RETHINKING CROSS-VIEW GEO-LOCALIZATION: A PROBABILISTIC MOTIVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization (CVGL) matches a street-view query, typically a panorama, to a geo-tagged satellite reference image of the same location. Existing methods primarily address the severe viewpoint gap between the two modalities, but pay much less attention to appearance variation caused by weather and daylight. We model a CVGL image as $x = \Phi(z, v, a, \xi)$, where Φ is an image formation function and z denotes location-preserving content, v viewpoint, a appearance, and ξ residual nuisance. This formulation clarifies the learning objective: preserve information about z while reducing sensitivity to the remaining factors. This probabilistic view motivates ProbInfoNCE, which replaces the shared temperature in InfoNCE with a learned pair-dependent scale to account for sample-dependent ambiguity during training. We also introduce CAV-Geo, a large-scale global dataset with appearance variation, with 256K street-view locations and 440K satellite-view images. CAV-Geo further includes a cross-continental split designed to evaluate out-of-distribution generalization. It is the largest CVGL benchmark with explicit weather and daylight variation, and its cross-continental protocol provides a substantially more challenging evaluation setting than existing standard splits. With the same DINOv3 initialization, ProbInfoNCE improves R@1 over InfoNCE by 5.1% on VIGOR cross-area and 13.6% on CVACT→CVUSA. Appearance augmentation with dual-matrix training and query-to-query alignment further improves R@1 on original VIGOR cross-area queries by 3.0% relative to ProbInfoNCE trained without appearance augmentation. All percentage gains in this paper are computed as ratios to their stated baselines.

1 INTRODUCTION

Cross-view geo-localization (CVGL) estimates the location of street-view images, usually panoramas, by matching them against a reference database of geo-tagged satellite imagery (Wilson et al., 2024). CVGL is usually optimized by contrastive objectives such as InfoNCE (van den Oord et al., 2018, Deuser et al., 2023), which bring matching latent embeddings closer while pushing unmatched pairs apart. Unlike same-view geo-localization (Berton et al., 2022), CVGL introduces a drastic geometric mismatch: a positive query-reference pair corresponds to the same place but is observed from radically different viewpoints, such as ground and aerial views. Beyond this viewpoint difference, a significant practical challenge arises from environmental domain shifts, yet most recent CVGL methods (Zhang et al., 2023b, 2026, Ye et al., 2024, Deuser et al., 2023, Zhu et al., 2021) assume both views are captured under ideal, clear-weather conditions. However, real-world ground query images are frequently subject to appearance variations due to weather and daylight. Consequently, the CVGL model must robustly match the distorted ground query to the reference satellite imagery despite these variations, which remains a primary hurdle for reliable real-world localization.

To reason about this challenge, we consider CVGL image formation in terms of location-preserving content, viewpoint, appearance, and other nuisance factors. Specifically, we write the image formation process as $x = \Phi(z, v, a, \xi)$ to provide a probabilistic motivation for our training design, where z denotes location-preserving content, v viewpoint, a appearance, and ξ residual nuisance. A positive query–reference pair shares a location, so its expected squared embedding distance measures the feature disagreement that retrieval must overcome. Under conditional independence, this distance decomposes into **query-side variation**, **reference-side variation**, and **cross-view mismatch** (Section 3.2). While the cross-view mismatch is the standard problem that CVGL handles, the other two terms are often overlooked. The query-side and reference-side variation can be due to v, a, ξ.

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.

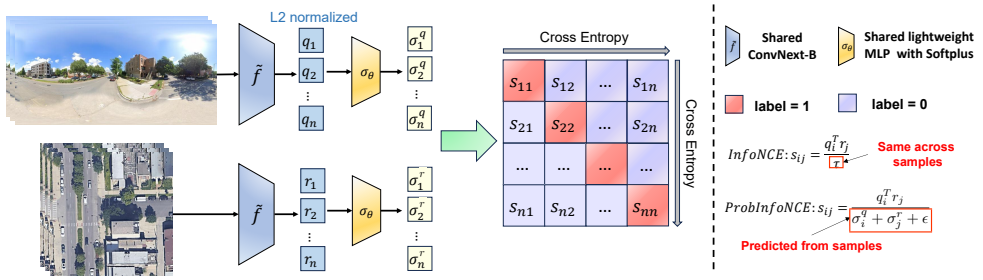


Figure 1: Illustration of our model structure (**left**) and comparison between InfoNCE and ProbInfoNCE (**right**). Standard InfoNCE applies cross-entropy losses to cosine similarities scaled by a learnable τ shared across all samples. We predict the scales in a sample-adaptive way via a lightweight MLP to handle query-side and reference-side variation.

For example, the variation in viewpoint v can be the result of the FOV of the camera or whether an equirectangular panorama or perspective crops are used (Zhang et al., 2026). The variation in appearance a alters texture and color statistics across weather or times of day. The residual nuisance ξ captures other domain-specific artifacts such as the camera sensor. None of these factors changes the location z, but each relates to the difficulty of retrieval.

This perspective clarifies the role of common methodologies. Hard negative mining (Deuser et al., 2023) encourages separation from confusable reference locations. Recent methods such as PanoBEV (Ye et al., 2024) and Geo^2 (Zhang et al., 2026) tackle the internal viewpoint variation of the query by introducing new views simulated by image transforms (details in Section 3.2). However, these approaches mainly handle variation through data sampling or deterministic transformations, rather than explicitly modeling the reliability of individual samples. This is important because many recent CVGL methods are optimized with contrastive objectives such as InfoNCE (van den Oord et al., 2018, Deuser et al., 2023), where a single temperature is shared by all query-reference pairs. Such a shared-temperature formulation does not explicitly account for sample-dependent ambiguity, even though some images may be intrinsically more ambiguous due to variations in viewpoint, appearance, or residual nuisance factors. For example, a query may be degraded by adverse weather, low illumination, or camera artifacts. Motivated by this observation, we propose **ProbInfoNCE**, *a sample-adaptive contrastive loss* that learns pair-dependent temperatures to account for query-side and reference-side ambiguity during training. As shown in Fig. 1, we retain deterministic query and reference embeddings and learn scalar scores to modulate their contrastive similarities. Specifically, a lightweight MLP with a softplus activation predicts individual scores, σ_i^q and σ_j^r, from the extracted features. $\sigma_i^q + \sigma_j^r + \varepsilon$ is used as the temperature of InfoNCE loss, where $\varepsilon > 0$ ensures numerical stability (details in Section 3.3).

The same perspective also highlights a crucial limitation of existing CVGL datasets. The query images are collected during the day under well-lit, clear weather conditions, resulting in a lack of variation in appearance a, whereas real-world applications usually encounter changes in weather and daylight. Clear-weather daytime evaluation alone does not establish robustness to weather and daylight changes. Collecting real observations of the same locations across these conditions is extremely challenging, requiring repeated, geographically aligned acquisition. Recent work (Zhang & Zhu, 2024) attempted to address this by applying simple synthetic augmentations to simulate environmental degradations, but failed to synthesize photorealistic ground imagery under severe weather conditions. Diffusion-based editing (Ho et al., 2020, Rombach et al., 2022, Black Forest Labs et al., 2025) can generate photorealistic weather and daylight variants. We assess whether the edits preserve location cues using visual inspection, SIFT, CLIP, and DINOv3 comparisons (Appendix E). We use these variants for appearance augmentation, providing same-location positives under different weather and daylight conditions to encourage robustness to query-side variation. Moreover, we curate a global dataset based on VIGOR (Zhu et al., 2021) and DRESS (Xia et al., 2025) to include 11 cities from all inhabited continents and introduce a cross-continental setting (see Tab. 1).

Contribution. (1) Probabilistic interpretation framework. We introduce a probabilistic framework for interpreting query-side variation, reference-side variation, and cross-view mismatch in CVGL (Section 3.2). It clarifies the roles of existing methods and motivates ProbInfoNCE and appearance augmentation. **(2) Sample-adaptive contrastive loss.** We propose ProbInfoNCE, which learns pair-dependent temperatures with deterministic embeddings. Compared with InfoNCE using the same DINOv3 initialization, ProbInfoNCE improves R@1 by **5.1**% on VIGOR cross-area and **13.6**% on CVACT→CVUSA (Section 5.2 and Appendix A). **(3) Appearance augmentation.** We propose appearance augmentation to introduce weather and daylight changes, enhancing R@1 on VIGOR cross-area by **3.0**% (Section 5.2). **(4) Global dataset with appearance variation.** Based

on VIGOR (Zhu et al., 2021) and DRESS (Xia et al., 2025), we introduce CAV-Geo, a global **C**ross-view **A**ppearance **V**ariation **Geo**-localization dataset with weather and daylight variation and a challenging cross-continental split. All percentage improvements are relative to the stated baseline: $(m_{\text{new}} - m_{\text{base}})/m_{\text{base}} \times 100\%$.

2 Related Works

Cross-view geo-localization. Cross-view geo-localization aims to identify, for a given query street-view image, the satellite image in a reference set that corresponds to the same location. Most prior work focuses on reducing the severe geometric gap between ground imagery and aerial imagery. Early methods exploit approximate azimuth correspondence between the two views (Shi et al., 2019, Yang et al., 2021, Shi et al., 2020, Zhang et al., 2023b, 2024, Zhu et al., 2021, 2022). For instance, SAFA (Shi et al., 2019) applies a polar transform to warp aerial images toward the panoramic domain and then aggregates multiple spatial embeddings, while GeoDTR (Zhang et al., 2023b, 2024) explicitly disentangles geometric layout descriptors from raw appearance features to improve same-area and cross-area generalization. Moreover, TransGeo (Zhu et al., 2022) and L2LTR (Yang et al., 2021) introduced the transformer architecture to capture the global correspondence between two views. Another line of work couples retrieval with cross-view generation: CDE (Toker et al., 2021) jointly learns satellite-to-street synthesis and localization in a multi-task architecture, showing that synthesis can regularize retrieval features. More recent methods simplify or enrich the retrieval pipeline in different ways. Sample4Geo (Deuser et al., 2023) relies on symmetric InfoNCE with hard negative sampling in a lightweight retrieval framework. Building upon that, PanoBEV (Ye et al., 2024) converts panoramas into bird's-eye-view representations and performs collaborative retrieval with both street-view and BEV branches. Geo^2 (Zhang et al., 2026) further leverages geometric foundation models such as VGGT (Wang et al., 2025) to build a shared geometry-aware latent space and jointly address geo-localization and bidirectional cross-view synthesis.

CVGL under appearance changes. Compared with geometric alignment, robustness to appearance shift remains much less explored. Zhang & Zhu (2024) provide a robustness benchmark by introducing 16 corruption types on CVUSA and CVACT, revealing substantial degradation under adverse weather and sensor noise. These benchmarks are valuable for robustness evaluation, but they are synthesized corruption benchmarks rather than semantically consistent edited image pairs. Recent advances in diffusion models make such editing increasingly practical. DDPM (Ho et al., 2020) and Stable Diffusion (Rombach et al., 2022) establish high-fidelity generative backbones. ControlNet (Zhang et al., 2023a) enables spatially conditioned generation from controls such as depth, edges, segmentation, and even in/out-painting. More recent methods (Li et al., 2024, Lyu et al., 2025, Li et al., 2025) improve conditional faithfulness and supervision quality for image-conditioned generation and editing. In addition, FLUX.1 Kontext (Black Forest Labs et al., 2025) unifies image generation and editing in a single flow-matching framework with strong context preservation. These advances suggest that modern diffusion-based editing is a more suitable tool than pattern-based corruption for introducing realistic weather and daylight variation while preserving scene structure.

Adaptive contrastive temperatures. Prior works learn input-dependent temperatures (Zhang et al., 2021) or individualized temperatures for each training example (Qiu et al., 2023). ProbInfoNCE uses an additive query–reference temperature for each pair, with a comparison to the method of Zhang et al. (2021) in Appendix A.

3 Method

3.1 Preliminary: InfoNCE

Following Sample4Geo (Deuser et al., 2023), many recent CVGL works (Ye et al., 2024, Mi et al., 2024, Xia et al., 2025, Zhang et al., 2026, 2025) adopt InfoNCE loss for training. In this section, we first introduce the definition of the commonly used InfoNCE loss for CVGL. Let $\tilde{f}$ be the shared pre-normalization encoder, instantiated as ConvNeXt-B (Liu et al., 2022) in Fig. 1, and define the final retrieval embedding map by

$$f(x) = \frac{\tilde{f}(x)}{\|\tilde{f}(x)\|_2}, \tag{1}$$

Let $q_i = f(x_i^q)$ and $r_i = f(x_i^r)$ denote the ℓ_2-normalized query and reference features extracted by the shared encoder. Standard symmetric InfoNCE (van den Oord et al., 2018, Radford et al., 2021) is

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2N} \sum_{i=1}^{N} \left(-\log \frac{\exp(q_i^\top r_i/\tau)}{\sum_{j=1}^{N} \exp(q_i^\top r_j/\tau)} - \log \frac{\exp(r_i^\top q_i/\tau)}{\sum_{j=1}^{N} \exp(r_i^\top q_j/\tau)} \right). \tag{2}$$

3.2 PROBABILISTIC MOTIVATION

We use $x = \Phi(z, v, a, \xi)$ to motivate sample-adaptive contrastive learning and appearance augmentation. We first characterize variation in query and reference observations.

