrain within a few metres of sea level. Across the 14,000 frames, AGL ranges from 104 to 1,508 m, with a median of 452 m (Table 6; top margin of Figure 4). AirLock covers the low-altitude range (median 397 m, down to 104 m), UAV-VisLoc and VETRA the middle range (medians 538 and 662 m), and USGS-RSCC is bimodal: PAC flies at about 280 m and Condor at about 1,370 m.

Table 6: UAV image statistics of the 14,000 sampled frames (3,500 per source). AGL: flight altitude above ground level. Footprint: geodesic area of the registered image footprint. †Footprints of 2,671 of the 3,500 frames are estimated from AGL (see text).

			AGL (m)				Footprint (10^3 m^2)			
Source	Country	Images	Min	Median	Mean	Max	Min	Median	Mean	Max
VETRA	Germany	3,500	206	662	668	1,480	3.5	141.2	147.3	702.6
AirLock	USA	3,500	104	397	333	462	3.3	118.2	99.2	270.2
UAV-VisLoc	China	3,500	220	538	520	860	6.2	191.1	191.2	629.0
USGS-RSCC†	USA	3,500	218	281	463	1,508	62.3	103.7	356.0	1,932.1
PAC		2,978	218	278	304	806	62.3	101.8	136.4	850.3
Condor		522	1,048	1,372	1,373	1,508	932.6	1,598.2	1,608.8	1,932.1
Total	3 countries	14,000	104	452	496	1,508	3.3	133.6	198.4	1,932.1

Ground footprint. The footprint is the geodesic area of the quadrilateral spanned by the four registered image corners. Footprints range from 3.3×10^3 to 1.9×10^6 m^2, with a median of 1.34×10^5 m^2 (right margin of Figure 4). Within each camera, footprint area grows with the square of AGL (Figure 4): each camera forms a line of slope two on log–log axes, offset by its field of view.

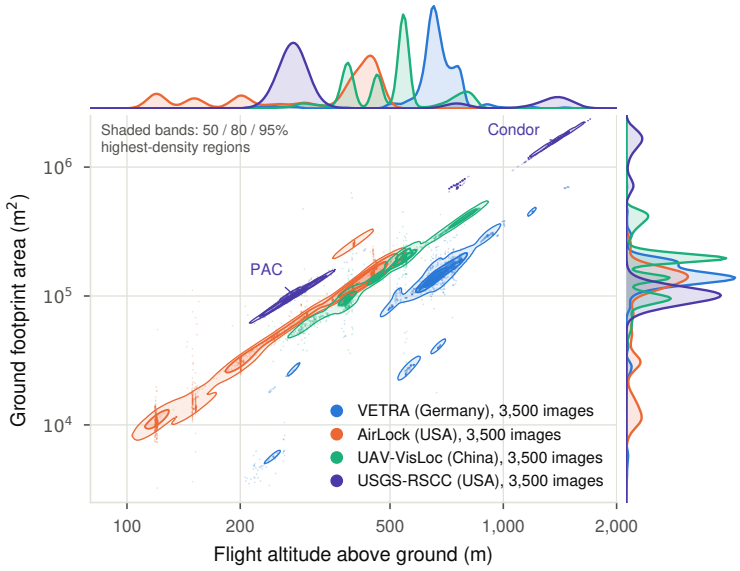


Figure 4: Flight altitude above ground level against ground footprint area for the 14,000 sampled frames (3,500 per source), on log axes. Shaded bands are the 50%, 80%, and 95% highest-density regions of each source; the margins show the altitude (top) and footprint (right) distributions. Each camera forms a line of slope two, since footprint area scales with the square of altitude. For USGS-RSCC, 2,671 footprints are estimated from altitude and therefore lie on the fitted PAC and Condor curves; the other 829 are measured.

B.2 Dataset Statistics

UAV-MapVQA contains 14,000 aligned UAV image–map pairs and 336K multiple-choice QA pairs derived from four public UAV datasets across three countries.

Image footprint and map density. We estimate each UAV image's ground coverage from the geo-coordinates of its registered corners and compute the footprint area from geodesic edge lengths. Footprint size varies substantially across source datasets, reflecting differences in flight altitude, camera intrinsics, and acquisition protocol. This variation directly affects the scale of observable entities and the degree of contextual map coverage available for grounding. Across all image–map pairs, we extract 3.34M OSM feature instances. The entity density is highly uneven, as expected from crowd-sourced maps: dense urban regions contain many roads, buildings, and amenities, while

rural or sparsely mapped areas provide fewer and sometimes incomplete entities. We intentionally preserve this heterogeneity rather than artificially equalizing it, because robust UAV–map grounding should operate under realistic variation in map completeness and local scene density.

Question distribution. The QA pairs span the eight tasks of Section 3.2, from cross-view grounding (T1–T3) to ego-centric (T4) and object-centric (T5–T8) spatial reasoning. To prevent the benchmark from being dominated by the easiest high-yield templates, each image generates three QA variants per task on average by shuffling the selected buildings, roads, vehicles, and distractors. This gives 42K questions per task and 336K QA pairs overall.

Answer and source balance. Source and task counts are balanced by construction: each source contributes 3,500 images, and each task contributes the same number of QA pairs. Answer letters are balanced within each task, and evaluation splits are stratified over both dataset sources and task types.

B.3 Human Verification

Human judgment enters the pipeline at four points: filtering registration outliers, verifying the cross-view correspondence pool, rating generated QA pairs, and providing an independent human answering baseline. The first three are performed by the same three coauthors, so that a consistent standard is applied across the whole pipeline; the fourth is deliberately performed by different, uninvolved people.

Stage 1: registration filtering by video replay. Because UAV images are sampled from continuous flight video, retained frames can be replayed in sequence together with their projected map overlay. Three coauthors watched this playback for every source and discarded frames where the automatic registration was visibly wrong—drift, an incorrect homography, or a misaligned basemap tile—beyond what the automatic inlier and footprint-plausibility checks already reject (Section 3.3; Figure 5 shows examples). This leaves the 14,000 released frames free of gross registration failures. To quantify what residual error remains, six annotators independently label ground control points (GCPs) for a stratified sample of 1,440 images (10.3% of the 14,000 retained frames) in QGIS (QGIS Development Team, 2026) and compare the manual and automatic transformations, pooled over a 64×64 grid per image (Table 7). The resulting few metres of error, from relief displacement or roof–footprint parallax, do not affect instance-level correspondence accuracy: building correspondences are established independently from SAM3 masks rather than by trusting the projected OSM polygon's position directly (Appendix B.5).

Stage 2: correspondence pool verification. The same three coauthors then inspected the OSM–SAM3 correspondence pool (Appendix B.5) that underlies T1–T3 and T5–T8, confirming accepted matches and discarding shifted or ambiguous pairs that the automatic acceptance rule had let through (Figure 6 shows examples). This human-verified pool is the one used to generate both the 3,200-question evaluation set (Test-3.2K) and the 12,800-question fine-tuning training split (Section 4.1): 16K questions in total, drawn from about 5,300 images since each image yields on average three QA variants per task (Section 3.3). We additionally audited 20% of these $\sim$5,300 images end-to-end and found 98% correctly matched, so every correspondence underlying our reported scores has been checked twice: once by the automatic agreement rule and once by a human annotator.

Stage 3: QA rating. The same three coauthors then audited 1,200 sampled QA pairs against four criteria—clarity, unique grounding in the joint image–map context, prompt correctness, and answer correctness—rating 92.73% *accept*, 3.32% *minor fix*, 1.58% *major fix*, and 2.37% *reject* (interface and criteria detailed in Section 3.4).

Stage 4: human answering baseline. To obtain an unbiased upper reference, six college students who were not involved in Stages 1–3—and therefore had no prior exposure to the construction pipeline or its failure modes—answered the human-reference questions reported alongside model accuracy (Table 2, Section 4.1).

