BEYOND TOP-1: GEOGRAPHIC ASSOCIATIONS FOR MULTI-REFERENCE CROSS-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale cross-view geo-localization typically follows a retrieval–localization paradigm: candidates are retrieved from an aerial gallery, and the Top-1 reference is selected for local localization. Yet an aerial gallery provides discrete observations of continuous geographic space, and several highly ranked candidates may jointly point to the same local region. Retaining only Top-1 severs geographic associations among candidates and discards complementary spatial localization evidence. We therefore propose GeoHypLoc, a framework for geographically associated multi-reference cross-view localization. It combines query visual support with geographic adjacency in the gallery to organize highly ranked candidates into a positively spatially associated reference set H^*. Preserving each reference's independent georeferencing, it maps their localization evidence into a common real-world geographic space and adaptively aggregates multi-reference evidence throughout coarse-to-fine localization. Finally, it combines region-level visual support with localization-derived spatial evidence to correct the localization response, shifting from single-reference matching to joint constraints over a geographic region. In complete retrieval–localization evaluation on VIGOR, GeoHypLoc improves Acc@3/5/10m over Sample4Geo+CCVPE by 6.60/5.90/4.23 percentage points on Same-area and 2.07/2.14/3.13 percentage points on Cross-area. These results show that organizing multi-reference localization evidence through geographic spatial associations among retrieved candidates can improve large-scale fine-grained cross-view geo-localization.

Keywords: cross-view geo-localization; geographic spatial association; reference-set organization; multi-reference joint localization; aerial image retrieval.

1 INTRODUCTION

Large-scale cross-view geo-localization must connect city-scale search with local, meter-level localization, and thus commonly follows a two-stage retrieval–localization pipeline (Jeon et al., 2026; Song et al., 2025). Most existing methods treat an aerial gallery as a collection of independent image samples: retrieval ranks images by visual similarity, after which Top-1 is selected for fine localization. Geographically, however, an aerial gallery discretely samples a continuous land surface. Neighboring tiles have explicit spatial adjacency, overlapping coverage, and regional continuity. The same real-world location may therefore be observed by several retrieved candidates. Reducing these candidates to a single Top-1 reference before fine localization directly discards their geographic relationships and complementary spatial evidence.

This observation suggests interpreting retrieval results not only as a visual similarity ranking, but also as observations distributed in real-world geographic space. For a given query, geographically neighboring references with high visual support jointly form a local structure of positive spatial association, which can be understood through spatial dependence and local spatial association (Tobler, 1970; MORAN, 1950; Anselin, 1995). The key is therefore not simply to increase k in Top-k, but to identify candidates that form a geographically consistent region and translate this regional spatial support into continuous localization constraints.

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.

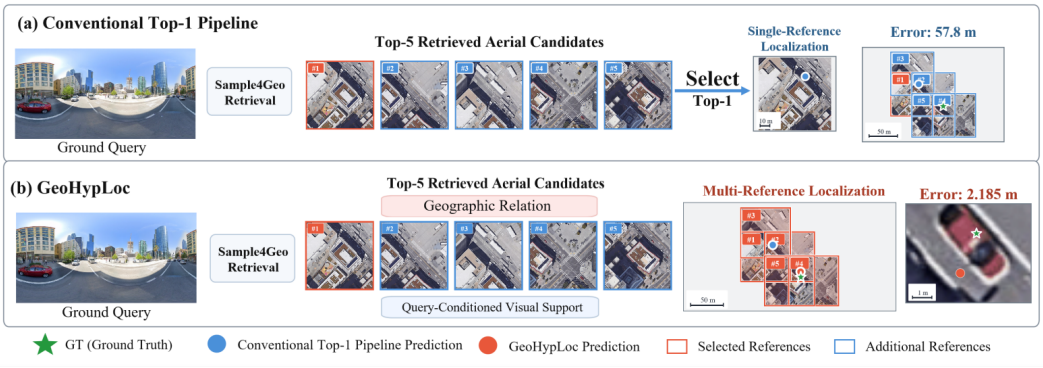

Figure 1: From Top-1 single-reference to multi-reference localization. (a) A conventional retrieval–localization pipeline selects the Top-1 aerial reference for single-reference localization. (b) GeoHypLoc retains multiple references based on geographic relationships and query visual support, and jointly uses their localization information.

This structure is common in VIGOR (Zhu et al., 2021): among Sample4Geo's Top-5 candidates, 75.50%/50.74% of queries in the Same/Cross test sets contain at least two Strong/Semi positive references, with an average of 2.05/1.56 positive references. Retrieved candidates thus often provide multiple observations of the same local geographic region rather than a single observation of the true location. This leads to our central question: how can we recover spatially consistent reference groups from Top-k without using ground-truth locations or positive-reference labels, and enable these independently georeferenced observations to jointly constrain subsequent fine localization?

Figure 1 contrasts the Top-1 single-reference pipeline with GeoHypLoc's multi-reference perspective using the same retrieved candidates. Visual ranking measures similarity to the query, whereas real-world positions reveal spatial proximity and coverage relationships among candidates. Combining these two sources of information allows references with common regional support to be retained before metric localization.

We introduce GeoHypLoc from a geographic perspective, reorganizing ranked retrieval candidates into regional localization hypotheses with geographic structure. First, real-world gallery connectivity and query visual support identify a positively spatially associated reference set H^* from Top-5. Second, preserving each reference's independent georeferencing, we map local localization evidence into a common real-world geographic space and adaptively aggregate valid observations at each location, allowing H^* to continually update the localization state throughout the coarse-to-fine process. Finally, retrieval-side regional visual support and localization-side spatial evidence correct the response. The central shift is from candidate-image fusion to joint constraints from multiple observations of the same geographic region.

This problem differs from existing multi-observation localization. Methods such as HLoc organize candidates using covisibility among same-view database images (Sarlin et al., 2019). Set-CVGL (Wu et al., 2026) and Slice-Loc (Zhang et al., 2026) primarily address multiple query-side observations or redundancy within a single aerial reference. GeoHypLoc instead concerns multiplicity on the aerial-reference side after retrieval: candidates have independent georeferencing and must first be related in real-world space before their localization evidence can be combined. Related work and implementation details are provided in the supplementary material.

Our three main contributions are as follows:

- We revisit the Top-1 paradigm through geographic spatial continuity. An aerial gallery provides discrete observations of a continuous land surface; retrieved references with high visual support and spatial proximity can jointly provide multiple observations of the target region. This perspective transforms independently ranked retrieval results into regional spatial hypotheses.
- We propose GeoHypLoc, which constructs H^* using gallery geographic topology and query visual support, aligns independently georeferenced multi-reference localization ev-

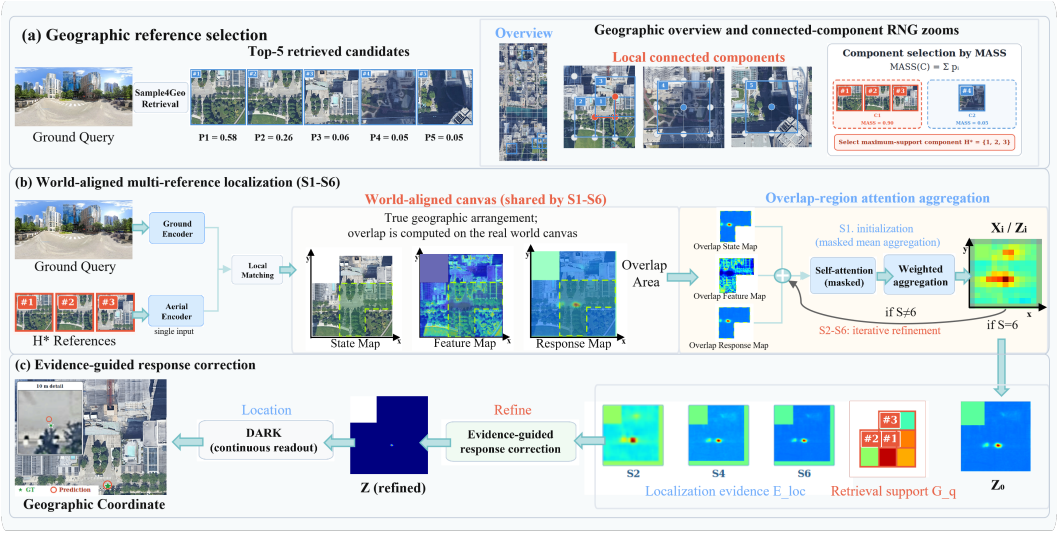

Figure 2: Overview of GeoHypLoc. (a) Top-5 candidates are organized into H^* using full-gallery geographic connectivity and query visual support. (b) Independently georeferenced references are aligned to a common world canvas and jointly update CCVPE's coarse-to-fine localization with position-dependent contributions. (c) Retrieval-side regional support and localization-side spatial evidence correct Z_0 before DARK decoding.

idence in a common real-world geographic space, and performs position-dependent joint updates throughout coarse-to-fine localization, shifting from single-reference matching to joint constraints over a geographic region.

- We validate this approach under VIGOR's complete retrieval–localization protocol. Relative to Sample4Geo+CCVPE, GeoHypLoc improves Acc@3/5/10m by 6.60/5.90/4.23 percentage points on Same-area and 2.07/2.14/3.13 on Cross-area. Reference-composition and response analyses explain the sources and limits of these gains.

2 METHOD

GeoHypLoc aims to transform the geographic regional structure implicit in retrieved Top-k candidates into spatial evidence that a fine localizer can continually exploit. The framework has three components (Figure 2): (i) using Sample4Geo (Deuser et al., 2023), identify a spatially consistent, positively spatially associated reference set H^* from Top-5; (ii) map references' local localization evidence into a common real-world geographic space using their independent georeferencing and jointly update it along CCVPE's coarse-to-fine path (Xia et al., 2024); and (iii) combine region-level query visual support and localization-derived spatial evidence to correct the final response. The method thus addresses the organization, alignment, and joint constraints of multiple observations in continuous geographic space, rather than generic multi-image fusion. Sections 2.1–2.3 describe these steps.

2.1 EXTRACTING A POSITIVELY SPATIALLY ASSOCIATED REFERENCE SET THROUGH GEOGRAPHIC CONNECTIVITY

Cross-view retrieval measures the visual similarity between the current query and each aerial reference, but high ranks alone do not establish whether candidates point to the same real-world region. When high query visual similarity coincides with real-world spatial proximity, candidates can form a spatial association structure that provides regional support. GeoHypLoc therefore retains the fixed retriever's original output and uses geographic connectivity in the aerial gallery to recover spatial associations among Top-5 candidates and identify the positively spatially associated reference set H^* supported by the query.