Let Z denote the geographic location variable shared by a positive query–reference pair. Given a specific location $Z = z$, the query observation and the reference observation are generated by two branch-specific sensing processes:

$$X^q = \Phi_q(z, N^q), \qquad X^r = \Phi_r(z, N^r), \tag{3}$$

where q, r denote query and reference, and $N^q = (V^q, A^q, \Xi^q)$ and $N^r = (V^r, A^r, \Xi^r)$ collect nuisance factors in the query and reference branches, respectively. Here V denotes structural viewpoint, A denotes appearance factors such as illumination and weather, and Ξ denotes residual domain-specific factors, such as sensor artifacts. Thus, a positive pair shares the same retrieval variable z but may differ substantially in viewpoint, appearance, and sensor.

Let $Y^q = f(X^q), Y^r = f(X^r)$ be the normalized embeddings. Ideally, all observations from the same location should collapse to a common location prototype:

$$f(\Phi_q(z, N^q)) \approx \psi(z), \qquad f(\Phi_r(z, N^r)) \approx \psi(z), \tag{4}$$

while prototypes of different locations should remain separated.

To characterize embedding variation, define the conditional mean and covariance of query and reference branches:

$$\begin{aligned} \mu_q(z) &= \mathbb{E}[Y^q \mid Z = z], & \mu_r(z) &= \mathbb{E}[Y^r \mid Z = z], \\ \Sigma_q(z) &= \mathrm{Var}(Y^q \mid Z = z), & \Sigma_r(z) &= \mathrm{Var}(Y^r \mid Z = z). \end{aligned} \tag{5}$$

Assuming $N^q \perp N^r \mid Z = z$, we obtain the following bias–variance decomposition:

$$\mathbb{E}\big[\|Y^q - Y^r\|_2^2 \mid Z = z\big] = \underbrace{\mathrm{tr}(\Sigma_q(z))}_{\text{query-side variation}} + \underbrace{\mathrm{tr}(\Sigma_r(z))}_{\text{reference-side variation}} + \underbrace{\|\mu_q(z) - \mu_r(z)\|_2^2}_{\text{cross-view mismatch}}. \tag{6}$$

This decomposition concerns positive-pair embedding distance, while retrieval also depends on separation from negative references. From this perspective, the shared temperature τ in Eq. (2) does not explicitly account for sample-dependent uncertainty, although contrastive training can affect all three terms in Eq. (6). Hard negative mining in Sample4Geo (Deuser et al., 2023) encourages separation from confusable references, while PanoBEV (Ye et al., 2024) and Geo^2 (Zhang et al., 2026) address viewpoint variation through geometric or viewpoint-based image transformations. However, appearance-induced query variation remains underexplored in CVGL. This perspective motivates two complementary choices: adapting contrastive scaling to individual samples and providing query images with appearance variation.

For the loss design, consider independent Gaussian perturbations around the deterministic unit-norm features:

$$s_i^q = q_i + \epsilon_i^q, \qquad \epsilon_i^q \sim \mathcal{N}(0, \Sigma_i^q), \tag{7}$$

$$s_j^r = r_j + \epsilon_j^r, \qquad \epsilon_j^r \sim \mathcal{N}(0, \Sigma_j^r), \tag{8}$$

where $q_i = f(x_i^q)$, $r_j = f(x_j^r)$, $q_i, r_j, s_i^q, s_j^r \in \mathbb{R}^d$, and d denotes the embedding dimension. The positive semidefinite covariance matrices $\Sigma_i^q, \Sigma_j^r \in \mathbb{R}^{d\times d}$ and auxiliary variables s_i^q, s_j^r describe perturbations for analysis only. The loss uses deterministic embeddings and learned scalar scores.

Because independent Gaussian noises add under subtraction, we have

$$s_i^q - s_j^r \sim \mathcal{N}\big(q_i - r_j,\ \Sigma_i^q + \Sigma_j^r\big), \tag{9}$$

which, under the independence assumption, implies

$$\mathbb{E}\big[\|s_i^q - s_j^r\|_2^2\big] = \|q_i - r_j\|_2^2 + \mathrm{tr}(\Sigma_i^q) + \mathrm{tr}\big(\Sigma_j^r\big). \tag{10}$$

Table 1: Comparison with existing datasets.

	CVUSA (Zhai et al., 2017)	CVACT (Liu & Li, 2019)	VIGOR (Zhu et al., 2021)	CVGlobal (Ye et al., 2024)	DRESS (Xia et al., 2025)	CAV-Geo
Query locations	44,416	128,334	105,214	134,233	174,934	255,875
Query images	44,416	128,334	105,214	134,233	174,934	511,750
Satellite images	44,416	128,334	90,618	134,233	422,760	439,846
Global	×	×	×	✓	✓	✓
Regions	Suburban	Urban	Urban	Urban	Urban + Suburban	Urban + Suburban
Weather	Sunny	Sunny	Sunny	Sunny	Sunny	Sunny + Rainy + Snowy + Foggy
Daylight	daytime	daytime	daytime	daytime	daytime	daytime + night

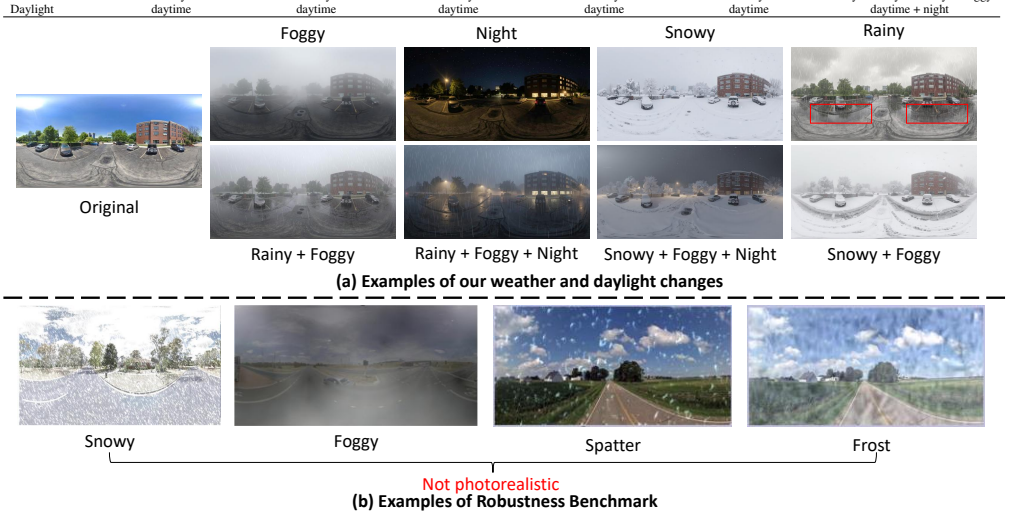


Figure 2: (a) Visualization of our weather and daylight changes. The images are photorealistic and preserve scene layouts. Red boxes show the reflection generated in the rainy case. (b) Examples from robustness benchmark (Zhang & Zhu, 2024). We use existing examples from this benchmark for the comparison.

Eq. (10) has the same second-moment form as Eq. (6) (see Appendix B): the query- and reference-side covariance traces add to the deterministic squared distance. Since q_i and r_j are unit norm, $\|q_i - r_j\|_2^2 = 2 - 2q_i^\top r_j$. We preserve the additive uncertainty structure in ProbInfoNCE with two nonnegative, feature-dependent learned surrogates whose sum defines a pairwise scale. The learned scales are optimized for retrieval and do not directly estimate the covariance traces.

3.3 ProbInfoNCE: Uncertainty-aware contrastive learning

The additive scale leads to the contrastive logit $q_i^\top r_j/\tau_{ij}$ under the isotropic Gaussian likelihood ratio against a zero-similarity reference (Appendix C). A lightweight softplus MLP predicts its nonnegative query and reference components, whose sum replaces the shared temperature in Eq. (2):

$$\tau_{ij} = \sigma_\theta(q_i) + \sigma_\theta(r_j) + \varepsilon, \qquad \varepsilon > 0, \tag{11}$$

where $\sigma_\theta(\cdot) \geq 0$ predicts a scalar uncertainty score. For brevity, $\sigma_\theta(q_i)$ and $\sigma_\theta(r_j)$ denote scores computed from the corresponding pre-normalization features $\tilde{f}(x_i^q)$ and $\tilde{f}(x_j^r)$. We define the ProbInfoNCE losses as

$$\ell_i^{q\to r} = -\log \frac{\exp(q_i^\top r_i/\tau_{ii})}{\sum_{j=1}^N \exp(q_i^\top r_j/\tau_{ij})}, \qquad \ell_i^{r\to q} = -\log \frac{\exp(r_i^\top q_i/\tau_{ii})}{\sum_{j=1}^N \exp(r_i^\top q_j/\tau_{ji})}. \tag{12}$$

Here the reverse direction uses τ_{ji} because the candidate pair is (q_j, r_i). The final objective is

$$\mathcal{L}_{\text{ProbInfoNCE}}(q,r) = \frac{1}{2N}\sum_{i=1}^N (\ell_i^{q\to r} + \ell_i^{r\to q}) + \frac{\lambda}{2N}\sum_{i=1}^N [\sigma_\theta(q_i) + \sigma_\theta(r_i)]\,. \tag{13}$$

$\sigma_\theta(q)$ and $\sigma_\theta(r)$ are learned jointly with the encoder as effective uncertainty scales for the contrastive logits. They do not directly estimate the conditional covariance terms in Eq. (6). The learned scales need not increase with appearance degradation (see Appendix G). The regularizer in Eq. (13) penalizes large scores. At inference, we rank reference candidates by $q_i^\top r_j/\tau_{ij}$, retaining the uncertainty predictor.

3.4 Training with appearance augmentation

We use appearance augmentation to provide same-location query variants under different weather and daylight conditions. Let q_a denote a batch of appearance-augmented queries, q_o their original counterparts, and r the reference batch. Matching indices correspond to the same geolocation. We consider the following training strategies (visualized in Fig. 9):

- **MixBatch.** Construct q by randomly partitioning indices $1, 2, ..., B$, where B is batch size, into two disjoint sets I and J with $|I| = |J|$. Then set $q[I] = q_a[I]$ and $q[J] = q_o[J]$ (in torch-like notation). Compute the loss as $\mathcal{L}_{Mix} = \mathcal{L}_{ProbInfoNCE}(q, r)$.
- **Dual matrix.** Compute $\mathcal{L}_{Dual} = \lambda_1 \mathcal{L}_{ProbInfoNCE}(q_o, r) + \lambda_2 \mathcal{L}_{ProbInfoNCE}(q_a, r)$.
- **Query-to-query (q2q) alignment.** $\mathcal{L}_{q2q} = \mathcal{L}_{ProbInfoNCE}(q_o, q_a)$.

Here, ProbInfoNCE adapts contrastive scaling to individual samples, while appearance augmentation provides supervision for appearance-induced variation in the query branch. MixBatch and Dual matrix encourage robustness to query-side variation by feeding additional training data to the model. Query-to-query (q2q) alignment directly encourages consistency between original and appearance-augmented query embeddings. We find that the dual matrix strategy with query-to-query alignment achieves the best performance, which supports the effectiveness of training for robustness to query-side variation. Therefore, our final training loss is:

$$\mathcal{L} = \mathcal{L}_{Dual} + \lambda_3 \mathcal{L}_{q2q}. \tag{14}$$

4 Dataset Curation

Raw Data Curation: To curate CAV-Geo, we first combine DRESS (Xia et al., 2025) and VIGOR (Zhu et al., 2021). We chose DRESS for its large scale and coverage of both urban and suburban areas. VIGOR is chosen because it is one of the most popular benchmarks. By doing so, in addition to our dataset, we also provide appearance variation data for VIGOR. Among the cities in VIGOR (Chicago, New York, Seattle, and San Francisco), only Chicago is also included in DRESS. We retain the VIGOR version of Chicago because its satellite images have higher resolution (640×640 in VIGOR, 224×224 in DRESS). In addition to resolution, VIGOR collects all images from Google Maps, but DRESS collects street-view images from Google Maps and satellite images from Esri World Imagery. To unify them, we re-collect the satellite images in DRESS (excluding Chicago) via Google Maps Static API based on the provided GPS, following the satellite-image collection process of VIGOR (Zhu et al., 2021). The comparison of our dataset with existing datasets is shown in Tab. 1.

Table 2: Quantitative results on VIGOR. Each block reports recall at top-K (R@K) and Hit Rate. † indicates Polar Transformation is applied. * means our reproduction. We mark PanoBEV and Geo^2 with gray and exclude them from ranking as they require additional inputs and models. The best results are **boldfaced** and the second-best results are underlined.

