BEYOND LOCAL REFINEMENT: WIDE-AREA PRECISE CROSS-VIEW LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

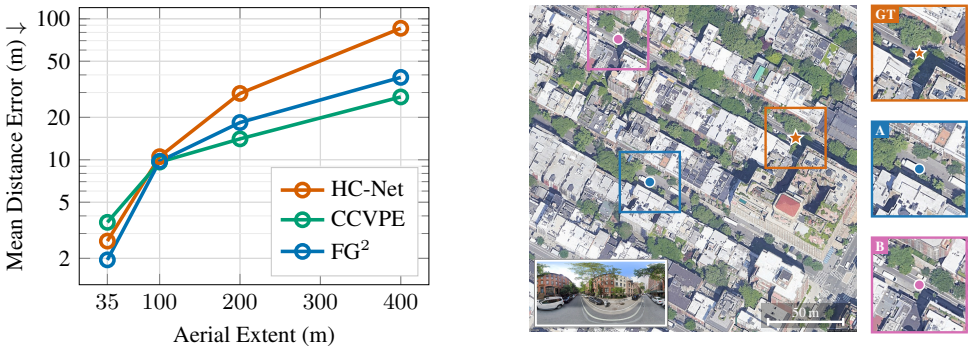
Fine-grained cross-view localization aims to precisely localize a ground camera against geo-referenced aerial imagery, but current methods typically assume a narrow geographic prior, substantially limiting their applicability when upstream localization remains coarse. To bridge this gap, we study this task over much larger aerial regions. However, this seemingly simple extension fundamentally changes the localization problem: the growing disparity in spatial extent between ground and aerial views makes reliable cross-view matching difficult to establish directly, while the increasing number of look-alike locations in larger regions introduces substantial visual ambiguity. To address these challenges, we propose RingMatch, a search-and-verification framework for wide-area cross-view localization. RingMatch first leverages predictive uncertainty to uncover plausible locations and progressively zooms into their surroundings to obtain increasingly fine-grained evidence, yielding a compact set of promising candidates. We then introduce the *depth ring*, a novel geometric representation that reveals subtle variations in the spatial layout around a location and serves as a view-shared spatial signature for an LVLM verifier to resolve residual ambiguity. To benchmark this setting, we construct VIGOR-Pro by composing VIGOR aerial tiles into contiguous regions with spatial extents of up to $400\,\mathrm{m}$. Experiments show that RingMatch maintains strong localization performance as the aerial region expands, consistently outperforming representative end-to-end and retrieval-based methods. Code is available at `https://anonymous.4open.science/r/ringmatch`.

1 INTRODUCTION

Determining where an image was captured from visual observations is a fundamental problem with broad applications in navigation, robotics, and public safety. At the global scale, image geo-localization provides a coarse location estimate for a ground image from geographic cues in the scene (Vivanco Cepeda et al., 2023; Li et al., 2025). To achieve finer localization, aerial imagery has been widely used as a reference owing to its broad coverage and rich scene details (Deuser et al., 2023). Within this paradigm, ongoing research has increasingly focused on the final stage of the overall pipeline, known as fine-grained cross-view localization (Zhu et al., 2021), which seeks to estimate the ground camera's precise metric position within a geo-referenced aerial image. By leveraging advances in cross-view representation learning (Xia et al., 2022; Fervers et al., 2023; Xia et al., 2024) and fine-grained spatial matching (Wang et al., 2023; Lentsch et al., 2023; Xia & Alahi, 2025), existing approaches have achieved notable improvements in metric localization accuracy.

Despite these advances, a critical gap remains before existing fine-grained cross-view localization methods can be reliably deployed in real-world applications. Specifically, current methods largely operate in a *local-refinement* regime, assuming that the camera has been narrowed down to a tightly bounded aerial region before estimating its precise position. On VIGOR (Zhu et al., 2021), a widely used benchmark, the commonly adopted setting restricts the camera to a prior area of only about 35×35 meters (Lentsch et al., 2023; Xia et al., 2024; Xia & Alahi, 2025). Such a narrow range requires the upstream localization system to reduce its uncertainty to a comparable scale. In practice, however, upstream localization from GPS/GNSS or image retrieval may remain coarse, and such a precise prior may not be available (Xia et al., 2022; Shi et al., 2023). This mismatch limits the integration of cross-view localization into practical visual localization systems.

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.



(a) Degradation over larger aerial regions.

(b) Ambiguity from look-alike locations.

Figure 1: (a) Existing fine-grained cross-view localization methods incur substantially larger errors as the aerial region expands from 35×35 m to 400×400 m. (b) A large aerial region can contain multiple look-alike locations compatible with the same ground observation (inset). In this example, locations • A and • B share similar appearance with the true location ($\star$). The main differences lie in subtle spatial details, such as building outlines and the arrangement of surrounding vegetation.

A natural way to accommodate coarser upstream localization priors is to enlarge the aerial region. However, this seemingly straightforward extension fundamentally changes the localization problem. As the aerial region expands, the ground observation covers only a small area within it, making it difficult to directly establish reliable cross-view matching. As shown in Figure 1a, expanding the aerial region can severely degrade existing methods. This growing spatial-extent disparity inherently shifts the task from local refinement to wide-area spatial search. Meanwhile, larger aerial regions often contain more recurring structures, creating multiple look-alike locations compatible with the same ground observation. Such locations often share similar appearance but differ in subtle details of their spatial layouts (see Figure 1b). Resolving this ambiguity requires recognizing and comparing these spatial cues across large viewpoint disparities, yet such fine-grained cross-view spatial reasoning remains challenging for current multimodal models (Jeong et al., 2026; Liu et al., 2026).

To address these challenges, we propose RingMatch, a search-and-verification framework for precise cross-view localization over large aerial regions. Rather than localizing directly over the full region, RingMatch first performs a *multi-hypothesis branch-and-zoom search* that explores the aerial region from coarse to fine. We use predictive uncertainty to expose distinct plausible hypotheses, retain them as separate search branches, and recursively zoom into their surroundings to access increasingly fine-grained evidence. By avoiding premature commitment under sparse cross-view correspondences, the search produces a compact shortlist of promising candidates.

Verification then focuses on resolving ambiguity among these look-alike candidates. To this end, we introduce the *depth ring*, a view-shared geometric representation that encodes camera-centered distances to surrounding scene surfaces across horizontal azimuths. Its azimuth-wise depth pattern captures subtle variations in the surrounding spatial layout, forming a distinctive location-specific signature that can be compared across ground and aerial views. We derive the ground ring from estimated depth and train an aerial ring predictor to reconstruct the corresponding ring from each candidate-centered aerial patch. A large vision-language model (LVLM) verifier then examines the overall ring patterns together with their visual context to select the final location.

To evaluate precise localization over large aerial regions, we construct VIGOR-Pro by automatically composing VIGOR aerial tiles into contiguous regions of approximately 100, 200, and 400 m. Across both same-area and cross-area evaluations, RingMatch consistently outperforms representative end-to-end and retrieval-based localization methods. At 400 m, RingMatch achieves mean localization errors of 12.36 m and 19.81 m on same-area and cross-area evaluation, respectively, compared with 15.53 m and 26.38 m for the strongest retrieval-based baseline. Beyond this, further analysis shows that our depth ring can serve as explicit spatial evidence for improving zero-shot cross-view spatial reasoning in frontier multimodal models such as GPT-5.6 Terra and Gemini-3.7 Flash, highlighting its potential for broader applications.

In summary, our main contributions are as follows:

- We study fine-grained cross-view localization beyond the conventional *local-refinement* regime, extending the task to substantially larger aerial regions. We further construct VIGOR-Pro to enable systematic evaluation under this setting.
- We propose a *multi-hypothesis branch-and-zoom search* that uses predictive uncertainty to preserve multiple plausible locations and progressively refine them, avoiding premature commitment under sparse cross-view correspondences.
- We introduce the *depth ring*, a view-shared geometric representation that captures the fine-grained spatial structure surrounding a location. It enables effective disambiguation among look-alike locations and serves as explicit spatial evidence for cross-view reasoning.

2 Problem Setup and Challenges

2.1 Problem Formulation

Following the standard problem formulation of fine-grained cross-view localization (Xia et al., 2022; 2024), we are given a ground panorama I_g and a geo-referenced aerial image I_a depicting an aerial region that contains the camera location. In practice, I_a can be obtained from an upstream localization system as a coarse geographic prior. Let Ω_a denote the image domain of I_a. Our goal is to estimate the camera position

$$\hat{\mathbf{p}} = (\hat{x}, \hat{y}) \in \Omega_a. \tag{1}$$

Unlike conventional fine-grained cross-view localization, we consider a broader aerial region to accommodate a coarser upstream localization prior. In this work, we consider the known-orientation setting, where the ground panorama has a known heading relative to the north-up aerial image, allowing us to isolate the challenges introduced by wide-area cross-view localization.

Since I_a is geo-referenced, let $\mathcal{G}_a : \Omega_a \to \mathbb{R}^2$ map each image coordinate to its corresponding geographic coordinate. The localization error can then be measured directly in meters:

$$e_{\text{geo}}(\hat{\mathbf{p}}, \mathbf{p}^\star) = d_{\text{geo}}(\mathcal{G}_a(\hat{\mathbf{p}}), \mathcal{G}_a(\mathbf{p}^\star)), \tag{2}$$

where $\mathbf{p}^\star \in \Omega_a$ denotes the ground-truth camera position and d_{geo} denotes the geographic distance.

2.2 What Challenges Does Wide-Area Cross-View Localization Pose?

Expanding the aerial region introduces challenges beyond simply enlarging the coordinate range. We identify two that are particularly critical.

Sparse ground–aerial correspondences. A ground observation captures only the local surroundings of the camera. As the aerial extent grows, the mismatch in spatial extent between the ground observation and the aerial region becomes increasingly pronounced, leaving only a small portion of the region with evidence that corresponds to the ground view. As shown in Figure 1a, expanding the aerial region can severely degrade existing end-to-end methods, particularly those relying on local cross-view alignment, including HC-Net (Wang et al., 2023) and FG^2 (Xia & Alahi, 2025). This degradation stems from the fact that reliable cross-view evidence exists only within an unknown local area of the aerial region, making direct localization over the full region inherently challenging. The task therefore shifts from local refinement to a wide-area spatial search problem.

Visual ambiguity in ground–aerial matching. When cross-view localization is extended to larger aerial regions, recurring scene layouts and visual patterns can give rise to multiple look-alike locations. As shown in Figure 1b, such look-alike locations can share similar semantic content and visual layouts, making them difficult to distinguish based on appearance alone. Resolving this ambiguity requires fine-grained cross-view spatial reasoning: the ability to recognize and compare subtle spatial differences under large viewpoint disparities, including spatial configuration, relative geometry, and structural organization of scene elements. Recent studies show that multimodal models struggle with spatial reasoning, particularly egocentric–allocentric transformations (Yang et al., 2025). These limitations become more pronounced across viewpoints, where models must establish spatial correspondence and preserve object and layout consistency (Jeong et al., 2026; Liu et al., 2026).

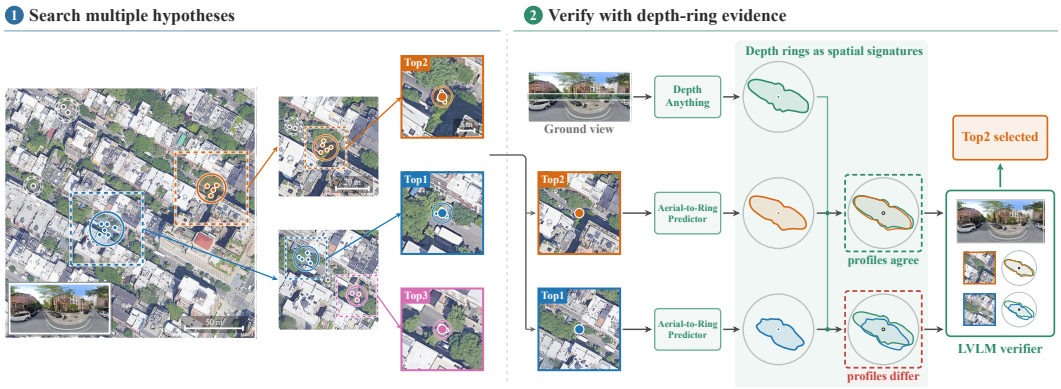


Figure 2: Overview of RingMatch. RingMatch first performs multi-hypothesis branch-and-zoom search over the aerial region, using predictive uncertainty to expose plausible hypotheses and recursively zoom into them to produce a compact candidate shortlist. It then resolves the remaining ambiguity with depth-ring-grounded verification. Ground observations and candidate aerial patches are mapped to depth rings that capture subtle details of the surrounding spatial layout. The correct candidate exhibits strong consistency with the ground observation, whereas a look-alike candidate does not. An LVLM verifier leverages this evidence to select the final location.