For query q, the fixed retriever returns the Top-5 candidate set $\mathcal{C}_q = \{r_i\}_{i=1}^{5}$, where s_i denotes the image-embedding similarity between the query and reference r_i, and c_i is its geographic position. To represent real-world spatial proximity, we precompute adjacency over the full aerial gallery using the relative neighborhood graph (RNG) definition (Toussaint, 1980). Let E denote the full-gallery RNG edge set and $d(\cdot,\cdot)$ the Euclidean distance between georeferenced gallery centers. Two gallery references u, v are connected if no third reference is closer to both than they are to each other:

$$(u,v) \in E \iff \nexists w \notin \{u,v\} : d(u,w) < d(u,v), \quad d(v,w) < d(u,v).$$

For the current Top-5, we retain only edges inherited from the full-gallery RNG; the resulting connected components define the candidate family $\mathcal{H}_q$. Let μ_q and σ_q denote the mean and population standard deviation of the current Top-5 similarities, respectively, and let $\epsilon = 10^{-6}$ be the lower bound on σ_q. We standardize the current Top-5 similarities and normalize them into relative query visual support:

$$z_i = \frac{s_i - \mu_q}{\max(\sigma_q, \epsilon)}, \quad p_i = \frac{\exp z_i}{\sum_{k \in \mathcal{C}_q} \exp z_k}, \quad H^* = \arg\max_{H \in \mathcal{H}_q} \sum_{r_i \in H} p_i.$$

We select the component with the largest cumulative support as H^*. Construction uses neither query ground-truth positions nor Strong/Semi labels, and each selected member retains its original aerial image and independent georeferencing. Appendix A gives the exact numerical normalization, tie-breaking rule, and footprint convention.

2.2 Multi-reference joint localization with spatial heterogeneity

References in H^* observe different windows of the continuous land surface. Their local feature grids have different geographic origins, so the same pixel index does not represent the same real-world location. Using each reference's georeferencing, we resample stage-wise matching responses and localization states onto a common world canvas, enabling cross-reference interaction at the same real-world geographic location. We then learn the relative effectiveness of different observations at different positions, rather than assigning a fixed weight to an entire reference image, explicitly modeling spatial heterogeneity in multi-reference localization evidence (Brunsdon et al., 1996).

We retain pretrained CCVPE's coarse-to-fine localization path (Xia et al., 2024), without redesigning the single-reference localizer. At each stage, references independently produce matching responses and localization states, followed by geographic alignment, same-position member interaction, and evidence aggregation; the aggregated shared state enters the next stage. Multi-reference information thus continually influences subsequent matching throughout S1–S6, rather than being fused only at the final coordinate level.

At S1, member-wise matching responses and descriptor states are mapped with each reference's georeferencing to the common world canvas. At S2–S6, current-scale aerial features are aligned before matching and combined with member-masked copies of the shared state; cross-reference interaction is restricted to references valid at the same world position. S1 uses validity-masked averaging, while S2–S6 estimate position-dependent contributions from current localization evidence (Appendix B). For valid members, matching responses, localization states, and coverage information are encoded into $e_{j,s}(x)$. Let $\mathcal{V}_s(x)$ denote the references with valid coverage at world position x. Self-attention yields interacted representations $h_{j,s}(x)$, and the shared scalar scoring head at stage s, denoted by g_s, produces

$$w_{j,s}(x) = \frac{\exp(g_s(h_{j,s}(x)))}{\sum_{k \in \mathcal{V}_s(x)} \exp(g_s(h_{k,s}(x)))}.$$

These weights vary with stage and world position and use only localization-side response, state, and coverage evidence, not retrieval similarity, rank, or Strong/Semi labels. Denote the aligned member response and localization state on the common world canvas by $\widetilde{R}_{j,s}(x)$ and $\widetilde{X}_{j,s}(x)$, respectively. The same weights aggregate aligned responses and states:

$$R_s(x) = \sum_{j \in \mathcal{V}_s(x)} w_{j,s}(x) \widetilde{R}_{j,s}(x), \quad X_s(x) = \sum_{j \in \mathcal{V}_s(x)} w_{j,s}(x) \widetilde{X}_{j,s}(x).$$

The aggregated response and normalized state are decoded and upsampled to form the next shared state. Repeating this process through S6 yields the base response Z_0. Appendix B specifies the exact state construction, validity masks, member representation, attention architecture, and recurrence.

2.3 Response correction with region-level query visual support and localization-derived spatial evidence

Section 2.2 enables multiple references in H^* to participate continuously in joint localization within a common real-world geographic space. Their position-dependent contributions are primarily determined by evidence internal to the localizer. Meanwhile, the retrieval results in Section 2.1 retain the query's visual support for the gallery region containing H^*. We therefore map both region-level query visual support and localization-derived spatial evidence into the common real-world geographic space to correct the base localization response.

Specifically, we summarize the concentration, relative dominance, and member overlap of reference contributions across S2–S6 into localization-derived spatial evidence. Separately, we select the Top-5 references in the same gallery city as H^*, denoting this set by $\mathcal{C}_q^{H^*}$. Projecting each reference r_i's relative query visual support p_i onto its real-world geographic footprint F_i yields region-level query visual support:

$$G_q(x) = \sum_{i \in \mathcal{C}_q^{H^*}} p_i \, \mathbf{1}[x \in F_i].$$

Localization-derived spatial evidence and region-level query visual support drive two residual branches that correct the base response:

$$Z = Z_0 + \Delta_{\mathrm{loc}} + \Delta_{\mathrm{ret}}.$$

We apply DARK (Zhang et al., 2020) directly to the corrected response, $q_D = \mathrm{DARK}(Z)$. Appendix C specifies both evidence encodings, residual branches, and coordinate decoding; Figure C.1 visualizes the intermediate evidence and response changes. Training additionally applies a peak-margin objective on Z that separates the ground-truth neighborhood from distant competing regions. Its exact formulation, metric conversion, loss weighting, and gradient boundaries are given in Appendix C.6.

3 Experiments

We evaluate GeoHypLoc through complete retrieval–localization performance, localization given a correct reference, and controlled comparisons of reference organization and localization configurations; Section 4 analyzes H^* and the base response Z_0.

3.1 Experimental setup

We evaluate on VIGOR (Zhu et al., 2021) with known orientation and complete test sets. Same-area uses 52,609/52,605 train/test queries; Cross-area trains on 51,520 New York/Seattle queries and tests on 53,694 unseen Chicago/San Francisco queries. We use fixed, officially pretrained Sample4Geo (Deuser et al., 2023) for full-gallery Top-5 retrieval and initialize localization from the corresponding official CCVPE weights (Xia et al., 2024). All newly trained configurations run five epochs and report the predetermined E5 checkpoint only; Appendix D gives complete optimization and evaluation settings. We report Acc@1/3/5/10m and median localization error for complete retrieval–localization evaluation; correct-reference evaluation additionally reports mean error. Metric distances use the city-specific conversion in Appendix C.5, and accuracy differences are reported in percentage points.

3.2 Complete-system performance and localization given a correct reference

3.2.1 Complete retrieval–localization performance

Table 1 compares VIGOR, CIPER, Sample4Geo+CCVPE, and full GeoHypLoc. VIGOR (Zhu et al., 2021) results are computed from official prediction files, while CIPER (Jeon et al., 2026) is evaluated with its released weights under the metrics above. Sample4Geo+CCVPE is the direct cascade baseline, feeding fixed Sample4Geo Top-1 (Deuser et al., 2023) to original CCVPE (Xia et al., 2024).

Table 1: Complete retrieval–localization performance on VIGOR.

Split	Method	Acc@1m ↑	Acc@3m ↑	Acc@5m ↑	Acc@10m ↑	Median ↓
Same	VIGOR	0.403	3.555	8.940	25.530	29.112
Same	CIPER	1.661	11.075	20.673	35.614	42.997
Same	Sample4Geo+CCVPE	30.750	65.148	73.970	82.057	1.685
Same	GeoHypLoc	**36.128**	**71.748**	**79.871**	**86.288**	**1.409**
Cross	VIGOR	0.071	0.730	1.928	6.185	1819.810
Cross	CIPER	0.251	2.116	4.516	8.653	1933.922
Cross	Sample4Geo+CCVPE	**18.310**	50.782	59.658	63.884	2.917
Cross	GeoHypLoc	16.972	**52.849**	**61.802**	**67.013**	**2.714**

Acc is in % and Median in m; bold denotes the best value within each protocol.

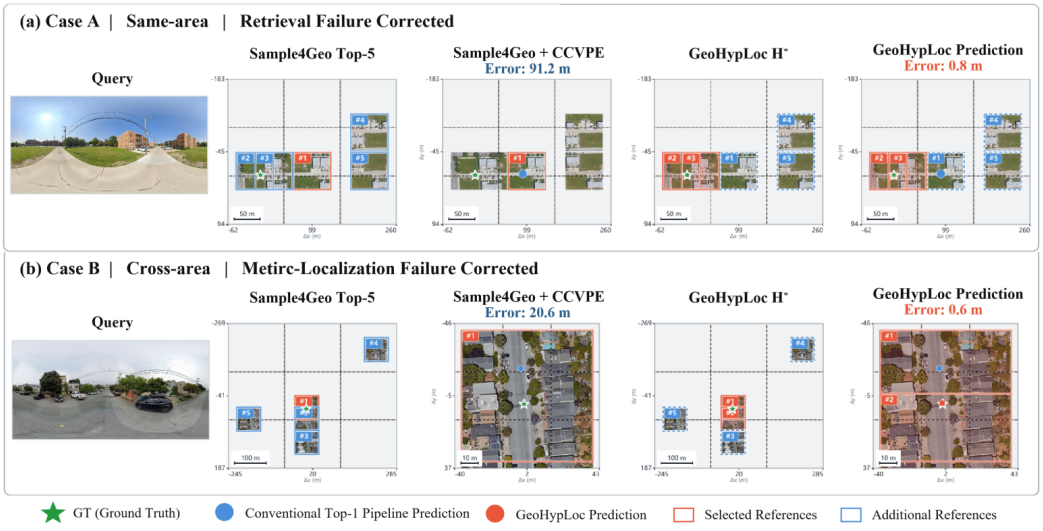

Figure 3: Representative improvements over Sample4Geo+CCVPE. Top: Top-1 coverage failure. Bottom: displaced single-reference localization despite coverage. Rows show the query, retrieved candidates, baseline, GeoHypLoc H^*, and final prediction.