Method	VIGOR Same					VIGOR Cross				
	R@1	R@5	R@10	R@1%	Hit Rate	R@1	R@5	R@10	R@1%	Hit Rate
SAFA† (Shi et al., 2019)	33.93	58.42	68.12	98.24	36.87	8.20	19.59	26.36	77.61	8.85
TransGeo (Zhu et al., 2022)	61.48	87.54	91.88	99.56	73.09	18.99	38.24	46.91	88.94	21.21
SAIG-D (Zhu et al., 2023)	65.23	88.08	-	99.68	74.11	33.05	55.94	-	94.64	36.71
GeoDTR (Zhang et al., 2023b)	56.51	80.37	86.21	99.25	61.76	30.02	52.67	61.45	94.40	30.19
GeoDTR+ (Zhang et al., 2024)	59.01	81.77	87.10	99.07	67.41	36.01	59.06	67.22	94.95	39.40
Sample4Geo (Deuser et al., 2023)	77.86	95.66	97.21	99.61	89.82	61.70	83.50	88.00	98.17	69.87
PanoBEV* (Ye et al., 2024)	80.07	96.84	98.03	99.77	91.98	67.20	87.31	90.77	98.51	76.39
Geo^2 (Zhang et al., 2026)	81.59	96.53	98.62	99.68	90.35	66.71	87.34	91.02	98.25	72.13
Ours	**81.92**	**97.46**	**98.35**	**99.72**	**94.13**	**70.10**	**89.60**	**92.39**	**98.55**	**79.73**

Appearance Variation: Notably, collecting a dataset under different weather and daylight conditions is extremely challenging, as these conditions cannot be directly controlled during real-world data collection. To obtain diverse query images under such conditions, we therefore employ FLUX.2 [klein] 9B (Black Forest Labs, 2026) to edit the query to include one or more of rainy, snowy, foggy, and night conditions (details in Appendix D). We select this model for its strong editing quality, publicly available weights, and efficient distilled inference, balancing image quality and computational cost at the scale of CAV-Geo. To identify potential failure cases that may change geolocation cues after editing, we visually inspect 100 examples each with the lowest SIFT (Lowe, 2004) Inlier ratio with RANSAC (Fischler & Bolles, 1981), CLIP similarity (Radford et al., 2021), and DINOv3 similarity (Siméoni et al., 2025). Among the inspected examples, we observed no changes to major geolocation cues, such as road layout or building structure. We provide visualizations in Fig. 2 and the similarity values with corresponding visualization in Appendix E as evidence. In Fig. 2 (a), the structures of the buildings and road are preserved, and the generation model can generate realistic reflections in the rainy case. The pattern-based corruption shown in Fig. 2 (b), proposed by Zhang & Zhu (2024), is not photorealistic. Note that we do not include appearance variation for the reference set because the reference set is supposed to be a **fixed offline database** with good quality, whereas the query is an online input.

Evaluation Protocols: Following prior work (Zhu et al., 2021, Xia et al., 2025, Ye et al., 2024), we define two evaluation protocols, **Same-Area** and **Cross-Continent**. Under the Same-Area setting, all query locations from all cities are randomly split into a training and testing set, resulting in 128,056 training query locations and 127,819 testing query locations. To benchmark the model in an extreme

Table 3: Quantitative results on CVUSA, CVACT Val, and CVACT Test in recall at top-K retrievals (R@K). † indicates Polar Transformation is applied. * indicates we reproduce the results based on its codebase. We mark PanoBEV and Geo^2 with gray and exclude them from ranking as they require additional inputs and models. The best results are **boldfaced** and the second-best results are underlined.

Method	CVUSA				CVACT Val				CVACT Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA† (Shi et al., 2019)	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
TransGeo (Zhu et al., 2022)	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR (Zhang et al., 2023b)	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
GeoDTR+ (Zhang et al., 2024)	95.05	98.42	98.92	99.77	87.76	95.50	96.50	98.32	67.75	90.15	92.73	98.53
SAIG-D† (Zhu et al., 2023)	96.34	99.10	99.50	99.86	89.06	96.11	97.08	98.89	67.49	89.39	92.30	96.80
Sample4Geo (Deuser et al., 2023)	98.68	99.68	99.78	99.87	90.35	96.61	97.53	98.78	71.51	92.42	94.45	98.70
PanoBEV* (Ye et al., 2024)	98.71	99.63	99.79	99.88	89.49	96.12	96.96	98.73	72.30	92.78	94.73	98.86
Geo^2 (Zhang et al., 2026)	98.83	99.72	99.79	99.91	94.36	97.41	97.97	99.05	75.08	94.89	95.77	99.01
Ours	**99.24**	**99.76**	**99.83**	**99.90**	**92.56**	**97.47**	**98.10**	**98.93**	**74.06**	**94.18**	**95.79**	**98.95**

scenario, we establish a challenging Cross-Continent protocol. We choose to train on cities outside North America (Johannesburg, London, Rio De Janeiro, Sydney, Taipei, and Tokyo) while testing in North American cities (Chicago, New York, San Francisco, Seattle, and Vancouver), totaling 129,094 training query locations and 126,781 testing query locations. Both partitions use satellite images acquired from Google Maps under the VIGOR collection protocol described above. More details can be found in Appendix F.

5 Experiments

Datasets, baselines, and evaluation metrics. We follow the most recent work (Zhang et al., 2026) to report the evaluation results on three existing datasets: CVUSA (Workman et al., 2015, Zhai et al., 2017), CVACT (Liu & Li, 2019), and VIGOR (Zhu et al., 2021), as well as our own CAV-Geo dataset. Details of the dataset can be found in Tab. 1. SAFA (Shi et al., 2019), TransGeo (Zhu et al., 2022), GeoDTR (Zhang et al., 2023b), SAIG (Zhu et al., 2023), Sample4Geo (Deuser et al., 2023), PanoBEV (Ye et al., 2024), and Geo^2 (Zhang et al., 2026) are selected as our baselines. We report recall accuracy at K (R@K, K = 1,5,10, and 1%) for all datasets and hit rate for VIGOR and our dataset. R@K checks whether the designated ground-truth satellite image is among the top K. For both VIGOR and CAV-Geo, Hit Rate is the percentage of queries for which no negative tile scores higher than the main ground-truth tile, with the three semi-positive tiles ignored.

Table 4: Cross-dataset results in R@K. CVUSA→CVACT means training on CVUSA and evaluating on CVACT, and similarly for CVACT→CVUSA. † indicates Polar Transformation is applied. * means our reproduction. We mark PanoBEV and Geo^2 with gray and exclude them from ranking as they require additional inputs and models. The best results are **boldfaced** and the second-best results are underlined.

Method	CVUSA→CVACT				CVACT→CVUSA			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA† (Shi et al., 2019)	30.40	52.93	62.29	85.82	21.45	36.55	43.79	69.83
TransGeo (Zhu et al., 2022)	37.81	61.57	69.86	89.14	17.45	32.49	40.48	69.14
GeoDTR (Zhang et al., 2023b)	43.72	66.99	74.61	91.83	29.85	49.25	57.11	82.47
GeoDTR+ (Zhang et al., 2024)	60.16	79.97	84.67	94.48	52.56	73.08	79.82	94.80
SAIG-D† (Zhu et al., 2023)	15.29	33.07	42.14	72.95	18.97	35.60	44.28	75.33
Sample4Geo (Deuser et al., 2023)	56.62	77.79	87.02	94.69	44.95	64.36	72.10	90.65
PanoBEV* (Ye et al., 2024)	65.79	86.17	90.33	96.70	29.30	50.61	59.99	86.70
Geo^2 (Zhang et al., 2026)	63.17	82.53	87.88	95.09	55.14	73.58	80.03	95.33
Ours	**65.65**	**86.36**	**90.27**	**96.89**	**55.72**	**76.07**	**82.82**	**96.23**

Training protocol. We follow the training protocol of Sample4Geo to train our models. However, we initialize ConvNeXt-B (Liu et al., 2022) with weights distilled from DINOv3 (Siméoni et al., 2025) and replace InfoNCE with ProbInfoNCE. Our ablation study in Section 5.2 shows that the performance is not solely due to DINOv3. We set λ_1, λ_2 in dual matrix to be 0.5, 0.5, and λ_3 in q2q to be 0.1 for our loss function shown in Eq. (14), and λ to 0.5 in Eq. (13). For CVUSA and CVACT, we use ProbInfoNCE on original queries, without FLUX-generated variants or q2q alignment. More details are provided in Appendix I.

5.1 Experimental results

Fair comparison. Since PanoBEV is partially open-sourced, we reproduce the results based on its codebase (details in Appendix H). It is important to note that PanoBEV (Ye et al., 2024) and Geo^2 (Zhang et al., 2026) require substantially more GFLOPs and model parameters and can be implemented in parallel with any existing methods. To ensure a fair comparison, we exclude PanoBEV and Geo^2 from direct ranking in standard benchmarks but keep them in the table for reference. **We additionally plug PanoBEV into our method for a comparison under the same setting** in Sec. 5.2. As Geo^2 requires distributed training on at least four A100/H100 GPUs, we are unable to reproduce it with our resources.

Quantitative results on standard benchmarks. Tables 2, 3, and 4 report results on VIGOR, CVUSA, CVACT, and cross-dataset transfer. Under the ranking protocol above, our method achieves

Table 5: Initial reference results on CAV-Geo. Appearance variation uses FLUX-edited queries with synthetic weather and daylight changes. Baselines are trained on ideal clear-weather daytime queries, while our model additionally uses appearance augmentation. * denotes our reproduction. The best results are **boldfaced**, excluding PanoBEV (gray) because it uses additional inputs and models.

Method	Original queries										Appearance variation									
	Same Area					Cross-Continent					Same Area					Cross-Continent				
	R@1	R@5	R@10	R@1%	Hit	R@1	R@5	R@10	R@1%	Hit	R@1	R@5	R@10	R@1%	Hit	R@1	R@5	R@10	R@1%	Hit
GeoDTR (Zhang et al., 2023b)	9.50	25.99	35.76	98.33	10.73	3.21	13.19	20.09	56.69	3.78	0.75	2.95	5.12	32.73	0.81	0.28	1.36	2.47	12.58	0.32
SAIG-S (Zhu et al., 2023)	8.51	22.51	31.28	96.97	9.13	1.10	4.37	7.77	25.33	1.21	0.61	2.72	4.54	28.45	0.66	0.06	0.36	0.66	4.63	0.07
SAIG-D (Zhu et al., 2023)	11.40	26.31	36.31	97.46	12.94	4.61	15.30	20.20	59.89	4.93	0.63	2.68	4.09	27.00	0.67	0.24	1.76	2.43	12.76	0.29
Sample4Geo (Deuser et al., 2023)	57.64	78.47	82.94	98.42	63.78	13.64	33.41	42.06	91.89	18.45	3.00	6.28	8.00	37.43	3.31	0.69	2.21	3.31	26.05	0.91
PanoBEV* (Ye et al., 2024)	61.79	81.91	85.95	98.95	68.11	18.50	41.77	50.53	93.37	25.91	3.62	7.63	9.75	48.61	3.94	0.91	3.04	4.50	37.87	1.25
Ours	**67.33**	**83.76**	**87.47**	**99.07**	**70.79**	**20.82**	**46.57**	**55.89**	**95.76**	**25.09**	**58.91**	**79.86**	**84.24**	**98.74**	**65.34**	**17.14**	**40.32**	**49.33**	**94.35**	**23.81**

Table 6: Comparison across benchmarks, model size, and GFLOPs. We abbreviate Same to S, Cross to C, Val to V, Test to T, and remove 'CV' from the data name, like CVUSA, to save space. GFLOPs is measured with a query (384×768) - reference (384×384) pair. GFLOPs of Geo^2 is estimated by simulating with VGGT and Sample4Geo. * means our reproduction.

Methods	VIGOR S	VIGOR C	CAV-Geo S	CAV-Geo C	USA	ACT V	ACT T	USA→ACT	ACT→USA	# Params	GFLOPs
Sample4Geo (Deuser et al., 2023)	77.86	61.70	57.64	13.64	98.68	90.35	71.51	56.62	44.95	88.6M	151.20
PanoBEV* (Ye et al., 2024)	80.07	67.20	61.79	18.50	98.71	89.49	72.30	**65.79**	29.30	2×88.6M	241.25
Geo^2 (Zhang et al., 2026)	81.59	66.71	-	-	98.83	**94.36**	**75.08**	63.17	55.14	1.05B	≈5000
Ours	**81.92**	**70.10**	**67.33**	**20.82**	**99.24**	92.56	74.06	65.65	**55.72**	89.1M	151.22

the highest R@1 across all settings. Relative to Sample4Geo (Deuser et al., 2023), R@1 improves by 13.6% on VIGOR cross-area and 24.0% on CVACT→CVUSA. Even compared with PanoBEV (Ye et al., 2024) and Geo^2 (Zhang et al., 2026), which require additional inputs and models, our method achieves higher R@1 on VIGOR, CVUSA, and CVACT→CVUSA, while remaining competitive on CVACT and CVUSA→CVACT. Together with the controlled ablations, these standard-benchmark results provide the primary evidence for our method's effectiveness.