3 METHODOLOGY

3.1 OVERVIEW OF RINGMATCH

Section 2.2 leaves two coupled difficulties: cross-view correspondences over a large aerial region are sparse, and coarse visual evidence may point to multiple look-alike locations. Rather than aligning the ground view to the full aerial region in one step, we propose RingMatch, a search-and-verification framework that first explores the aerial region for plausible locations and then disambiguates them using fine-grained spatial cues, as illustrated in Figure 2. *Search* explores the aerial region from coarse to fine, retaining multiple plausible hypotheses and progressively zooming into their surroundings to produce a candidate shortlist. *Verification* then distinguishes the remaining look-alike candidates. Since raw appearance is difficult to compare reliably across the ground–aerial viewpoint gap, we represent each view with a camera-centered *depth ring*, a view-shared geometric representation that captures fine-grained spatial structure around a location. An LVLM verifier then leverages this evidence together with visual context to select the most consistent candidate.

3.2 MULTI-HYPOTHESIS BRANCH-AND-ZOOM SEARCH

Sparse cross-view correspondences make a single prediction unreliable over a large aerial region. We therefore maintain multiple plausible hypotheses rather than committing prematurely. We represent the search as a tree, where each node corresponds to a location hypothesis and its associated search region. Starting from the root covering the full aerial region, we use predictive uncertainty to branch each node into multiple child hypotheses and zoom into their surroundings. Repeating this branch-and-zoom process progressively refines plausible locations while preserving alternative search paths. The resulting leaf hypotheses are ranked to produce a shortlist for verification.

Uncertainty-guided branching. Let node v correspond to a search region $R_v \subseteq \Omega_a$, with $R_{\text{root}} = \Omega_a$. We feed the aerial crop $I_a^{R_v}$ and I_g to a localizer parameterized by θ, which predicts a camera coordinate $\mathbf{u} \in R_v$. To expose multiple plausible modes, we draw M stochastic predictions

$$\mathcal{U}_v = \{\mathbf{u}_{v,m}\}_{m=1}^{M}, \qquad \mathbf{u}_{v,m} \sim p_\theta(\mathbf{u} \mid I_g, I_a^{R_v}). \tag{3}$$

We then apply single-linkage clustering (Gower & Ross, 1969) to group nearby predictions. Each retained cluster $C_{v'} \subseteq \mathcal{U}_v$ defines a child node v', with its centroid as the corresponding hypothesis for a separate branch. Branching is therefore adaptive: concentrated predictions produce few branches, whereas multiple separated modes produce more. In this way, the search expands according to predictive uncertainty, allocating additional branches only when distinct hypotheses emerge.

Recursive zooming. For each retained hypothesis, we form a search region centered on it according to a predefined zoom schedule. Progressively narrowing the search region allows the localizer to access increasingly fine-grained evidence around each hypothesis. We repeat uncertainty-guided branching within each region, continuing this branch-and-zoom cycle until the final zoom level.

Hypothesis scoring and candidate shortlist. We score each hypothesis by aggregating the cluster sizes along its search path, based on the intuition that a promising hypothesis should remain well supported as the search progresses. For a hypothesis $\mathbf{h}$ at level ℓ with path $(v_0, v_1, \ldots, v_\ell)$, we define $S(\mathbf{h}) = \sum_{t=1}^{\ell} |C_{v_t}|$, where C_{v_t} is the cluster that produced node v_t. To bound the search cost, we rank hypotheses by S after each level and retain at most B for further expansion. At the final level, the top-k leaf hypotheses $\{\mathbf{h}_i\}_{i=1}^{k}$ form the candidate shortlist for verification.

3.3 Depth-Ring-Grounded Verification

Search returns a shortlist of k plausible locations, but residual ambiguity may remain among them. As noted in Section 2.2, distinguishing them requires fine-grained spatial reasoning across large viewpoint disparities. To resolve this ambiguity, we introduce the *depth ring*, a view-shared geometric representation that encodes the estimated distance from a location to surrounding scene surfaces across horizontal azimuths. By expressing fine-grained spatial structure from both views in a shared representation, depth rings enable verification based on geometric consistency rather than direct appearance comparison across the viewpoint gap.

Depth ring as a spatial signature of a place. For a location $\mathbf{p}$, we define its depth ring as the distance from $\mathbf{p}$ to the first scene surface encountered along each horizontal direction:

$$r_{\mathbf{p}}(\phi) = \inf\{\rho > 0 : \mathbf{p} + \rho\mathbf{d}(\phi) \text{ reaches a scene surface}\}, \qquad \phi \in [0, 2\pi), \tag{4}$$

where ϕ denotes the azimuth, and $\mathbf{d}(\phi) = (\cos\phi, \sin\phi)^\top$ is the corresponding unit vector in the horizontal plane. Sampling the ring at A azimuths gives $\mathbf{r}_{\mathbf{p}} = [r_{\mathbf{p}}(\phi_1), \ldots, r_{\mathbf{p}}(\phi_A)] \in \mathbb{R}^A$.

Although one-dimensional, the depth ring faithfully captures subtle variations in the surrounding spatial layout (see Figure 3). Its azimuth-wise depth pattern therefore forms a distinctive spatial signature of a location. As shown in Figure 2, the aerial ring at the true location closely agrees with the ground ring, whereas a look-alike location with a different surrounding layout produces an inconsistent profile. Since the depth ring is indexed by azimuth, cross-view comparison requires a common directional reference. Under the known-orientation setting, ground and aerial rings can be directly aligned; for unknown headings, the comparison can be extended by first estimating and compensating for the relative yaw.

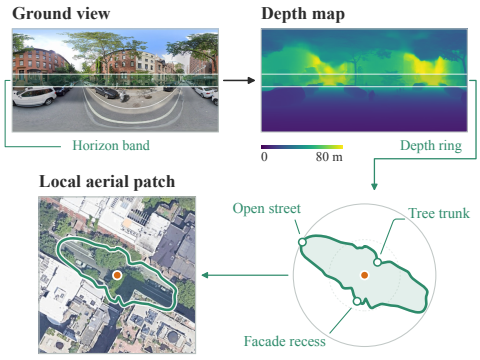


Figure 3: A depth ring summarizes azimuth-wise distances to surrounding scene surfaces in the horizontal plane. Its shape reveals subtle variations in the surrounding spatial layout, forming a compact spatial signature for cross-view comparison. The ring is derived from the ground observation and overlaid on the aerial patch only for visualization.

This representation is conceptually related to classical isovist-based descriptions of surrounding visibility (Benedikt, 1979) and LiDAR place descriptors that compactly encode surrounding geometry (Kim & Kim, 2018; Lu et al., 2022). In our setting, however, no such range scan is directly available, so the depth rings must be inferred from the ground and aerial observations.

Constructing rings from both views. *Ground-to-ring.* The ground panorama observes the surrounding scene, with its horizontal coordinates naturally corresponding to azimuths. We use a pretrained metric depth model (Yang et al., 2024) to estimate a depth map D_g from I_g. For each azimuth ϕ_a, we then pool depth values within a narrow vertical band $\mathcal{B}(\phi_a)$ around the horizon:

$$r_g(\phi_a) = \mathrm{Pool}\left\{D_g(x, y) \mid (x, y) \in \mathcal{B}(\phi_a)\right\}, \qquad a = 1, \ldots, A. \tag{5}$$

Aerial-to-ring. Unlike the ground panorama, the aerial view does not directly provide azimuth-wise depth around a location. We therefore learn an aerial-to-ring predictor f_φ that maps a local aerial patch $I_a^{\mathbf{P}}$ together with its specified location $\mathbf{p}$ to the corresponding azimuth-wise depth ring:

$$\hat{\mathbf{r}}_a(\mathbf{p}) = f_\varphi\left(I_a^{\mathbf{P}}, \mathbf{p}\right). \tag{6}$$

We implement f_φ as azimuth-wise metric depth estimation. Training requires no additional annotation: for each ground–aerial pair with a known camera position, we construct the ground ring using Eq. (5) and use it to supervise the aerial prediction at the corresponding location. Architectural and training details are provided in Appendix C.2.

LVLM-based candidate verification. Although depth rings provide explicit spatial evidence, direct numerical comparison can be brittle under local depth errors. We therefore use a learned LVLM verifier to compare overall ring patterns rather than relying on point-wise agreement. For each candidate $\mathbf{h}_i$, we align the ground ring $\mathbf{r}_g$ and predicted aerial ring $\hat{\mathbf{r}}_a(\mathbf{h}_i)$ in a shared polar coordinate system, and provide them together with the ground panorama and the corresponding candidate aerial patch to the verifier. It then selects the most plausible candidate from $\{\mathbf{h}_i\}_{i=1}^k$ as the final prediction.

3.4 Training Strategy for RingMatch

Multi-scale training for search. We use a single cross-view localizer p_θ across all levels of branch-and-zoom, requiring it to operate over aerial regions at different scales. We therefore aggregate ground–aerial training pairs across scales and jointly train the localizer. Leveraging the visual understanding and flexible generation capabilities of LVLMs, we implement p_θ within an LVLM framework rather than using conventional cross-view localization models (Xia et al., 2024; Xia & Alahi, 2025). Given a ground panorama and an aerial image, the model is trained to generate the camera coordinate in text form using the ground-truth coordinate as supervision.

Hard-candidate training for verifier. After training the search model, we run multi-hypothesis search on the ground–aerial training pairs and use the retained hypotheses as hard candidates for verifier training. The verifier is fine-tuned to output the selected candidate in text form, using the candidate closest to the ground-truth location as supervision. Because these candidates are drawn from the hypotheses retained after search, they naturally preserve the residual visual ambiguity, exposing the verifier to challenging cases similar to those encountered at inference.

4 Experiments

4.1 Setup

Dataset and evaluation protocol. Existing benchmarks for fine-grained cross-view localization are mainly designed for narrow geographic priors, where localization is performed within a relatively small aerial region. To evaluate precise localization over larger aerial regions, we construct VIGOR-Pro from VIGOR (Zhu et al., 2021) by composing neighboring aerial tiles into contiguous mosaics while preserving the original ground panoramas and camera locations. No additional imagery or manual annotation is introduced. We construct three scales with approximate side lengths of 100, 200, and 400 m, corresponding to increasingly coarse geographic priors. Following the standard VIGOR protocol, we evaluate under both *same-area* and *cross-area* settings. The former uses all cities for both training and testing, while the latter trains on two cities and evaluates on the remaining two unseen cities. All experiments use the known-orientation setting, and we report mean and median localization errors in meters following prior work (Xia & Alahi, 2025). Detailed construction procedures and dataset statistics are provided in Appendix B.

Implementation details. The search model is initialized from Qwen3-VL-4B-Instruct (Bai et al., 2025) and jointly trained on 50, 100, 200, and 400 m regions to support localization across varying aerial extents. The additional 50 m regions are generated from the 100 m training regions and used only for multi-scale training. During inference, we follow the $\{400, 200, 100, 50\}$ m zoom schedule starting from each evaluation scale. Unless specified otherwise, we draw $M = 16$ predictions per node at temperature 1 and use a pruning budget of $B = 4$. After the final search stage, we retain the top-2 hypotheses for verification. Ground depth rings are constructed with Depth Anything

Table 1: Main results on VIGOR-Pro under same-area and cross-area settings across increasingly large aerial regions. Mean and median localization errors are reported in meters. The best results are highlighted in bold, and the second-best results are underlined.

Area	Method	100 m		200 m		400 m	
		Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓
Same-area	*End-to-End*						
	HC-Net (Wang et al., 2023)	10.52	3.70	29.54	10.53	85.62	58.49
	CCVPE (Xia et al., 2024)	9.63	2.11	14.07	2.64	27.90	3.48
	FG^2 (Xia & Alahi, 2025)	9.76	4.63	18.38	8.14	38.41	17.00
	Qwen3-VL-4B	5.36	2.27	12.16	3.52	23.03	3.93
	Retrieval-Based						
	Sample4Geo (Deuser et al., 2023)	23.84	20.85	22.80	16.85	23.84	15.89
	Sample4Geo+FG^2	13.35	5.17	12.80	3.68	<u>15.53</u>	4.72
	CIPER (Jeon et al., 2026)	18.69	10.48	25.78	8.61	40.79	8.65
	Our Search-Based Framework						
	Single-Hypothesis	5.00	2.00	10.29	2.27	21.42	2.46
	RingMatch (Search-only)	<u>4.32</u>	**1.65**	<u>8.03</u>	<u>1.82</u>	15.60	<u>1.92</u>
	RingMatch	**4.02**	**1.65**	**6.57**	**1.79**	**12.36**	**1.88**
Cross-area	*End-to-End*						
	HC-Net (Wang et al., 2023)	14.02	5.10	44.12	18.49	121.71	110.38
	CCVPE (Xia et al., 2024)	19.29	3.58	41.65	4.67	94.18	7.69
	FG^2 (Xia & Alahi, 2025)	10.77	5.22	21.68	10.23	62.21	35.56
	Qwen3-VL-4B	7.12	3.87	17.25	5.45	34.64	6.16
	Retrieval-Based						
	Sample4Geo (Deuser et al., 2023)	25.02	21.63	27.78	18.01	34.80	17.25
	Sample4Geo+FG^2	14.00	4.86	17.66	3.36	26.38	4.78
	CIPER (Jeon et al., 2026)	21.57	12.46	40.41	13.85	80.84	22.59
	Our Search-Based Framework						
	Single-Hypothesis	5.79	2.48	12.90	2.29	29.19	2.40
	RingMatch (Search-only)	<u>4.95</u>	<u>1.95</u>	<u>11.01</u>	<u>1.99</u>	<u>23.99</u>	<u>2.09</u>
	RingMatch	**4.81**	**1.94**	**9.31**	**1.96**	**19.81**	**2.04**

V2 (Yang et al., 2024), and the aerial ring predictor is trained with ground-derived ring supervision. The verifier is initialized from the search model and further fine-tuned on hard candidates generated under the same search configuration. Further implementation details are provided in Appendix C.