GeoHypLoc uses the same retriever but constructs H^* from Top-5 and performs multi-reference localization, response correction, and DARK decoding. Relative to Sample4Geo+CCVPE, GeoHypLoc improves Same-area Acc@1/3/5/10m by 5.38/6.60/5.90/4.23 percentage points and reduces Median from 1.685 to 1.409 m. On Cross-area it improves Acc@3/5/10m by 2.07/2.14/3.13 points and reduces Median from 2.917 to 2.714 m, while Acc@1m decreases by 1.34 points. Figure 3 shows two baseline failures: Top-1 misses ground-truth coverage or yields displaced localization despite coverage.

3.2.2 Localization given a correct reference

Table 2 isolates metric localization from retrieval errors using the official Strong reference. SliceMatch and CCVPE values are quoted from their papers (Lentsch et al., 2023; Xia et al., 2024); all methods use only ground and aerial imagery. At test time, both GeoHypLoc rows use the same single Strong reference, $q_0 = \mathrm{DARK}(Z_0)$, and no response correction. Strong-only trains on the Strong reference, whereas GeoHypLoc trains with the complete aligned H^*; the comparison jointly reflects training reference composition and organization (Appendix D.3). Relative to GeoHypLoc (Strong-only), multi-reference training reduces Same-area Mean/Median from 3.338/1.439 to 3.279/1.296 m and Cross-area Mean from 4.782 to 4.543 m, while Cross Median increases from 1.742 to 1.876 m. Thus the same single-reference test condition shows improved average error with a cross-area median trade-off; this comparison does not isolate composition from joint training.

Table 2: Localization performance given a correct reference on VIGOR.

Method / training configuration	Same Mean ↓	Same Median ↓	Cross Mean ↓	Cross Median ↓
SliceMatch	5.18	2.58	5.53	2.55
CCVPE	3.60	1.36	4.97	**1.68**
GeoHypLoc (Strong-only)	3.338	1.439	4.782	1.742
GeoHypLoc	**3.279**	**1.296**	**4.543**	1.876

All methods use correct references and known orientation. Mean/Median are in m; bold marks best. Both GeoHypLoc rows test on one Strong reference with DARKZ_0 and no correction..

Table 3: Reference organization and joint localization configurations under actual retrieval conditions.

(a) Same-area

Configuration	Acc@1m↑	Acc@3m↑	Acc@5m ↑	Acc@10m↑	Median ↓
Base: Sample4Geo+CCVPE	30.750	65.148	73.970	82.057	1.685
A: Rank-matched independent	35.744	70.843	78.504	84.642	1.394
B: H^* independent	35.776	70.739	78.276	84.302	**1.393**
C: H^* joint localization	35.748	70.440	78.010	84.150	1.435
D: Full GeoHypLoc	**36.128**	**71.748**	**79.871**	**86.288**	1.409

(b) Cross-area

Configuration	Acc@1m ↑	Acc@3m ↑	Acc@5m ↑	Acc@10m ↑	Median ↓
Base: Sample4Geo+CCVPE	**18.310**	50.782	59.658	63.884	2.917
A: Rank-matched independent	15.898	50.320	58.724	63.612	2.960
B: H^* independent	15.871	49.773	57.763	62.413	3.028
C: H^* joint localization	17.577	**52.931**	60.800	65.383	**2.668**
D: Full GeoHypLoc	16.972	52.849	**61.802**	**67.013**	2.714

A/B use independent single-reference training; C/D use complete-H^* joint localization. Base/A/B test on Top-1, C/D on H^*. C and D use DARK on Z_0 and Z, respectively. Acc is in % and Median in m.

3.3 Comparing reference organization and joint localization configurations

Table 3 compares reference composition, joint localization, and response processing under fixed Sample4Geo retrieval. Base uses Sample4Geo Top-1 with original CCVPE. A independently trains on the first $K_q = |H_q^*|$ Top-5 references by retrieval rank, whereas B independently trains on the H^* members; Base/A/B all receive Top-1 at test time. C uses the complete H^* jointly at training and testing and decodes $\mathrm{DARK}(Z_0)$; D adds response correction and decodes $\mathrm{DARK}(Z)$. D/Base correspond to GeoHypLoc/Sample4Geo+CCVPE in Table 1; Appendix D.4 gives the exact protocols. A versus B changes only the independently trained reference composition. Replacing rank-matched references with H^* does not improve this Top-1 test configuration: Acc@3/5/10m changes by $-0.10/-0.23/-0.34$ points on Same and $-0.55/-0.96/-1.20$ on Cross. Section 4.1 separately examines the composition of H^*. B to C introduces complete-H^* joint localization at both training and testing. Cross-area Acc@3/5/10m improves by 3.16/3.04/2.97 points and Median drops from 3.028 to 2.668 m, while Same-area changes are slightly negative. Because training organization and test inputs change together, this comparison cannot isolate an individual module. C to D isolates response correction under the same multi-reference localization and DARK decoder. It improves Acc@5/10m by 1.86/2.14 points on Same and 1.00/1.63 on Cross, with trade-offs at stricter Cross-area thresholds and Median.

4 Analysis

We analyze two questions suggested by Section 3: whether H^* concentrates geographically relevant references and how the structure of the base response Z_0 relates to localization difficulty and correction gains.

Table 4: Reference composition under different selection strategies.

Split	Reference set	Mean K	Any Positive	Positive Fraction	Multi Positive	Strong hit
Same	Raw Top-1	1.000	93.632	93.632	0.000	77.852
Same	Raw Top-5	5.000	98.215	41.058	75.497	95.668
Same	Rank-matched	2.327	95.124	78.032	60.289	91.505
Same	H^*	2.327	94.521	87.044	73.590	92.735
Cross	Raw Top-1	1.000	74.824	74.824	0.000	61.796
Cross	Raw Top-5	5.000	88.673	31.153	50.736	83.499
Cross	Rank-matched	1.847	77.562	64.774	35.064	72.299
Cross	H^*	1.847	75.860	71.718	47.365	73.172

All composition metrics except Mean K are percentages over the complete test set.

Table 5: Grouped localization results and response-correction gains for the multi-reference base response.

Split	Response group	Queries	Share	Z_0 Acc@5m	Z_0 Acc@10m	Z_0 Median	Refinement Gain@5m	Refinement Gain@10m
Same	Remaining	41,659	79.19%	87.9186	92.35	1.174 m	+0.890	+0.847
Same	Strong-secondary	10,946	20.81%	40.2978	52.95	8.593 m	+5.555	+7.053
Cross	Remaining	33,712	62.79%	79.0757	81.91	1.760 m	+0.445	+1.164
Cross	Strong-secondary	19,982	37.21%	29.9670	37.51	32.654 m	+1.942	+2.417

Acc/Share are percentages, Median is m, and Gain in percentage points; groups partition each complete test set.

4.1 Composition and geographic coverage of the positively spatially associated reference set

Under fixed Sample4Geo retrieval, we compare Raw Top-1, Raw Top-5, Rank-matched Top-K_q, and H^*. Rank-matched and H^* use the same per-query member count $K_q = |H_q^*|$, isolating reference composition. Table 4 reports Mean K, AnyPositive, PositiveFraction, MultiPositive, and Strong hit, treating Strong/Semi references as positives; these discrete statistics differ from footprint coverage. At matched member counts, H^* raises PositiveFraction over Rank-matched by 9.012/6.944 points and MultiPositive by 13.301/12.301 on Same/Cross, while AnyPositive decreases by 0.603/1.702 points. Geographic organization therefore concentrates positives and more often retains multiple positives, although it does not improve every hit statistic. Reference identity and geographic coverage are not equivalent. When Strong is absent but other H^* members still cover the ground truth, Same/Cross Acc@10m remains 60.35%/59.90%; when the ground truth lies outside the footprint union, Acc@1/3/5/10m is zero on both test sets. Thus overlapping references can retain localization support even without the official Strong reference.

4.2 Base-response structure and response-correction gains

We next relate the multi-reference base response Z_0 to localization difficulty and subsequent response-correction gains, and compare strong-secondary-peak prevalence with the single-reference baseline. Within valid coverage, let ρ be the response ratio of the strongest local secondary peak at least 2 m from the primary peak. Queries with $\rho \geq 0.25$ form Strong-secondary; the rest form Remaining. Groups are fixed from Z_0 before correction, and Remaining does not imply strict unimodality. Table 5 uses the complete-H^* model of row C in Table 3. Base predictions are $q_0 = \mathrm{DARK}(Z_0)$, and correction gain is the within-group accuracy change to $q_D = \mathrm{DARK}(Z)$ using the same decoder. These diagnostics do not decompose the C-to-D difference. Strong-secondary queries comprise 20.81%/37.21% of Same/Cross under complete-H^* localization, versus 20.00%/32.15% for the single-reference Base. This comparison does not isolate a reference-count effect. Strong-secondary queries are markedly harder and benefit more from correction. Their Same/Cross base Acc@10m is 52.95%/37.51% versus 92.35%/81.91% for Remaining, with Medians 8.593/32.654 m versus 1.174/1.760 m. Their correction gains at Acc@5/10m reach 5.555/7.053

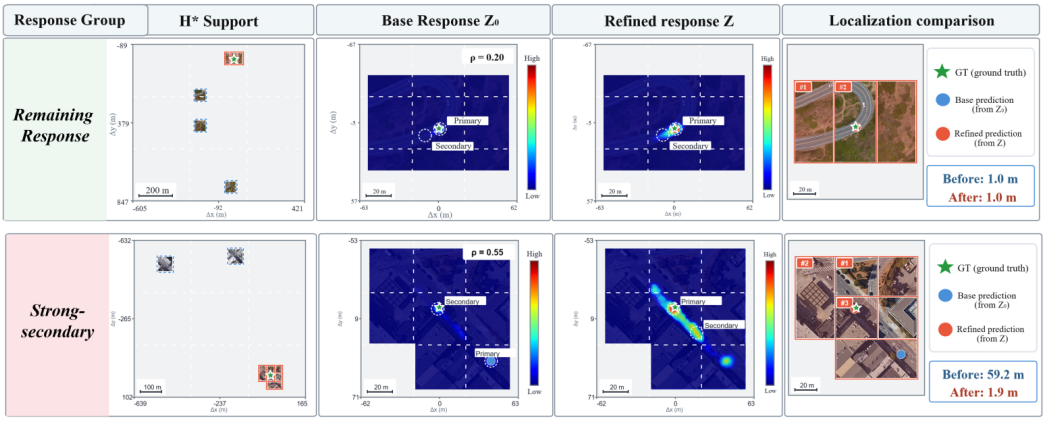

Figure 4: Multi-reference response structure and correction. Rows show Remaining and Strong-secondary examples; columns compare H^* support, base response Z_0, corrected response Z, and localization. Both predictions use DARK, and base responses mark ground truth, primary peak, and the valid secondary peak.

points on Same and 1.942/2.417 on Cross, far above Remaining. Strong secondary peaks are therefore statistically associated with larger base errors and greater scope for correction; these results do not establish that they cause the errors or that a particular peak is directly suppressed. Figure 4 visualizes a Remaining case and a Strong-secondary case under a fixed H^*, comparing Z_0, corrected Z, and the corresponding DARK predictions.