CAV-Geo: initial reference results. For CAV-Geo, we report reproduced baselines alongside our results to establish initial references for future research (Tab. 5). Our primary method comparisons use the established benchmarks above. The original-query results highlight the difficulty of the dataset even under ideal conditions. Sample4Geo achieves only 57.64 R@1 in the same-area setting, below its 61.70 on VIGOR cross-area. The cross-continental setting is especially challenging, with several existing methods achieving less than 10% R@1. Our model reaches 67.33 R@1 on same-area and 20.82 on cross-continent, while the low absolute recall underscores the difficulty of cross-continental generalization.

Appearance variation diagnostic. The appearance variation results in Tab. 5 illustrate the difficulty that models trained under ideal clear-weather daytime conditions face when handling weather and daylight changes. On the same-area split, Sample4Geo drops from 57.64 R@1 on original queries to 3.00 on FLUX-edited queries. Our model, trained with appearance augmentation, achieves 58.91 R@1 on edited queries. Together with the controlled ablations in Section 5.2, these results highlight the importance of appearance augmentation for robustness to such changes.

Qualitative visualization. Fig. 3 compares retrievals on original and FLUX-edited queries. In the edited examples, Sample4Geo, trained only on clear-weather queries, misses the correct location. Some top-ranked candidates share the edited query's broad colors (white for snow or dark blue for night) despite different road layouts, illustrating its sensitivity to appearance changes. Additional examples are in Appendix J.

Overall, our method achieves state-of-the-art performance on most datasets and is competitive with, and in several cases outperforms, heavier models while requiring substantially less computation, as shown in Tab. 6. Our method adds negligible computational cost compared with Sample4Geo, while PanoBEV and Geo^2 require substantially more GFLOPs.

5.2 Ablation Study

Effectiveness of components and incorporating PanoBEV. Tab. 7 ablates our components and PanoBEV on the VIGOR cross split. Within each loss setting, the Original and MixBatch rows compare training with and without appearance augmentation under the same initialization, isolating its effect. Row ① uses Sample4Geo (Deuser et al., 2023) with the DINOv3 distilled backbone. We find:

- ① versus ⑧ shows that DINOv3 alone does not explain our improvement.

Table 7: Ablation studies on loss, training strategy, and BEV transform. † means our reproduction with DINOv3 initialization.

Method	Strategy	VIGOR Cross R@1	R@5	R@10	R@1%	Hit	VIGOR Cross Appearance Change R@1	R@5	R@10	R@1%	Hit
InfoNCE†	Original ①	64.74	86.18	90.00	98.50	73.74	12.21	22.68	27.19	56.47	13.82
	MixBatch ②	65.26	86.35	90.24	98.52	74.16	61.50	83.41	87.86	98.16	69.64
	Dual matrix ③	**66.90**	**87.36**	**90.96**	**98.62**	**75.58**	**62.91**	**84.20**	**88.49**	**98.22**	**70.93**
	Dual matrix w/ q2q ④	65.57	86.14	89.99	98.38	74.10	61.80	83.39	87.83	98.02	69.84
ProbInfoNCE	Original ⑤	68.07	88.72	91.89	**98.71**	75.64	13.82	24.44	28.83	55.19	15.45
	MixBatch ⑥	68.48	88.85	91.98	98.60	78.25	63.01	84.31	88.47	97.95	71.45
	Dual matrix ⑦	69.26	89.42	92.23	98.57	78.99	64.01	84.80	88.66	97.92	72.35
	Dual matrix w/ q2q ⑧	**70.10**	**89.60**	**92.39**	98.55	**79.73**	**66.69**	**87.05**	**90.58**	**98.17**	**75.83**
BEV transform	PanoBEV† (Ye et al., 2024) ⑨	68.77	88.39	91.30	**98.81**	77.95	10.21	20.77	24.38	56.39	12.02
	Dual matrix w/ q2q ⑩	**73.12**	**91.19**	**93.69**	98.76	**82.63**	**68.09**	**87.83**	**91.24**	**98.34**	**77.03**

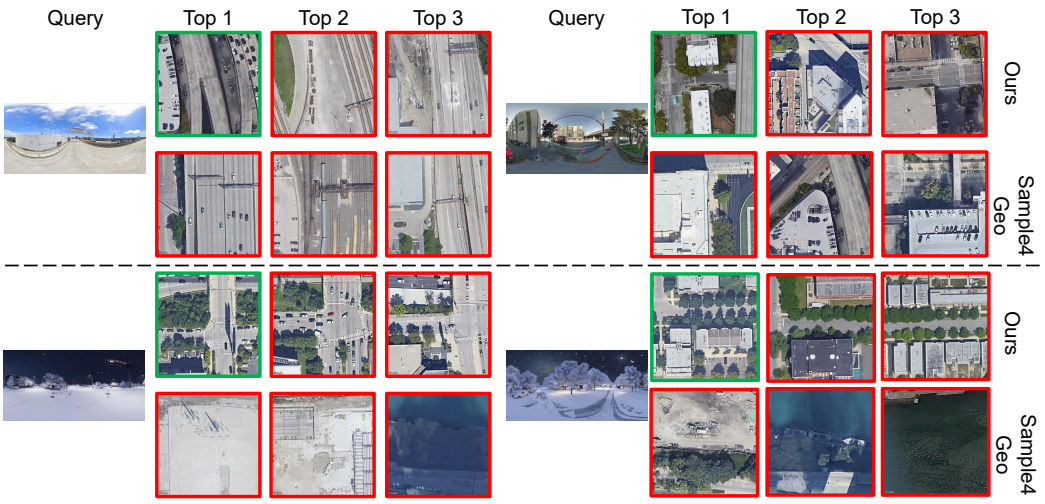


Figure 3: Qualitative Comparisons. Positives are in green, negatives are in red.

- Across matched strategies, ProbInfoNCE yields 3.5–6.9% relative R@1 gains over InfoNCE (rows ①–⑧). Further loss ablations appear in Appendix A.
- ②–④ outperform ①, and ⑥–⑧ outperform ⑤, on the test split without appearance changes. Appearance augmentation thus also improves performance on original clear-weather queries, consistent with Eq. (6).
- Query-to-query (q2q) alignment raises Dual-matrix R@1 with ProbInfoNCE (⑦ vs. ⑧) and lowers it with InfoNCE (③ vs. ④). This pattern is consistent with adaptive scaling helping q2q alignment.
- PanoBEV (Ye et al., 2024) is compatible with our method (⑨ vs. ⑩). With the same initialization, our training protocol raises its R@1 by 6% relative. The BEV transform is not our contribution, so we report it only as an ablation.

6 Conclusion

We motivate ProbInfoNCE through a probabilistic view of CVGL that distinguishes location-preserving content from viewpoint, appearance, and residual nuisance factors. ProbInfoNCE is an uncertainty-aware contrastive loss that adaptively accounts for sample-dependent ambiguity during training. We further introduce CAV-Geo, a large-scale global benchmark with explicit weather and daylight variation and a challenging cross-continental split. Experiments across standard and proposed benchmarks show that combining ProbInfoNCE with appearance augmentation improves robustness and generalization under both cross-view and appearance shifts.

AI USE STATEMENT

We used FLUX to generate weather and daylight variations of street-view images for CAV-Geo, as described in Appendix D. We also used GPT6-Astra as an image-pair evaluator of geolocation preservation, as reported in Appendix E. The manuscript reports quantitative similarity checks and visual inspection of edited images, together with the limitations of these checks. We used GPT-6 to polish the writing and OpenAI Codex to assist with manuscript editing and LaTeX formatting. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Cross-view geo-localization can aid localization, disaster response, and map updating, but can also expose private locations and enable surveillance, stalking, or unauthorized tracking. We discuss these risks in Appendix M. CAV-Geo includes generatively edited street-view images. The checks for changes to geolocation cues and their limitations are described in Appendices E and L. The source-data terms and access arrangements are described in Appendix F.

REPRODUCIBILITY STATEMENT

Section 5 describes the experimental setup. Appendices D and F document image editing and dataset splits, and Appendix I specifies optimization, architecture, and hardware settings. Appendices B and C provide supporting derivations, and Appendix H documents the PanoBEV reproduction.

REFERENCES

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4878–4888, 2022.

Black Forest Labs. FLUX.2 [klein]: Towards Interactive Visual Intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

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

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2024.

Ming Li, Xin Gu, Fan Chen, Xiaoying Xing, Longyin Wen, Chen Chen, and Sijie Zhu. SuperEdit: Rectifying and facilitating supervision for instruction-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19206–19215, 2025.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5624–5633, 2019.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

Zonglin Lyu, Ming Li, Xinxin Liu, and Chen Chen. CPO: Condition preference optimization for controllable image generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. ConGeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, pp. 214–230. Springer, 2024.

Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *Machine Learning for Intelligent Transportation Systems Workshop, NIPS*, 2018.

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28389–28421. PMLR, 2023. URL https://proceedings.mlr.press/v202/qiu23a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 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*, 32:10090–10100, 2019.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4064–4072, June 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6488–6497, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132(4):1350–1392, 2024.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3961–3969, 2015.

Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, et al. Cross-view geo-localization with panoramic street-view and VHR satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29009–29020. Curran Associates, Inc., 2021.

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*, pp. 74–90. Springer, 2024.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 867–875, July 2017.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023a.

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*, 2021.

Qingwang Zhang and Yingying Zhu. Benchmarking the robustness of cross-view geo-localization models. In *European Conference on Computer Vision*. Springer, 2024.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3480–3488, 2023b.

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.

Xiaohan Zhang, Tavis Shore, Chen Chen, Oscar Mendez, Simon Hadfield, and Safwan Wshah. VICI: VLM-instructed cross-view image-localisation. In *Proceedings of the 3rd International Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*, 2025.

Yancheng Zhang, Xiaohan Zhang, Guangyu Sun, Zonglin Lyu, Safwan Wshah, and Chen Chen. Geo^2: Geometry-guided cross-view geo-localization and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19432–19442, 2026.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.

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.

Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023.

Appendix

The appendix is structured as follows:

- Section A: Additional ablation studies.
- Section B: Proof of Eq. 10.
- Section C: Additional probabilistic motivation for our loss.
- Section D: Details on image editing.
- Section E: Evidence to show strong similarity between original and edited queries.
- Section F: More details of the CAV-Geo Dataset.
- Section G: Understanding the predicted uncertainty σ.
- Section H: Details on reproducing PanoBEV (Ye et al., 2024).
- Section I: More details on the training setup.
- Section J: Additional visual examples for our dataset and visual comparisons.
- Section K: No-change FLUX editing control.
- Section L: Limitations and future work.
- Section M: Broader impact.

A Additional Ablation Studies

Our method follows the probabilistic motivation in Section 3.2 and is related to temperature scaling. Specifically, Zhang et al. (2021) learn input-dependent temperatures as uncertainty scores. Qiu et al. (2023) learn an individualized temperature for each training example. Neumann et al. (2018) calibrate classification confidence by scaling logits rather than modifying a contrastive objective. Among them, the method of Zhang et al. (2021) is the closest temperature-scaling baseline. We implement its input-dependent scaling with our architecture using only a query-side scaler, yielding a single-sided temperature baseline. For this comparison, both methods use the same DINOv3-distilled ConvNeXt-B initialization, original and edited training queries, dual-matrix and q2q strategy, hard-negative sampling, optimization schedule, and VIGOR cross-area evaluation protocol. We also ablate the regularization term in Eq. (13) on the VIGOR cross-area split in Tab. 8.

Table 8: Ablation study on the regularization term and temperature scaling.

Method	VIGOR Cross				
	R@1	R@5	R@10	R@1%	Hit rate
Temperature scaling (Zhang et al., 2021)	66.27	87.17	90.85	**98.57**	75.70
Ours	**70.10**	**89.60**	**92.39**	98.55	**79.73**
Ours w/o reg	68.24	88.47	91.68	**98.57**	78.25

Inference scoring. We evaluate each trained ProbInfoNCE model using either cosine similarity or the pair-dependent score, keeping the model fixed within each setting. Tab. 9 reports R@1 for the nine settings in Tab. 6. Pair-dependent scoring gives higher R@1 in all nine settings.

Table 9: Inference scoring ablation. R@1 (%) under the nine evaluation settings of Tab. 6.

Score	VIGOR S	VIGOR C	CAV-Geo S	CAV-Geo C	USA	ACT V	ACT T	USA→ACT	ACT→USA
Cosine $q_i^\top r_j$	81.26	69.33	66.33	20.55	99.11	92.08	73.49	64.84	54.29
Pair-dependent $q_i^\top r_j / \tau_{ij}$	**81.92**	**70.10**	**67.33**	**20.82**	**99.24**	**92.56**	**74.06**	**65.65**	**55.72**

DINOv3 initialization and ProbInfoNCE. Tab. 10 compares Sample4Geo, an InfoNCE model with DINOv3 initialization, and our method on CVUSA and CVACT. All three use original queries without appearance augmentation. The DINOv3-initialized InfoNCE model improves R@1 over Sample4Geo, and our method further improves R@1 in all five settings.

