X MARKS THE SPOT: DIFFUSION FOR IMAGE GEOLOCALIZATION WITH LOCATION REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image geolocalization predicts the location at which a photo was taken. Accurate geolocalization requires discriminating between both physically and semantically adjacent areas on Earth. However, prior diffusion-based approaches predict locations by generating coordinates, which are semantically unstable over space. We avoid this limitation by learning geolocalization behavior in the pretrained representation spaces of location encoders, which provide both positional *and* semantic understanding. To make this space usable for coordinate prediction, we introduce a **Location Autoencoder (LAE)**, which pairs a frozen location encoder with a learned decoder mapping embeddings back to coordinates with kilometer-level accuracy worldwide. In the LAE space, we train **DIG**, a latent diffusion model that transports noise to location embeddings conditioned on image features. DIG consistently outperforms coordinate-based generative models and sets a new state of the art on three geolocalization benchmarks: OSV-5M, Im2GPS3k, and YFCC4k. Notably, DIG is the first generative model to outperform retrieval-based methods, and further improves retrieve-and-rerank systems when used as a prior. Source material will be released upon acceptance.

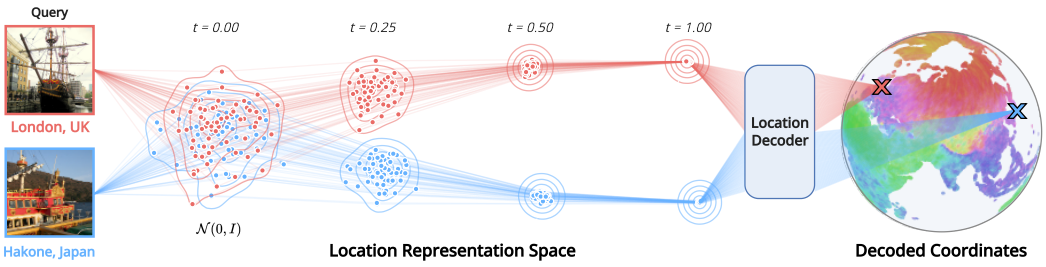


Figure 1: We propose **D**iffusion for **I**mage **G**eolocalization (**DIG**) in the location representation space. Conditioned on the image, a rectified flow transports noise to location embeddings, which a learned decoder maps to coordinates on the globe.

1 INTRODUCTION

Locating where a photo was taken supports numerous investigative tasks, from forensic applications to human trafficking (Yokota et al., 2020; Bamigbade et al., 2024), media verification (Tahmasebzadeh et al., 2023), and species range mapping (Mac Aodha et al., 2019; Cole et al., 2023). Yet many photographs are never geotagged (Bamigbade et al., 2024), making predictive geolocalization necessary to recover location information from only visual content. Progress in geolocalization performance has advanced dramatically from pre-deep learning (Hays & Efros, 2008) to large transformer architectures (Seo et al., 2018) and retrieval augmented approaches (Haas et al., 2024).

The problem remains challenging because a photograph is often ambiguous. An image in a forest or of a suburban street can be consistent with many places across the globe (Dufour et al., 2025).

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.

Nonetheless, classification- and retrieval-based methods predict a single location from a predefined grid cell (Vo et al., 2017; Seo et al., 2018) or a reference gallery (Vivanco Cepeda et al., 2023; Gadi et al., 2026), respectively. Diffusion-based models remove these constraints by learning a probability distribution over continuous coordinates conditioned on the image (Dufour et al., 2025; Wang et al., 2025). These generative approaches denoise coordinates on the sphere (Dufour et al., 2025) or a spherical harmonic encoding of the location (Wang et al., 2025), adopting diffusion spaces that respect the geometry of the Earth. However, since these spaces encode position alone, two locations are similar only if they are nearby. Imagery behaves differently: a pine forest in Canada looks like a pine forest in Siberia, and a city block looks nothing like farmland a few kilometers away.