5 Conclusion

We address a key discontinuity in retrieval–localization: Top-k retrieval already provides multiple georeferenced candidates, but conventional pipelines reduce them to Top-1 before fine localization, preventing further use of their spatial relationships and overlapping evidence. GeoHypLoc organizes Top-k into a positively spatially associated reference set H^*, lets multiple references jointly update coarse-to-fine localization states in common world coordinates, and corrects the final response using retrieval-side regional support and localization-side spatial evidence. On VIGOR, this design improves Sample4Geo+CCVPE at 3–10 m thresholds on both Same-area and Cross-area. Reference-composition and response analyses respectively validate the motivations of spatially consistent reference organization and competing-response correction.

Current validation is limited to VIGOR with known orientation, a fixed Sample4Geo retriever, and Top-5 candidates; multi-reference processing also adds computational cost. Future work may examine different retrievers, candidate counts, and gallery structures, and further improve localization response modeling and coordinate readout. The results support retaining and jointly using geographically associated aerial references after retrieval, while broader settings are needed to establish the method's scope. Figure E.1 presents representative failures.

AI Use Statement

Generative AI tools were used to assist with paper organization and language editing, LaTeX/document preparation, code writing and debugging, experimental protocol checks, and figure/layout preparation. The authors reviewed AI-assisted content and verified the reported results against the underlying experimental artifacts and code. The authors take responsibility for all claims, equations, references, code, and figures in this paper.

REFERENCES

Luc Anselin. Local indicators of spatial association—lisa. *Geographical Analysis*, 27 (2):93–115, 1995. doi: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1995.tb00338.x.

Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996. doi: https://doi.org/10.1111/j.1538-4632.1996.tb00936.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1996.tb00936.x.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16801–16810, 2023. doi: 10.1109/ICCV51070.2023.01545.

Mingtao Feng, Fenghao Tian, Jianqiao Luo, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Semantic ambiguity modeling and propagation for fine-grained visual cross view geo-localization. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, volume 39 of *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2978–2986, United States, April 2025. Association for the Advancement of Artificial Intelligence (AAAI). doi: 10.1609/aaai.v39i3.32305. Publisher Copyright: © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 39th Annual AAAI Conference on Artificial Intelligence, AAAI 2025 ; Conference date: 25-02-2025 Through 04-03-2025.

Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. C-bev: Contrastive bird's eye view training for cross-view image retrieval and 3-dof pose estimation, 2023. URL https://arxiv.org/abs/2312.08060.

Yurim Jeon, Dongseong Seo, and Seung-Woo Seo. Ciper: A unified framework for cross-view image-retrieval and pose-estimation, 2026. URL https://arxiv.org/abs/2606.05011.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

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

Songlian Li, Min Hu, Xiongwu Xiao, and Zhigang Tu. Patch similarity self-knowledge distillation for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5091–5103, 2024. doi: 10.1109/TCSVT.2023.3336844.

P. A. P. MORAN. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 06 1950. ISSN 0006-3444. doi: 10.1093/biomet/37.1-2.17. URL https://doi.org/10.1093/biomet/37.1-2.17.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. *Spatial-aware feature aggregation for cross-view image based geo-localization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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, 2025. URL https://arxiv.org/abs/2505.07622.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240, 1970. doi: 10.2307/143141. URL https://www.tandfonline.com/doi/abs/10.2307/143141.

Godfried T. Toussaint. The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, 12(4):261–268, 1980. doi: 10.1016/0031-3203(80)90066-7.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geolocalization using a correlation-aware homography estimator. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cgiP4cMBP9.

Qiong Wu, Panwang Xia, Lei Yu, Yi Liu, Mingtao Xiong, Liheng Zhong, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yi Wan. Set-cvgl: A new perspective on cross-view geolocalization with unordered ground-view image sets. *ISPRS Journal of Photogrammetry and Remote Sensing*, 233:328–345, 2026. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2026.01.037. URL https://www.sciencedirect.com/science/article/pii/S0924271626000456.

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

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. Visual cross-view metric localization with dense uncertainty estimates. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 90–106, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19842-7.

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

Zimin Xia, Chenghao Xu, and Alexandre Alahi. Loc^2: Interpretable cross-view localization via depth-lifted local feature matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2ciXKn2UlS.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091–7100, 2020. doi: 10.1109/CVPR42600.2020.00712.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geolocalization via learning disentangled geometric layout correspondence. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25457. URL https://doi.org/10.1609/aaai.v37i3.25457.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10419–10433, 2024. doi: 10.1109/TPAMI.2024.3443652.

Yongjun Zhang, Mingtao Xiong, Yi Wan, and Gui-Song Xia. Cross-view localization via redundant sliced observations and a-contrario validation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 236:421–437, 2026. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2026.04.014. URL https://www.sciencedirect.com/science/article/pii/S0924271626001863.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1162–1171, June 2022.

SUPPLEMENTARY MATERIAL

This supplementary material specifies exact reference-set construction, geographically aligned multi-reference localization, response correction, and DARK coordinate decoding, together with implementation definitions for training, evaluation, and response analysis.

SUPPLEMENTARY RELATED WORK

Cross-view image retrieval Cross-view image retrieval ranks large galleries by similarity between a ground query and georeferenced aerial or satellite imagery. SAFA (Shi et al., 2019) strengthens representations through spatial-aware feature aggregation; TransGeo (Zhu et al., 2022) models cross-view features with Transformers; GeoDTR (Zhang et al., 2023) and GeoDTR+ (Zhang et al., 2024) learn geometric layout representations; Sample4Geo (Deuser et al., 2023) improves retrieval discrimination through hard-negative sampling; and PaSS-KD (Li et al., 2024) performs self-knowledge distillation using patch similarity. VIGOR (Zhu et al., 2021) allows multiple overlapping aerial windows to cover the same ground location, demonstrating that valid references need not be unique. Retrieval methods primarily address reference representation and ranking. GeoHypLoc additionally preserves retrieved references' independent georeferencing and uses their geographic associations to organize subsequent metric localization.

Fine-grained cross-view metric localization Fine-grained cross-view localization estimates continuous position or pose within a given local aerial reference or region. CVML (Xia et al., 2022) models dense localization uncertainty, while SliceMatch (Lentsch et al., 2023) and CCVPE (Xia et al., 2024) construct matching responses with geometric or orientation information. HC-Net (Wang et al., 2023) uses correlation-aware homography estimation; FG^2 (Xia & Alahi, 2025) and Loc^2 (Xia et al., 2026) recover position or pose from fine-grained local correspondences. GeoHypLoc complements these methods' local matching capabilities: building on officially pretrained CCVPE, it studies how retrieved aerial windows with independent georeferencing can jointly participate in stage-wise localization.

Unified large-scale fine-grained geo-localization Unified frameworks connect city-scale retrieval with local metric localization. UnifyGeo (Song et al., 2025) performs fine-grained matching and reranking of retrieval candidates, then localizes relative to a selected aerial reference. CIPER (Jeon et al., 2026) integrates retrieval and three-degree-of-freedom pose estimation in one architecture, with pose prediction primarily conditioned on the highest-similarity reference. C-BEV (Fervers et al., 2023) and semantic ambiguity modeling (Feng et al., 2025) address matching ambiguity through spatial representations and uncertainty propagation, respectively. GeoHypLoc organizes candidates with high query visual similarity and geographic proximity into a positively spatially associated reference set, allowing multiple members to continually provide localization evidence in a common geographic space.

A EXACT CONSTRUCTION OF THE POSITIVELY SPATIALLY ASSOCIATED REFERENCE SET

A.1 FULL-GALLERY GEOGRAPHIC GRAPH AND INDUCED SUBGRAPH

The fixed, officially pretrained Sample4Geo retriever (Deuser et al., 2023) returns query q's actual Top-5 candidates $\mathcal{C}_q = \{r_i\}_{i=1}^{5}$ from the full aerial gallery. Each reference has a georeferenced center c_i. Adjacency over the complete gallery is precomputed using the RNG definition (Toussaint, 1980), with the same condition as in Section 2.1: an edge (u, v) is added when no third gallery node is closer to both u and v than they are to each other.

For the current candidates, only edges already present in the full-gallery RNG are extracted to form the induced subgraph $G_{\mathrm{RNG}}[\mathcal{C}_q]$; an RNG is not rebuilt on the current Top-5. Connected components form the candidate reference-set family $\mathcal{H}_q$. Gallery geographic connectivity defines the groups' spatial structure, and query visual support determines which group is selected.

A.2 Query visual support and reference-set selection

Let s_i be the embedding similarity between the query and candidate r_i. Standardization uses only the current Top-5 list, whose mean and population standard deviation are

$$\mu_q = \frac{1}{5}\sum_{i=1}^{5} s_i, \qquad \sigma_q = \left[\frac{1}{5}\sum_{i=1}^{5}(s_i - \mu_q)^2\right]^{1/2}. \tag{A.1}$$

As defined in Section 2.1, standardized scores and relative query visual support are

$$z_i = \frac{s_i - \mu_q}{\max(\sigma_q, \epsilon)}, \qquad p_i = \frac{\exp(z_i)}{\sum_{k\in\mathcal{C}_q}\exp(z_k)}. \tag{A.2}$$

The standard deviation is lower-bounded by $\epsilon = 10^{-6}$, so the denominator is $\max(\sigma_q, \epsilon)$. For exponential normalization, we first subtract the list maximum $z_{\max}$, then clip the exponential's input to $[-50, 0]$ to control its numerical range. Candidate supports are normalized within the same retrieval list. We sum member supports for each connected component and select the reference set with the largest total:

$$\mathrm{MASS}(H) = \sum_{r_i\in H} p_i, \qquad H^* = \arg\max_{H\in\mathcal{H}_q} \mathrm{MASS}(H). \tag{A.3}$$

If support totals are exactly equal, we deterministically select the component whose best-ranked member has the higher retrieval rank. p_i expresses relative query visual support within the current retrieval list, not a calibrated probability of correct localization. Reference-set construction uses no query ground-truth location, Strong/Semi identity, or test label; these annotations are used only for the offline composition statistics in Section 4.1.