Table 10: Ablation of DINOv3 initialization and ProbInfoNCE on CVUSA and CVACT. We report R@1 (%) using only original queries, without appearance augmentation.

Method	CVUSA	CVACT Val	CVACT Test	CVUSA→CVACT	CVACT→CVUSA
Sample4Geo (Deuser et al., 2023)	98.68	90.35	71.51	56.62	44.95
InfoNCE + DINOv3 init.	98.86	91.27	72.89	60.12	49.03
Ours	**99.24**	**92.56**	**74.06**	**65.65**	**55.72**

B Proof of Eq. (10)

We show that the auxiliary uncertainty model yields the same second-moment identity as Eq. (6). Let

$$s_i^q = q_i + \epsilon_i^q, \qquad s_j^r = r_j + \epsilon_j^r, \tag{15}$$

where

$$\mathbb{E}[\epsilon_i^q] = 0, \qquad \mathbb{E}[\epsilon_j^r] = 0, \qquad \mathrm{Var}(\epsilon_i^q) = \Sigma_i^q, \qquad \mathrm{Var}(\epsilon_j^r) = \Sigma_j^r. \tag{16}$$

Assume that ϵ_i^q and ϵ_j^r are independent, or more generally uncorrelated. Then

$$\mathrm{Cov}(\epsilon_i^q, \epsilon_j^r) = 0. \tag{17}$$

Thus,

$$\mathbb{E}[s_i^q] = q_i, \qquad \mathbb{E}[s_j^r] = r_j, \tag{18}$$

and

$$\mathrm{Var}(s_i^q) = \Sigma_i^q, \qquad \mathrm{Var}(s_j^r) = \Sigma_j^r. \tag{19}$$

Applying $\mathbb{E}\|U\|_2^2 = \|\mathbb{E}U\|_2^2 + \mathrm{tr}(\mathrm{Var}(U))$ to $U = s_i^q - s_j^r$ gives

$$\mathbb{E}\big[\|s_i^q - s_j^r\|_2^2\big] = \mathrm{tr}(\Sigma_i^q) + \mathrm{tr}\big(\Sigma_j^r\big) + \|q_i - r_j\|_2^2 \tag{20}$$

which proves Eq. (10).

C Probabilistic Motivation for ProbInfoNCE

We provide an additional probabilistic motivation for the proposed uncertainty-normalized contrastive objective. Let $q_i, r_j \in \mathbb{R}^d$ denote the deterministic query and reference embeddings, respectively. We assume that both embeddings are ℓ_2-normalized, i.e., $\|q_i\|_2 = \|r_j\|_2 = 1$. Their stochastic counterparts are modeled as

$$s_i^q = q_i + \epsilon_i^q, \qquad s_j^r = r_j + \epsilon_j^r, \tag{21}$$

where ϵ_i^q and ϵ_j^r are independent zero-mean noises as in Section 3.2.

Following Eq. (10), the expected squared distance between the stochastic embeddings can be written as

$$\mathbb{E}\left[\|s_i^q - s_j^r\|_2^2\right] = \|q_i - r_j\|_2^2 + \mathrm{tr}(\Sigma_i^q) + \mathrm{tr}(\Sigma_j^r). \tag{22}$$

We summarize the embedding uncertainty by learned scalar scores. Since these scores are learned and are used only as an effective contrastive uncertainty scale, any fixed global factor, such as the embedding dimension d, can be absorbed into the uncertainty parameterization. Therefore, we define the effective pairwise uncertainty as

$$\tau_{ij} = \sigma_\theta(q_i) + \sigma_\theta(r_j) + \varepsilon, \qquad \varepsilon > 0, \tag{23}$$

where $\sigma_\theta(\cdot)$ is a learnable uncertainty predictor. In this auxiliary model, $\sigma_\theta(\cdot)$ serves as a learned scalar variance surrogate, not a direct estimate of embedding covariance.

Define the stochastic difference variable

$$\Delta_{ij} = s_i^q - s_j^r. \tag{24}$$

For a scoring interpretation, we use the local isotropic Gaussian approximation

$$\Delta_{ij} \sim \mathcal{N}\left(q_i - r_j, \tau_{ij} I_d\right). \tag{25}$$

This form uses τ_{ij} as the effective per-dimension uncertainty. If one instead starts from a raw trace-level uncertainty, the corresponding covariance would be $(\tau_{ij}^{\mathrm{tr}}/d) I_d$. Because τ_{ij} is learned, this fixed factor can be absorbed by redefining $\tau_{ij} = \tau_{ij}^{\mathrm{tr}}/d$.

To score the compatibility of a pair, we evaluate the density of Δ_{ij} at zero:

$$\log p(\Delta_{ij} = 0 \mid q_i, r_j) = -\frac{d}{2}\log(2\pi\tau_{ij}) - \frac{1}{2}(q_i - r_j)^\top (\tau_{ij} I_d)^{-1}(q_i - r_j) \tag{26}$$

$$= C - \frac{d}{2}\log\tau_{ij} - \frac{1}{2\tau_{ij}}\|q_i - r_j\|_2^2, \tag{27}$$

where $C = -\frac{d}{2}\log(2\pi)$ is independent of the candidate index j.

Since q_i and r_j are unit-normalized,

$$\|q_i - r_j\|_2^2 = \|q_i\|_2^2 + \|r_j\|_2^2 - 2q_i^\top r_j = 2 - 2q_i^\top r_j. \tag{28}$$

Substituting Eq. (28) into Eq. (27) gives

$$\log p(\Delta_{ij} = 0 \mid q_i, r_j) = C - \frac{d}{2}\log\tau_{ij} - \frac{1}{2\tau_{ij}}\left(2 - 2q_i^\top r_j\right) \tag{29}$$

$$= C + \frac{q_i^\top r_j}{\tau_{ij}} - \frac{1}{\tau_{ij}} - \frac{d}{2}\log\tau_{ij}. \tag{30}$$

Eq. (30) contains one similarity-dependent term, $\frac{q_i^\top r_j}{\tau_{ij}}$, and two uncertainty-only terms, $-\frac{1}{\tau_{ij}}$ and $-\frac{d}{2}\log\tau_{ij}$. Since τ_{ij} depends on the candidate r_j, these uncertainty-only terms are also candidate-dependent and do not cancel in a softmax over reference candidates. Therefore, directly using the full Gaussian log-density as the contrastive logit would allow the uncertainty predictor to rank candidates through uncertainty-only effects, rather than only modulating the reliability of the semantic similarity.

To isolate the similarity-dependent evidence, we define the matching evidence as a likelihood ratio against a zero-similarity baseline with the same uncertainty level. Specifically, let $p_0(\Delta_{ij} = 0 \mid \tau_{ij})$ denote the density under a baseline pair whose uncertainty is τ_{ij} but whose cosine similarity is zero, i.e., $q_i^\top r_j = 0$. From Eq. (30), the baseline log-density is

$$\log p_0(\Delta_{ij} = 0 \mid \tau_{ij}) = C - \frac{1}{\tau_{ij}} - \frac{d}{2}\log\tau_{ij}. \tag{31}$$

The log-likelihood ratio is therefore

$$\log \frac{p(\Delta_{ij} = 0 \mid q_i, r_j)}{p_0(\Delta_{ij} = 0 \mid \tau_{ij})} = \frac{q_i^\top r_j}{\tau_{ij}}. \tag{32}$$

This likelihood-ratio formulation removes the candidate-dependent uncertainty-only terms by construction while preserving the similarity-dependent term normalized by pairwise uncertainty. The choice of baseline motivates the score but does not follow uniquely from the expected-distance identity or establish calibrated matching probabilities. We therefore define the ProbInfoNCE matching score as

$$g_{ij}^{q\to r} = \frac{q_i^\top r_j}{\tau_{ij}}. \tag{33}$$

The conditional probability that reference r_j matches query q_i is modeled by a softmax over all reference candidates:

$$p_\theta(j \mid q_i) = \frac{\exp(g_{ij}^{q\to r})}{\sum_{k=1}^{N}\exp(g_{ik}^{q\to r})}. \tag{34}$$

Maximizing the likelihood of the positive reference r_i yields the query-to-reference loss

$$\ell_i^{q\to r} = -\log\frac{\exp(q_i^\top r_i/\tau_{ii})}{\sum_{j=1}^{N}\exp(q_i^\top r_j/\tau_{ij})}. \tag{35}$$

The reverse direction is defined analogously. For the pair (r_i, q_j), the corresponding uncertainty is

$$\tau_{ji} = \sigma_\theta(q_j) + \sigma_\theta(r_i) + \varepsilon. \tag{36}$$

Thus, the reference-to-query loss is

$$\ell_i^{r\to q} = -\log \frac{\exp(r_i^\top q_i/\tau_{ii})}{\sum_{j=1}^N \exp(r_i^\top q_j/\tau_{ji})}. \tag{37}$$

Finally, including the uncertainty regularizer in Eq. (13), the bidirectional ProbInfoNCE loss is

$$\mathcal{L}_{\text{ProbInfoNCE}} = \frac{1}{2N}\sum_{i=1}^N (\ell_i^{q\to r} + \ell_i^{r\to q}) + \frac{\lambda}{2N}\sum_{i=1}^N [\sigma_\theta(q_i) + \sigma_\theta(r_i)]. \tag{38}$$

D IMAGE EDITING

Our dataset contains cities worldwide, including Johannesburg, London, Rio, Sydney, Taipei, Tokyo, Chicago, New York, San Francisco, Seattle, and Vancouver. Depending on the GPS location, we do not include snowy settings for Taipei, Johannesburg, Rio, San Francisco, and Sydney. We have the following choices:

Cities with snow edits:

- rain
- fog
- night
- rain + fog
- rain + night
- fog + night
- rain + fog + night
- snow
- snow + fog
- snow + night
- snow + fog + night

Cities without snow edits:

- rain
- fog
- night
- rain + fog
- rain + night
- fog + night
- rain + fog + night

We sample the choices with equal probability and let the editing model change the weather or time of day. For each input query, we provide FLUX.2 [klein] 9B with a text prompt specifying the sampled conditions and retain all default inference settings, including CFG and the number of steps. We do not manually override other generation parameters. The final distribution of weather conditions is shown in Tab. 11.

Table 11: Per-city and partition-level weather condition distribution. Values are percentages.

City / Partition	N	Rain	Fog	Night	Rain+ Fog	Rain+ Night	Fog+ Night	Rain+ Fog+ Night	Snow	Snow+ Fog	Snow+ Night	Snow+ Fog+ Night
Chicago	25,479	8.79	9.35	9.49	8.85	8.82	8.96	9.28	9.12	8.85	9.16	9.33
London	22,383	8.69	9.42	9.34	8.89	8.85	8.95	9.23	9.24	8.87	9.11	9.40
Johannesburg	21,697	14.08	14.13	14.14	14.58	14.02	14.93	14.11	0.00	0.00	0.00	0.00
Seattle	23,751	8.82	9.41	9.39	8.80	8.83	8.97	9.23	9.17	8.93	9.11	9.33
Taipei	21,541	14.08	14.13	14.16	14.58	14.00	14.93	14.13	0.00	0.00	0.00	0.00
New York	27,769	8.93	9.36	9.52	8.85	8.85	8.92	9.27	9.10	8.91	9.14	9.15
Rio	20,365	14.11	14.10	14.17	14.63	13.99	14.98	14.03	0.00	0.00	0.00	0.00
Vancouver	21,567	8.62	9.37	9.34	8.92	8.88	9.03	9.25	9.25	8.87	9.05	9.43
Sydney	21,707	14.08	14.13	14.14	14.58	14.02	14.93	14.12	0.00	0.00	0.00	0.00
Tokyo	21,401	8.64	9.38	9.29	8.91	8.89	9.04	9.27	9.23	8.86	9.06	9.43
San Francisco	28,215	14.21	14.33	13.97	14.49	14.05	14.71	14.23	0.00	0.00	0.00	0.00
North America	126,781	10.00	10.47	10.45	10.11	10.00	10.24	10.37	7.12	6.91	7.09	7.23
Non-North America	129,094	12.25	12.52	12.51	12.66	12.27	12.93	12.45	3.13	3.01	3.08	3.19
Overall	**255,875**	**11.14**	**11.51**	**11.49**	**11.40**	**11.14**	**11.60**	**11.42**	**5.11**	**4.94**	**5.07**	**5.19**

E EVIDENCE FOR GEOLOCATION

Table 12: SIFT inlier ratios, CLIP feature similarity, and DINOv3 patch-feature similarity.