Location encoders learn representations with precisely this structure. They map a latitude–longitude pair to a high-dimensional embedding through a positional encoding and a network trained to align with co-located imagery (Vivanco Cepeda et al., 2023; Klemmer et al., 2025; Cher et al., 2026a). The positional encoding resolves fine spatial detail, and the alignment maps visually similar environments to nearby embeddings, even when they are far apart on the globe (Klemmer et al., 2025; Cher et al., 2026a). In image generation, recent work shows that frozen, pretrained vision foundation models can serve directly as diffusion spaces, achieving state-of-the-art visual quality (Zheng et al., 2026; Tong et al., 2026). We explore whether the same holds for geolocalization: *can pretrained location representations serve as a diffusion space for generative geolocalization?*

Predicting geolocations by generating location embeddings requires an inverse mapping from representations back to coordinates. To this end, we introduce *Location Autoencoders* (LAEs), which pair a frozen pretrained location encoder with a learned decoder that recovers original locations. In the LAE latent space, we train DIG, a diffusion model that generates a location embedding conditioned on image features (Figure 1). We evaluate this approach with three location encoders, GeoCLIP, SatCLIP, and TTE. Our contributions are as follows:

- We propose **DIG**, a latent diffusion model for image geolocalization in the location representation space. Using a simple architecture and denoising objective, we show that changing the generation space from coordinates to location embeddings improves geolocalization accuracy at every scale.
- We introduce **Location Autoencoders (LAEs)**, a pairing of pretrained location encoders with learned coordinate decoders that reconstruct to kilometer-level accuracy across the globe, enabling coordinate prediction.
- We show that DIG sets a new state of the art on OSV-5M, Im2GPS3k, and YFCC4k, and its samples can further serve as a prior to improve retrieve-and-rerank system at every distance threshold.

2 Related Works

Global Visual Geolocalization. Image geolocalization models can primarily be categorized as classification-, retrieval-, or MLLM-based. Classification methods (Weyand et al., 2016; Seo et al., 2018; Muller-Budack et al., 2018; Pramanick et al., 2022; Clark et al., 2023; Haas et al., 2024) partition the Earth into cells and predict one. Retrieval methods (Haas et al., 2023; Vivanco Cepeda et al., 2023; Gadi et al., 2026; Jia et al., 2025) match the query against a geotagged database and return the top candidate's location. MLLM-based methods (Li et al., 2024; Zhou et al., 2024; Li et al., 2025; Jin et al., 2026; Jia et al., 2026a; Li et al., 2026; Yu et al., 2026; Jia et al., 2026b; Le-Duc et al., 2026) reason with multimodal large language models, often with search and map tools. However, classification and retrieval are bounded by the geographic distribution of their cells or reference database, while MLLM-based methods require multi-step inference through large models.

Recently, generative approaches have offered an alternative. Rather than predicting a single point, they capture visual ambiguity by modeling a distribution over locations. Dufour et al. (2025) uses Riemannian flow matching in coordinates on the sphere while LocDiff (Wang et al., 2025) diffuses in multi-scale spherical harmonics. Both trail strong retrieval-based methods, which we attribute to the diffusion spaces being defined by position alone. We instead propose a diffusion model that operates in *semantic* location representation spaces.

Location Representation Learning. Location encoders map a coordinate to a high-dimensional representation through a positional encoding (Mac Aodha et al., 2019; Mai et al., 2020; 2023; Rußwurm et al., 2024) followed by a neural network. The network is trained to predict downstream

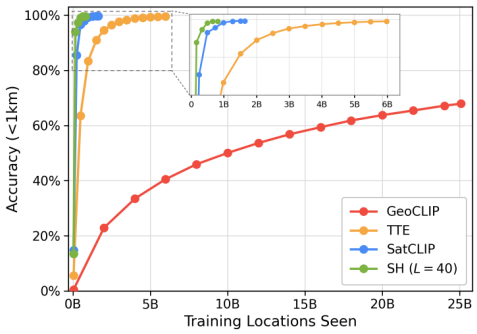


Figure 2: **Decoding different encoders.**

Table 1: **Coordinate reconstruction.** Fraction of 1M land coordinates localized within distances. (a) Choice of location encoder. (b) Analytic refinement on GeoCLIP decoder.

Encoder	<1 km	<25 km
SH ($L = 10$)	99.80	100.00
SH ($L = 20$)	99.73	100.00
SH ($L = 40$)	99.74	100.00
GeoCLIP	67.91	100.00
SatCLIP	99.66	100.00
TTE	99.60	100.00

(a) Encoder choice.