B.4 UAV-to-Map Registration Accuracy

Table 7: Registration error of the automatic UAV-to-map registration against manual GCP georeferencing from the Stage 1 GCP study (Appendix B.3; 1,440 images, 64×64 grid per image, pooled). Left: error quantiles; right: fraction of grid points with error below each threshold.

Median	P25	P75	P95	<1 m	<3 m	<5 m	<7 m	<9 m
1.9 m	1.3 m	2.6 m	5.4 m	18%	78%	93%	96%	99%



Figure 5: Examples of UAV-to-map registration from the automatic registration and verification pipeline.

B.5 Cross-View Correspondence Pool

Why. The ground truth of UAV-MapVQA requires knowing which building in the UAV image corresponds to which building on the map: T1–T3 are built from such pairs, and T5–T8 use their per-image offset to lift vehicles onto the map. Registration alone is not enough. Residual registration error, OSM digitization at ground level, and roof–footprint parallax can shift projected footprints onto neighbouring buildings in dense areas, and OSM may omit visible structures. We therefore keep only building pairs confirmed from both views.

How. **(a)** The OSM building polygons inside the UAV footprint are **(b)** projected into the UAV image with the registration homography (Figure 6, top). **(c)** SAM3 segments buildings from the text prompt "building", independently of OSM. **(d)** The offset between polygons and masks is largely shared across the image, so we remove it with a global transform estimated by mask cross-correlation and refined by RANSAC. Corrected polygons and masks are then matched 1-to-1 by Hungarian assignment on IoU. A match is kept only if it is unambiguous and consistent: IoU ≥ 0.5, a margin ≥ 0.2 over the runner-up, a confident SAM3 detection, agreement with a second SAM3 mask prompted by the corrected OSM box, and no contact with the image border. Unmapped buildings, buildings cut by the border, and look-alike neighbours are excluded.



Figure 6: The cross-view correspondence pool. **Top:** the four steps on one image; in (d), red dashed outlines are matches that were not kept. **Bottom:** two pool examples per source dataset. Maps are north-up; the dashed outline is the UAV footprint and the solid edge is the top of the UAV image. UAV images are shown as captured. The same number and colour mark one OSM–SAM3 pair.

C TASK EXAMPLES

We show two examples for each of the eight tasks (Section 3.2). Each example is a card with four parts. **(i) Model inputs:** the map and the UAV image exactly as the model receives them in the *Visual* version, with the randomized visual prompts (marker colour, fill, line width, marker shape and label style vary per question); the UAV image is the same in the *Tabular* version. **(ii) Prompt and answer:** the task-specific part of the prompt, the options and the ground truth; the shared head and answer instruction are given below. **(iii) Tabular version:** the map image is replaced by a serialized table of the same map window (MAP DATA), and each map-side visual prompt becomes text: a row ID for a building or road, or a coordinate for a location. The card gives the prompt text of this version and the table rows the question refers to; the coloured letters in the first column are added for the reader and are not part of the table. **(iv) Debug view:** a panel that is never shown to the model and explains the answer from the registration and the verified SAM3↔OSM pairs. Common elements are the UAV footprint on the map (dashed; the solid edge is the top edge of the UAV image, and the dot is its top-left corner), option outlines in one colour per letter, the correct option marked with ✓, the vehicle lifted to the map (★), and a zoom on it from the UAV image. Complete prompts of both versions, full tables and records of all examples are in the supplementary material.

Shared design rules. Every question requires information that exists in only one of the two views, so map-only or UAV-only inputs should perform at chance. Ground truth is derived from the UAV-to-map registration and from cross-view-verified SAM3↔OSM building pairs (Appendix B.5), and every decision margin is at least three times the per-image registration error. Closed-way building targets and candidates must be at least $100\,\mathrm{m}^2$. Answer letters are balanced, visual prompts (marker colour, fill, line width, marker shape, label font, and placement) are randomized per question under a CIELAB contrast constraint, and UAV images are shown as captured and never rotated to north-up. Per-task construction rules that make each task non-trivial are given below, at the start of each subsection.

Prompt structure. In the Visual version the map (Image 1) and the UAV image (Image 2) are followed by the text below; in the Tabular version the UAV image is followed by the second head, the task text, the tables and the options.

Visual head (T1–T3): You are given two images of the same area. Image 1 is an OpenStreetMap map, drawn north-up (north is the top of the map). Image 2 is an aerial UAV (drone) photo; it can be rotated arbitrarily relative to the map and covers only part of the map area.
Visual head (T4–T8): . . . Image 2 is an aerial UAV (drone) photo; it is not necessarily north-up (its orientation relative to the map is unknown) and it covers only part of the map area.
Tabular head: You are given a map and an aerial UAV (drone) photo of the same area. The map is given as text (MAP DATA below): tables of the OpenStreetMap entities in a north-up square map window. The UAV photo is not necessarily north-up (its orientation relative to the map is unknown) and it covers only part of the map area.
⟨*task text and question, as in each card*⟩ [Tabular only: **MAP DATA** (tabular map, replaces the map image). Frame: north-up square map window, $S\,\mathrm{m} \times S\,\mathrm{m}$. Coordinates (x,y) in bbox/centroid columns are normalized 0–1, origin at the top-left (north-west) corner; x increases eastward, y increases southward. . . . Tables *Roads / Railways / Waterways* (id, name, category, tags, length_m, bbox, centroid, orientation), *Buildings / Areas* (id, name, category, tags, area_m2, shape_type, aspect_ratio, bbox, centroid) and *Topology: junctions* (feature_a_id, feature_a_name, feature_b_id, feature_b_name, at).]
Options: A. ⟨·⟩ B. ⟨·⟩ C. ⟨·⟩ D. ⟨·⟩ (one per line) **Tail:** Answer with the option letter (A, B, C or D) only. (T3: Answer with all correct option letters, separated by commas (e.g. "A, C").)

The tables are serialized from the OSM features that are rendered into the map image, in the normalized pixel frame of that image. Tags are restricted to a whitelist of semantic keys (road class, speed limit, lanes, building type, address, land use, . . .), lengths and areas are measured inside the map window, and the drivable ways of one named road form one row, so that every T7 option is one row. Nothing derived from the UAV image or its footprint enters the tables; for T3 they are serialized from the edited map.

C.1 Task Overview: Tabular Version

Figures 7 and 8 show the eight questions of Figure 3 in the Tabular version. The UAV images, UAV-side visual prompts and answers are the same; the map image is replaced by the MAP DATA tables of the same window, and map-side visual prompts become row IDs or coordinates in the question and options.