Baselines. We compare RingMatch against representative end-to-end and retrieval-based localization methods, and further report controlled variants within our search-based framework. (1) *End-to-end localization* directly predicts the camera position from a ground–aerial pair. This category includes representative cross-view localization methods, such as HC-Net (Wang et al., 2023), CCVPE (Xia et al., 2024), and FG^2 (Xia & Alahi, 2025). We additionally report direct prediction results from our fine-tuned Qwen3-VL-4B search model. (2) *Retrieval-based localization* treats the original aerial tiles composing each VIGOR-Pro mosaic as the retrieval gallery. We adopt Sample4Geo (Deuser et al., 2023), a Sample4Geo+FG^2 cascade, and CIPER (Jeon et al., 2026). Sample4Geo uses the center of the top-1 retrieved tile as the prediction; Sample4Geo+FG^2 further refines the location within that tile using FG^2, a recent strong method for local refinement, while CIPER jointly performs retrieval and local pose estimation. (3) *Our search-based framework* progressively explores and zooms into the aerial region. Within this framework, we introduce Single-Hypothesis, an ablation variant that uses the same search model as RingMatch but retains only one hypothesis per level. Each existing end-to-end baseline is retrained independently at every VIGOR-Pro scale using its official configuration. For Sample4Geo and CIPER, we use the released checkpoints and follow their official inference procedures. Further details are provided in Appendix D.

4.2 Main Results

Localization over increasingly large aerial regions. Table 1 shows that conventional end-to-end localizers degrade sharply as the aerial region expands. On same-area evaluation, the mean error of CCVPE rises from 9.63 m at 100 m to 27.90 m at 400 m, with even larger increases for HC-Net and FG^2. In contrast, RingMatch achieves the lowest errors across all scales, reaching $12.36/1.88$ m mean/median error at 400 m, compared with $15.53/4.72$ m for the strongest retrieval-based baseline.

Table 3: Verifier ablation on VIGOR-Pro under same-area evaluation. Mean localization error (m) is reported on all test samples (ALL) and the verification subset (VFY). "Search init." denotes initialization from the trained search model.

	Verifier components			100 m		200 m		400 m	
	Local patch	Depth ring	Search init.	ALL	VFY	ALL	VFY	ALL	VFY
Search-only	–	–	–	4.32	22.59	8.03	41.64	15.60	70.58
w/ Verifier	✓	–	–	4.60	26.89	8.67	48.11	17.05	81.06
	✓	✓	–	<u>4.26</u>	<u>21.67</u>	<u>7.36</u>	<u>34.77</u>	<u>13.89</u>	<u>58.17</u>
	✓	✓	✓	**4.02**	**17.59**	**6.57**	**26.69**	**12.36**	**47.12**

Cross-area generalization. The advantage of RingMatch persists when generalizing to unseen geographic areas. At 400 m, RingMatch achieves mean/median errors of $19.81/2.04$ m, outperforming the strongest retrieval-based baseline Sample4Geo+FG^2 with $26.38/4.78$ m. This demonstrates the robustness of RingMatch to geographic distribution shifts.

Broader regions demand multi-hypothesis exploration. Compared with Single-Hypothesis, RingMatch (Search-only) reduces the same-area mean error by 13.6%, 22.0%, and 27.2% at 100, 200, and 400 m, respectively. The increasing gain shows that premature commitment becomes more costly as the aerial region expands, underscoring the need for multi-hypothesis exploration.

Depth-ring-grounded verification corrects costly mis-selections. Compared with search alone, verification further reduces the mean error at 400 m from 15.60 to 12.36 m under same-area evaluation and from 23.99 to 19.81 m under cross-area evaluation. In contrast, the corresponding median errors change only slightly. This suggests that verification primarily corrects costly mis-selections among plausible look-alike candidates rather than refining already accurate predictions.

4.3 Analysis

Depth rings enable stronger zero-shot cross-view spatial reasoning. Table 2 evaluates whether depth rings can help frontier multimodal models resolve visual ambiguity across ground–aerial views without task-specific fine-tuning. With visual evidence alone, GPT-5.6 Terra and Gemini-3.7 Flash remain limited in recognizing fine-grained spatial differences, with GPT-5.6 Terra performing only slightly above chance. Ring-only already achieves 67.29% and 69.42% accuracy at 200 and 400 m, showing that depth rings carry discriminative spatial information. Providing depth-ring evidence improves GPT-5.6 Terra by 10.65 and 6.74 points and Gemini-3.7 Flash by 3.22 and 2.73 points. Notably, Gemini with depth-ring evidence further surpasses Ring-only, suggesting that depth rings serve not only as geometric descriptors for direct cross-view matching, but also as explicit spatial evidence for improving cross-view reasoning.

Table 2: Verification accuracy (%) on VIGOR-Pro same-area verification samples. FT denotes task-specific fine-tuning. Ring-only selects the candidate with the higher Pearson correlation between its predicted aerial depth ring and the ground depth ring.

Method	FT	200 m	400 m
Random	–	50.00	50.00
Ring-only	–	67.29	69.42
GPT-5.6 Terra	–	56.93	59.47
w/ depth ring	–	**67.58** (+10.65)	**66.21** (+6.74)
Gemini-3.7 Flash	–	66.70	72.56
w/ depth ring	–	**69.92** (+3.22)	**75.29** (+2.73)
RingMatch	✓	81.92	82.81

Depth-ring evidence improves candidate verification. Table 3 isolates the verifier components. A verifier using visual patches alone fails to improve over search and can even degrade localization, confirming the difficulty of resolving look-alike candidates from appearance alone. Adding depth-ring evidence reduces the mean error on verification samples from 81.06 to 58.17 m at 400 m, while initialization from the trained search model further lowers it to 47.12 m, showing the complementarity of explicit spatial evidence and cross-view representations learned during search.

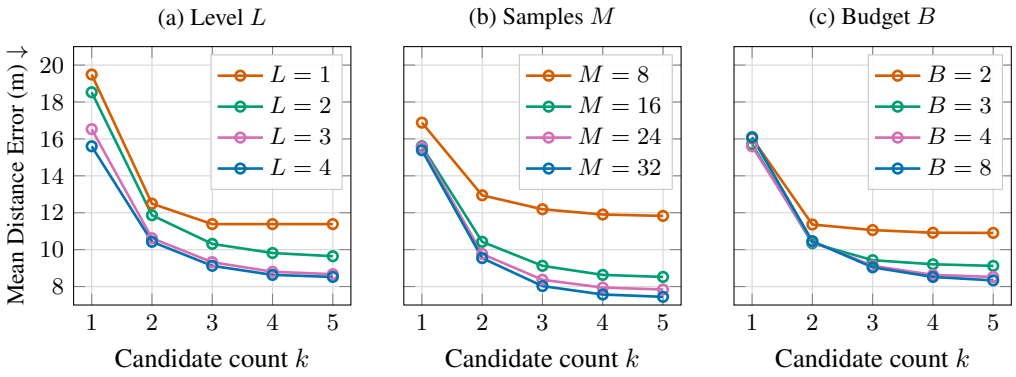


Figure 4: Analysis of multi-hypothesis search on VIGOR-Pro under same-area evaluation at the 400 m scale. We report top-k oracle mean localization error when varying (a) search level L, (b) sampling count M, and (c) pruning budget B, where the oracle selects the closest location among the top-k candidates. Unless otherwise varied, we use $L = 4$, $M = 16$, and $B = 4$.

Table 4: Generalization beyond VIGOR to four unseen cities at the 200 m scale.

Method	London		Tokyo		São Paulo		Cape Town		Overall	
	Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓	Mean ↓	Median ↓
FG^2	14.98	7.24	36.26	20.87	30.91	17.13	23.52	12.41	26.42	13.41
Sample4Geo+FG^2	9.47	2.22	24.38	4.72	19.09	3.92	14.14	2.82	16.77	3.20
RingMatch (Search-only)	9.65	1.23	21.64	1.04	19.95	1.49	9.57	1.02	15.20	1.19
RingMatch	**8.58**	**1.21**	**17.42**	**1.00**	**18.48**	**1.45**	**7.89**	**0.98**	**13.09**	**1.16**

Multi-hypothesis search preserves promising locations. To isolate search quality from verification, Figure 4 reports top-k oracle mean error on VIGOR-Pro under same-area evaluation at 400 m. With the default setting ($L = 4$, $M = 16$, $B = 4$), increasing k from 1 to 2 reduces the error from 15.60 to 10.43 m, while $k = 5$ further reaches 8.52 m. The second candidate captures 73.0% of the total gain from $k = 1$ to $k = 5$, supporting a compact top-2 shortlist for verification.

Sensitivity to search parameters. Search quality improves with deeper exploration and sufficient stochastic sampling. At $k = 2$, increasing L from 1 to 4 reduces the oracle error from 12.49 to 10.43 m, while increasing M from 8 to 16 reduces it from 12.94 to 10.43 m, with diminishing gains at larger M. Performance is comparatively insensitive to the pruning budget when enough branches are retained, with $B = 4$ and 8 giving similar results. More results are provided in Appendix G.

Generalization beyond VIGOR. To examine whether the observed gains extend beyond VIGOR, we further collect evaluation data from four cities outside the benchmark—London, Tokyo, São Paulo, and Cape Town—and evaluate at the 200 m scale. For all methods, we directly use models trained under the VIGOR same-area setting without target-city adaptation. As shown in Table 4, RingMatch consistently outperforms both baselines across all four cities, achieving 13.09/1.16 m mean/median error overall compared with 16.77/3.20 m for Sample4Geo+FG^2. These results show that the gains of RingMatch extend beyond the geographic and domain characteristics of VIGOR.

5 Conclusion

We study fine-grained cross-view localization beyond the conventional local-refinement regime to better accommodate the coarse localization priors. To address the sparse cross-view correspondences and visual ambiguity arising over larger aerial regions, RingMatch combines multi-hypothesis branch-and-zoom search with depth-ring-grounded verification, preserving plausible locations while exploiting fine-grained spatial structure for disambiguation. Experiments on VIGOR-Pro show consistent improvements across increasing aerial scales. Further analysis shows that depth rings provide explicit spatial evidence for cross-view reasoning in frontier multimodal models.

AI USE STATEMENT

In this work, we used generative AI tools for implementation-related assistance, including code editing, debugging, visualization implementation, and table formatting based on author-designed specifications. We also used generative AI tools to assist with manuscript preparation, including language editing, rewriting, translation, improving clarity, and refining the presentation of technical descriptions throughout the manuscript. Additionally, generative AI tools were used to assist with literature search and organizing relevant information during manuscript preparation.

We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, proposing or refining research hypotheses, designing the methodology or experimental protocols, cleaning or reformatting datasets, interpreting experimental results, or drawing scientific conclusions. The proposed methods, experimental settings, result interpretations, and scientific conclusions were developed and verified by the authors. The use of generative AI tools for formulating mathematical claims, writing proofs, and conducting qualitative or thematic data analysis is not applicable to this work.

All AI-assisted manuscript content was reviewed by the authors. AI-assisted code was inspected and tested for correctness, and literature identified with the assistance of AI tools was verified against the original sources. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide anonymized code, training scripts, and evaluation scripts at `https://anonymous.4open.science/r/ringmatch`. Detailed descriptions of the VIGOR-Pro construction, implementation details, training configurations, and evaluation protocols are provided in Appendices B–C.

REFERENCES

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

Michael L. Benedikt. To take hold of space: Isovists and isovist fields. *Environment and Planning B: Planning and Design*, 6(1):47–65, 1979.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16847–16856, 2023.

Yunus Talha Erzurumlu, Jiyong Kwag, and Alper Yilmaz. Just zoom in: Cross-view geo-localization via autoregressive zooming. *arXiv preprint arXiv:2603.25686*, 2026.

Ruijie Fan, Junyan Ye, Qi Zhu, and Weijia Li. One query, many scales: Sparse mixture-of-experts for efficient hierarchical cross-view geo-localization. *arXiv preprint arXiv:2608.01060*, 2026.

Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21621–21631, 2023.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64, 1969.

Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7258–7267, 2018.

Yurim Jeon, Dongseong Seo, and Seung-Woo Seo. CIPER: A unified framework for cross-view image-retrieval and pose-estimation. *arXiv preprint arXiv:2606.05011*, 2026.

Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, Dongyoon Hahm, Woogyeol Jin, Juheon Choi, and Kimin Lee. Learning multi-view spatial reasoning from cross-view relations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Pengyue Jia, Yiding Liu, Xiaopeng Li, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in neural information processing systems*, 37:53198–53221, 2024.

Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4802–4809, 2018. doi: 10.1109/IROS.2018.8593953.

Ted Lentsch, Zimin Xia, Holger Caesar, and Julian F. P. Kooij. SliceMatch: Geometry-guided aggregation for cross-view pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17225–17234, 2023.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 29222–29233, 2024.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 69058–69085, 2025.

Ruixun Liu, Lingyu Zhang, Lanxuan Xue, Kaiyu Li, Bowen Fu, and Xiangyong Cao. Cvsbench: A comprehensive benchmark for cross-view spatial reasoning and dreaming. In *European Conference on Computer Vision*, pp. 496–514. Springer, 2026.

Sha Lu, Xuecheng Xu, Huan Yin, Zexi Chen, Rong Xiong, and Yue Wang. One RING to rule them all: Radon sinogram for place recognition, orientation and translation estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2778–2785, 2022. doi: 10.1109/IROS47612.2022.9981308.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Juhye Park, Wooju Lee, Dasol Hong, Changki Sung, Youngwoo Seo, Dongwan Kang, and Hyun Myung. VIRD: View-invariant representation through dual-axis transformation for cross-view pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33693–33702, 2026.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, 2021.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17010–17020, 2022.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21459–21469. IEEE, 2023.

Zhuo Song, Ye Zhang, Kunhong Li, Longguang Wang, and Yulan Guo. A unified hierarchical framework for fine-grained cross-view geo-localization over large-scale scenarios. *arXiv preprint arXiv:2505.07622*, 2025.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 2621–2630, 2017.

Wei Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Siliang Tang, Jun Xiao, and Yueting Zhuang. CrossView Suite: Harnessing cross-view spatial intelligence of MLLMs with dataset, model and benchmark. *arXiv preprint arXiv:2605.18621*, 2026.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36:5301–5319, 2023.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M. de Melo, Jieneng Chen, and Alan L. Yuille. Spatial457: A diagnostic benchmark for 6D spatial reasoning of large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet – photo geolocation with convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.

Zimin Xia and Alexandre Alahi. FG^2: Fine-grained cross-view localization by fine-grained feature matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6362–6372, 2025.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision (ECCV)*, pp. 90–106, 2022. doi: 10.1007/978-3-031-19842-7_6.

Zimin Xia, Olaf Booij, and Julian F. P. Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(5):3813–3831, 2024. doi: 10.1109/TPAMI.2023.3346924.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-BEV co-retrieval network. In *European Conference on Computer Vision (ECCV)*, 2024.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3640–3649, 2021. doi: 10.1109/CVPR46437.2021.00364.

Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer is all you need for cross-view image geo-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1162–1171, 2022.

A Related Work

Image geo-localization and cross-view geo-localization. Visual geo-localization spans a broad range of spatial extents. Global methods infer geographic locations through classification, retrieval, or joint image–coordinate representations (Weyand et al., 2016; Vo et al., 2017; Vivanco Cepeda et al., 2023), while recent LVLM-based systems exploit semantic cues and geographic knowledge for reasoning-based image geo-localization (Li et al., 2024; Jia et al., 2024; Li et al., 2025). When geo-referenced aerial imagery is available, cross-view geo-localization matches ground observations against an aerial gallery to identify the corresponding geographic location (Hu et al., 2018; Shi et al., 2019; Zhu et al., 2022; Deuser et al., 2023; Ye et al., 2024). These methods provide broad geographic coverage and can serve as an upstream stage that narrows the localization prior.

Fine-grained cross-view localization. A complementary line of work focuses on precise metric localization within a local aerial reference. Existing methods predict continuous 2D positions or 3-DoF poses using dense pose distributions, geometric transformations, correspondence matching, or pose-conditioned features (Xia et al., 2022; Fervers et al., 2023; Lentsch et al., 2023; Wang et al., 2023; Xia et al., 2024; Xia & Alahi, 2025; Park et al., 2026). While these approaches achieve high metric accuracy, they largely operate in a local-refinement regime whose localization range is determined by the supplied aerial prior. On VIGOR (Zhu et al., 2021), commonly used protocols assume only a small neighborhood around the camera (Lentsch et al., 2023; Xia et al., 2024; Xia & Alahi, 2025). Our work instead studies precise cross-view localization when this upstream prior remains much broader.

Bridging global and fine-grained localization. Recent systems increasingly connect these two stages. UnifyGeo integrates retrieval, re-ranking, and metric localization, while CIPER jointly performs city-scale retrieval and local 3-DoF pose estimation (Song et al., 2025; Jeon et al., 2026). Just Zoom In and GeoMoE instead perform hierarchical coarse-to-fine geographic narrowing (Erzurumlu et al., 2026; Fan et al., 2026). Their retrieval and narrowing stages are complementary to ours and can provide the coarse geographic prior assumed by our setting. Although hierarchical narrowing is related to our progressive search, these hierarchical methods primarily navigate predefined discrete cells or gallery entries. In contrast, we directly localize continuous camera coordinates within a contiguous aerial region, using predictive uncertainty to retain multiple plausible hypotheses and progressively refine them. Our goal is therefore to extend precise cross-view localization to broader geographic priors rather than replace the upstream localization stage.

Cross-view geometry and spatial representations. Across ground–aerial cross-view tasks, geometry has long been used to bridge the viewpoint gap through polar transformations, bird's-eye-view projections, homography estimation, geometry-guided matching, and point–plane correspondence (Shi et al., 2019; Shi & Li, 2022; Lentsch et al., 2023; Wang et al., 2023; Ye et al., 2024; Xia & Alahi, 2025; Park et al., 2026). These designs support cross-view matching or pose estimation by making the two views more geometrically compatible. Our depth ring instead provides a view-shared geometric interface that brings spatial structure from both views into a common comparison space, exposing subtle layout differences among look-alike locations. It is conceptually related to vantage-point representations such as isovists and LiDAR place descriptors such as Scan Context and RING (Benedikt, 1979; Kim & Kim, 2018; Lu et al., 2022). Unlike LiDAR descriptors, our image-only setting provides no direct range measurements. We therefore derive ground rings from estimated depth and learn to predict corresponding aerial rings from overhead imagery using ground-derived supervision.

LVLMs for geolocation and spatial reasoning. LVLMs have recently been explored for image geo-localization through semantic and geographic reasoning (Li et al., 2024; Jia et al., 2024; Li et al., 2025). However, fine-grained spatial reasoning remains challenging for current multimodal models (Wang et al., 2025; Yang et al., 2025). Recent cross-view studies further highlight challenges in establishing correspondence and maintaining spatial consistency across viewpoints (Jeong et al., 2026; Wang et al., 2026; Liu et al., 2026). Rather than relying on LVLMs to infer cross-view geometry directly from raw imagery, we use their predictive uncertainty for multi-hypothesis search and provide explicit depth-ring evidence for resolving look-alike locations.

B VIGOR-PRO DATASET CONSTRUCTION

VIGOR-Pro turns an existing VIGOR panorama–aerial pair into a large-area localization example. The construction changes only the aerial context and coordinate frame, while preserving the original panorama and camera location. It introduces neither newly collected imagery nor manual location labels. Figure 5 illustrates the overall construction process.

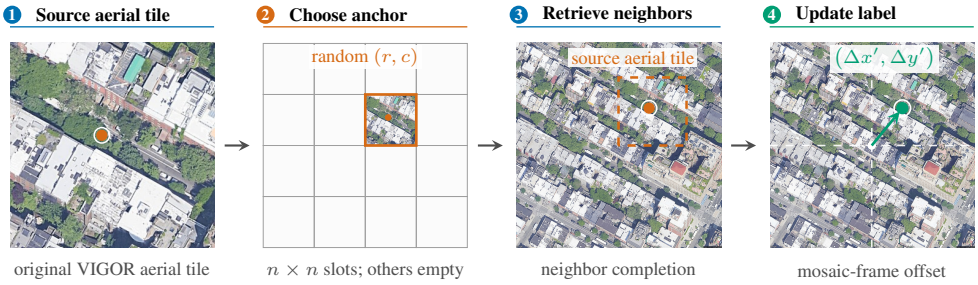


Figure 5: VIGOR-Pro construction retains the original camera location while enlarging its aerial context. The source tile is assigned to a sampled lattice cell, geographic neighbors complete the overlapping mosaic, and the camera offset is transformed into the mosaic frame. The empty grid in Step 2 denotes logical tile slots; completed tiles overlap by half their width.

B.1 CONSTRUCTION PROTOCOL

Tile lattice and mosaic geometry. VIGOR aerial tiles are 640×640 pixels with 50% overlap, hence a 320-pixel lattice stride. An $n \times n$ mosaic has native side length

$$L_n = 640 + 320(n-1). \tag{7}$$

Tiles are pasted in a deterministic order at this stride. In overlap regions, later tiles overwrite earlier pixels; no interpolation or blending is applied. Thus every output pixel is copied from an original VIGOR tile.

Geographic retrieval and rejection. Each VIGOR query is associated with one positive and three semi-positive aerial tiles. We independently use each of these four tiles as a source anchor. For a query–source pair, the source tile is placed at a uniformly sampled lattice cell $(r, c) \in \{0, \dots, n-1\}^2$. Expected centers for the remaining cells are obtained by stepping on the city-specific VIGOR lattice. A spatial index returns the nearest tile, which is accepted only when the geodesic distance between its center and the expected center is below $10\,\mathrm{m}$. A mosaic is valid only if all n^2 cells pass this test. The query is retained when at least one of its four source anchors yields a complete mosaic.

Coordinate transformation. Let $(\Delta x, \Delta y)$ denote the camera offset from the native aerial tile center, and let $q = (n-1)/2$. With image x increasing eastward and image y increasing southward, the offset in the native mosaic frame is

$$\Delta x' = \Delta x + 320(q - c), \qquad \Delta y' = \Delta y + 320(r - q). \tag{8}$$

We express the target in an image-size-independent $[0, 1000]^2$ coordinate system as

$$x = 1000\left(\frac{1}{2} - \frac{\Delta x'}{L_n}\right), \qquad y = 1000\left(\frac{1}{2} + \frac{\Delta y'}{L_n}\right). \tag{9}$$

The inverse used by evaluation is

$$\Delta x' = L_n\left(\frac{1}{2} - \frac{x}{1000}\right), \qquad \Delta y' = L_n\left(\frac{y}{1000} - \frac{1}{2}\right). \tag{10}$$

Equations (9) and (10) form an exact round trip before decimal rounding. The center of a mosaic maps to $(500, 500)$. Algorithm 1 summarizes the complete construction procedure, including geographic retrieval, mosaic validation, stitching, and coordinate transformation.

Algorithm 1: Construct one query's VIGOR-Pro examples

Input: panorama I_g; source tuples $\{(T_j, \Delta x_j, \Delta y_j)\}_{j=1}^{4}$; grid size n.
Output: zero or more complete mosaic–label pairs.
for source anchor $j = 1, \ldots, 4$ **do**
  Sample anchor cell (r_j, c_j) uniformly.
  For every cell, compute its expected geographic center and retrieve the nearest city tile.
  Reject anchor j if any center discrepancy is at least 10 m.
  Otherwise stitch at stride 320 and transform the label with Eqs. (8)–(9).
end for
Return all valid pairs; discard I_g only if none remain.

B.2 Benchmark Scales

We instantiate $n \in \{2, 5, 10\}$, yielding native mosaic resolutions $L_n \in \{960, 1920, 3520\}$ pixels. Using the city-specific ground sampling distance (GSD) reported by SliceMatch (Lentsch et al., 2023), the mean side lengths are approximately 106 m, 213 m, and 390 m, which we refer to as the 100, 200, and 400 m scales, respectively. The city-specific GSD values and corresponding physical side lengths are reported in Table 5, and representative examples are shown in Figure 8.

Dataset split. The benchmark inherits the official VIGOR query split. Neighbor retrieval is performed only within the query city and does not alter the original training/test city partition. All mosaics originating from the same panorama remain in the same split.

Dataset statistics. Table 6 reports statistics after complete-mosaic filtering. *Original panoramas* counts ground panoramas in the source VIGOR annotations, and *original pairs* counts their associated panorama–tile pairs before mosaic construction. Because each panorama has four source tiles, the 105,214 original panoramas yield 420,856 original pairs at every scale. *Complete mosaics* counts original pairs for which all n^2 aerial tiles are recovered. The completion rate is the number of complete mosaics divided by the number of original pairs. *Retained panoramas* counts unique ground panoramas with at least one complete mosaic.

Table 5: Aerial-region side lengths induced by the source GSD and native mosaic geometry. Values are side lengths in meters.