Decoder	<1 km	GFLOPs
GeoCLIP	67.91	0.076
+1 LM	99.10	0.208
+2 LM	99.67	0.341
+3 LM	99.79	0.473
+4 LM	99.85	0.605
+5 LM	99.89	0.737

(b) Analytic refinement.

targets (Mac Aodha et al., 2019; Cole et al., 2023; Cher et al., 2026b) or contrastively aligned with co-located ground-level (Vivanco Cepeda et al., 2023) or overhead imagery (Dhakal et al., 2024; Klemmer et al., 2025; Cher et al., 2026a; Dhakal et al., 2025; Liu et al., 2026). These learned representations are shown to be semantically richer than raw coordinates or the positional encoding alone, and outperform both as features for downstream prediction (Klemmer et al., 2025; Dhakal et al., 2025). In these works, the embedding is an input, either a feature for a predictor or a condition for a generative model (Sastry et al., 2026; Wei et al., 2026). We instead treat it as an output space: we invert the encoder to recover coordinates and generate directly in its embedding space.

Representations for Generation. Image generation models have progressed remarkably to achieve high visual fidelity (Rombach et al., 2022; Peebles & Xie, 2023; Ma et al., 2024; Baade et al., 2026). Recent works (Yu et al., 2024; Singh et al., 2025; Shi et al., 2026; Yao et al., 2025) align diffusion features with pretrained vision representations during training for improved generation. RAE (Zheng et al., 2026; Singh et al., 2026) goes further, diffusing directly in the vision representation space. Subsequent works adopt this paradigm and achieve improved performance on novel-view synthesis (Jang et al., 2026) and video generation (Guo et al., 2026). We extend this generative framework by diffusing in pretrained location representations for image geolocalization.

3 Recovering Coordinates from Location Representations

Diffusion for image geolocalization in a location representation space requires a means of recovering coordinates from the generated embeddings. Inspired by RAE (Zheng et al., 2026), we construct a **Location Autoencoder (LAE)** that pairs a frozen location encoder E with a learnable coordinate decoder D_ψ that maps embeddings back to coordinates.

Inverting location representations. Location encoders share a common structure: a fixed positional encoding (e.g., spherical harmonics (Rußwurm et al., 2024)) that counteracts the spectral bias of neural networks (Rahaman et al., 2019; Tancik et al., 2020), followed by a learned network. The inverse problem, recovering coordinates from an embedding, has received less attention. Because the learned network is non-linear, inversion has been viewed as difficult (Wang et al., 2025). The output space, however, is compact. Every location lies on the sphere $\mathcal{S}^2$, so densely sampling coordinates can approximate it. We therefore learn a decoder $D_\psi : \mathbb{R}^d \to \mathcal{S}^2$ to recover coordinates from location embeddings via direct regression. We instantiate this network as a simple residual MLP, trained with the objective to minimize geodesic distance between predicted and true locations,

$$\mathcal{L}_{\text{dist}} = \mathbb{E}_\ell \left[d_{\mathcal{S}^2}\big(D_\psi(x),\, \ell\big) \right], \tag{1}$$

where $\ell \in \mathcal{S}^2$ is a location sampled uniformly over land, $x = E(\ell)$ is the location embedding, and $d_{\mathcal{S}^2}$ is the great-circle distance in kilometers, as seen in Figure 3. Full architecture and training details are in App. A.1. In Figure 2 and Table 1, we evaluate the coordinate reconstruction accuracy within distance thresholds (< 1km, < 25km) over 1 million randomly held out land coordinates.

Decoding difficulty depends on the encoder. Figure 2 and Table 1a compare decoders trained to recover coordinates from spherical harmonics (SH) and three learned location encoders (GeoCLIP,

SatCLIP, TTE). All decoders recover coarse position easily (100% at 25 km) but differ sharply at the kilometer scale. SH and SatCLIP exceed 99% after 0.5B and 0.9B training locations, TTE after 4.3B, whereas GeoCLIP is at 67.91% after 25B with diminishing returns (<1 point per additional 2B). We attribute this to the spatial frequency of the encoding. SH is band-limited, SatCLIP is a smooth network on that basis, and TTE interpolates over learned sites, so coordinate inversion is smooth. GeoCLIP's random Fourier features encode far higher frequencies, forcing the decoder to learn a correspondingly fine-grained inverse, which requires orders of magnitude more training.