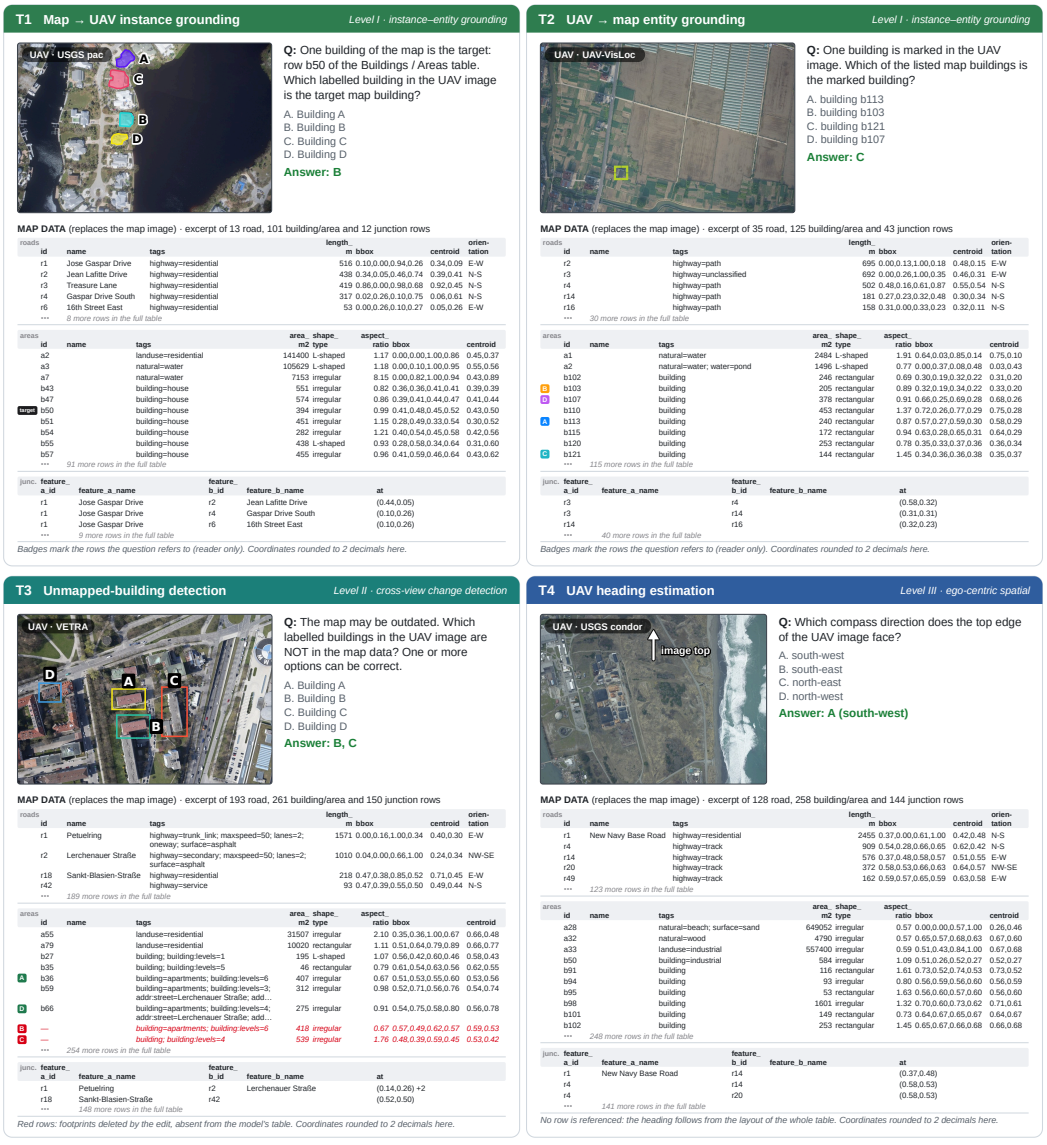


Figure 7: The questions of Figure 3 (left, T1–T4) in the Tabular version. Each card shows the UAV image with the question's visual prompts, the Tabular question, the options, the answer and an excerpt of the MAP DATA tables: the rows the question refers to plus nearby context rows, with coordinates rounded to two decimals (three in the model's tables). The model receives the full tables. Badges next to the rows and the red rows (T3: footprints deleted by the edit, absent from the model's table) are for the reader only.

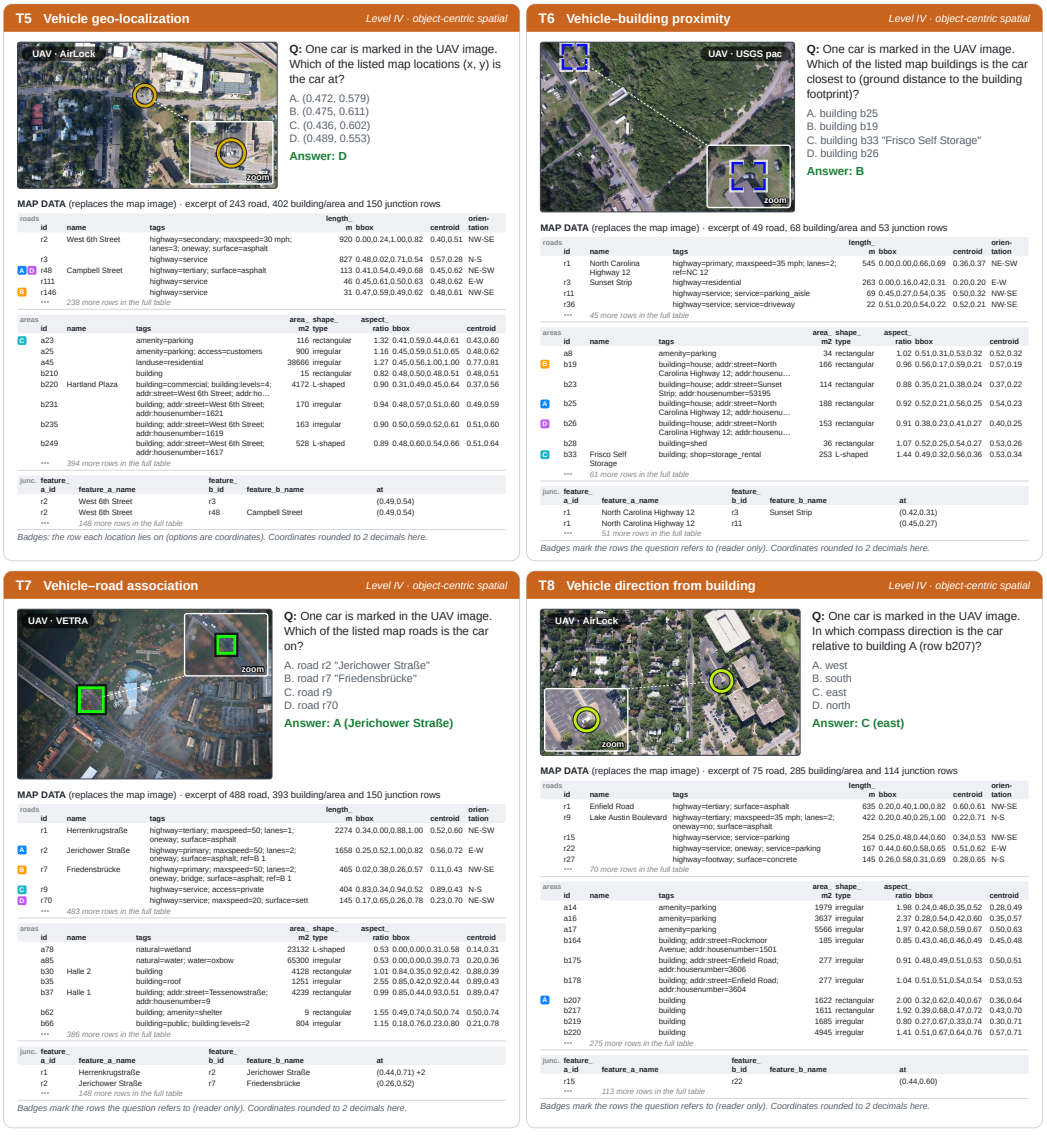
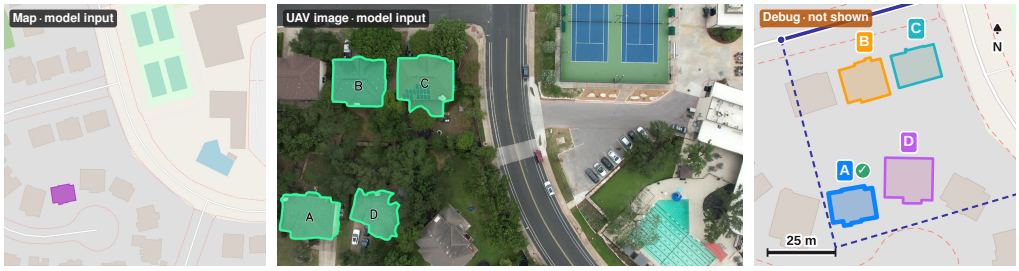


Figure 8: The questions of Figure 3 (right, T5–T8) in the Tabular version, laid out as in Figure 7. Badges mark the rows the options refer to (T5: the row each candidate location lies on) and are for the reader only; the inset is a zoom on the marked car.