Average Inlier Ratio	Weighted Average Inlier Ratio	CLIP similarity	DINOv3 similarity
0.5999	0.7619	0.7660	0.8111

We include the average inlier ratio, weighted average inlier ratio for SIFT+RANSAC, as well as CLIP feature similarity and DINOv3 patch-feature similarity between the original queries and edited queries in our CAV-Geo dataset in Tab. 12. The average inlier ratio is defined as

$$\left(\frac{\sum_{i=1}^{N} \frac{\text{number of inliers in pair } i}{\text{number of matches in pair } i}}{N} \right)$$

The weighted inlier ratio is defined as

$$\left(\frac{\sum_{i=1}^{N} \text{number of inliers in pair } i}{\sum_{i=1}^{N} \text{number of matches in pair } i} \right)$$

Both CLIP and DINOv3 similarities remain high after editing. CLIP (Radford et al., 2021) assesses broad semantic similarity, while our DINOv3 (Siméoni et al., 2025) comparison uses patch features to assess local visual similarity. The high DINOv3 patch-feature similarity supports local visual consistency between original and edited queries. The weighted average inlier ratio is much higher than the simple average, which means that for the pairs with more points of interest (more confidence), the inlier ratio turns out to be higher, also representing strong structural similarity.

In addition to quantitative numbers, we include additional visualizations to show that editing does not change the structure in Fig. 4 (random selection), Fig. 5 (lowest SIFT inlier ratio), Fig. 6 (lowest CLIP similarity), and Fig. 7 (lowest DINOv3 similarity). Though there are some hallucinated patterns in transient objects, such as light sources (in red boxes) and changing the face direction of pedestrians (in red circles), we did not observe any structural changes altering the geolocation cues.

MLLM judge. As there is no existing method to detect whether the image editing changes geolocation cues, we use GPT6-Astra via Codex to directly inspect all 255,875 original-edited query pairs in CAV-Geo. We ask the MLLM "Do the two provided images describe the same geolocation? Think carefully with the structures and layouts. Output 1 if yes and 0 otherwise." We average the response value from GPT6-Astra and get a value > 0.9999, meaning that GPT6-Astra consistently believes the image after editing still describes the same location. This agreement is an auxiliary consistency check on original-edited pairs and does not measure false-positive rates on mismatched locations.

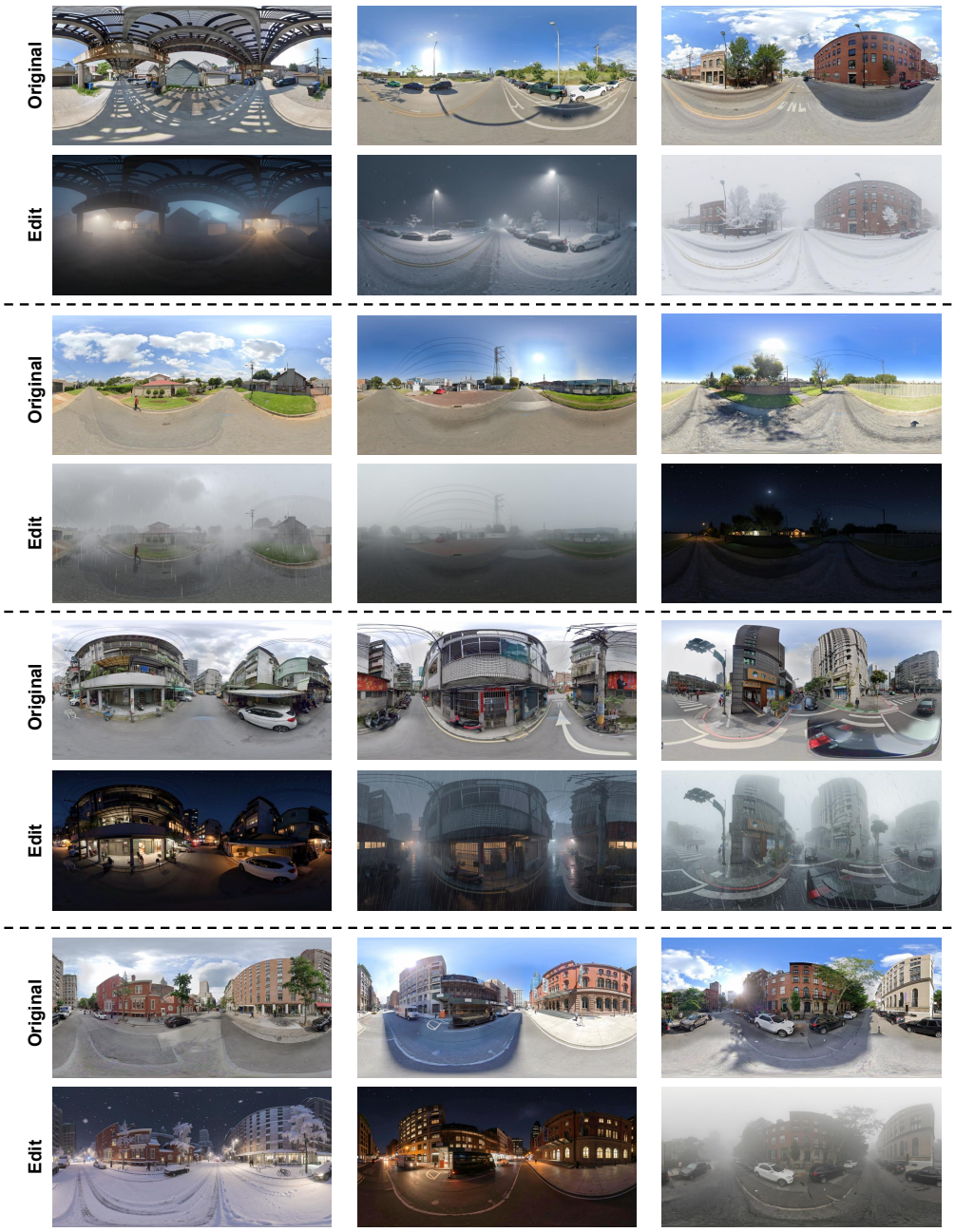


Figure 4: Additional comparisons of original street-view images and edited street-view images with random selection. Overall, we did not observe any geolocation cues, such as building and road structures, being changed.



Figure 5: Additional comparisons of original street-view images and edited street-view images with the lowest SIFT inlier ratio (all 0). Overall, we did not observe any changes in geolocation cues, such as building and road structures.

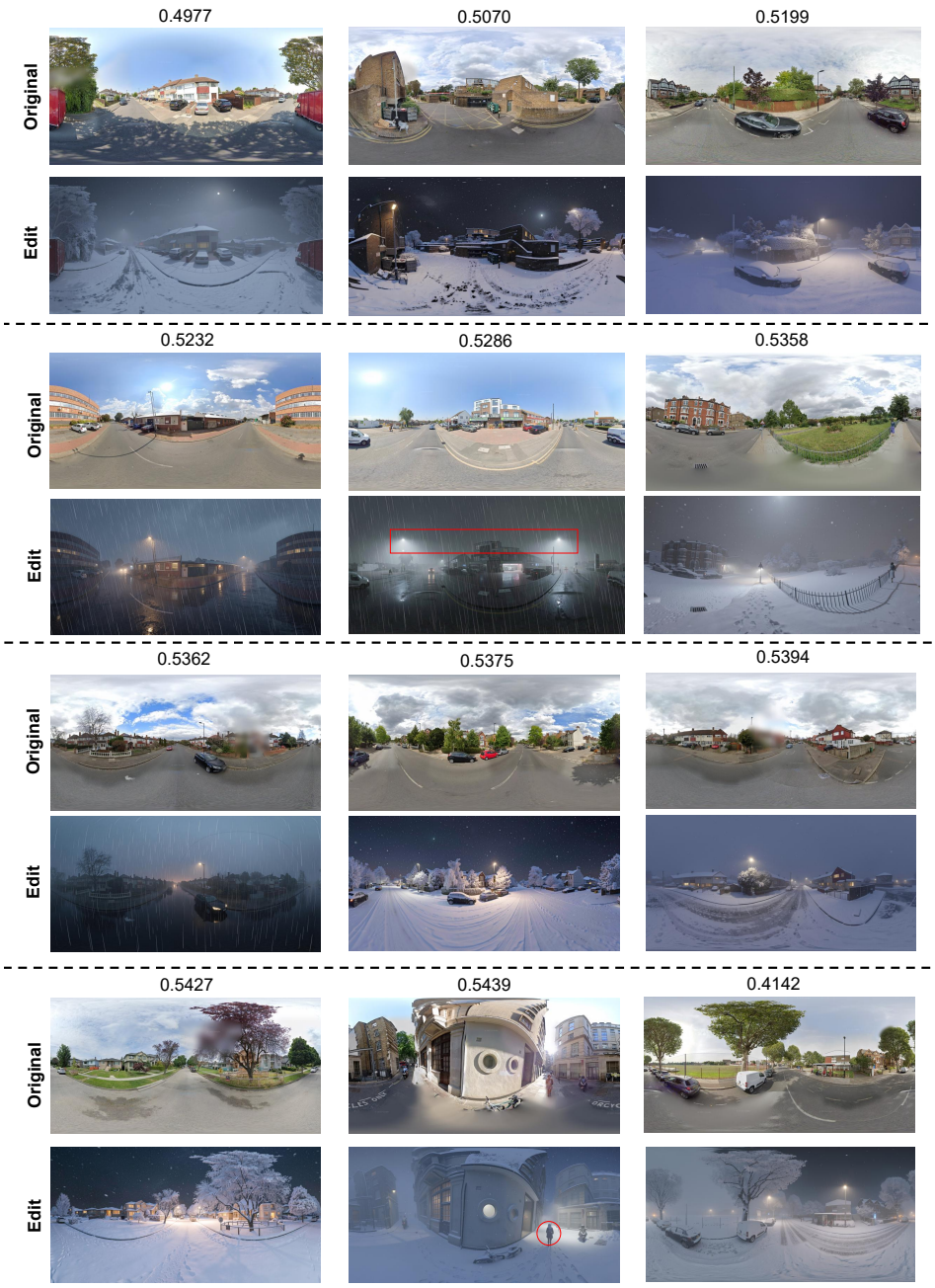


Figure 6: Additional comparisons of original street-view images and edited street-view images with the lowest CLIP similarity. Numbers indicate CLIP similarity. The red box marks a hallucinated light source. The red circle marks a case where FLUX changes a person from facing toward the camera to facing away. However, there are no structural differences. Overall, we did not observe any geolocation cues, such as building and road structures, being changed.



Figure 7: Additional comparisons of original street-view images and edited street-view images with the lowest DINOv3 similarity. Numbers indicate DINOv3 similarity. Overall, we did not observe any geolocation cues, such as building and road structures, being changed.

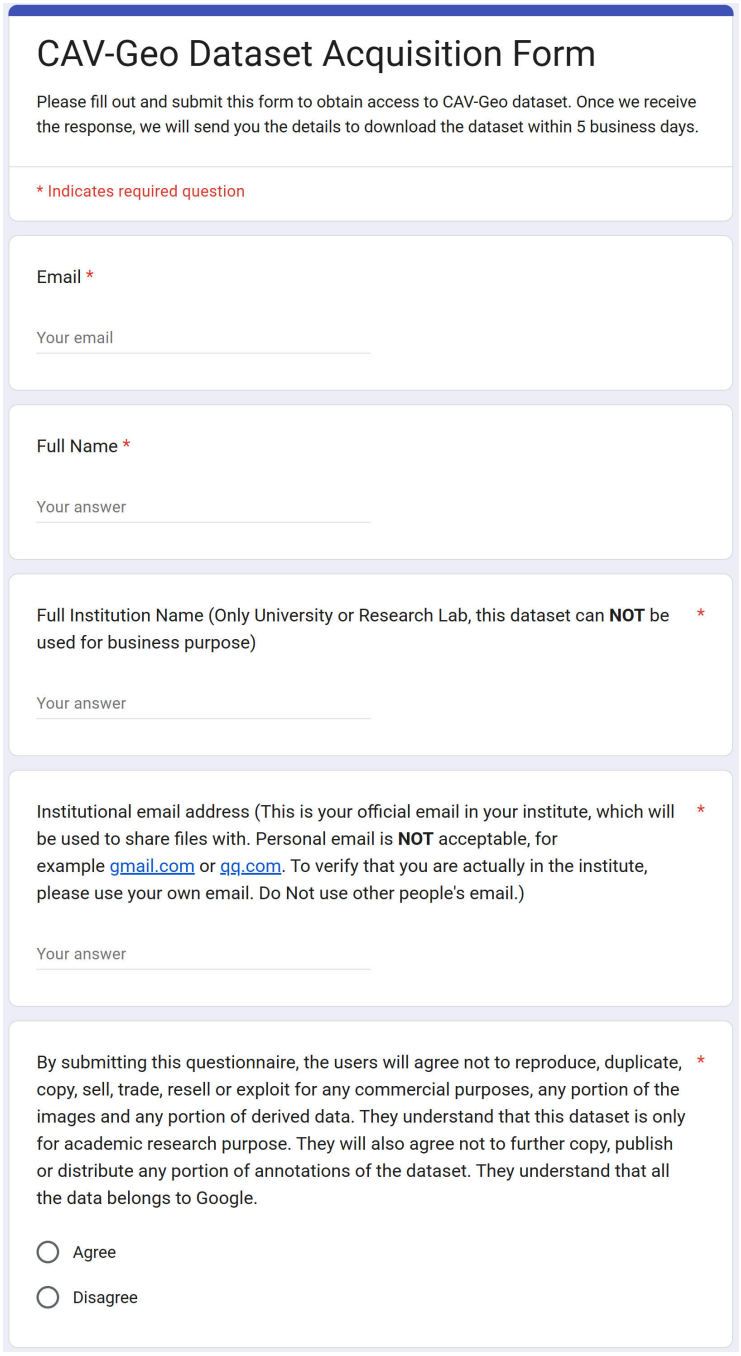
F DATASET DETAILS

Table 13: Detailed training and testing split statistics of the proposed CAV-Geo dataset. "NA" is short for "North American".