City	GSD (m/pixel)	100 m	200 m	400 m
New York	0.113248	108.72	217.44	398.63
Seattle	0.100817	96.78	193.57	354.88
San Francisco	0.118141	113.42	226.83	415.86
Chicago	0.111262	106.81	213.62	391.64
Mean		106.43	212.87	390.25

Table 6: VIGOR-Pro geometry and construction statistics after complete-mosaic filtering.

Scale	Grid $n \times n$	Native resolution	Mean side (m)	Original panoramas	Original pairs	Complete mosaics	Completion rate (%)	Retained panoramas
100 m	2×2	960^2	106	105,214	420,856	417,375	99.17	105,214
200 m	5×5	1920^2	213	105,214	420,856	400,740	95.22	104,098
400 m	10×10	3520^2	390	105,214	420,856	372,532	88.52	101,053

C Implementation Details

This section describes the training and inference details of the multi-scale search model, depth-ring predictor, and learned verifier, including their supervision, inputs, and optimization settings.

C.1 SEARCH MODEL: TRAINING AND INFERENCE

Supervision format. Each supervised example contains exactly two images in a fixed order—the ground panorama followed by the aerial region—and one user/assistant turn. The template below gives both roles; the two image inputs are inserted in this order.

User

```
<image><image>
Given a ground-level panoramic image and a satellite image, please
predict the coordinates [x, y] of the ground-level camera location in
the satellite image.
```

Assistant

```
[x, y]
```

The assistant target is the bare string `[x, y]`, with each value rounded to one decimal place in the normalized coordinate system of Eq. (9); no explanation or auxiliary token is appended. Keeping the same coordinate grammar at every scale permits one model to operate throughout the tree.

Multi-scale training set. Separate search models are trained for same-area and cross-area evaluation using their respective training splits. The training mixture contains direct VIGOR-Pro examples with approximately 100, 200, and 400 m coverage. We additionally introduce a 50 m scale to provide finer-grained supervision for the search stage. These examples are derived by cropping $1/2$-scale regions from the 100 m training examples. To simulate realistic search refinement, the crop center is displaced from the ground-truth location using independent zero-mean Gaussian jitter with a standard deviation equal to 0.1 of the crop half-size. The jitter is truncated to ensure that the ground-truth location remains inside the crop whenever the parent label lies within the image. At an image boundary, the crop is translated inward rather than reduced in size. Let (a_x, a_y, b_x, b_y) denote the crop box in the normalized parent-image coordinates, and (x, y) the ground-truth camera location. The in-crop label is $x' = 1000(x - a_x)/(b_x - a_x)$ and $y' = 1000(y - a_y)/(b_y - a_y)$.

Optimization. We initialize from Qwen3-VL-4B-Instruct. During training, only the language model parameters are updated, while the vision tower and multimodal projector remain frozen. All models are trained on 8× NVIDIA A100 (80GB) GPUs. Table 7 summarizes the optimization configuration.

Table 7: Search-model training configuration.

Configuration	Value	Configuration	Value
Initialization	Qwen3-VL-4B-Instruct	Optimizer	AdamW
Trainable	Language model	Batch size	256
Frozen	Vision, projector	Learning rate	1×10^{-5}
Training scales	50/100/200/400 m	Model input	$480^2/512^2/512^2/768^2$ px
Training epochs	2	LR schedule	Cosine
Warmup ratio	0.1		

Branch-and-zoom inference. Table 8 summarizes the branch-and-zoom schedule used at each evaluation scale. At each node, we draw M textual coordinate samples. The coordinates are clustered by single linkage in the normalized crop-local coordinate system, using a Euclidean distance threshold of 50. Clusters with fewer than two members are discarded, and at most the three largest clusters are retained. Each retained cluster is represented by its centroid, which defines a child hypothesis. After expanding all active nodes at each level, we rank the resulting hypotheses using the path-aggregated cluster-size score defined in Section 3.2 and retain at most B for further expansion. For each retained hypothesis, we form the corresponding child search region and extract its aerial crop directly from the original mosaic. If the crop would exceed an image boundary, the search region is translated inward without changing its size. To express predictions from different search

Table 8: Branch-and-zoom inference schedule. The search-region schedule denotes the approximate side length of the aerial region presented at each level.

Initial scale	Search-region schedule (m)	Model-input size (px)
100 m	$100 \rightarrow 50$	$512 \rightarrow 480$
200 m	$200 \rightarrow 100 \rightarrow 50$	$512 \rightarrow 512 \rightarrow 480$
400 m	$400 \rightarrow 200 \rightarrow 100 \rightarrow 50$	$768 \rightarrow 512 \rightarrow 512 \rightarrow 480$

regions in a common coordinate frame, we record each region as (a_x, a_y, b_x, b_y) in the normalized root-mosaic coordinate system. A local prediction $\mathbf{u} = (u_x, u_y)$ is then mapped back to the root frame as

$$\mathcal{T}(\mathbf{u}) = \left(a_x + \frac{u_x}{1000}(b_x - a_x), a_y + \frac{u_y}{1000}(b_y - a_y)\right). \tag{11}$$

After the final level, the leaf hypotheses are ranked by the same score, and the top-k are retained for verification.

C.2 Depth Ring: Construction, Learning, and Inference

Ground panorama to depth ring. VIGOR panoramas are oriented with the center facing north. We sample the depth ring at A uniformly spaced clockwise azimuths $\phi_a = 360^\circ(a-1)/A$, starting from the leftmost panorama direction as $\phi = 0^\circ$. Since Depth Anything V2 is designed for perspective images, we first apply equirectangular-to-perspective (E2P) projection to render four 512×512 rectilinear views with view directions $\psi_j \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, pitch 0°, and horizontal field of view 90°. We then apply Depth Anything V2 Metric Outdoor Large to each view independently. Eq. (5) is implemented by mapping each perspective-view column back to its panorama azimuth. For a view of width W_f and focal length $f = W_f/[2\tan(45^\circ)]$, column x of view j corresponds to

$$\phi_j(x) = \left[\psi_j + \tan^{-1}\left(\frac{x - W_f/2}{f}\right)\right] \bmod 360^\circ. \tag{12}$$

For each target azimuth ϕ_a, we pool the estimated metric depth values in the corresponding horizon band to obtain $r_g(\phi_a)$ as in Eq. (5). For robustness to isolated depth artifacts, each perspective-view column is represented by the median estimated depth within a ± 8-row band around the horizontal center. After mapping the samples from all four views to panorama azimuths using Eq. (12), we aggregate them into the A uniformly spaced azimuths of the depth ring, averaging samples assigned to the same azimuth. The Metric Outdoor Large model uses an 80 m depth range, preserving a common nominal metric scale across panoramas. Figure 6 illustrates the conversion from the panorama to the four perspective views and the resulting depth ring construction.

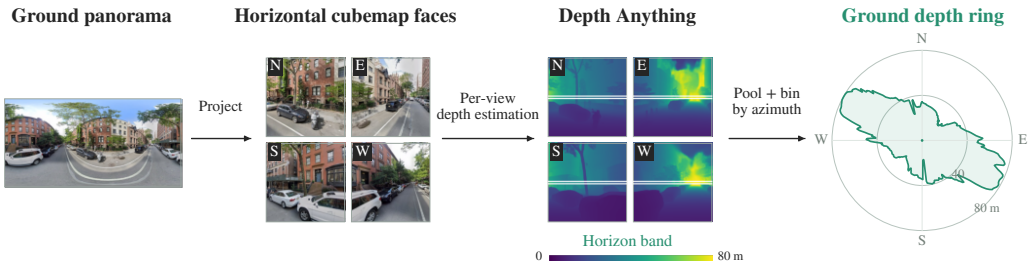


Figure 6: Ground depth-ring construction. The panorama is projected into four rectilinear views using equirectangular-to-perspective (E2P) projection. Depth Anything V2 Metric Outdoor Large estimates depth for each view, from which we pool the highlighted horizon band and map the resulting samples to their world azimuths. Samples from all views are then aggregated into the A uniformly spaced azimuths of the ground depth ring.

Training recipe of the aerial-to-ring predictor. Figure 7 illustrates the architecture of the aerial-to-ring predictor. It uses a frozen DINOv2-Base encoder (Oquab et al., 2023) with multi-level

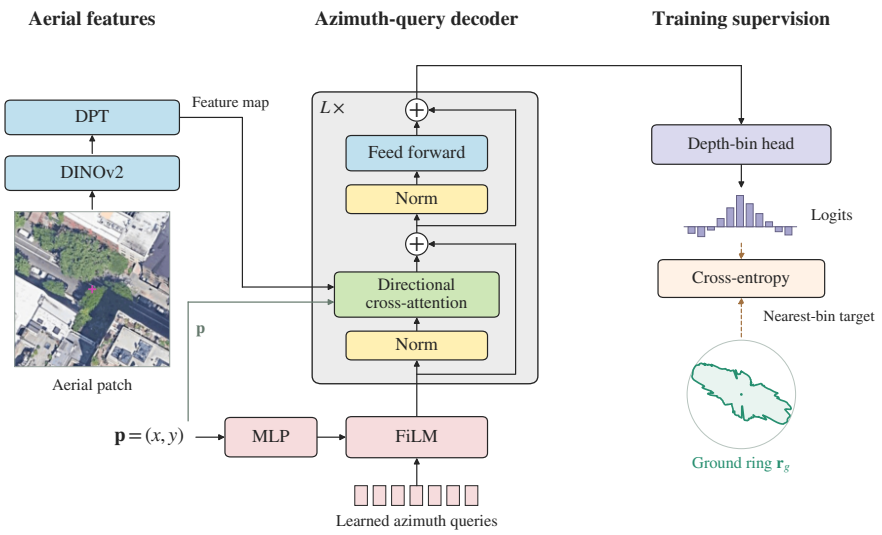


Figure 7: Architecture of the aerial-to-ring predictor. DINOv2 and DPT encode the local aerial patch into a dense feature map. The candidate location $\mathbf{p}$ conditions the A azimuth queries through FiLM modulation. For each azimuth, a directional beam is sampled from $\mathbf{p}$, and the corresponding query attends only to features along that beam. Each updated query predicts a metric-depth class, producing the aerial depth ring over all azimuths. During training, the predictor is supervised by the ground depth ring extracted from the corresponding panorama.

Table 9: Aerial-to-ring model and training configuration.

Configuration	Value	Configuration	Value
Model			
Aerial patch	640^2	FFN width	1,024
Azimuth samples A	360	Depth classes K	80
Metric depth range	$[0, 80]$ m	Backbone	frozen DINOv2-B
DPT output width	256	Cross-attn. layers / heads	4 / 8
Training			
Training loss	azimuth-wise CE	Optimizer	AdamW
Batch size	64	Learning rate	1×10^{-4}
Weight decay	0.05	Warmup	500 steps
LR schedule	cosine	Training epochs	1

features from transformer blocks 3, 6, 9, and 12, which are fused by a DPT decoder (Ranftl et al., 2021) into a dense aerial feature map. It maintains A learned azimuth queries. The candidate coordinate $\mathbf{p}$, represented in the normalized local-patch coordinate system, is encoded by an MLP and used to condition these queries through feature-wise linear modulation (FiLM) (Perez et al., 2018), while its mapped feature-map location serves as the origin for directional sampling. For each azimuth ϕ_a, we sample a thin directional beam from the aerial feature map starting at $\mathbf{p}$ to exclude irrelevant context outside that direction. Each azimuth query is then updated by four cross-attention blocks using only the features sampled along its corresponding beam. The prediction head maps each query to logits $z_{a,k}$ over K metric-depth classes corresponding to uniformly spaced depth values in $[0, 80]$ m. Each ground-ring value is quantized to its nearest metric-depth class and used as the cross-entropy target. The predictor is supervised by the ground depth ring constructed above and trained with azimuth-wise cross-entropy.

Training uses the VIGOR-Pro training sets at the 100, 200, and 400 m scales, with separate predictors trained for same-area and cross-area evaluation using their respective training splits. Each panorama–aerial pair provides one example. From each aerial mosaic, we extract a 640×640 local

patch around the annotated camera location. At image boundaries, the crop is shifted inward while preserving its size, and the camera coordinate is updated accordingly.

At inference, the same aerial preprocessing is applied to each search hypothesis. For each azimuth, we select the highest-scoring depth class and use its corresponding depth value as the prediction.

Table 9 summarizes the training and model configuration of the aerial-to-ring predictor.

C.3 Verifier: Hard-Candidate Training and Inference

Hard-candidate training set. Separate verifiers are trained for same-area and cross-area evaluation using their respective training splits. Hard candidates from the 100, 200, and 400 m scales are combined for verifier training. For each ground–aerial pair in the training split, we run the same branch-and-zoom procedure used at test time, with $M = 16$ stochastic predictions per node, sampling temperature 1, pruning budget $B = 4$, and the corresponding zoom schedule. We then take the two highest-scoring hypotheses at the final search level. We discard examples for which fewer than two valid hypotheses are available or the two candidates are separated by less than 10 m. The latter removes near-duplicate alternatives that provide little useful selection supervision. For each retained pair, the two hypotheses are randomly assigned to Candidate A and Candidate B, preventing the verifier from exploiting candidate order. The candidate with the smaller Euclidean distance to the ground-truth camera position is used as the supervision target. Importantly, the competing candidate is generated by the search model itself rather than independently sampled, exposing the verifier to plausible hard alternatives that remain after search.