C.2 LEVEL I: T1 MAP→UAV INSTANCE GROUNDING

One building is marked on the map; the model selects the corresponding instance among four buildings labelled in the UAV image. In the Tabular version the marked building is named by its row ID. All four options are real, distinct buildings verified in both views, with area rank and centrality rank balanced among them so heuristics such as "pick the largest" or "pick the most central" are at chance; options are at least 34 px apart on the map and at least 4.5% of the image width apart in the UAV image.

T1 · Example 1 AirLock · Austin, Texas, USA · map window 185 m × 185 m · T1_0001



Visual One building is marked on the map. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which labelled building in the UAV image is the building marked on the map?
Options A. Building A B. Building B C. Building C D. Building D **Answer: A**

Tabular MAP DATA replaces the map image (16 road/rail/water, 36 building/area and 14 junction rows; excerpt: 4 of 36 building/area rows). One building of the map is the target: the row with ID 383155354 in the Buildings / Areas table. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which labelled building in the UAV image is the target map building?

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
target	383155354		BUILDINGS	building; addr:street=Chictora Cove; addr:housenumber…	193	irregular	1.22	0.177,0.686,0.274,0.765	0.226,0.725
	383155363		BUILDINGS	building; addr:street=Chictora Cove; addr:housenumber…	283	irregular	0.82	0.059,0.690,0.161,0.814	0.110,0.753
	383155351		BUILDINGS	building; addr:street=Chictora Cove; addr:housenumber…	272	irregular	1.06	0.290,0.630,0.386,0.721	0.339,0.673
	383155372		BUILDINGS	building; addr:street=Chictora Cove; addr:housenumber…	141	irregular	0.55	0.000,0.618,0.067,0.740	0.022,0.680

Debug Verified OSM footprints of the four UAV options, on the clean map (colour = letter). A is the building marked on the map. SAM3–OSM IoU of the pairs: A 0.69, B 0.89, C 0.81, D 0.56; answer area rank 2/4, centrality rank 4/4 (both stratified).

T1 · Example 2 UAV-VisLoc · China · map window 646 m × 646 m · T1_0023



Visual One building is marked on the map. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which labelled building in the UAV image is the building marked on the map?
Options A. Building A B. Building B C. Building C D. Building D **Answer: B**

Tabular MAP DATA replaces the map image (8 road/rail/water, 22 building/area and 5 junction rows; excerpt: 4 of 22 building/area rows). One building of the map is the target: the row with ID 1231418151 in the Buildings / Areas table. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which labelled building in the UAV image is the target map building?

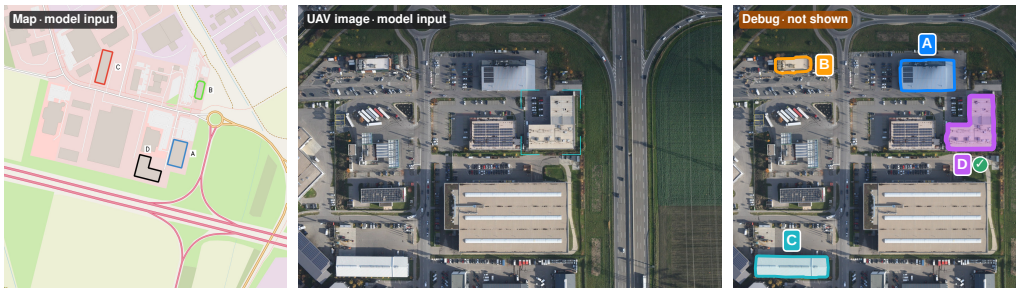
	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
target	1231418151		BUILDINGS	building	2885	rectangular	1.70	0.289,0.505,0.401,0.571	0.345,0.538
	1231418152		BUILDINGS	building	2844	rectangular	1.79	0.286,0.426,0.400,0.489	0.343,0.457
	1231418153		BUILDINGS	building	8602	rectangular	1.57	0.277,0.601,0.465,0.720	0.371,0.661
	1231418150		BUILDINGS	building	2868	rectangular	1.77	0.416,0.505,0.527,0.568	0.471,0.537

Debug Verified OSM footprints of the four UAV options, on the clean map (colour = letter). B is the building marked on the map. SAM3–OSM IoU of the pairs: A 0.92, B 0.93, C 0.93, D 0.82; answer area rank 2/4, centrality rank 2/4 (both stratified).

C.3 LEVEL I: T2 UAV→MAP ENTITY GROUNDING

One building is marked in the UAV image; the model selects the corresponding map entity among four labelled map buildings. In the Tabular version the options are four rows of the building table. Distractor rules are as in T1 (Appendix C.2); no T2 question reuses a T1 query pair.

T2 · Example 1 VETRA · Germany · map window 615 m × 615 m · T2_0229



Visual One building is marked in the UAV image. Four buildings are marked and labelled A, B, C and D on the map.
Question: Which labelled building on the map is the building marked in the UAV image?
Options A. Building A B. Building B C. Building C D. Building D **Answer: D**

Tabular MAP DATA replaces the map image (106 road/rail/water, 109 building/area and 136 junction rows; excerpt: 4 of 109 building/area rows). One building is marked in the UAV image. Four map buildings are listed as options A, B, C and D (rows of the Buildings / Areas table). **Question:** Which of the listed map buildings is the building marked in the UAV image?
Options A. building "Stadtbäcker Scharold" (ID: 536455330) B. building "McDonald's" (ID: 256301836) C. building "Hofner Hebetechnik GmbH" (ID: 242679234) D. building "Finstral Studio" (ID: 400512848)

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	536455330	Stadtbäcker Scharold	BUILDINGS	building=commercial; shop=bakery; addr:…	1748	irregular	0.70	0.577,0.474,0.648,0.575	0.613,0.525
B	256301836	McDonald's	BUILDINGS	building; amenity=fast_food; addr:stree…	606	irregular	0.57	0.673,0.267,0.710,0.333	0.691,0.300
C	242679234	Hofner Hebetechnik GmbH	BUILDINGS	building; addr:street=Winterbruckenweg…	1588	rectangular	0.52	0.320,0.157,0.384,0.280	0.352,0.219
D	400512848	Finstral Studio	BUILDINGS	building=industrial; addr:street=Winter…	1918	irregular	0.98	0.460,0.528,0.555,0.625	0.504,0.582

Debug SAM3 masks of the buildings paired with the four map options, on the clean UAV image (colour = letter). D is the building marked in the UAV image. SAM3–OSM IoU of the pairs: A 0.92, B 0.93, C 0.92, D 0.91; answer area rank 4/4, centrality rank 1/4 (both stratified).

T2 · Example 2 USGS · Gulf coast of Florida, USA · map window 506 m × 506 m · T2_0157



Visual One building is marked in the UAV image. Four buildings are marked and labelled A, B, C and D on the map.
Question: Which labelled building on the map is the building marked in the UAV image?
Options A. Building A B. Building B C. Building C D. Building D **Answer: A**

Tabular MAP DATA replaces the map image (98 road/rail/water, 80 building/area and 150 junction rows; excerpt: 4 of 80 building/area rows). One building is marked in the UAV image. Four map buildings are listed as options A, B, C and D (rows of the Buildings / Areas table). **Question:** Which of the listed map buildings is the building marked in the UAV image?
Options A. building (ID: 243116326) B. building "Surf Style" (ID: 523898194) C. building (ID: 243116347) D. building "Gulfview Hotel on the Beach" (ID: 192113277)

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	243116326		BUILDINGS	building	660	irregular	0.80	0.837,0.379,0.900,0.457	0.870,0.419
B	523898194	Surf Style	BUILDINGS	building; shop=clothes	1309	irregular	0.84	0.694,0.343,0.777,0.441	0.736,0.392
C	243116347		BUILDINGS	building	526	irregular	0.65	0.618,0.367,0.665,0.439	0.641,0.404
D	192113277	Gulfview Hotel on the Beach	BUILDINGS	building; tourism=hotel; addr:street=So…	791	irregular	0.79	0.444,0.466,0.504,0.541	0.473,0.502

Debug SAM3 masks of the buildings paired with the four map options, on the clean UAV image (colour = letter). A is the building marked in the UAV image. SAM3–OSM IoU of the pairs: A 0.79, B 0.95, C 0.88, D 0.77; answer area rank 2/4, centrality rank 4/4 (both stratified).