		Same-Area		Cross-Continent	
		Number	City	Number	City
Train	Reference	439,846	All	277,742	Outside NA
	Query Locations	128,056	All	129,094	
Test	Reference	439,846	All	162,104	NA
	Query Locations	127,819	All	126,781	

Satellite-image collection and the definitions of the main-positive and three semi-positive reference tiles follow exactly the VIGOR protocol (Zhu et al., 2021). Table 13 details the distribution of training and testing samples across the same-area and cross-continent settings. Notably, each query location has an image pair: a clear-weather ground query image and its weather-augmented counterpart. Recall that cities outside North America are Johannesburg, London, Rio De Janeiro, Sydney, Taipei, and Tokyo, while North American cities are Chicago, New York, San Francisco, Seattle, and Vancouver. This setup supports evaluation on both original clear-weather queries and queries with synthetic weather and daylight changes, as shown in Table 5. Note that we use the Non-NA $\rightarrow$ NA as the default cross-continent split because it creates a challenging out-of-distribution test.

We use Google Maps Platform content under the Google Maps Platform Terms of Service and Google Maps Platform Service Specific Terms, rather than an open-source dataset license. We will provide two ways to make the dataset accessible. First, the exact API calls and editing code with a fixed random seed to reproduce will be provided. Second, we will follow VIGOR (Zhu et al., 2021) to follow the constraints of Google Maps Platform. We will create a direct application form so that a research-only, non-commercial academic institute can apply with the agreement of no distribution (see Fig. 8). It is important to notice that images downloaded from the Google Maps Platform API belong to Google as their assets.



CAV-Geo Dataset Acquisition Form

Please fill out and submit this form to obtain access to CAV-Geo dataset. Once we receive the response, we will send you the details to download the dataset within 5 business days.

* Indicates required question

Email *

Your email

Full Name *

Your answer

Full Institution Name (Only University or Research Lab, this dataset can **NOT** be used for business purpose) *

Your answer

Institutional email address (This is your official email in your institute, which will be used to share files with. Personal email is **NOT** acceptable, for example gmail.com or qq.com. To verify that you are actually in the institute, please use your own email. Do Not use other people's email.) *

Your answer

By submitting this questionnaire, the users will agree not to reproduce, duplicate, copy, sell, trade, resell or exploit for any commercial purposes, any portion of the images and any portion of derived data. They understand that this dataset is only for academic research purpose. They will also agree not to further copy, publish or distribute any portion of annotations of the dataset. They understand that all the data belongs to Google. *

- Agree
- Disagree

Figure 8: Google form for application. Images downloaded from Google Maps Platform API belong to Google as their assets.

G Understanding the Predicted Uncertainty σ

The query score σ_q in Eq. 11 is learned through the retrieval objective and should not be equated with semantic difficulty or appearance-change severity. To examine its relation to retrieval outcomes, we group VIGOR cross-area queries by top-1 correctness. The mean σ_q is 0.039 for correct retrievals and 0.043 for failures, approximately 10% higher for the latter. This association supports interpreting σ_q as retrieval-relevant uncertainty, while editing severity, pixel statistics, and retrieval ambiguity need not vary together.

H Reproducing PanoBEV

PanoBEV (Ye et al., 2024) open-sourced its CVACT training and evaluation code, as well as the validation split dataset. **We reproduce the CVACT validation results based on its weights and code, without any modifications by us**. We follow its BEV transform code to transform queries in CVUSA to evaluate CVACT→CVUSA based on the dataloader provided by Sample4Geo (Deuser et al., 2023) and weights provided by PanoBEV.

For other training data, since PanoBEV does not open-source the code and weights, we modify the CVACT version to adapt it to other datasets such as CVUSA and VIGOR. We train the models following their settings. We base our reproduction on commit "30afb334a34d0441a807634bbe2c3a8b90958f93" from July 7, 2025. To verify the correctness of our reproduction, we test the BEV-only case in the VIGOR cross split. We find that R@1 is 45.75, while PanoBEV reports 44.01. This indicates that our reproduction of the BEV-satellite retrieval is **at least as good as the results reported in PanoBEV**. We also verify the street-view-to-satellite retrieval by evaluating weights provided by Sample4Geo (Deuser et al., 2023), which turns out to be correct. In addition, using the provided CVACT weights, we are able to reproduce the results in CVUSA in our codebase with negligible differences. All these indicate the correctness of reproduction. Then we strictly follow their code to add the similarity scores from street-view-satellite and BEV-satellite retrieval, resulting in R@1 = 67.20, as reported in Tab. 2.

I Training Details

We use AdamW (Loshchilov & Hutter, 2019) with label smoothing of 0.1 and a batch size of 128. ConvNeXt-B models initialized with DINOv3-distilled weights are trained for 80 epochs with a learning rate of 10^{-4}. For the ImageNet-pretrained ConvNeXt-B used by Sample4Geo, we use a learning rate of 10^{-3} and train for 40 epochs. We evaluate the final-epoch checkpoint, including in cross-dataset transfer. In our experiments, applying the DINOv3 training schedule (10^{-4}, 80 epochs) to the ImageNet-pretrained Sample4Geo model reduces VIGOR R@1 from 77.86 to 77.11 on same-area and from 61.70 to 61.25 on cross-area. We therefore retain the initialization-specific schedules. The controlled InfoNCE and ProbInfoNCE comparisons with the same DINOv3 initialization use the same 80-epoch schedule. Hard negative sampling is performed every 4 epochs. We follow the data augmentation implementation in Sample4Geo. We set $\varepsilon = 10^{-6}$ in Eq. (11) for numerical stability. The lightweight MLP has the following architecture:

```
nn.Sequential(
            nn.Linear(1024, 512),
            nn.ReLU(),
            nn.Linear(512, 1),
            nn.Softplus()
        )
```

The input to the MLP is the average-pooled feature before L_2-normalization. The MLP has about 0.5M parameters, which is negligible compared to an 88.6M-parameter CNN (ConvNeXt-B). The models are trained with a single NVIDIA H100 GPU with bfloat16 precision. The training strategy is visualized in Fig. 9.

The results from our training runs reported in the tables are single-run values rather than averages across seeds. Due to computational constraints, we repeated only the final ProbInfoNCE model with dual-matrix and q2q training on VIGOR cross-area across five training seeds. Its mean R@1 is 70.13,

with a standard deviation of 0.13 percentage points. The reported R@1 of 70.10 is a single-run result, not this five-seed mean.

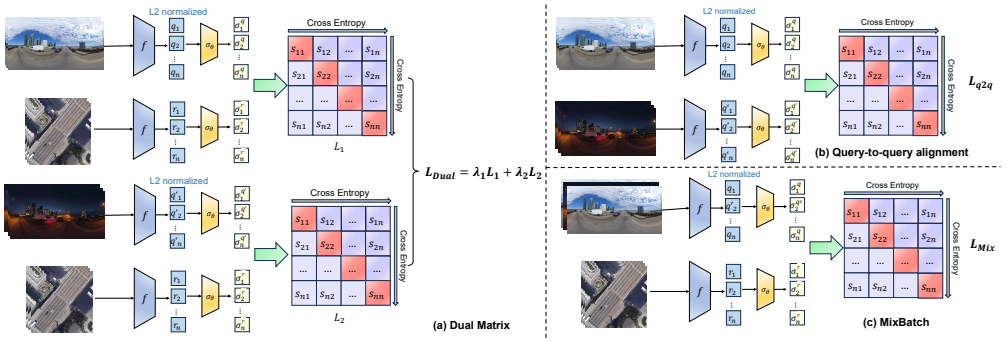


Figure 9: Visualization of our training strategy. (a) Dual Matrix. (b) Query-to-query (q2q) alignment. (c) MixBatch.

J Additional Qualitative Evaluation

We include additional qualitative comparisons in Fig. 10 to illustrate Sample4Geo's sensitivity to appearance changes when trained only on clear-weather queries. For night cases, it tends to rank dark satellite images very high to fit the darkness. In the daytime, if there is snow or fog that results in a large portion of white background, it tends to rank white satellite images very high to fit the white background. We also include qualitative evaluations on clean queries in Fig. 11.

The failure cases are shown in Fig. 12. When the query contains no significant landmarks, the retrievals may fail due to the existence of satellite images with very similar patterns.

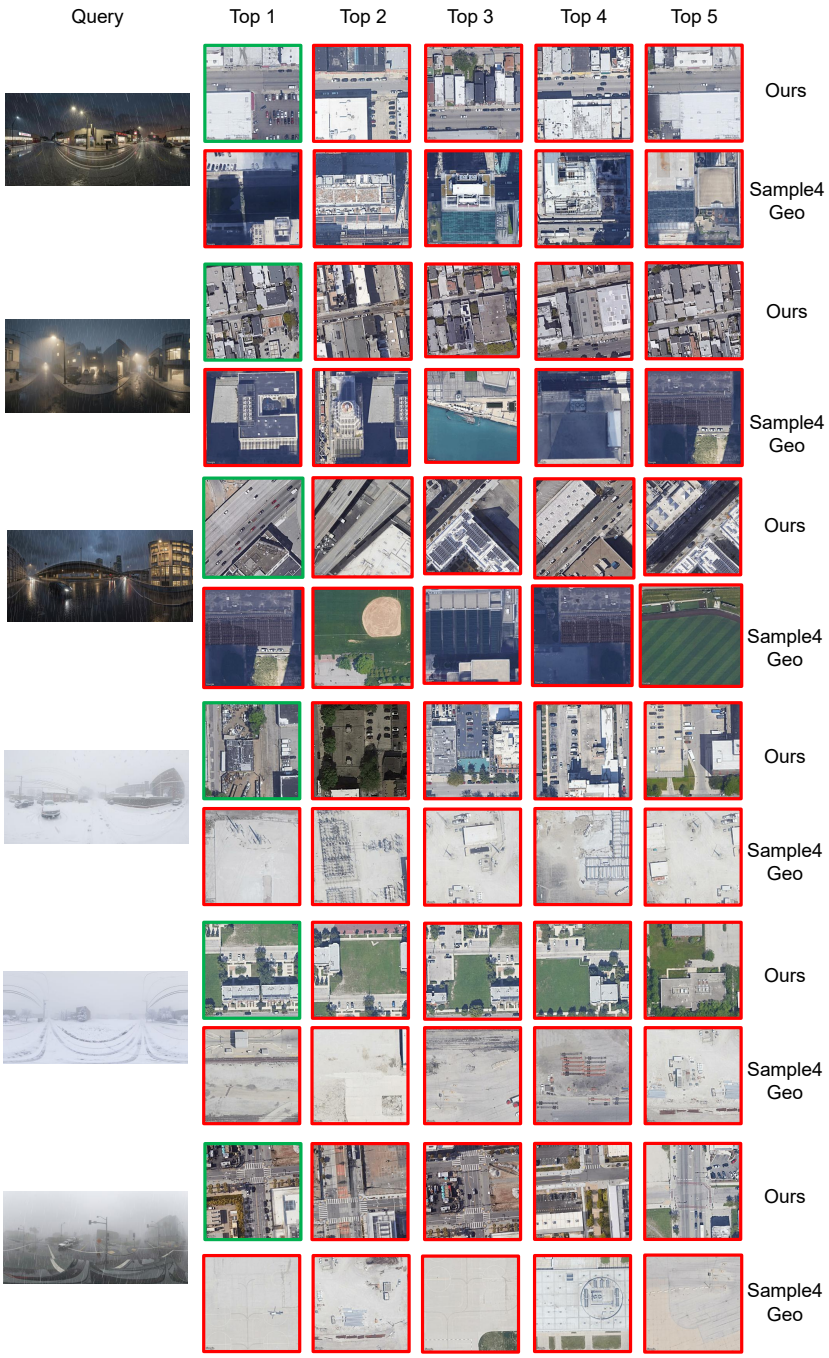


Figure 10: Additional qualitative results. In edited night, snow, and fog queries, Sample4Geo can rank satellite images with similar broad colors highly, illustrating its sensitivity to appearance changes. Correct results are in green, and incorrect results are in red.