A.3 Geographic footprint convention

Each reference retains its original aerial image and independent georeferencing. After mapping its center to normalized aerial-tile coordinates c_j^q as defined in Eq. (B.1), one original tile corresponds to one coordinate unit. Its geographic footprint is

$$F_j = \left[c_{j,x}^q - \frac{1}{2}, c_{j,x}^q + \frac{1}{2}\right] \times \left[c_{j,y}^q - \frac{1}{2}, c_{j,y}^q + \frac{1}{2}\right]. \tag{A.4}$$

This square represents the reference image's real-world geographic coverage, distinct from the pixel dimensions after network input resizing. The selected region is the union of member footprints, $\bigcup_{j\in H^*} F_j$. A discrete reference-identity hit and inclusion of the ground truth in this union are different tests; connectivity in the gallery RNG is also not directly equivalent to geometric properties of the footprint union. Appendix B defines the common canvas and validity masks; Section C.5 specifies city-specific metric scales.

Figure A.1 separately marks the identity of the official Strong reference and the geographic coverage of selected members, explaining the distinction between the two tests in Section 4.1. Even if Strong is absent from H^*, the union of other overlapping members' footprints may still contain the ground truth. Missing reference identity and missing geographic support should therefore be assessed separately. Figure 2 describes construction of H^*.

B Geographically aligned multi-reference joint localization

This appendix specifies the multi-stage localization path, coordinate mapping, valid coverage handling, and position-dependent contribution estimation in Section 2.2. Let H_q^* denote the current query's reference set and K_q its member count; we abbreviate it to H^* when the query is clear. The world canvas depends only on selected aerial references and their gallery georeferencing, not on query ground truth or Strong/Semi annotations.



Figure A.1: Representative cases of H^* geographic support. (a) The official Strong reference is retained, and the ground truth lies within the selected footprint union. (b) Strong is not retained, but other H^* members still cover the ground truth. (c) The ground truth lies outside the selected footprint union, indicating missing geographic support. Red borders denote H^* members, and the red region denotes their coverage union. In (b), a gray dashed box marks the unselected official Strong reference; green stars indicate ground truth. The figure distinguishes reference-identity hits from geographic coverage rather than repeating reference-set construction.

B.1 Member-wise CCVPE localization path

GeoHypLoc retains the coarse-to-fine localizer and decoder architecture of officially pretrained CCVPE. The ground panorama passes through the ground encoder once, and six stage-specific descriptor mappings produce $d_1^{\mathrm{grd}}, \ldots, d_6^{\mathrm{grd}}$. Each aerial reference in H^* is processed independently by the shared aerial encoder, producing multiscale features $A_{j,s}$. Later stages take the corresponding aerial skip features from these outputs without re-encoding the input images.

S1 constructs aerial descriptor features $X_{j,1}$ on each reference's local grid and matches them with d_1^{grd} to obtain $R_{j,1}$, then aligns both to the common world canvas. At S2–S6, the previous decoded and upsampled shared state is expanded over members, masked by each member's valid coverage, and concatenated channel-wise with current-scale aerial skip features already aligned to the same world grid. A convolutional mapping shared among members at that stage produces the localization state $\widetilde{X}_{j,s}$, which is matched with d_s^{grd} to obtain $\widetilde{R}_{j,s}$. The localization state is a multichannel feature formed before matching; it and the matching response are used together for subsequent member-evidence encoding and aggregation.

Table B.1: City-specific affine coefficients for geographic alignment.

City	Diagonal coefficient
New York	640.2926235139126
Seattle	640.297117647543
San Francisco	640.1139133849562
Chicago	640.022584833235

Geographic alignment acts on member-wise responses and descriptor features after matching in S1, but on aerial skip features before matching in S2–S6. Every stage aggregates valid members after matching. S1 uses masked averaging; S2–S6 additionally perform same-position member interaction before aggregation. Aggregated responses and states jointly pass through the decoder to form the next shared input. Multiple references thus jointly update the shared state throughout localization; Section B.6 specifies the recurrence.

B.2 World canvas, coordinate mapping, and valid coverage

The same local feature index in different references need not denote the same real-world position. To give cross-reference comparisons a consistent physical meaning, we map localization evidence onto a common world canvas using each reference's georeferencing. Let $c_j \in \mathbb{R}^2$ be member j's gallery geographic center, and let A_c, b_c be the city-specific affine coefficients used by the VIGOR localization implementation. Taking the mean member center as the origin of reference, normalized aerial-tile coordinates are defined by

$$\bar{c} = \frac{1}{K_q}\sum_{j=1}^{K_q} c_j, \qquad c_j^q = (c_j - \bar{c} - b_c)A_c^{-1}. \tag{B.1}$$

One normalized coordinate unit corresponds to one original VIGOR aerial tile. Network inputs are resized to 512×512, but the metric scale remains determined by the original 640×640 tile (Section C.5). A ground-truth geographic position x^* is converted to $q^* = (x^* - \bar{c} - b_c)A_c^{-1}$ only for supervision; it does not enter canvas construction. A member footprint is a unit square in this coordinate system:

$$F_j = \left[c_{j,x}^q - \frac{1}{2}, c_{j,x}^q + \frac{1}{2}\right] \times \left[c_{j,y}^q - \frac{1}{2}, c_{j,y}^q + \frac{1}{2}\right]. \tag{B.2}$$

City affine coefficients are provided by `CROSS_GALLERY_ONLY_AFFINE.json`, constructed using only aerial gallery metadata. Same-area and Cross-area use the same four-city parameters. All cities satisfy $A_c = a_c I_2$ and $b_c = (0, 0)$; Table B.1 lists the values. Normalized coordinates q correspond to geographic coordinates $x = qA_c + \bar{c} + b_c$. These affine coefficients are used for coordinate alignment; metric errors are still computed using κ_c in Section C.5.

The canvas origin is the coordinate-wise lower bound of all footprints:

$$o_q = \left(\min_j c_{j,x}^q - \frac{1}{2}, \min_j c_{j,y}^q - \frac{1}{2}\right). \tag{B.3}$$

The required normalized extents along the two axes are

$$s_x = \max\left(\max_j c_{j,x}^q + \frac{1}{2} - o_{q,x}, 1\right), \quad s_y = \max\left(\max_j c_{j,y}^q + \frac{1}{2} - o_{q,y}, 1\right). \tag{B.4}$$

To satisfy decoder size constraints, canvas width and height are rounded up to multiples of 64:

$$W = 64\left\lceil \frac{512 s_x}{64} \right\rceil, \qquad H = 64\left\lceil \frac{512 s_y}{64} \right\rceil. \tag{B.5}$$

Padding extends only the upper coordinate bounds, preserving o_q without recentering the reference set. The padded horizontal and vertical ranges are $[o_{q,x}, o_{q,x}+W/512]$ and $[o_{q,y}, o_{q,y}+H/512]$. For $K_q = 1$, an exact single-reference fast path uses $H = W = 512$ and $o_q = (-0.5, -0.5)$, preserving the original single-reference geometry. S1–S6 represent the same physical canvas at different grid resolutions. Their member-aggregation grids are $(H/64, W/64)$, $(H/32, W/32)$, $(H/16, W/16)$, $(H/8, W/8)$, $(H/4, W/4)$, and $(H/2, W/2)$, respectively; the final decoder outputs a response on an $H \times W$ grid. Each stage samples both axes uniformly over these ranges, including both endpoints. The physical canvas extent remains fixed as stage resolution changes.

For world position $x = (x_x, x_y)$, the local coordinates within member j's unit square and the sampling coordinates are

$$u = x_x - c^q_{j,x} + \frac{1}{2}, \quad v = x_y - c^q_{j,y} + \frac{1}{2}, \qquad g_x = 2u - 1, \quad g_y = 2v - 1. \tag{B.6}$$

S1 member responses and descriptor features, and S2–S6 aerial skip features, are mapped to the corresponding world-canvas resolution by bilinear resampling. Binary valid-support masks use nearest-neighbor resampling. All sampling uses zero padding and `align_corners=True`. The effective validity mask is the intersection of the geometric conditions $0 \le u \le 1$, $0 \le v \le 1$ and the resampled native valid-support mask, denoted $M_{j,s}(x)$. S1 response and state mappings are

$$\widetilde{R}_{j,1} = \mathcal{W}_{j,1}(R_{j,1}), \qquad \widetilde{X}_{j,1} = \mathcal{W}_{j,1}(X_{j,1}),$$
$$\mathcal{V}_s(x) = \{j \mid M_{j,s}(x) = 1\}. \tag{B.7}$$

The valid-member set $\mathcal{V}_s(x)$ in Eq. (B.7) applies to all six stages. At S2–S6, the mapping acts on current-scale aerial skip features, yielding $\widetilde{A}_{j,s} = \mathcal{W}_{j,s}(A_{j,s})$. The resulting member states $\widetilde{X}_{j,s}$ and responses $\widetilde{R}_{j,s}$ are already on the common canvas and are not remapped from local grids to the world grid. Tildes consistently denote member quantities in common coordinates.

Cross-reference interaction occurs only among valid members at the same world position. Interpolation and padding outside footprints do not constitute valid localization evidence. Geographic alignment changes coordinate representation without introducing a query-location prior.

Each stage in Figure 2(b) uses a common canvas with origin o_q and the same physical extent; padding extends only upper coordinate bounds. S1 geographically aligns local responses and descriptor features and applies masked averaging. S2–S6 first align aerial skip features to the current world grid, then combine them with member-masked copies of the shared state to construct member states for matching. References interact only at jointly valid coverage of the same real-world location. Decoder updates to aggregated results drive the next stage.

B.3 Masked-mean aggregation in S1

The coarsest stage S1 applies masked averaging only to valid members: when $|\mathcal{V}_1(x)| > 0$, $w_{j,1}(x) = 1/|\mathcal{V}_1(x)|$ for $j \in \mathcal{V}_1(x)$. The same weights aggregate responses and states, as in Eq. (B.10). A single valid member passes through its response and state exactly; with no valid member, there is no contribution and the position remains invalid. S1 does not directly use learned contributions, stabilizing the initial shared localization state and avoiding arbitrary member preferences from random initialization at the coarsest scale.