C.4 Level II: T3 Unmapped-Building Detection

Four buildings are labelled in the UAV image; the OSM footprints of $k \in \{1, 2, 3\}$ of them (and 0–2 unmarked decoys) are deleted from the vector map, which is then re-rendered or re-serialized. The model selects all labelled buildings that are missing from the map. Unmarked decoy deletions prevent answering via empty-area cues; each letter is positive in 50% of questions (exact-set chance $1/14$).

T3 · Example 1 AirLock · Austin, Texas, USA · map window 205 m × 205 m · T3_0078



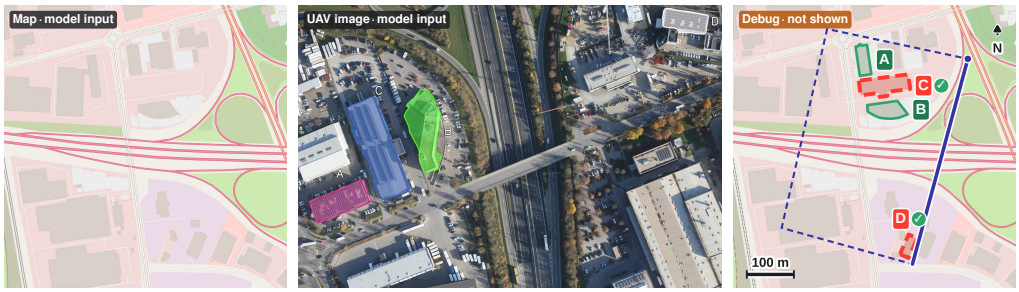
Visual The map may be outdated: some buildings that exist in the UAV image may be missing from the map. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which of the labelled buildings are NOT shown on the map (unmapped buildings)? One or more options can be correct.
Options A. Building A B. Building B C. Building C D. Building D **Answer: A**

Tabular MAP DATA replaces the map image (18 road/rail/water, 78 building/area and 23 junction rows; excerpt: 3 of 78 building/area rows; red rows were deleted by the edit and are *not* in the table). The map may be outdated: some buildings that exist in the UAV image may be missing from the map data. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which of the labelled buildings are NOT in the map data (unmapped buildings)? One or more options can be correct.

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	*382602041*		*BUILDINGS*	*building; addr:street=Waterston Avenue; addr:housenum…*	*124*	*irregular*	*0.82*	*0.383,0.547,0.449,0.628*	*0.414,0.585*
B	382395239		BUILDINGS	building; addr:street=Waterston Avenue; addr:housenum…	173	irregular	0.88	0.569,0.421,0.649,0.511	0.610,0.465
C	382395243		BUILDINGS	building; addr:street=Waterston Avenue; addr:housenum…	256	irregular	0.97	0.481,0.355,0.594,0.471	0.536,0.413
D	382395245		BUILDINGS	building; addr:street=Palma Plaza; addr:housenumber=1…	199	irregular	0.99	0.602,0.186,0.700,0.285	0.649,0.236

Debug Original map before the synthetic edit: the OSM footprints of A (red) were deleted and the map re-rendered with the same renderer; B, C and D (green) were kept. 2 unmarked decoy buildings (grey, hatched) were deleted as well, so an empty lot on the map is not by itself evidence for an option.

T3 · Example 2 VETRA · Germany · map window 600 m × 600 m · T3_0410



Visual The map may be outdated: some buildings that exist in the UAV image may be missing from the map. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which of the labelled buildings are NOT shown on the map (unmapped buildings)? One or more options can be correct.
Options A. Building A B. Building B C. Building C D. Building D **Answer: C, D**

Tabular MAP DATA replaces the map image (134 road/rail/water, 87 building/area and 150 junction rows; excerpt: 2 of 87 building/area rows; red rows were deleted by the edit and are *not* in the table). The map may be outdated: some buildings that exist in the UAV image may be missing from the map data. Four buildings are marked and labelled A, B, C and D in the UAV image. **Question:** Which of the labelled buildings are NOT in the map data (unmapped buildings)? One or more options can be correct.

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	93133975		BUILDINGS	building; addr:street=Daimlerstraße; ad…	1570	irregular	0.50	0.432,0.137,0.488,0.249	0.460,0.194
B	93133978		BUILDINGS	building; addr:street=Welserstraße; add…	2295	irregular	2.28	0.473,0.338,0.617,0.401	0.544,0.371
C	*93133987*		*BUILDINGS*	*building=warehouse; addr:street=Welsers…*	*3458*	*irregular*	*2.27*	*0.443,0.244,0.625,0.324*	*0.534,0.283*
D	*242107128*	*Schlosserei Bickebach*	*BUILDINGS*	*building=industrial; addr:street=Wernhe…*	*921*	*rectangular*	*0.73*	*0.589,0.812,0.651,0.895*	*0.620,0.853*

Debug Original map before the synthetic edit: the OSM footprints of C and D (red) were deleted and the map re-rendered with the same renderer; A and B (green) were kept. No decoy building was deleted in this question (0–2 per question).

C.5 Level III: T4 UAV Heading Estimation

Given the plain map (or its table) and the plain UAV image, the model selects the compass direction that the top edge of the UAV image faces. The options are always four directions 90° apart. Ground truth follows from the registration homography at the image centre; oblique frames and headings more than 20° from a principal direction are excluded, leaving a margin of at least 70° between the answer and any distractor, and images are drawn round-robin over the eight heading groups so "closest to north" scores only 24.2%.

T4 · Example 1 USGS · Gulf coast of Florida, USA · map window 525 m × 525 m · T4_0401



Visual **Question:** Which compass direction does the top edge of the UAV image face? (i.e., moving from the bottom of the UAV image towards its top, which direction on the map are you heading?)
Options A. north-east B. south-east C. north-west D. south-west **Answer: C (north-west)**

Tabular MAP DATA replaces the map image (42 road/rail/water, 345 building/area and 57 junction rows; excerpt: 4 of 42 road/rail/water rows). *(question text as in the Visual version)*

id	name	category	tags	length_m	bbox	centroid	orientation
11317309,…	Excelsior	ROADS	highway=residential; access=destinati…	979	0.275,0.289,0.994,0.954	0.670,0.571	NE-SW
48181200	Midnight Pass Road	ROADS	highway=secondary; maxspeed=35 mph; l…	582	0.190,-0.000,0.660,1.000	0.458,0.528	NW-SE
769577335,…	Casa Blanca Street	ROADS	highway=residential; oneway; access=d…	412	-0.000,0.167,0.329,0.491	0.221,0.274	NE-SW
11320299,…	Bay Oaks	ROADS	highway=residential; access=destinati…	398	0.282,-0.000,0.647,0.212	0.411,0.131	E-W

Debug The top edge of the UAV image faces 314°, 0° from north-west; the options (compass rose) are 90° apart. Heading spread across the frame 1.9°; registration error 0.7 m.

T4 · Example 2 UAV-VisLoc · China · map window 911 m × 911 m · T4_0204



Visual **Question:** Which compass direction does the top edge of the UAV image face? (i.e., moving from the bottom of the UAV image towards its top, which direction on the map are you heading?)
Options A. west B. north C. south D. east **Answer: A (west)**

Tabular MAP DATA replaces the map image (27 road/rail/water, 10 building/area and 34 junction rows; excerpt: 2 of 27 road/rail/water rows). *(question text as in the Visual version)*

id	name	category	tags	length_m	bbox	centroid	orientation
1107935612,…	连霍高速	ROADS	highway=motorway; lanes=2; oneway; re…	1403	0.367,-0.000,1.000,0.459	0.622,0.178	NW-SE
37096900	黄羚 - 山丹	ROADS	highway=trunk; ref=G307	962	0.092,0.605,1.000,1.000	0.475,0.886	E-W

Debug The top edge of the UAV image faces 270°, 0° from west; the options (compass rose) are 90° apart. Heading spread across the frame 0.4°; registration error 1.1 m.