Multimodal input and supervision. Each verifier example contains six images in a fixed order: (1) the ground panorama; (2) the full aerial region with Candidates A and B marked; (3–4) local aerial patches around Candidates A and B; and (5–6) candidate-specific depth-ring comparisons. For each candidate, the ground depth ring and the predicted aerial ring are rendered together in a common polar representation. The two candidate-specific comparisons use the same radial scale, enabling direct comparison of their geometric consistency. The candidate coordinates in the root aerial region are additionally provided in the normalized coordinate system. The supervision target contains only `\boxed{A}` or `\boxed{B}`, without an auxiliary rationale or coordinate-regression target. We provide the complete prompt in Figure 13.

Optimization and inference. We initialize the verifier from the trained search model, retaining the cross-view representations learned during multi-scale search training. As in search-model training, the vision tower and multimodal projector are frozen, and only the language-model parameters are updated. All models are trained on 8× NVIDIA A100 (80GB) GPUs. Table 10 summarizes the optimization configuration.

At inference, branch-and-zoom search first returns the two highest-scoring final hypotheses. We apply verification only when the two candidates are at least 10 m apart; otherwise, the top-ranked search hypothesis is retained as the final prediction. For verified samples, we construct the corresponding verifier input. The verifier performs greedy decoding with temperature 0 and selects Candidate A or B. The selected hypothesis, expressed in the root aerial coordinate frame, is returned directly as the final localization result.

Table 10: Verifier training configuration.

Configuration	Value	Configuration	Value
Initialization	Search model	Training epochs	2
Trainable	Language model	Optimizer	AdamW
Frozen	Vision, projector	Learning rate	1×10^{-5}
Batch size	256	LR schedule	Cosine
Warmup ratio	0.1	Candidate shortlist	2
Candidate crop	640×640		

D Baseline Adaptation and Evaluation Protocols

Common evaluation protocol. For each VIGOR-Pro scale and split, all methods are evaluated on the same retained examples and receive the same underlying geographic prior defined by the corresponding aerial mosaic. End-to-end and search-based methods operate on the complete contiguous mosaic, whereas retrieval-based methods search over the original VIGOR tiles composing the same region. All predictions are mapped to the native mosaic coordinate frame before evaluation and converted to metric errors using the city-specific GSD in Table 5.

End-to-end localization. HC-Net (Wang et al., 2023), CCVPE (Xia et al., 2024), and FG^2 (Xia & Alahi, 2025) are retrained independently at the 100, 200, and 400 m scales of VIGOR-Pro. At each scale, the complete aerial mosaic is provided as input and resized according to the preprocessing of the corresponding official implementation. Camera labels and model predictions are transformed consistently between the resized input and the native mosaic coordinate frame. Apart from these scale-dependent adaptations, we follow the official architectures, objectives, training recipes, and inference settings. For Qwen3-VL-4B, we directly evaluate our fine-tuned search model by using its single-pass prediction without applying the search procedure.

Retrieval-based localization. For each VIGOR-Pro example, the retrieval gallery consists only of the original VIGOR tiles composing the corresponding mosaic, rather than the full VIGOR gallery. This preserves the native tile scale used by the retrieval models while restricting the search to the same geographic region available to the other localization paradigms. For Sample4Geo (Deuser et al., 2023), we use the officially released checkpoint and inference configuration, taking the center of the top-ranked tile as the localization prediction. For Sample4Geo+FG^2, the same retrieved tile is further processed by an FG^2 model fine-tuned on the original VIGOR dataset for within-tile localization, and the resulting prediction is mapped to the native mosaic coordinate frame according to the tile position. For CIPER (Jeon et al., 2026), we likewise use the officially released checkpoint and inference configuration on the same gallery, and map the pose prediction associated with the selected aerial tile to the native mosaic coordinate frame.

Our search-based framework. Single-Hypothesis uses the same trained search model and inference configuration as RingMatch, but retains only the highest-scoring hypothesis at each search level for subsequent expansion. We also report RingMatch (Search-only), which follows the same search procedure but directly selects the top-ranked hypotheses without applying verification.

E Experimental Details for Additional Analyses

E.1 Zero-Shot Cross-View Spatial Reasoning

We evaluate zero-shot cross-view spatial reasoning on the VIGOR-Pro same-area verification samples at the 200 and 400 m scales. Following the verification protocol in Appendix C.3, the two highest-scoring final search hypotheses form Candidates A and B, and samples with candidate separation below 10 m are excluded. The candidate closer to the ground-truth camera location is treated as the correct answer. We evaluate GPT-5.6 Terra and Gemini-3.7 Flash without task-specific fine-tuning. For each model, we compare two input settings on the same candidate pairs. In the visual-only setting, the model receives the ground panorama, the full aerial region with both candidates marked, and a local aerial patch centered on each candidate, using the prompt in Figure 14. The depth-ring setting additionally provides the ground ring paired with the predicted aerial ring for each candidate, together with the corresponding instructions in Figure 13. We report candidate selection accuracy. As a direct geometric baseline, Ring-only selects the candidate whose predicted aerial ring has the higher Pearson correlation with the ground ring. We also report the verifier in RingMatch, which is fine-tuned for this task, on the same samples for reference.

E.2 Generalization Beyond VIGOR

Dataset construction. We construct an additional evaluation set of 4,000 panorama–aerial pairs from four cities outside the VIGOR benchmark: London, Tokyo, São Paulo, and Cape Town, with 1,000 examples per city. Ground panoramas are collected from Google Street View, sampled at

least 75 m apart, and converted to equirectangular images with the central viewing direction pointing north. Aerial imagery is collected through the Google Maps Static API. Following the same 5×5 mosaic geometry used at the 200 m VIGOR-Pro scale, each aerial region is composed of 25 overlapping 640×640 satellite images queried at zoom level 20 with 50% overlap, yielding a 1920×1920 mosaic. To avoid center bias, the camera-containing tile is evenly distributed across the 25 grid positions, with 40 examples per position in each city. No examples from these cities are used for task-specific training.

Evaluation protocol. We evaluate FG^2, Sample4Geo+FG^2 and RingMatch without target-city fine-tuning or adaptation, using models trained under the VIGOR same-area setting. Localization errors are measured in the native mosaic coordinate system and converted to meters using the corresponding city-specific ground sampling distances, which are summarized in Table 11.

Table 11: Aerial-region side lengths induced by the city-specific GSD and the native mosaic geometry.

City	GSD (m/pixel)	Side length (m)
London	0.092883	178.33
Tokyo	0.121224	232.75
São Paulo	0.136834	262.72
Cape Town	0.123851	237.79
Mean		227.90

F Qualitative Analysis of Depth Rings

Figure 9 provides qualitative examples of depth rings for cross-view matching. Across diverse scene layouts, aerial rings predicted near the ground-truth camera location preserve the characteristic spatial structure of the corresponding ground ring, including its dominant directions, relative extents, and local shape variations. In contrast, rings predicted at incorrect locations exhibit clear structural discrepancies, even when the candidate aerial regions contain similar roads or surrounding buildings. These examples illustrate how the depth ring maps ground and aerial observations into a shared representation of surrounding spatial layout, providing discriminative geometric evidence beyond direct cross-view appearance matching.

G Extended Analysis of Multi-Hypothesis Search

Figures 10–12 extend the search analysis in Figure 4 to all VIGOR-Pro scales under both same-area and cross-area evaluation.

Search level. Increasing the number of search levels becomes increasingly important as the aerial region expands. At 400 m and $k = 2$, increasing L from 1 to 4 reduces the mean oracle error from 12.49 to 10.43 m on same-area and from 19.26 to 15.64 m on cross-area; the corresponding median errors decrease from 3.05 to 1.76 m and from 4.68 to 1.99 m. The same trend holds at smaller scales, but the gains become more pronounced over broader regions, supporting the need for iterative zooming when the initial localization prior is coarse.

Sampling count. Increasing M provides the largest gain from 8 to 16, with further sampling yielding diminishing improvements. At 400 m and $k = 2$, the same-area mean error decreases from 12.94 to 10.43 m and further to 9.54 m at $M = 32$, while cross-area error decreases from 18.29 to 15.64 m and 15.17 m, respectively. Median errors change much less and are not consistently improved by larger M, suggesting that additional samples primarily help recover difficult cases by exposing alternative plausible modes.

Pruning budget. A small budget of $B = 2$ degrades search quality, particularly at larger scales, whereas $B = 4$ and 8 yield similar top-2 performance across both evaluation settings. Larger budgets become more beneficial mainly when considering deeper candidate lists. These results indicate that once enough competing branches are retained, search is comparatively insensitive to the exact pruning budget, supporting our default choice of $B = 4$ for the compact top-2 shortlist used by the verifier.

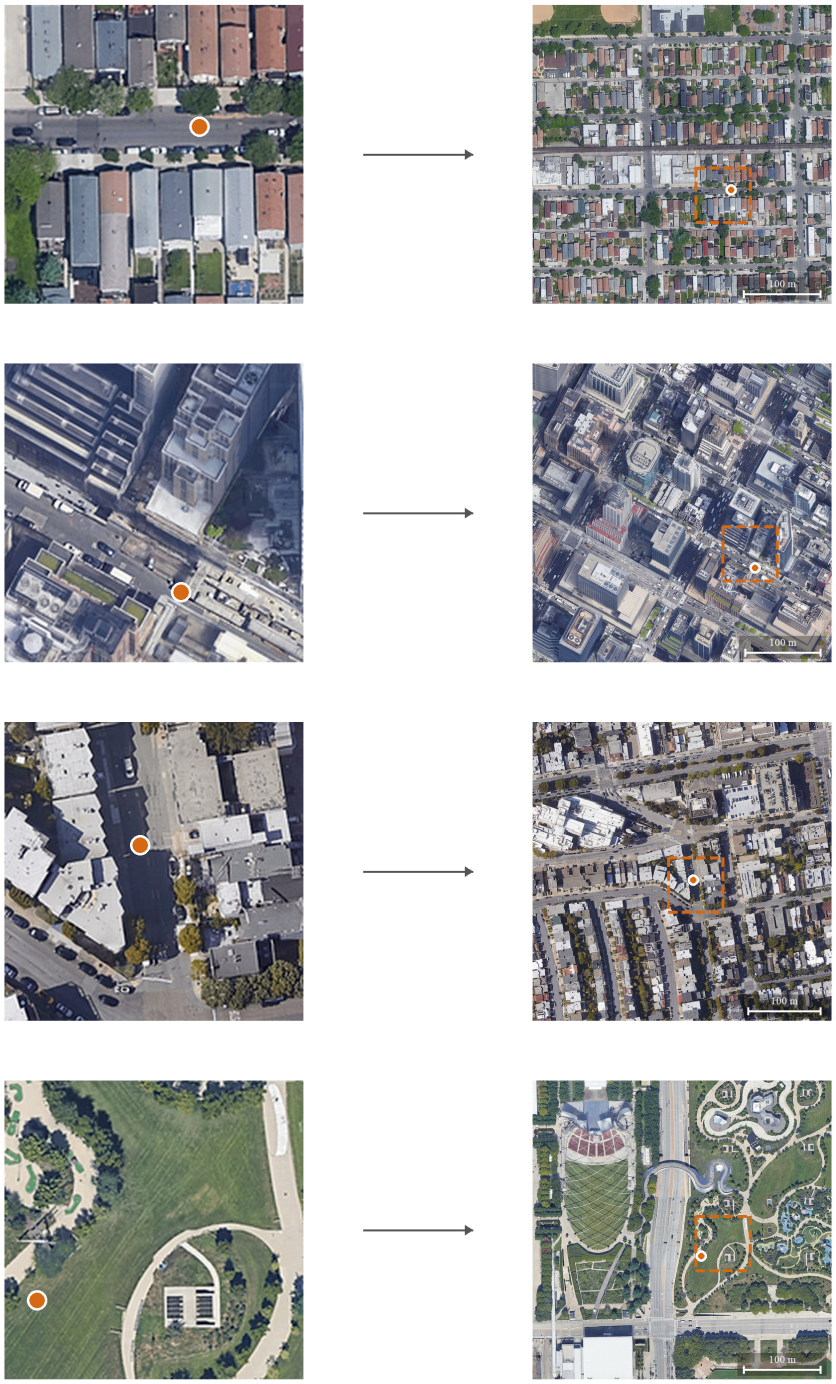
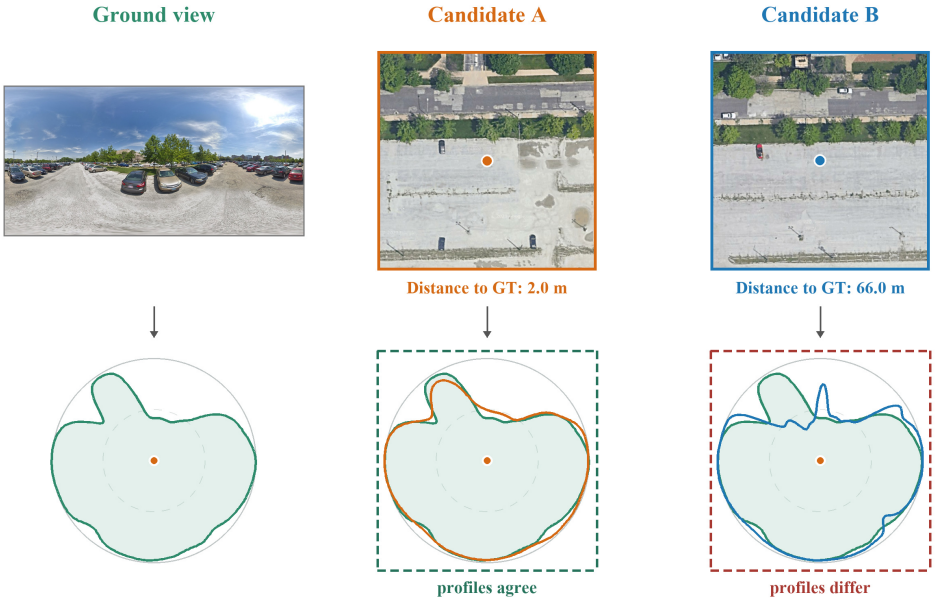
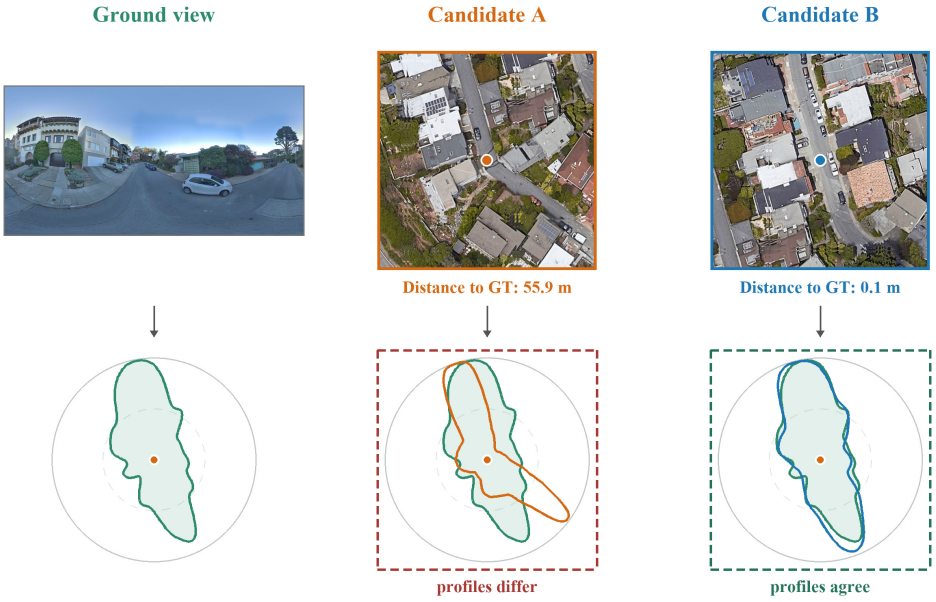