S1 averaging produces a response R_1 and state X_1 on the common canvas, both on an $H/64 \times W/64$ grid. Concatenating R_1 and normalized X_1 channel-wise and applying the first decoder update and upsampling yields the shared state $\widehat{X}_1$ on an $H/32 \times W/32$ grid for S2. This update preserves the full geographic extent of the common canvas, as specified in Eq. (B.11).

B.4 Member representations

At S2–S6, position-dependent contributions are jointly estimated from member matching responses, localization states, and coverage information. The localization state $\widetilde{X}_{j,s}$ is produced by the current-stage shared convolution described in Sections B.1 and B.6 and is also used for matching with the ground descriptor. Member representations use this convolutional feature directly; states are not

regenerated from response maps. A learnable 3×3 convolution first embeds single-channel local response patches from $\widetilde{R}_{j,s}$ into eight-dimensional features. Let $v_j = \widetilde{X}_{j,s}(x)$ be the member state vector and $\bar{v}$ the mean state of valid members at the same position. States with arbitrary channel counts are compressed into six statistics:

$$[\mathrm{mean}(v_j), \mathrm{std}(v_j), \mathrm{RMS}(v_j), \mathrm{maxabs}(v_j), \cos(v_j, \bar{v}), \mathrm{RMS}(v_j - \bar{v})]. \tag{B.8}$$

The same-position mean $\bar{v}$ averages all valid members, including the current member j; it is not a leave-one-out mean. All six statistics in Eq. (B.8) are computed over the channel vector at the current world position; std and cos denote the standard-deviation and cosine-similarity operations used directly by the implementation. Mean, standard deviation, root mean square (RMS), and maximum absolute value characterize state magnitude; cosine similarity and RMS difference characterize its relationship to the same-position mean state. LayerNorm, a $6 \to 8$ linear mapping, and a Gaussian error linear unit (GELU) transform these six statistics into an eight-dimensional state feature. Two scalars are added: valid coverage fraction in a 3×3 neighborhood and valid-member count normalized by the maximum count of five. Concatenating these features yields an 18-dimensional vector, which is linearly projected to a 16-dimensional member representation $e_{j,s}(x)$. Retrieval similarity, retrieval rank, and Strong/Semi identity do not enter the contribution network.

B.5 Position-dependent member interaction

Set interaction is performed only at world positions with at least two valid members. Given $\{e_{j,s}(x)\}_{j \in \mathcal{V}_s(x)}$, two-head self-attention with embedding dimension 16 is followed by a residual connection and LayerNorm, a $16 \to 32 \to 16$ feed-forward network (FFN) with GELU, a second residual connection and LayerNorm, and a shared scalar scoring head g_s. Attention models relative localization evidence within the member set at a common real-world position, rather than spatial attention between image patches (Lee et al., 2019). Let $h_{j,s}(x)$ denote the interacted representation. Contribution weights are normalized by masked softmax over valid members:

$$w_{j,s}(x) = \frac{\exp\{g_s(h_{j,s}(x))\}}{\sum_{k \in \mathcal{V}_s(x)} \exp\{g_s(h_{k,s}(x))\}}, \quad j \in \mathcal{V}_s(x), \quad s = 2, \ldots, 6. \tag{B.9}$$

The scalar scoring head is zero-initialized, making S2–S6 exactly equivalent to uniform weighting over valid members at the start of training. Supervision subsequently learns contributions that vary with position and stage. With only one valid member, set interaction is bypassed and its weight is one; with none, there is no contribution. Invalid members do not participate in normalization.

B.6 Synchronized response/state aggregation and recursive localization

The same $w_{j,s}(x)$ simultaneously aggregates geographically aligned matching responses and localization states:

$$R_s(x) = \sum_{j \in \mathcal{V}_s(x)} w_{j,s}(x) \widetilde{R}_{j,s}(x), \quad X_s(x) = \sum_{j \in \mathcal{V}_s(x)} w_{j,s}(x) \widetilde{X}_{j,s}(x). \tag{B.10}$$

To distinguish aggregated states from decoded shared inputs, let X_s denote the stage-s aggregated state and $\widehat{X}_s$ the shared state after decoding and upsampling. The aggregated response and normalized aggregated state are concatenated channel-wise and passed jointly to the original CCVPE decoder:

$$\widehat{X}_s = D_s(\mathrm{Concat}_C[R_s, \mathrm{Norm}(X_s)]). \tag{B.11}$$

Here, Concat_C denotes channel concatenation, Norm is CCVPE's original feature normalization, and D_s is the stage-specific decoder update and upsampling operator. In S1, this operation corresponds to `deconv6`; subsequent stages follow the original decoder sequence. The shared state's grid becomes finer across stages while its physical canvas extent remains fixed.

For stage $s \in \{2, \ldots, 6\}$, each member's current-scale aerial skip features are first aligned to the target grid:

$$\widetilde{A}_{j,s} = \mathcal{W}_{j,s}(A_{j,s}). \tag{B.12}$$

The previous shared state $\widehat{X}_{s-1}$ is expanded along the member dimension and masked by $M_{j,s}$. This expands the tensor from $[B, C_{\mathrm{dec}}, h_s, w_s]$ to $[B, K_q, C_{\mathrm{dec}}, h_s, w_s]$, where B is batch size and h_s, w_s are the common-grid dimensions at the current stage. Each branch retains the same complete canvas, but only its valid coverage contributes. Each member's shared-state copy is concatenated with $\widetilde{A}_{j,s}$ channel-wise on the same world grid. Shared convolution followed by reapplication of the validity mask produces the localization state:

$$\widetilde{X}_{j,s} = M_{j,s} \odot \phi_s\Big(\mathrm{Concat}_C[M_{j,s} \odot \widehat{X}_{s-1}, \widetilde{A}_{j,s}]\Big). \tag{B.13}$$

ϕ_s is the current-stage convolutional mapping shared among members; $\odot$ applies the spatial mask with broadcasting across channels. Both inputs have spatial size $h_s \times w_s$. Concatenation produces $[C_{\mathrm{dec}} + C_{A,s}, h_s, w_s]$, which convolution maps to the required $[C_s, h_s, w_s]$. Each member state is then matched with the same stage-specific ground descriptor:

$$\widetilde{R}_{j,s} = \mathrm{Match}_s(d_s^{\mathrm{grd}}, \widetilde{X}_{j,s}). \tag{B.14}$$

The localization state $\widetilde{X}_{j,s}$ exists before matching. Its matching response $\widetilde{R}_{j,s}$, the state itself, and coverage information jointly enter member-evidence encoding and contribution estimation in Sections B.4–B.5. The same contribution weights aggregate responses and states via Eq. (B.10); Eq. (B.11) jointly updates them into the next shared input. Member-wise aerial skips provide current-scale detail, while the shared state carries previously fused multi-reference information. This continues through S6, whose decoder update restores the $H \times W$ grid; a final output convolution then produces Z_0.

C Response correction and DARK coordinate decoding

C.1 Multi-stage localization evidence

Localization-derived spatial evidence E_{loc} summarizes position-dependent reference contributions from S2–S6 separately from retrieval similarity. At position x, let $n_s(x) = \sum_j \mathbf{1}[w_{j,s}(x) > 0]$ be the valid-member count at stage s, and sort contributions in descending order as $w_{(1),s}(x), w_{(2),s}(x), \ldots$. The five-channel statistics vector is

$$S_s(x) = (m_s(x), d_s(x), c_s(x), \nu_s(x), o_s(x)) \in \mathbb{R}^5. \tag{C.1}$$

The maximum contribution is $m_s = w_{(1),s}$, and the Top-2 margin is $d_s = w_{(1),s} - w_{(2),s}$. Entropy confidence is computed from normalized contributions:

$$\bar{w}_{j,s} = \frac{w_{j,s}}{\max(\sum_k w_{k,s}, 10^{-12})}, \qquad H_s = -\sum_j \bar{w}_{j,s} \log \max(\bar{w}_{j,s}, 10^{-12}). \tag{C.2}$$

$$c_s = \begin{cases} \mathrm{clip}_{[0,1]}(1 - H_s/\log n_s), & n_s \geq 2, \\ 1, & n_s < 2, \end{cases}$$

$$\nu_s = \mathrm{clip}_{[0,1]}(n_s/5), \qquad o_s = \mathbf{1}[n_s \geq 2]. \tag{C.3}$$

The five channels describe contribution concentration, relative dominance, entropy confidence, normalized member count, and overlap status, respectively. With one valid member, its contribution is one and the second-largest contribution is set to zero, yielding $S_s = (1, 1, 1, 0.2, 0)$. With no valid member, $S_s = (0, 0, 1, 0, 0)$, and the position remains excluded by the validity mask. This convention introduces no virtual member. Each stage has an independent lightweight encoder $\mathcal{E}_s$: a $5 \to 8$ pointwise (1×1) convolution, SiLU, and an $8 \to 1$ pointwise convolution, with no activation after the output convolution. Let $U_s = \mathcal{E}_s(S_s)$. Bilinear interpolation resizes it to the final $H \times W$ world grid with `align_corners=True`. Stage outputs are aggregated using learned weights:

$$\alpha_s = \frac{\exp \eta_s}{\sum_{t=2}^{6} \exp \eta_t}, \qquad E_{\mathrm{loc}}(x) = \sum_{s=2}^{6} \alpha_s U_s(x). \tag{C.4}$$

All stage parameters η_s are initialized to zero, so initial stage weights are $1/5$. Stage weights α_s combine evidence across scales, whereas member contributions $w_{j,s}(x)$ aggregate references at the same position within a stage; these are different operations.