C.6 LEVEL IV: T5 VEHICLE GEO-LOCALIZATION

A vehicle is marked in the UAV image and four locations are marked on the map; the model selects the vehicle's position. In the Tabular version the locations are given as coordinates in the frame of the tables. Vehicles (here and in T6–T8) are SAM3 detections kept only if car-like in ground-projected size (5–16 m^2) with confidence at least 0.5. Distractor locations lie on drivable roads or parking areas, at least $\max(20\,m, 5\times$ registration error) from the answer and from each other, so road semantics alone cannot identify the answer.

T5 · Example 1 AirLock · Austin, Texas, USA · map window 431 m × 431 m · `T5_ex1`



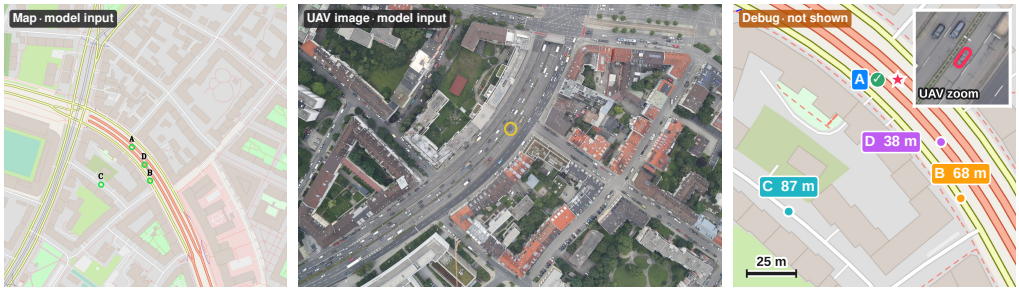
Visual One car is marked in the UAV image. Four locations are marked and labelled A, B, C and D on the map. **Question:** Which labelled location on the map is the marked car at?
Options A. Location A B. Location B C. Location C D. Location D **Answer: C**

Tabular MAP DATA replaces the map image (76 road/rail/water, 287 building/area and 102 junction rows; excerpt: 2 of 76 road/rail/water rows). One car is marked in the UAV image. Four map locations are listed as options A, B, C and D (x, y in the frame of the map tables). **Question:** Which of the listed map locations is the marked car at?
Options A. location (x, y) = (0.409, 0.606) B. location (x, y) = (0.491, 0.493) C. location (x, y) = (0.439, 0.556) D. location (x, y) = (0.495, 0.377)

	id	name	category	tags	length_m	bbox	centroid	orientation
A B C	160839711, . . .	Charlotte Street	ROADS	highway=residential; surface=asphalt	176	0.408,0.176,0.656,0.609	0.503,0.441	NE-SW
D	15410415, . . .	West 11th Street	ROADS	highway=residential; maxspeed=25 mph. . .	360	-0.000,0.099,0.879,0.734	0.549,0.468	NW-SE

Debug Marked car lifted onto the map through the registration (★), at option C; the badges give the distance of each distractor from the car. Distractors lie on drivable roads and parking areas, at least 20 m apart; A and B lie on the car's road (Charlotte Street). Registration error 1.7 m.

T5 · Example 2 VETRA · Germany · map window 503 m × 503 m · `T5_ex2`



Visual One car is marked in the UAV image. Four locations are marked and labelled A, B, C and D on the map. **Question:** Which labelled location on the map is the marked car at?
Options A. Location A B. Location B C. Location C D. Location D **Answer: A**

Tabular MAP DATA replaces the map image (424 road/rail/water, 354 building/area and 150 junction rows; excerpt: 2 of 424 road/rail/water rows). One car is marked in the UAV image. Four map locations are listed as options A, B, C and D (x, y in the frame of the map tables). **Question:** Which of the listed map locations is the marked car at?
Options A. location (x, y) = (0.452, 0.505) B. location (x, y) = (0.516, 0.624) C. location (x, y) = (0.342, 0.637) D. location (x, y) = (0.496, 0.567)

	id	name	category	tags	length_m	bbox	centroid	orientation
A B D	100307487, . . .	Tegernseer Landstraße	ROADS	highway=trunk; maxspeed=60; lanes=2. . .	1889	0.163,-0.000,0.733,1.000	0.437,0.461	NW-SE
C	4981748	Zasingerstraße	ROADS	highway=residential; maxspeed=30; sur. . .	136	0.224,0.557,0.434,0.724	0.342,0.643	NW-SE

Debug Marked car lifted onto the map through the registration (★), at option A; the badges give the distance of each distractor from the car. Distractors lie on drivable roads and parking areas, at least 20 m apart; B and D lie on the car's road (Tegernseer Landstraße). Registration error 1.4 m.

C.7 LEVEL IV: T6 VEHICLE–BUILDING PROXIMITY

A vehicle is marked in the UAV image; the model selects, among four labelled map buildings, the one nearest to the vehicle (ground distance to the footprint). The answer is the nearest of *all* OSM buildings, not only the options, closer than the runner-up by at least $\max(4\,\text{m}, 3\times \text{error})$ and by a factor of 1.5, and agrees with the nearest instance mask in the UAV image; area and centrality ranks are stratified as in T1.

T6 · Example 1 USGS · Gulf coast of Florida, USA · map window 541 m × 541 m · T1_0445



Visual One car is marked in the UAV image. Four buildings are labelled A, B, C and D on the map. **Question:** Which labelled building on the map is the marked car closest to (ground distance to the building footprint)?
Options A. Building A B. Building B C. Building C D. Building D **Answer: C**

Tabular MAP DATA replaces the map image (29 road/rail/water, 95 building/area and 30 junction rows; excerpt: 4 of 95 building/area rows). One car is marked in the UAV image. Four map buildings are listed as options A, B, C and D (rows of the Buildings / Areas table). **Question:** Which of the listed map buildings is the marked car closest to (ground distance to the building footprint)?
Options A. building (ID: 523216887) B. building (ID: 523211996) C. building (ID: 523211974) D. building (ID: 523212000)

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	523216887		BUILDINGS	building; addr:street=Captains Place; addr:housenumbe…	801	irregular	0.61	0.355,0.498,0.407,0.585	0.377,0.540
B	523211996		BUILDINGS	building; addr:street=Half Moon Walk; addr:housenumbe…	497	irregular	0.99	0.414,0.355,0.462,0.403	0.438,0.377
C	523211974		BUILDINGS	building; addr:street=Gordon Drive; addr:housenumber=…	1454	irregular	0.67	0.230,0.383,0.296,0.482	0.268,0.429
D	523212000		BUILDINGS	building=house; addr:street=Half Moon Walk; addr:hous…	517	irregular	1.16	0.661,0.350,0.726,0.407	0.692,0.378

Debug Lifted car (★) with its ground distance to each footprint (badges). C is the nearest of all OSM buildings, 47 m closer than the runner-up (registration error 1.4 m), and also the nearest SAM3 mask in the UAV image.