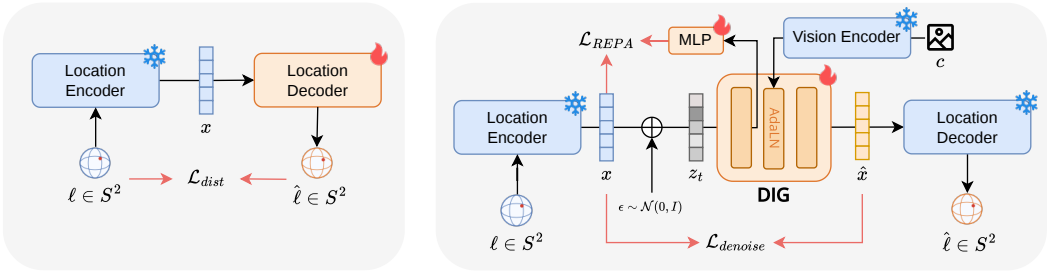
Analytic refinement. Instead of training the GeoCLIP decoder on billions more locations, we exploit that the encoder E is known, cheap, and differentiable, and use it to refine decoder estimates that are already correct at a coarse scale. Starting from $\hat{\ell}_0 = D_\psi(x)$, we minimize the encoder residual $\|x - E(\ell)\|^2$ with Levenberg–Marquardt (LM) steps (Levenberg, 1944; Marquardt, 1963),

$$\hat{\ell}_{k+1} = \hat{\ell}_k \oplus \left(J_k^\top J_k + \lambda_k \operatorname{diag}(J_k^\top J_k)\right)^{-1} J_k^\top \left(x - E(\hat{\ell}_k)\right), \qquad J_k = \frac{\partial E}{\partial \ell}\Big|_{\hat{\ell}_k} \in \mathbb{R}^{d\times 2}, \tag{2}$$

where J_k is computed exactly by forward-mode automatic differentiation and $\oplus$ applies the 2-D step along the sphere. Exact details are in App A.1.2. We find this refinement drastically improves fine-grained reconstruction at a small cost (Table 1b). Notably, a single refinement step raises the GeoCLIP decoder's accuracy at 1 km from 67.91% to 99.10%. We use five refinement steps for our GeoCLIP decoder in all subsequent experiments.

4 Latent Diffusion with Location Autoencoders

In this section, we present our generative model, **D**iffusion for **I**mage **G**eolocalization (**DIG**). As seen in Figure 3, DIG is conditioned on imagery and transports noise to pretrained location embeddings. Using LAEs, generated embeddings are decoded back to original coordinates for prediction.



(a) LAE Pretraining **(b) Diffusion for Image Geolocalization**

Figure 3: **Overview of our approach.** (a) Location Autoencoder (LAE) pretraining: a decoder learns to recover coordinates from the embeddings of a frozen location encoder. (b) A latent diffusion model (DIG), conditioned on the image, generates location embeddings, which the frozen LAE decoder maps back to coordinates. We leave out timestep conditioning for visual simplicity.

4.1 Diffusion for Image Geolocalization

Denoising objective. Let $x = E(\ell) \in \mathbb{R}^d$ be the embedding of a location $\ell \in \mathcal{S}^2$ under a location encoder E, so that $x \sim p_{\text{data}}$, and let c be the image condition. We follow rectified flow (Liu et al., 2022; Lipman et al., 2022) and define a linear path between noise $\epsilon \sim p_{\text{noise}} = \mathcal{N}(0, I)$ and data,

$$z_t = t\,x + (1-t)\,\epsilon, \qquad t \in [0,1], \qquad v = \tfrac{\mathrm{d}z_t}{\mathrm{d}t} = x - \epsilon. \tag{3}$$

Following JiT (Li & He, 2026), our network $\hat{x}_\theta(z_t, t, c)$ directly predicts the clean embedding (x-prediction) rather than the velocity. Unlike JiT which converts this prediction to a velocity and regresses v (v-loss), we follow EDM (Karras et al., 2022) and keep the loss in the prediction space (x-loss) and train with the denoising objective:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x,\,c,\,t,\,\epsilon}\left[\left\|\hat{x}_\theta(z_t, t, c) - x\right\|^2\right]. \tag{4}$$

Representation alignment. Singh et al. (2026) demonstrates how representation alignment (REPA (Yu et al., 2024)) is complementary to diffusing in the same representation space for image generation. We apply the same idea and attach a projection head h_ϕ to the features f_k of intermediate block k, aligning them with the location representation x:

$$\mathcal{L}_{\text{REPA}} = -\,\mathbb{E}_{x,\,c,\,t,\,\epsilon}\big[\mathrm{sim}\big(h_\phi(f_k),\,x\big)\big], \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ is a similarity function. The final training loss is

$$\mathcal{L} = \mathcal{L}_{\text{denoise}} + \lambda\,\mathcal{L}_{\text{REPA}}. \tag{6}$$