Figure 8: Examples of VIGOR-Pro at the 400 m scale. Each row compares an original VIGOR aerial tile (left) with the corresponding VIGOR-Pro 10×10 mosaic (right). The orange point denotes the ground-camera location, and the dashed orange box indicates the footprint of the original tile within the enlarged aerial region.

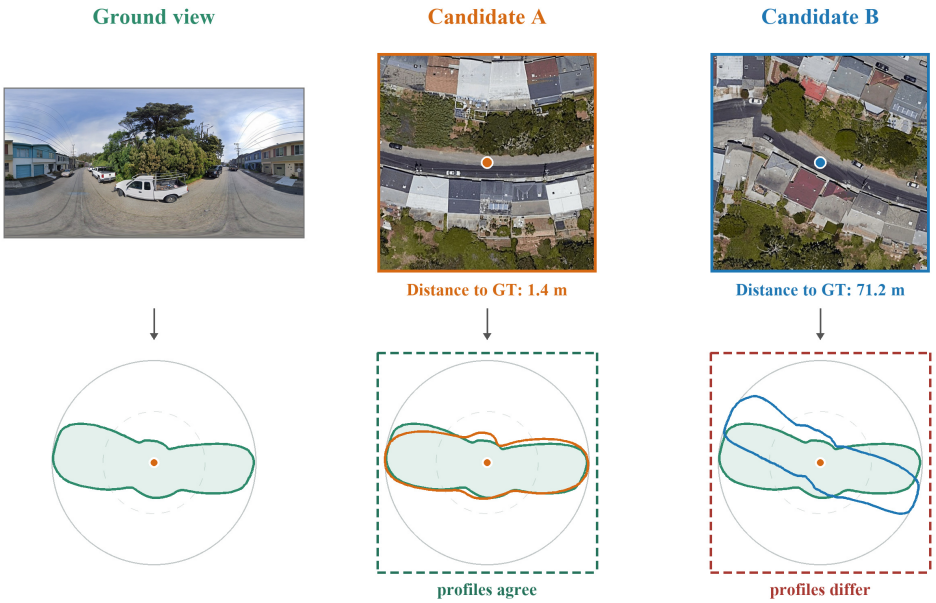


(a)

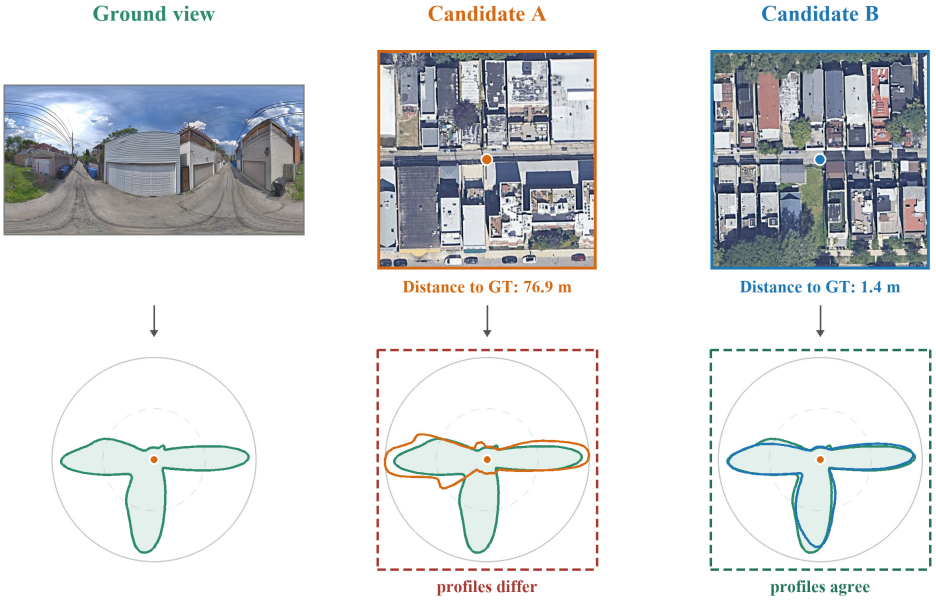


(b)

Figure 9: Qualitative examples of depth rings for cross-view matching. Each panel compares the ground-view depth ring with predicted aerial depth rings at two candidate locations. Candidates closer to the ground-truth camera location exhibit stronger correspondence in surrounding spatial layout, while incorrect candidates show clear geometric discrepancies.

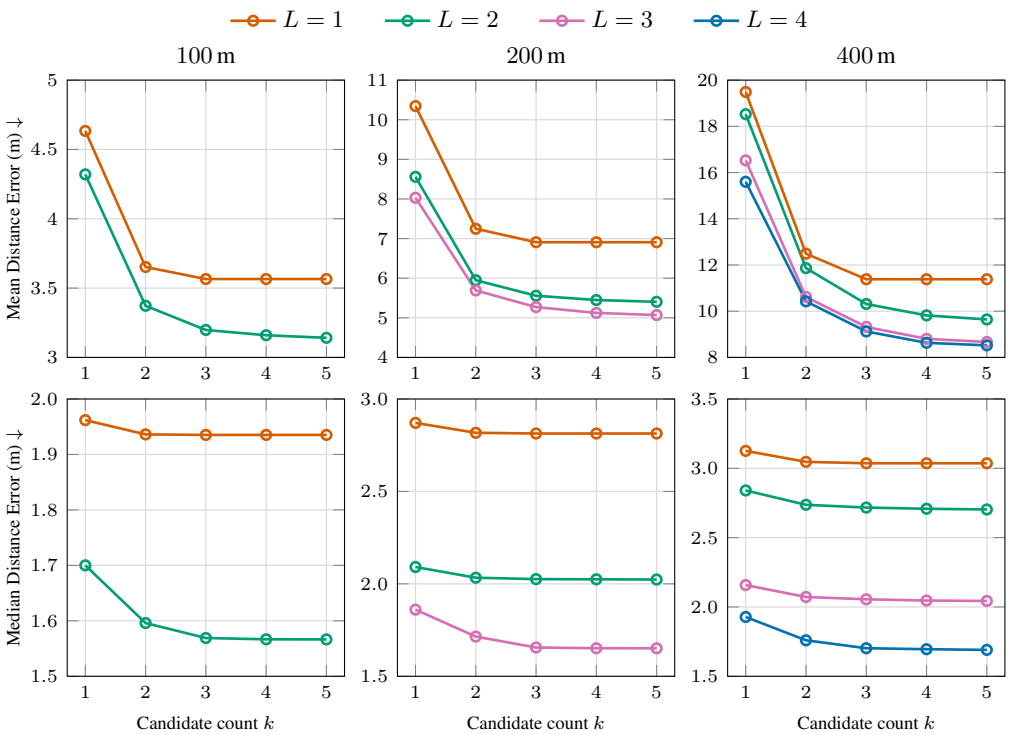


(c)

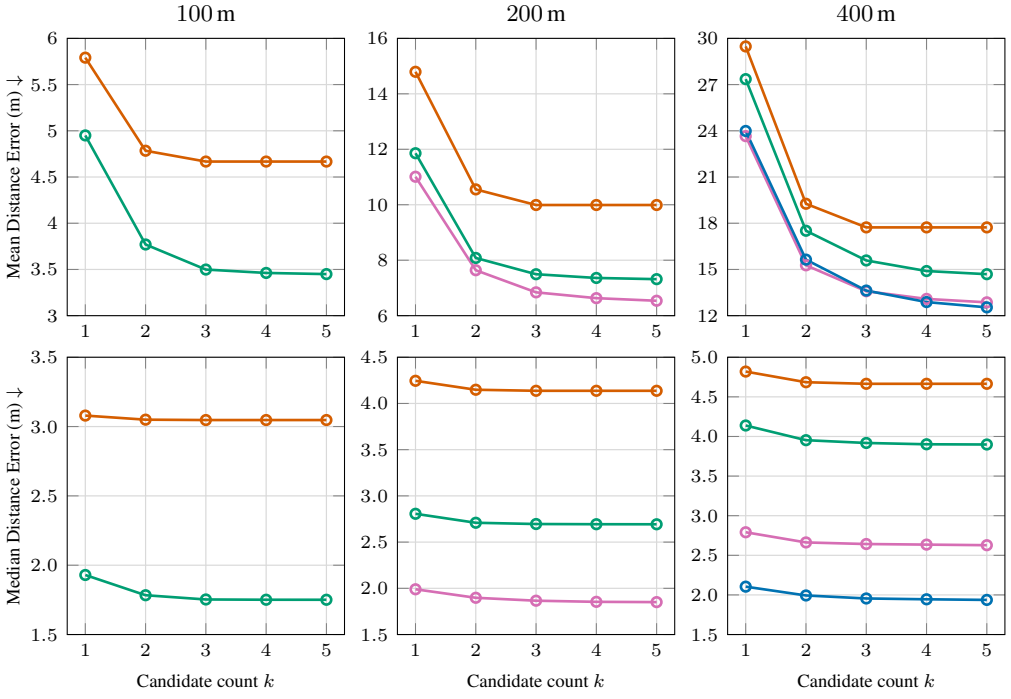


(d)

Figure 9: Qualitative examples of depth rings for cross-view matching (continued).