T6 · Example 2 AirLock · Austin, Texas, USA · map window 708 m × 708 m · T1_0024



Visual One car is marked in the UAV image. Four buildings are labelled A, B, C and D on the map. **Question:** Which labelled building on the map is the marked car closest to (ground distance to the building footprint)?
Options A. Building A B. Building B C. Building C D. Building D **Answer: D**

Tabular MAP DATA replaces the map image (797 road/rail/water, 559 building/area and 150 junction rows; excerpt: 4 of 559 building/area rows). One car is marked in the UAV image. Four map buildings are listed as options A, B, C and D (rows of the Buildings / Areas table). **Question:** Which of the listed map buildings is the marked car closest to (ground distance to the building footprint)?
Options A. building (ID: 99868630) B. building "Seaholm Residences" (ID: 516271460) C. building (ID: 99868617) D. building (ID: 135049606)

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	99868630		BUILDINGS	building=office; building:levels=1	1373	irregular	1.08	0.683,0.253,0.754,0.318	0.717,0.285
B	516271460	Seaholm Residences	BUILDINGS	building; building:levels=29; addr:stre…	2772	irregular	0.74	0.367,0.395,0.451,0.509	0.409,0.452
C	99868617		BUILDINGS	building=apartments; addr:street=West A…	1211	irregular	0.60	0.442,0.275,0.495,0.363	0.468,0.317
D	135049606		BUILDINGS	building; amenity=parking; access=priva…	3623	rectangular	1.12	0.430,0.116,0.537,0.213	0.482,0.165

Debug Lifted car (★) with its ground distance to each footprint (badges). D is the nearest of all OSM buildings, 38 m closer than the runner-up (registration error 1.6 m), and also the nearest SAM3 mask in the UAV image.

C.8 Level IV: T7 Vehicle–Road Association

A vehicle is marked in the UAV image; the model selects, among four highlighted map roads, the road the vehicle is on. The vehicle lies on the carriageway, aligned within 25° and 30 m from other roads (excluding junctions); width, length, and class are stratified with a same-class distractor, so "pick the widest road" scores only 25.2%.

T7 · Example 1 VETRA · Germany · map window 504 m × 504 m · T2_0019



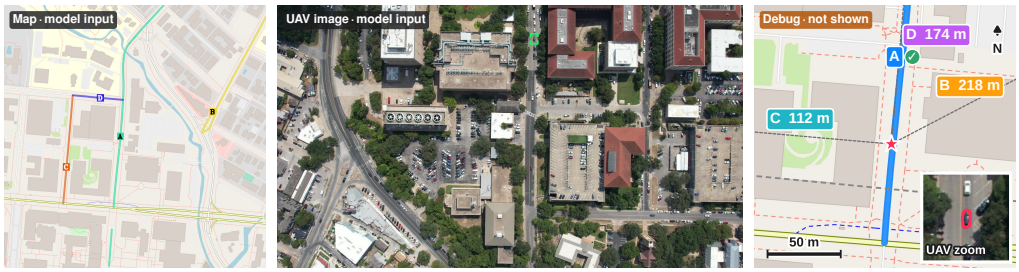
Visual One car is marked in the UAV image. Four roads are highlighted and labelled A, B, C and D on the map. **Question:** Which labelled road on the map is the marked car on?
Options A. Road A B. Road B C. Road C D. Road D **Answer: B (Sankt-Blasien-Straße)**

Tabular MAP DATA replaces the map image (193 road/rail/water, 263 building/area and 150 junction rows; excerpt: 4 of 193 road/rail/water rows). One car is marked in the UAV image. Four map roads are listed as options A, B, C and D (rows of the Roads / Railways / Waterways table). **Question:** Which of the listed map roads is the marked car on?
Options A. road "Georg-Brauchle-Ring" (ID: 4300524,4300523,95800607,146396400) B. road "Sankt-Blasien-Straße" (ID: 23540976) C. road "Spiridon-Louis-Ring" (ID: 1143276742,1279888068,17919189,17919030) D. road "Birnauer Straße" (ID: 311413661,89360845)

	id	name	category	tags	length_m	bbox	centroid	orientation
A	4300524,…	Georg-Brauchle-Ring	ROADS	highway=trunk; maxspeed=60; lanes=2…	287	-0.000,0.292,0.290,0.315	0.139,0.303	E-W
B	23540976	Sankt-Blasien-Straße	ROADS	highway=residential	218	0.473,0.377,0.849,0.515	0.709,0.450	E-W
C	1143276742,…	Spiridon-Louis-Ring	ROADS	highway=unclassified; maxspeed=30; la…	312	-0.000,0.379,0.394,0.658	0.265,0.538	NW-SE
D	311413661,…	Birnauer Straße	ROADS	highway=residential; maxspeed=30; sur…	284	0.480,0.504,1.000,0.721	0.697,0.632	NE-SW

Debug Lifted car (★) on road B (Sankt-Blasien-Straße, residential); the badges give the distances to the other option roads. The car is on the carriageway, aligned with the road and on the SAM3 road mask, and every other drivable road is at least 30 m away (11 drivable roads visible).

T7 · Example 2 AirLock · Austin, Texas, USA · map window 588 m × 588 m · T2_0034



Visual One car is marked in the UAV image. Four roads are highlighted and labelled A, B, C and D on the map. **Question:** Which labelled road on the map is the marked car on?
Options A. Road A B. Road B C. Road C D. Road D **Answer: A (Speedway)**

Tabular MAP DATA replaces the map image (440 road/rail/water, 204 building/area and 150 junction rows; excerpt: 4 of 440 road/rail/water rows). One car is marked in the UAV image. Four map roads are listed as options A, B, C and D (rows of the Roads / Railways / Waterways table). **Question:** Which of the listed map roads is the marked car on?
Options A. road "Speedway" (ID: 25703280,124953127,571500826,124953126,37749597,37749598,939765764) B. road "Duval Street" (ID: 126015162,126015163,655283373,126015164) C. road "Wichita Street" (ID: 25775203) D. road "East 27th Street" (ID: 328089344,15385824)

	id	name	category	tags	length_m	bbox	centroid	orientation
A	25703280,…	Speedway	ROADS	highway=tertiary; surface=asphalt	569	0.401,0.000,0.610,0.991	0.476,0.310	N-S
B	126015162,…	Duval Street	ROADS	highway=tertiary; maxspeed=30 mph; la…	362	0.755,0.044,1.000,0.484	0.825,0.368	NE-SW
C	25775203	Wichita Street	ROADS	highway=unclassified; oneway; surface…	243	0.226,0.343,0.262,0.754	0.247,0.519	N-S
D	328089344,…	East 27th Street	ROADS	highway=unclassified; surface=asphalt	114	0.262,0.343,0.454,0.360	0.362,0.352	E-W

Debug Lifted car (★) on road A (Speedway, tertiary); the badges give the distances to the other option roads. The car is on the carriageway, aligned with the road and on the SAM3 road mask, and every other drivable road is at least 30 m away (14 drivable roads visible).

C.9 Level IV: T8 Vehicle Direction from Building

A vehicle is marked in the UAV image and one building is labelled A on the map; the model answers in which compass direction the vehicle lies relative to building A. T8 composes T1 (ground building A), T4 (image heading), and bearing estimation: the vehicle is 3–40 m from A with bearing at least 22.5° from any sector boundary, so answering from image directions alone, ignoring the true heading, is wrong in roughly three quarters of cases.

T8 · Example 1 — UAV-VisLoc · China · map window 483 m × 483 m · `T3_0094`



Visual One car is marked in the UAV image. One building is labelled "A" on the map. **Question:** In which compass direction is the marked car located relative to building A?
Options A. west B. north C. south D. east **Answer: A (west)**

Tabular MAP DATA replaces the map image (24 road/rail/water, 74 building/area and 40 junction rows; excerpt: 3 of 74 building/area rows). One car is marked in the UAV image. Building A is the row with ID 721133900 in the Buildings / Areas table. **Question:** In which compass direction is the marked car located relative to building A?

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	721133900		BUILDINGS	building	638	rectangular	0.73	0.652,0.358,0.708,0.436	0.680,0.397
	721133902		BUILDINGS	building	927	irregular	1.28	0.625,0.434,0.715,0.504	0.674,0.470
	721133896		BUILDINGS	building	487	rectangular	0.50	0.648,0.277,0.691,0.363	0.669,0.320

Debug Compass sectors (±45°) around the centroid of building A: the car (★) is at bearing 286°, 17 m from A, i.e. west. The top edge of the UAV image faces 169°, so reading the UAV image as if it were north-up would give "east".