C.2 Retrieval-derived region-level query visual support

Using the relative query visual support p_i from Section 2.1, we retain only Top-5 references in the same gallery city as H^* to form $\mathcal{C}_q^{H^*}$. Projecting support onto each reference's real-world geographic footprint F_i gives

$$G_q(x) = \sum_{i \in \mathcal{C}_q^{H^*}} p_i \mathbf{1}[x \in F_i]. \tag{C.5}$$

The implementation clips cumulative support to $[0, 1]$ to preserve its numerical range. City information comes from gallery metadata of retrieved aerial images, not from query ground truth. p_i expresses relative visual support within a retrieval list and is not a calibrated probability. Unlike E_{loc}, G_q comes from the frozen retriever's visual support for geographic regions rather than the localizer's internal member contributions. Let $\mathcal{V}$ be the valid positions on the final world canvas. Define

$$g(x) = \log \max(G_q(x), 10^{-12}), \quad \widetilde{g}(x) = g(x) - \max_{y \in \mathcal{V}} g(y), \quad \bar{g}(x) = \mathrm{clip}_{[-4,0]}(\widetilde{g}(x)/6). \tag{C.6}$$

C.3 Two residual branches for response correction

First, the base response Z_0 is converted to log probabilities over the valid region and used to construct a bounded input:

$$\ell_0(x) = \log \frac{\exp Z_0(x)}{\sum_{y \in \mathcal{V}} \exp Z_0(y)}, \qquad \bar{\ell}_0(x) = \mathrm{clip}_{[-2,0]}(\ell_0(x)/6). \tag{C.7}$$

The localization-evidence and retrieval-support branches take $h_{\mathrm{loc}} = (\bar{\ell}_0, \tanh E_{\mathrm{loc}})$ and $h_{\mathrm{ret}} = (\bar{\ell}_0, \tanh E_{\mathrm{loc}}, \bar{g})$, respectively. Each branch comprises a 1×1 convolution, SiLU, a 3×3 depthwise convolution, SiLU, and a 1×1 output convolution. Both have eight intermediate channels and one output channel, with two and three input channels, respectively. Their residuals are

$$\Delta_{\mathrm{loc}}(x) = \beta E_{\mathrm{loc}}(x) + F_{\mathrm{loc}}(h_{\mathrm{loc}}(x)),$$

$$\Delta_{\mathrm{ret}}(x) = \gamma \bar{g}(x) + F_{\mathrm{ret}}(h_{\mathrm{ret}}(x)). \tag{C.8}$$

$$Z(x) = Z_0(x) + \Delta_{\mathrm{loc}}(x) + \Delta_{\mathrm{ret}}(x). \tag{C.9}$$

β, γ, and both branches' output layers are zero-initialized, so $Z = Z_0$ at training start. Residuals at invalid canvas positions are set to zero, and subsequent decoding remains constrained by the validity mask. Training uses the localization supervision and peak-margin constraint in Section 2.3, detailed in Section C.6.

Figure C.1 relates intermediate evidence to response changes. E_{loc} derives from multi-stage member-contribution statistics; G_q projects retrieval scores onto aerial reference footprints. Both evidence types jointly construct correction residuals using the inputs in Section C.3. Comparing Z_0, Z, and their DARK predictions for the same query, reference set, and geographic extent illustrates relationships among evidence, responses, and location estimates.

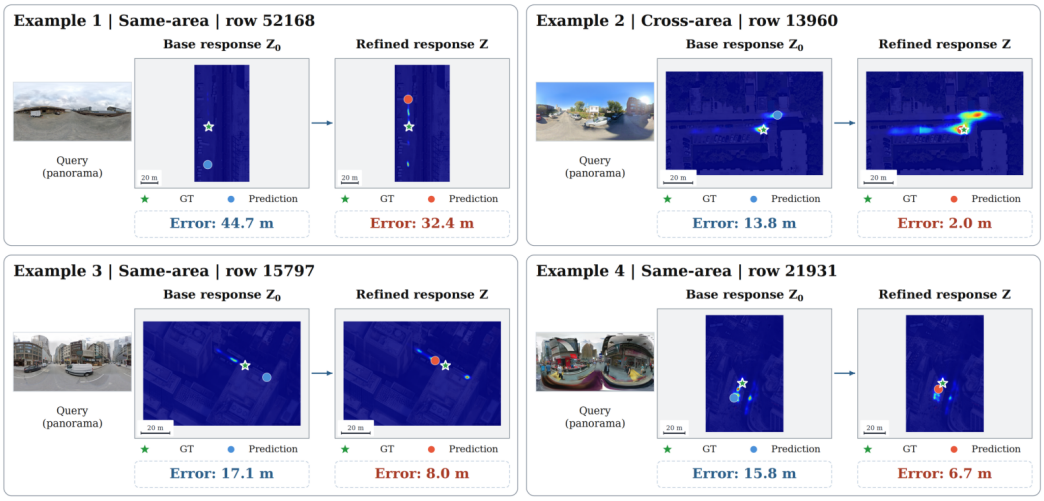

Figure C.1: Effects of response-correction evidence. Base and corrected responses, ground truth, predicted locations, and metric errors are compared for the same query. Predictions are obtained from $\mathrm{DARK}(Z_0)$ and $\mathrm{DARK}(Z)$, respectively. The figure focuses on intermediate evidence and response changes.

Table C.1: Metric coordinate conversion for VIGOR's four cities.

City	s_c (m/original pixel)	κ_c (m/normalized tile unit)
New York	0.113248	72.47872
Seattle	0.100817	64.52288
San Francisco	0.118141	75.61024
Chicago	0.111262	71.20768

C.4 DARK COORDINATE DECODING

DARK (Zhang et al., 2020) is applied directly to the corrected response Z to obtain the final predicted position q_D. The decoder first smooths the spatial probability map with a 5×5 Gaussian kernel of standard deviation one, lower-bounds probabilities by 10^{-12} before taking logarithms, and initializes the peak at the maximum smoothed log response. A Hessian-based offset is used only if this position is at least two pixels from the response-map boundary, the Hessian determinant is finite with absolute value greater than 10^{-9}, and the computed offset is finite. Each offset component is clipped to $[-1, 1]$ response pixels. If any condition fails, the initial peak position is retained. The same decoder is used for base and corrected responses:

$$q_0 = \mathrm{DARK}(Z_0), \qquad q_D = \mathrm{DARK}(Z). \tag{C.10}$$

C.5 METRIC COORDINATE CONVERSION

Coordinates q are measured in normalized aerial-tile units. For city c, let s_c be the meters-per-pixel scale of the original aerial image. Then $\kappa_c = 640 s_c$ is the number of meters per normalized tile unit. Here, 640 is the original aerial tile's pixel dimension; resizing network inputs to 512×512 does not change this physical metric definition.

For query q, define the metric localization error as a function of a predicted position $\hat{q}$:

$$e_q(\hat{q}) = \kappa_c \|\hat{q} - q^*\|_2. \tag{C.11}$$

C.6 Objective and gradient boundaries

We retain CCVPE's (Xia et al., 2024) localization cross-entropy, six-stage InfoNCE matching supervision, and orientation supervision, adding the main paper's peak-margin constraint:

$$\mathcal{L}_{\mathrm{CCVPE}} = \mathcal{L}_{\mathrm{CE}} + \frac{10^4}{6}\sum_{s=1}^{6}\mathcal{L}_{\mathrm{InfoNCE},s} + 10\mathcal{L}_{\mathrm{ori}},$$
$$\mathcal{L}_{\mathrm{loc}} = \mathcal{L}_{\mathrm{CCVPE}} + 0.3\mathcal{L}_{\mathrm{peak}}. \tag{C.12}$$

Each of the six InfoNCE losses is computed within the current device's batch, treating locations with target response greater than 10^{-2} as valid positive supervision. If a stage has no valid positive supervision in that batch, its loss returns a differentiable zero that preserves the computation graph and is still included in the six-stage average. Queries are not dropped and the whole batch is not skipped. This handling applies to empty InfoNCE supervision.

The peak-margin loss compares positive and negative regions only within the valid canvas. Locations within 1 m of ground truth form the positive region; those at least 3 m away form the negative region. The intervening 1–3 m annulus is excluded from both peak computations:

$$p^+ = \max_{x\in\mathcal{V}:\,\kappa_c\|x-q^*\|_2\le 1\,\mathrm{m}} Z(x), \quad p^- = \max_{x\in\mathcal{V}:\,\kappa_c\|x-q^*\|_2\ge 3\,\mathrm{m}} Z(x),$$
$$\mathcal{L}_{\mathrm{peak}} = \mathrm{softplus}(0.5 + p^- - p^+). \tag{C.13}$$

Distances are converted to meters as in Section C.5. The peak-margin loss is computed directly on the corrected response Z, and gradients propagate through both correction branches into the localization path.

The Sample4Geo retriever, Top-5 lists, full-gallery RNG, and discrete selection of H^* are outside backpropagation. The localizer is initialized from official CCVPE weights for the corresponding protocol. Position-dependent contributions and response-correction modules are optimized by the localization objective. The contribution scoring head starts from uniform valid-member weights, and both response residuals start from zero output.

DARK decodes responses into predicted coordinates and is outside backpropagation. Training optimizes the localization path and correction modules using CCVPE supervision and the response-domain peak-margin loss; the retrieval-support map G_q remains fixed. Appendix D provides the complete schedule and optimization settings.

D Training and evaluation details

D.1 Fixed and trainable computation graphs

The Sample4Geo retriever, its Top-5 candidates, the full-gallery RNG, and discrete construction of H^* remain fixed and outside backpropagation. The CCVPE localizer is initialized from official pretrained weights for the corresponding evaluation protocol and participates in localization training. The position-dependent reference-contribution module and response-correction branches are newly trainable components. Appendix C.6 specifies losses and gradient boundaries.

D.2 Common training protocol

All newly trained configurations use five epochs, and formal evaluation uses only the predetermined epoch-5 checkpoint (E5). Intermediate validation or test performance is not used for model selection. Table D.1 summarizes optimization, augmentation, and evaluation settings; Section D.4 distinguishes the reference inputs and output paths of each configuration.

Multi-reference joint training uses only queries with at least one valid member whose selected footprints include the ground-truth position. Coverage is tested on both normalized local-coordinate axes using the numerical tolerance $[-10^{-7}, 1+10^{-7}]$. Each epoch is one traversal of the eligible queries by a sampler bucketed by member count and canvas size. Last-batch completion and distributed step-count padding may repeat some queries; an epoch therefore does not mean exactly one

Table D.1: Common training and evaluation settings. Learning rates apply only to modules optimized in the corresponding configuration.