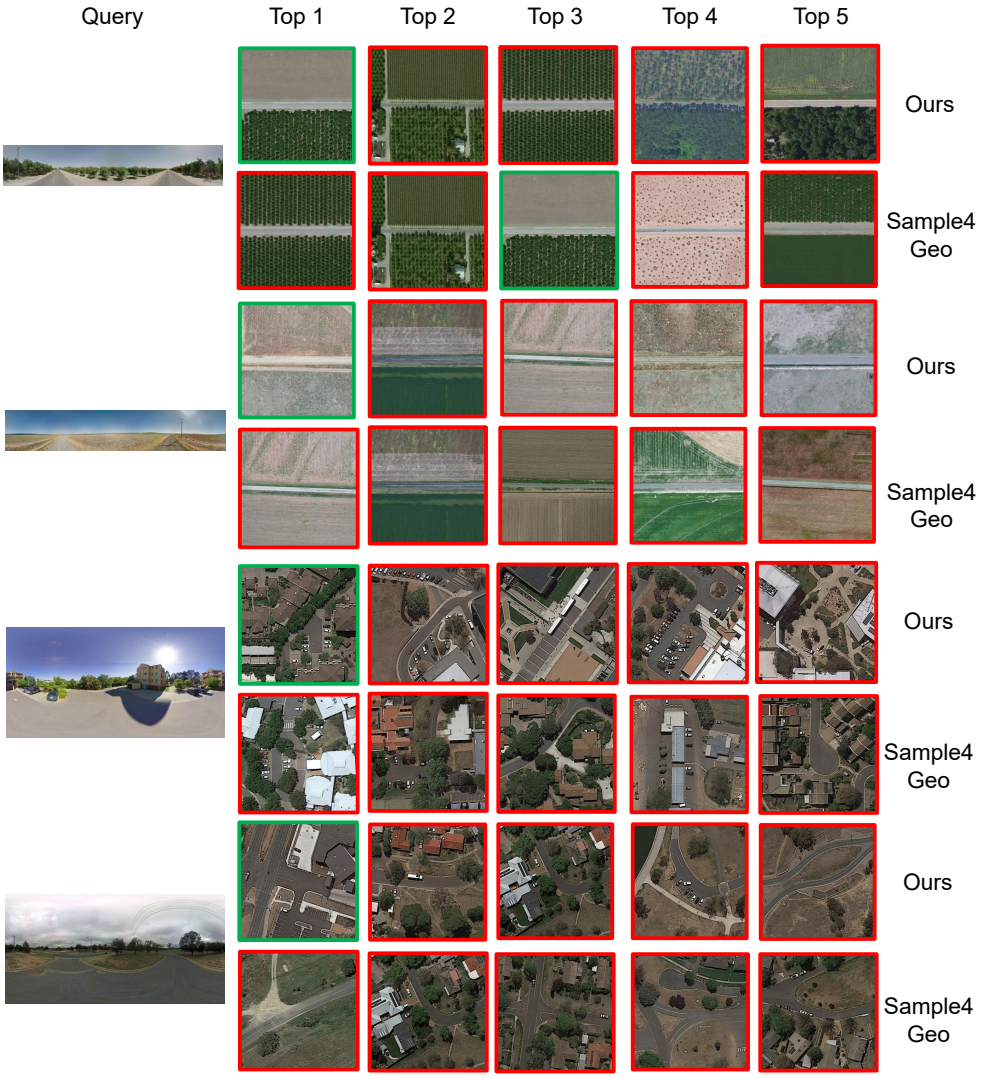


Figure 11: Additional qualitative results with clean queries. Correct results are in green, and incorrect results are in red.

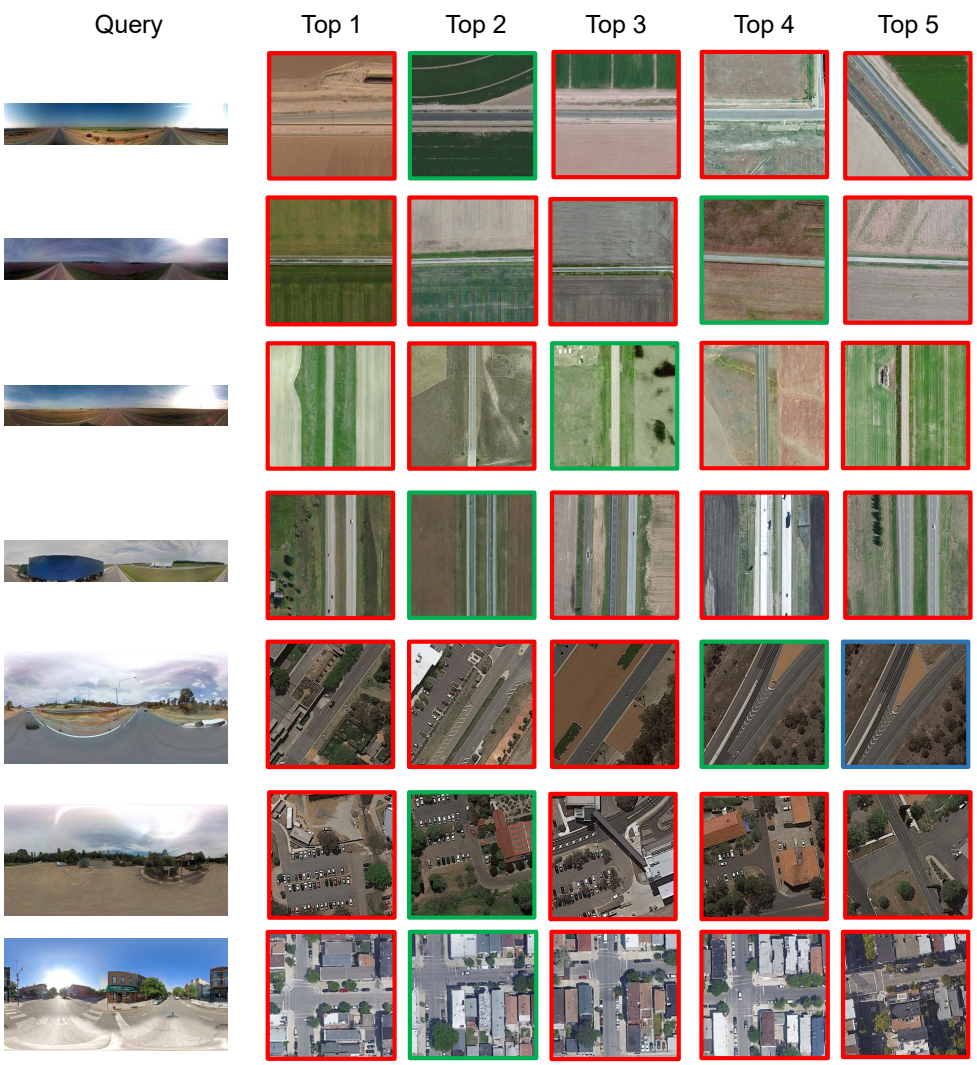


Figure 12: Failure cases of our method. When the query contains no apparent landmarks, the retrievals fail on satellite images with very similar patterns. Correct results are in green, and incorrect results are in red.

K NO-CHANGE FLUX EDITING CONTROL

To check whether the FLUX generation process itself causes the large drop on weather- and daylight-edited queries, we edit the original clear-weather VIGOR queries with FLUX.2 [klein] 9B (Black Forest Labs, 2026) using the prompt *"do not change anything"*. The outputs pass through the full image-editing model while retaining the original clear-weather condition. We evaluate the same Sample4Geo (Deuser et al., 2023) model on original and no-change edited queries. Tab. 14 shows only small R@1 changes, from 77.86 to 77.71 on VIGOR Same and from 61.70 to 61.21 on VIGOR Cross. Thus, FLUX processing without a weather or daylight change cannot by itself explain the much larger degradation under appearance changes. This control tests the full editing process, not VAE encoding and decoding alone.

Table 14: Sample4Geo evaluated on original clear-weather VIGOR queries and FLUX edits prompted with "do not change anything".

Method	VIGOR Same					VIGOR Cross				
	R@1	R@5	R@10	R@1%	Hit Rate	R@1	R@5	R@10	R@1%	Hit Rate
Original clear query	77.86	95.66	97.21	99.61	89.82	61.70	83.50	88.00	98.17	69.87
FLUX no-change edit	77.71	95.58	97.14	99.61	89.64	61.21	83.31	87.82	98.15	69.41

L LIMITATIONS AND FUTURE WORK

Despite the strong performance, there are limitations due to current technology and our methods.

Limitations of current technology.

1. There is currently no reliable automatic detector that can determine whether an image-editing model has altered geographically discriminative cues, such as road layout, building structure, or other persistent scene geometry. To reduce this risk, we visually inspected the 100 examples with the lowest SIFT+RANSAC inlier ratio, the 100 examples with the lowest CLIP similarity, the 100 examples with the lowest DINOv3 similarity, and 100 randomly sampled examples. We did not observe cases in which major geographically discriminative cues were changed. Nevertheless, subtle editing artifacts may still exist, and developing automatic detectors for geolocation-preserving image editing remains an important direction for future work.
2. Generating and evaluating every combination of weather and daylight conditions for every query is computationally expensive, as it would require multiple edited versions of each query and substantially increase storage, training, and inference costs. Although image editing can be made deterministic by fixing random seeds, generating the full dataset remains costly. In our pipeline, the distilled FLUX.2 [klein] 9B model (Black Forest Labs, 2026) requires days to weeks of H100 computation to produce appearance-edited images at the scale of CAV-Geo. Therefore, we generate one appearance-edited counterpart for each original query and leave exhaustive multi-condition generation as future work.
3. To our knowledge, there is no public CVGL dataset that provides paired ground-view queries of the same locations captured under real weather and daylight changes at meaningful scale. Such data would be valuable for validating synthetic appearance variation and benchmarking real-world robustness, but collecting it is extremely challenging because it requires repeated geo-aligned street-view acquisition across different times, seasons, and weather conditions. Future efforts to collect even small-scale real paired weather/daylight benchmarks would provide an important complement to CAV-Geo.

Limitations of our method. ProbInfoNCE does not decouple v, a, ξ at the algorithmic level. It handles the general variation in the feature space where everything is mixed up. The decoupling is performed at the data level, such as appearance variation introduced by us, BEV transforms proposed by Ye et al. (2024), and the equirectangular-to-perspective transformation used by Zhang

et al. (2026). Decoupling these variations or identifying other types of variation will be beneficial for future research.

Future work. We are not able to properly include seasonal changes, but instead integrate winter as snowy weather. The reason is that the seasonal behavior is very different from city to city, country to country, and continent to continent. In some cities, such as those in Southern California, we cannot even tell the seasonal change without feeling the temperature. In some cities, such as New York City, there are golden and red maple leaves in the fall, but this is a specific characteristic of maples. Therefore, we are unable to include seasonal change in a systematic and realistic way at this point. We leave this as potential future work to enrich the appearance variation.

We do not apply weather or daylight variations to the reference satellite images because this setting is less aligned with practical cross-view geo-localization. In real retrieval systems, the reference set is usually a **fixed offline database**: satellite images are collected, curated, and indexed in advance, and are expected to be relatively stable and high-quality. In contrast, the query image is captured by the user at test time, and is therefore naturally affected by uncontrolled weather, illumination, season, and time-of-day changes. Thus, the main practical robustness challenge lies on the query side rather than the reference side.

Moreover, adverse conditions are not symmetric across ground and satellite views. Rain, fog, snow, and nighttime directly affect the visibility, texture, and color distribution of ground-view images, while satellite images are top-down and typically selected to avoid severe degradation. Applying the same synthetic appearance edits to satellite references may therefore introduce unrealistic artifacts and confound the evaluation, making it measure robustness to artificial reference corruption rather than realistic query-side appearance variation.

Training with edited satellite references is a possible extension, but it would introduce additional engineering and loss-balancing complexity. If both queries and references have original and augmented versions, the objective may need to balance terms such as $\mathcal{L}(q^o, r^o)$, $\mathcal{L}(q^a, r^o)$, $\mathcal{L}(q^o, r^a)$, $\mathcal{L}(q^a, r^a)$, $\mathcal{L}(q^a, q^o)$, and $\mathcal{L}(r^a, r^o)$. This requires further tuning of loss weights. Due to resource constraints, we focus on the practically motivated query-side appearance variation and leave reference-side augmentation as a direction for future work.

M Broader Impact

The proposed CVGL model may have positive societal impacts by improving visual localization in applications such as GPS retrieval, disaster response, map updating, and emergency localization. For example, cross-view matching can provide an additional localization cue when GPS signals are unavailable, noisy, or unreliable, and may help emergency responders quickly identify locations from ground-view images after disasters. It can also support map maintenance by matching ground-level observations with aerial imagery to detect outdated or changed geographic information.

However, the model may also introduce negative societal risks if misused. Since CVGL aims to infer geographic location from visual content, it could be used to identify sensitive or private locations from personal images, including homes, workplaces, hospitals, shelters, or other places where individuals may not expect location disclosure. It could also be misused for surveillance, stalking, or unauthorized tracking, especially when applied to images shared online. These potential risks are inherent to the CVGL task rather than specific to our method. Following common practice in this research area, we adopt a controlled dataset-release process for academic and research-only, non-commercial use, with access governed by a user agreement that restricts redistribution and improper use. We also encourage future work on privacy-preserving CVGL, misuse detection, and deployment safeguards for real-world localization systems.