T8 · Example 2 — AirLock · Austin, Texas, USA · map window 431 m × 431 m · `T3_0062`



Visual One car is marked in the UAV image. One building is labelled "A" on the map. **Question:** In which compass direction is the marked car located relative to building A?
Options A. west B. east C. south D. north **Answer: C (south)**

Tabular MAP DATA replaces the map image (148 road/rail/water, 194 building/area and 150 junction rows; excerpt: 3 of 194 building/area rows). One car is marked in the UAV image. Building A is the row with ID 382559228 in the Buildings / Areas table. **Question:** In which compass direction is the marked car located relative to building A?

	id	name	category	tags	area_ m2	shape_ type	aspect_ ratio	bbox	centroid
A	382559228		BUILDINGS	building; addr:street=West 6th Street; addr:housenumb…	1557	irregular	0.76	0.334,0.374,0.460,0.538	0.399,0.456
	382559661		BUILDINGS	building	57	irregular	0.89	0.341,0.419,0.363,0.444	0.351,0.432
	382559480		BUILDINGS	building	17	rectangular	1.20	0.360,0.397,0.374,0.409	0.367,0.403

Debug Compass sectors (±45°) around the centroid of building A: the car (★) is at bearing 198°, 15 m from A, i.e. south. The top edge of the UAV image faces 74°, so reading the UAV image as if it were north-up would give "east".

D Evaluation Protocol

D.1 Metric and Scoring

All tasks are evaluated using *multiple-choice accuracy*:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\text{pred}_i = \text{label}_i] \tag{1}$$

where pred_i is the predicted option letter(s) and label_i is the ground truth. Tasks require either **single-answer** selection (one letter) or **multi-answer** selection: T3 (unmapped-building detection) asks for all labelled buildings that are missing from the map and is the only multi-answer task; all other tasks have exactly one correct option.

A prediction is correct if it equals the ground-truth letter, or for T3 the exact ground-truth set; invalid predictions, refusals, truncated responses, and over-length prompts are scored incorrect, with their rate reported per model. Requests failing with rate-limit, server, or timeout errors, or an empty message, are retried with exponential backoff up to five times. Every response is logged with question ID, model, provider, finish reason, and token usage, so all scores can be recomputed from these logs without re-querying the models. Prompts are versioned in `prompts_config.json`, and evaluation uses stratified sampling with a fixed seed (42) to select n questions per task from each source; full scripts and configurations are in the code release.

D.2 Request Construction

The shared heads, task-specific text, and answer-option format of all eight templates are given in Appendix C, instantiated for two examples per task. Each question is a single user turn with no system prompt or in-context examples: the Visual version contains the map (Image 1), the UAV image (Image 2), and the prompt text; the Tabular version contains the UAV image and the prompt text, including the MAP DATA tables (Appendix C). Both images are resized to 640 px on the longest side, preserving aspect ratio, and sent as base64 data. The prompt text is the dataset record's question with its last line replaced by a zero-shot chain-of-thought instruction (Kojima et al., 2022): "`Think step by step, then end with Answer: <letter>.`" for single-answer tasks, and the analogous multi-letter instruction for T3. Figure 9 shows a complete request. For the heading-aligned maps of Section 4.3, the north-up sentence is replaced by one stating that the map is rotated and north is given by the north arrow (Visual) or the stated bearing (Tabular).

[Image 1: map] *[Image 2: UAV image]*

```
You are given two images of the same area.  Image 1 is an OpenStreetMap map, drawn
north-up (north is the top of the map).  Image 2 is an aerial UAV (drone) photo; it can
be rotated arbitrarily relative to the map and covers only part of the map area.
One building is marked on the map.  Four buildings are marked and labelled A, B, C and D
in the UAV image.
Question:  Which labelled building in the UAV image is the building marked on the map?
Options:
A. Building A
B. Building B
C. Building C
D. Building D
Think step by step, then end with Answer:  <letter>.
```

Figure 9: A complete T1 request in the Visual version. Only the bold line differs from the dataset record.

D.3 Model Serving and Decoding

Serving. All models are queried through one client that uses the OpenAI-compatible chat-completions format, so every model receives the same messages. The seven proprietary models and Minimax-01 are accessed through the OpenRouter API, with each model pinned to a single upstream provider and fallback routing disabled; Qwen VL Max and Qwen VL Plus are accessed through the OpenAI-compatible endpoint of Alibaba Cloud Model Studio. Qwen2.5-VL-7B and Qwen2.5-VL-32B are served locally in bfloat16 with vLLM on NVIDIA A100 (40 GB) GPUs, with the context extended to 128K tokens by YaRN rope scaling as documented for these models, so that

long Tabular prompts are not truncated. GeoChat (Kuckreja et al., 2024) is evaluated on the Visual version only, with its released inference code, `llava_v1` template, and preprocessing (padding to a square and resizing to 504 px). Because it accepts a single image, the map and the UAV image are placed side by side on one canvas under the headers "Image 1: map (north-up)" and "Image 2: UAV", so the question text is unchanged.

Decoding. Models without built-in reasoning decode greedily (temperature 0) with a budget of 4,096 output tokens. Models with built-in reasoning (GPT-5, GPT-5.5, and the Claude models) run with reasoning enabled at medium effort, a total budget of 16,384 tokens shared by reasoning and the visible answer, and the provider's default temperature, since these models do not support greedy decoding with reasoning enabled. Hidden reasoning is never parsed; only the final visible message is scored. GeoChat decodes greedily with up to 1,024 new tokens.

D.4 Answer Parsing

The parser is deterministic and identical for all models. After removing Markdown emphasis, it takes the text after the *last* occurrence of `Answer:` (case-insensitive) up to the end of that line, so a model may revise an earlier answer while reasoning, and collects the standalone uppercase letters A–D in it ("C", "(C)", and "Building C" all yield C). A response without an answer line is accepted only if it consists solely of option letters and separators (e.g., "C" or "A, D"). A single-answer prediction is valid only if exactly one distinct letter is found, so hedged answers such as "B or D" are invalid; a T3 prediction is the set of letters found and must be non-empty. The reasoning text is never searched for letters, because chain-of-thought responses mention every option while comparing them.

D.5 Supplementary Results

Test-0.8K scores of all models. Proprietary models are scored on Test-0.8K only. For comparison, Table 8 reports the other models on both Test-3.2K and its nested Test-0.8K subset; the two agree within the sampling error of 800 questions.

Table 8: Overall accuracy (%) of the models evaluated on Test-3.2K (3,200 questions) and on its nested Test-0.8K subset (800 questions), which the proprietary models use.

	Visual		Tabular	
Model	**3.2K**	**0.8K**	**3.2K**	**0.8K**
Qwen VL Max	34.2	32.9	33.1	32.4
Minimax-01	34.3	33.5	35.6	36.3
Qwen VL Plus	25.4	26.1	24.9	22.9
Qwen2.5-VL-32B	23.0	23.4	31.8	34.4
Qwen2.5-VL-7B	23.8	23.3	24.2	25.1
GeoChat	20.9	20.6	–	–
Qwen2.5-VL-32B FT	57.2	57.1	54.1	54.0
Qwen2.5-VL-7B FT	47.5	48.0	45.7	46.1

Single-view ablations. To test whether questions can be answered from one view, we withhold the map (UAV-only) or the UAV image (map-only, Visual raster or Tabular), keeping the question and options unchanged. Table 9 shows that every model, including the fine-tuned one, stays near chance: no model exceeds it by more than 0.7 points. Refusals, which are scored as wrong, can push scores below chance.

Table 9: Single-view ablations, overall accuracy (%). UAV-only withholds the map; map-only withholds the UAV image. Proprietary models on Test-0.8K, Qwen on Test-3.2K. **Chance**: expected accuracy of uniform random guessing among the options, 22.8% overall (Table 2).

Model	**UAV-only**	**Map-only (Visual)**	**Map-only (Tabular)**
GPT-5	19.1	20.4	23.1
Claude Opus 4.7	21.0	20.5	18.1
Qwen2.5-VL-32B	21.7	23.5	19.8
Qwen2.5-VL-32B FT	21.7	22.5	23.2
Chance	22.8	22.8	22.8