Item	Setting
Epochs and model selection	5 epochs; predetermined E5; no intermediate-epoch selection
Random seeds	Training seed 17; independent augmentation sampler seed 17
Optimizer	Adam; $\beta = (0.9, 0.999)$; weight decay 0
Learning rates and schedule	CCVPE: 1×10^{-5}; reference contribution: 2×10^{-4}; response correction: 1×10^{-4}; no learning-rate schedule
Gradient clipping	Maximum global gradient norm 5
Numerical precision	FP32 for training and evaluation; automatic mixed precision (AMP) disabled
Batch normalization	Fixed BatchNorm running statistics; trainable affine parameters
Training orientation	Complete 20-bin orientation representation
Training batches and resources	Adaptive bucketing by member count and canvas size; per GPU: maximum batch size 2, at most 10 aerial images, and at most 1,800,000 canvas pixels
Data loading	1 worker per GPU for both training and formal evaluation
Member order	Randomly shuffled during multi-reference training
Member dropout	Attempt to drop at most 1 member with probability 0.50; retain at least 1 member
Photometric augmentation	Brightness, contrast, and saturation independently sampled from $U[0.92, 1.08]$
Gaussian blur	Probability 0.10; blur radius sampled from $U[0.10, 0.80]$
Geometric augmentation	No random cropping, flipping, rotation, or translation
Retriever and candidate budget	Fixed officially pretrained Sample4Geo; full-gallery Top-5
Formal evaluation	Known orientation; batch size 2 per GPU; no test-time augmentation
Complete test population	Same-area: 52,605 queries; Cross-area: 53,694 queries

occurrence of every query. This trainability filter applies only during training. Formal evaluation retains all test queries in each protocol.

All retained aerial members for a query share one aerial photometric transformation; the ground panorama uses an independent photometric transformation. After member dropout, at least one member must remain, and the valid grid must contain a position within 1 m of ground truth, tested with a 10^{-6} m numerical tolerance. Otherwise, the pre-dropout member set is restored. In distributed training, if dropout on any device fails the supervision condition, all devices synchronously restore their pre-dropout sets for that step. All augmentation is disabled during formal evaluation.

D.3 Correct-reference evaluation protocol

Table 2 directly provides each query's official Strong aerial reference, excluding retrieval errors. GeoHypLoc (Strong-only) and GeoHypLoc use the same architecture: the former trains with one Strong reference, while the latter performs geographically aligned joint training with the complete H^*. At test time, both receive the same official Strong reference and, with known orientation, directly decode the base response using $q_0 = \mathrm{DARK}(Z_0)$, without multi-reference joint inference or response correction. Both are evaluated over the corresponding protocol's complete test population. The comparison assesses the combined effect of training reference composition and organization.

D.4 Exact definitions of Base, A, B, C, and D

Table D.2 corresponds to the actual retrieval conditions in Section 3.3. Let $K_q = |H_q^*|$. Training references in A and B are used as independent single-reference samples; C and D receive the entire reference set jointly. Base, A, and B receive only Sample4Geo Top-1 at test time, whereas C and D receive the complete H^*. All inputs come from the same fixed retrieval results.

D corresponds to GeoHypLoc in Table 1, and Base to Sample4Geo+CCVPE. A versus B compares reference composition under independent training. B versus C changes both training organization

Table D.2: Training, test, and output protocols for the configurations compared in the main paper.

Configuration	Training references	Training form
Base	Original official CCVPE pretraining configuration	Original model retained
A: Rank-matched independent	First K_q references from actual Top-5 by retrieval rank	Independent single-reference training
B: H^* independent	All members of H^*	Independent single-reference training
C: H^* joint localization	Complete H^*	Geographically aligned multi-reference joint training
D: Full GeoHypLoc	Complete H^*	Geographically aligned multi-reference joint training

Configuration	Test input	Localization	Response correction	Decoding and output
Base	Sample4Geo Top-1	Single-reference	Disabled	Original CCVPE output
A: Rank-matched independent	Sample4Geo Top-1	Single-reference	Disabled	Direct decoding of base response
B: H^* independent	Sample4Geo Top-1	Single-reference	Disabled	Direct decoding of base response
C: H^* joint localization	Complete H^*	Multi-reference joint localization	Disabled	$\mathrm{DARK}(Z_0)$
D: Full GeoHypLoc	Complete H^*	Multi-reference joint localization	Enabled	$\mathrm{DARK}(Z)$

and test input, so their difference cannot be attributed solely to a training mechanism or submodule. C and D both use DARK, applied to the base response Z_0 and corrected response Z, respectively.

E Definitions for supplementary analyses

E.1 Strong secondary peaks and response-correction gains

Within valid coverage, we compute log-softmax of the base response Z_0 in FP32 and exponentiate it to obtain the probability response P_0, without DARK's Gaussian smoothing. The valid position with maximum probability is the primary peak; ties are broken by the first index in row-major order. Candidates are examined in descending probability order, and the highest-probability local peak at least 2 m from the primary peak is selected as the strongest secondary peak. Let x_1, x_2 denote their positions. Define

$$\rho = \frac{P_0(x_2)}{P_0(x_1)}. \tag{E.1}$$

Queries with $\rho \geq 0.25$ form Strong-secondary; the rest form Remaining. Grouping uses only the base response before correction and remains fixed in subsequent comparisons. Remaining is not equivalent to a strictly unimodal response.

Local peaks are detected within a circular neighborhood of radius 1 m, comparing probabilities only in its intersection with valid coverage; neighborhoods are truncated at boundaries. A candidate is a local peak if its probability is no less than the neighborhood maximum minus 10^{-15}, allowing tied maxima on plateaus. No additional non-maximum suppression is used. If no local secondary peak satisfies the separation condition, its probability and ρ are set to zero, its distance is undefined, and the query is assigned to Remaining. When computing the ratio, the primary-peak denominator is lower-bounded by 10^{-300}.

For a fixed query group Q, let $q_0 = \mathrm{DARK}(Z_0)$ and $q_D = \mathrm{DARK}(Z)$ and use the main paper's metric error definition. Response-correction gain is

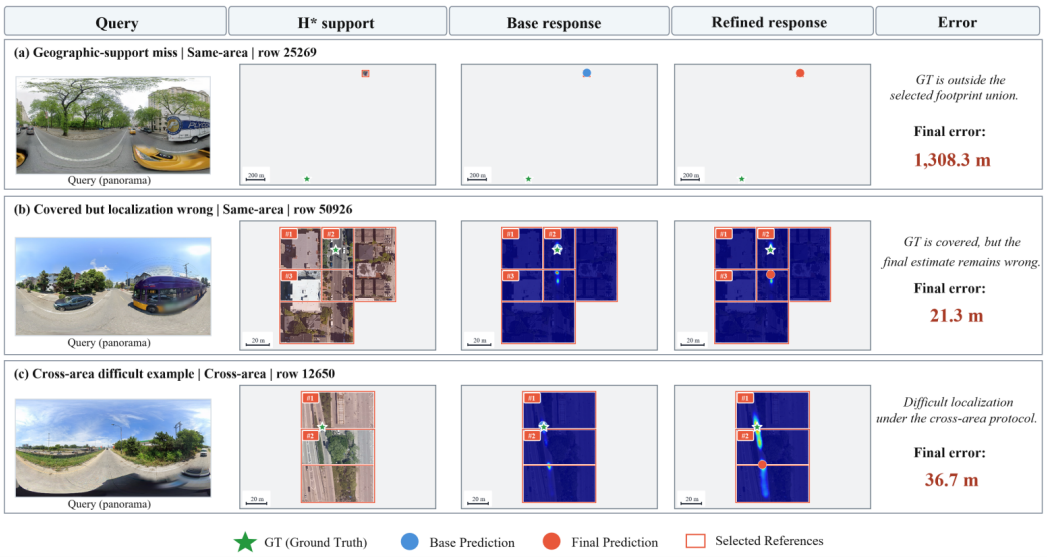

Figure E.1: Representative GeoHypLoc failures. (a) Missing geographic support: ground truth lies outside H^*'s footprint union. (b) Localization failure despite coverage: ground truth is inside the union, but the final prediction is substantially displaced. (c) Difficult Cross-area examples that fail strict distance thresholds. Each case juxtaposes the ground query, H^* geographic support, ground truth, final prediction, and metric error, with Z_0 and Z shown where corresponding response records are available. All final predictions use $q_D = \mathrm{DARK}(Z)$. Green stars mark ground truth; coverage and predictions are distinguished by the legend.

$$\mathrm{Gain}_{\mathrm{corr}}\ @d = \frac{100}{|Q|} \sum_{q \in Q} \left[\mathbf{1}\{e_q(q_D) \le d\} - \mathbf{1}\{e_q(q_0) \le d\} \right]. \tag{E.2}$$

Here, d is a metric distance threshold and gain is in percentage points. Both responses use the same DARK decoder; groups are determined from the base response before correction and remain fixed. This metric measures the fixed model's net accuracy change within each query group. It is not used as a numerical decomposition of configuration differences in Table 3.

E.2 Representative localization-stage examples

Two examples from complete test outputs illustrate localization error changes after response correction. For a Same-area test record (row index 47943), metric error decreases from 149.499 m to 0.413 m; for a Cross-area record (row index 47393), it decreases from 133.632 m to 2.899 m. These values correspond to $q_0 = \mathrm{DARK}(Z_0)$ and $q_D = \mathrm{DARK}(Z)$, respectively. They illustrate potentially substantial error changes after correction, without providing causal evidence that a particular secondary peak is directly suppressed.

E.3 Representative failure cases

Figure E.1 illustrates limitations through geographic coverage, localization error despite coverage, and strict cross-area thresholds. When selected references do not cover the ground truth, the input region lacks the target location. Even within the footprint union, the final prediction may deviate from the target. Under Cross-area, a prediction near the target may still fail a strict distance threshold. These cases explain representative failures; they are neither mutually exclusive failure categories nor frequency statistics.

Table F.1: Computational cost under consistent inference conditions. Latency is in ms; peak allocated memory is in MiB.

Protocol	Model	Mean latency	P50	P90	Peak memory	Latency ratio	Memory ratio
Same-area	Official CCVPE	27.03	27.21	28.11	329.3	1.00×	1.00×
Same-area	GeoHypLoc	50.52	41.59	77.40	2500.6	1.87×	7.59×
Cross-area	Official CCVPE	24.31	24.51	25.45	329.3	1.00×	1.00×
Cross-area	GeoHypLoc	42.39	40.10	59.96	2260.4	1.74×	6.86×

F COMPUTATIONAL COST

Inference is measured consistently on a single RTX 4090 in FP32 with batch size one, 30 warmup iterations, and CUDA synchronization. Timing includes model forward computation and coordinate decoding but excludes data loading, image decoding, and host-to-GPU transfers. Peak memory denotes peak allocated GPU memory. Within each protocol, relative costs use official CCVPE (Xia et al., 2024) measured over the same timing scope as the baseline.