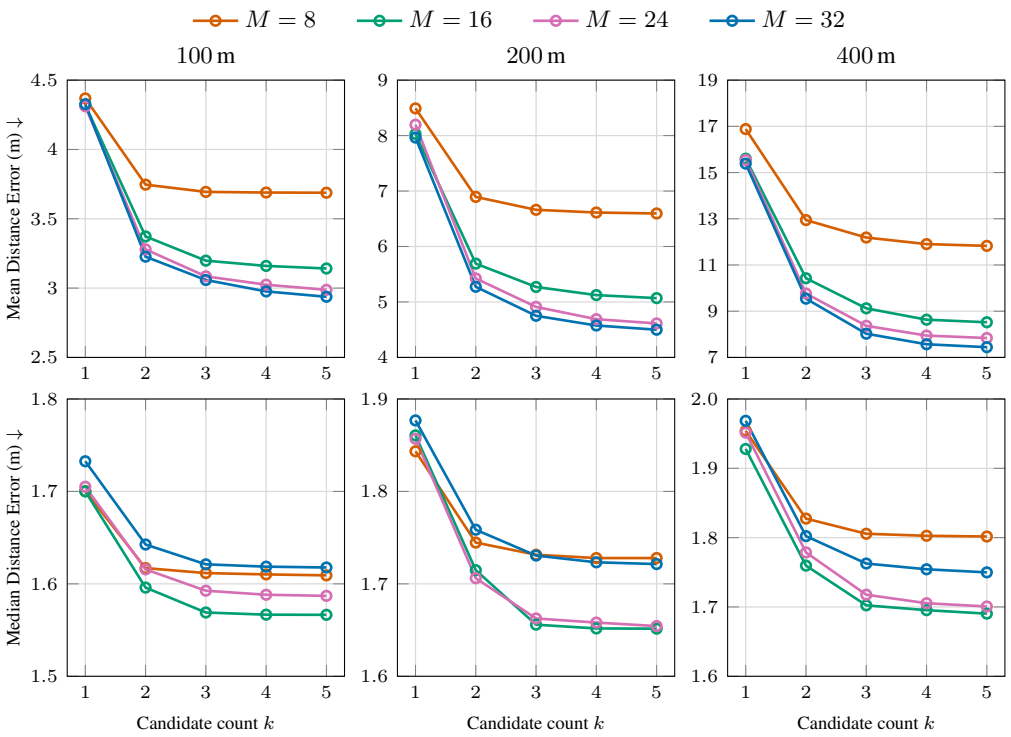


(a) Same-area

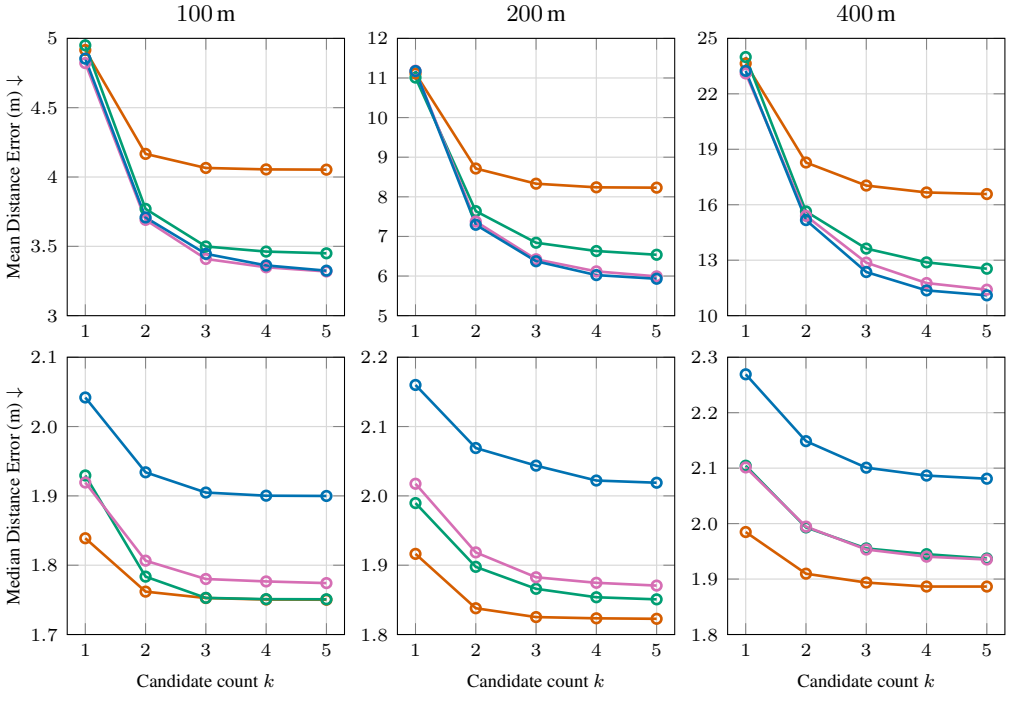


(b) Cross-area

Figure 10: Sensitivity to search level L. Columns correspond to the 100, 200, and 400 m scales, which contain two, three, and four search levels, respectively; rows report mean and median top-k oracle error. Results are reported for both same-area and cross-area evaluation at all three scales.

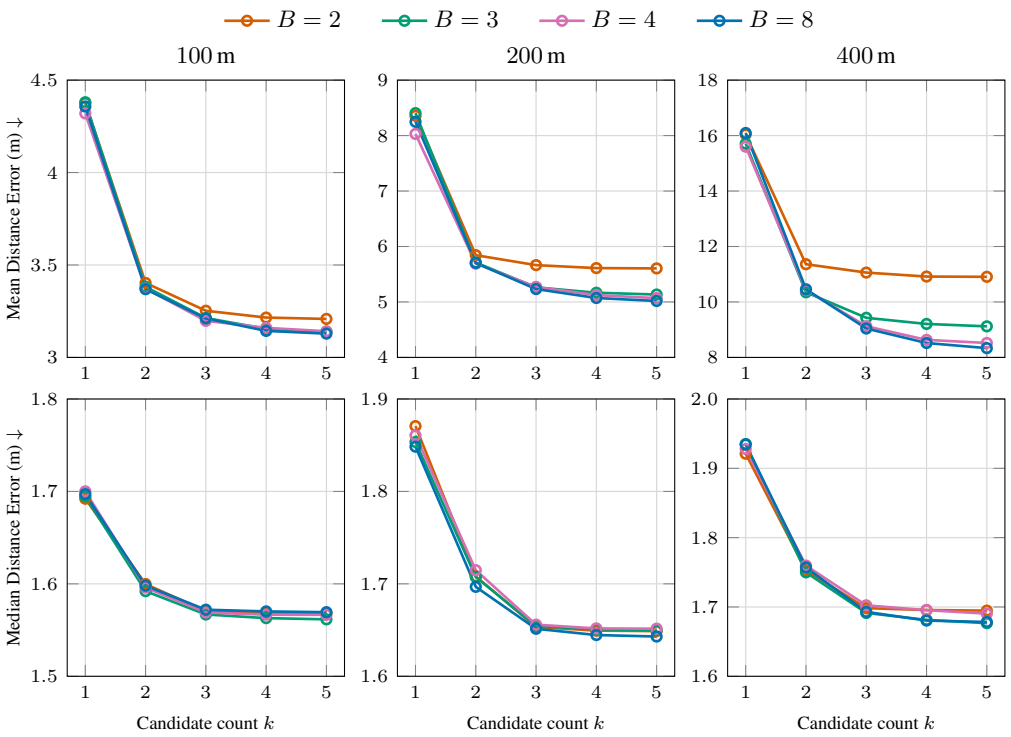


(a) Same-area

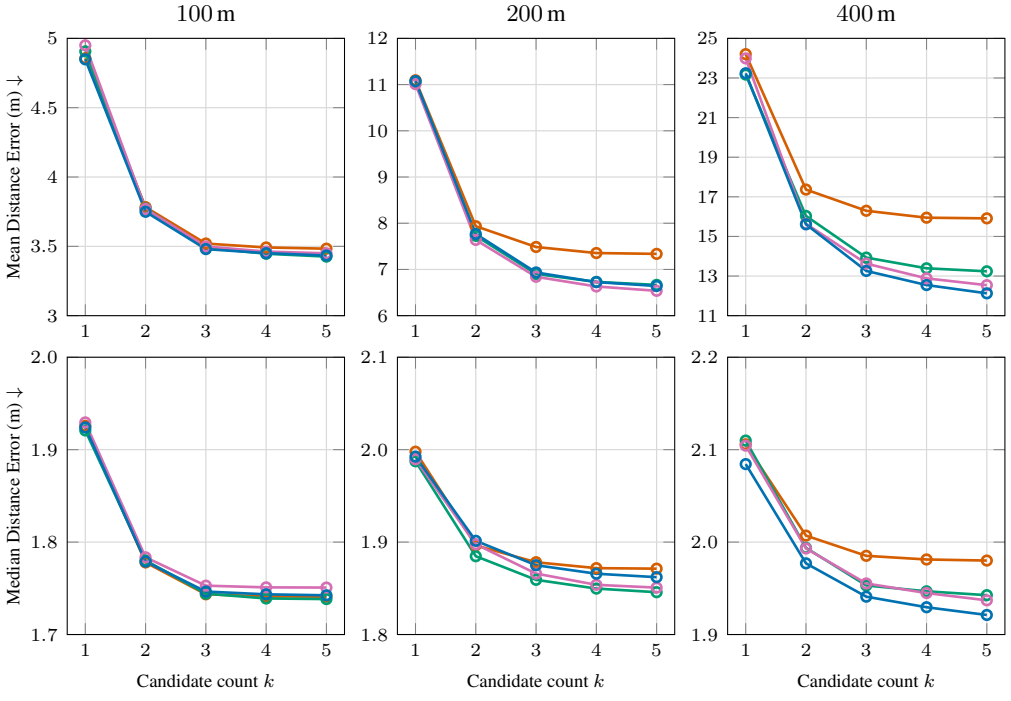


(b) Cross-area

Figure 11: Sensitivity to sampling count M. Columns correspond to the 100, 200, and 400 m scales; rows report mean and median top-k oracle error. Results are reported for both same-area and cross-area evaluation at all three scales.



(a) Same-area



(b) Cross-area

Figure 12: Sensitivity to pruning budget B. Columns correspond to the 100, 200, and 400 m scales; rows report mean and median top-k oracle error. Results are reported for both same-area and cross-area evaluation at all three scales.

User

<image><image><image><image><image><image>
You are given six images:
- Image 1: a ground-level 360° panoramic view from the unknown camera location.
- Image 2: the full satellite image with two candidate camera locations marked.
- Image 3: a 640×640 satellite crop centered on Candidate A (red marker).
- Image 4: a 640×640 satellite crop centered on Candidate B (blue marker).
- Image 5: a polar plot comparing two depth rings:
  * BLACK --- depth ring derived from Image 1 (the panorama).
  * RED --- depth ring predicted for Candidate A's satellite crop.
- Image 6: the same comparison for Candidate B:
  * BLACK --- the same panorama depth ring as in Image 5.
  * BLUE --- depth ring predicted for Candidate B's satellite crop.

Candidate positions (normalized to [0, 1000]):
- A (Red): $[x_A, y_A]$
- B (Blue): $[x_B, y_B]$

Determine which candidate location better matches the camera position of the panoramic image.

Compare the visual surroundings of each candidate in the satellite images with what is visible in the panorama. Consider:
- Road patterns: intersections, lane markings, crosswalks, road width and curvature
- Buildings: color, size, shape, roof type, relative position to the road
- Vegetation: trees, their size and arrangement, proximity to the road
- Landmarks: parking lots, bridges, special structures, fences, walls
- Spatial layout: which side of the road features appear on, relative distances between objects
- Depth-ring agreement (Images 5--6): the correct candidate's colored curve should track the BLACK panorama curve more closely than the other candidate's does. Compare the overall envelope and the azimuths of sharp transitions. If NEITHER colored curve follows the black curve, the depth ring is uninformative for this sample and you should rely on Images 1--4.

Use Image 2 for global context, Images 3--4 for local detail around each candidate, and Images 5--6 for cross-view geometric consistency. Focus on features that clearly distinguish the two candidates. Prioritize strong, unambiguous visual evidence over minor differences.

Think step by step, then provide your final answer as \boxed{A} or \boxed{B}.

Figure 13: Prompt for verification with depth-ring inputs.

User

<image><image><image><image>
You are given four images:
- Image 1: a ground-level 360° panoramic view from the unknown camera location.
- Image 2: the full satellite image with two candidate camera locations marked.
- Image 3: a 640×640 satellite crop centered on Candidate A (red marker).
- Image 4: a 640×640 satellite crop centered on Candidate B (blue marker).

Candidate positions (normalized to [0, 1000]):
- A (Red): [x_A, y_A]
- B (Blue): [x_B, y_B]

Determine which candidate location better matches the camera position of the panoramic image.

Compare the visual surroundings of each candidate in the satellite images with what is visible in the panorama. Consider:
- Road patterns: intersections, lane markings, crosswalks, road width and curvature
- Buildings: color, size, shape, roof type, relative position to the road
- Vegetation: trees, their size and arrangement, proximity to the road
- Landmarks: parking lots, bridges, special structures, fences, walls
- Spatial layout: which side of the road features appear on, relative distances between objects

Use Image 2 for global context and Images 3--4 for local detail around each candidate. Focus on features that clearly distinguish the two candidates. Prioritize strong, unambiguous visual evidence over minor differences.

Think step by step, then provide your final answer as \boxed{A} or \boxed{B}.

Figure 14: Prompt for verification without depth-ring inputs.