Sampling and classifier-free guidance. Sampling starts from Gaussian noise $z_0 \sim \mathcal{N}(0, I)$, and transports it to a clean embedding by integrating the flow ODE $\mathrm{d}z_t/\mathrm{d}t = v$ from $t = 0$ to $t = 1$. At each step the network predicts the clean embedding $\hat{x}_\theta(z_t, t, c)$, which we convert to a velocity $\hat{v} = \frac{\hat{x}_\theta(z_t,t,c)-z_t}{1-t}$, and take an Euler step $z_{t+\Delta t} = z_t + \Delta t\,\hat{v}$ over N steps. For classifier-free guidance (Ho & Salimans, 2022), the network is evaluated with and without the condition and the prediction is extrapolated with guidance scale s,

$$\hat{x} = \hat{x}_\theta(z_t, t, \varnothing) + s\,\big(\hat{x}_\theta(z_t, t, c) - \hat{x}_\theta(z_t, t, \varnothing)\big). \tag{7}$$

4.2 Implementation

Model architecture. Unlike image features, a location representation is a single vector in $\mathbb{R}^d$. There is no sequence to attend over, so we drop tokenization and attention altogether and adopt a simpler residual MLP architecture with D blocks and hidden width w. Following Dufour et al. (2025), each block is conditioned through adaptive layer normalization (AdaLN). The conditioning vector is the sum of a sinusoidal embedding of t and a linear projection of the image condition c. Inspired by MM-DiT (Esser et al., 2024), we build c from two frozen image encoders and concatenate their `[CLS]` tokens, here StreetCLIP (Haas et al., 2023) and DINOv3 (Siméoni et al., 2025). For CFG, the condition is replaced by a zero vector with probability $p_{\text{drop}} = 0.1$ during training.

5 Experiments

5.1 Experimental Protocol

Training and inference details. We train B and XL models whose hidden widths and depths follow DiT (Peebles & Xie, 2023) (Tab. 11). All models operate on location embeddings from a location encoder E, standardized per dimension with training-set statistics, and are conditioned on frozen StreetCLIP and DINOv3 features that are precomputed. The REPA head is attached to block $k = 8$ with $\lambda = 0.5$. We use AdamW (Loshchilov & Hutter, 2017) with learning rate 8×10^{-4} with weight decay 0.05. Main results use the XL model trained for 1M steps (and an extended 10M step model) following Dufour et al. (2025). Ablation experiments use the B model trained for 400k steps. All models train on a single NVIDIA H100 with batch size 1024. At inference we take 50 Euler steps; guidance scales, step schedule, and hyperparameters are given in App. A.2.1.

Datasets. We evaluate on three image geolocalization benchmarks. **OSV-5M** (Astruc et al., 2024) provides 4.9M street-level training images and a 210K-image test set separated from training by a 1 km buffer; we train and test on the official splits. **Im2GPS3k** (Vo et al., 2017) (2,997 photos) and **YFCC4k** (Vo et al., 2017) (4K photos from YFCC-100M (Thomee et al., 2016)) are Flickr benchmarks. For these we train on MP16-Pro (Jia et al., 2024), a 4.1M-image geotagged subset of YFCC-100M. To compare against generative methods trained at scale, we also train on a subset of YFCC-100M, following the filtering protocol from Dufour et al. (2025) and additionally removing test photos of both Flickr benchmarks, yielding 44M images. We denote this subset as YFCC-44M.

Baselines. We compare to deterministic (Haas et al., 2023; Astruc et al., 2024; Muller-Budack et al., 2018; Gadi et al., 2026; Weyand et al., 2016; Seo et al., 2018; Haas et al., 2024; Vivanco Cepeda et al., 2023; Zhou et al., 2024; Jia et al., 2024; 2025) and generative models (Dufour et al., 2025; Wang et al., 2025). On OSV-5M, baselines are either zero-shot or trained on official splits. For Im2GPS3k and YFCC4k, most baselines train on Flickr-derived geotagged datasets, such as MP-16, MP16-Pro, and YFCC. Full training source breakdown is in Table 17.

Hybrid approach. While DIG does not use a reference database, it is composable with retrieve-and-rerank systems (Jia et al., 2025). In this hybrid approach, DIG samples serve both as additional candidates and as a prior that rewards nearby database candidates. We apply this to our GeoCLIP model, denoted (+ GeoRanker); see App. A.2.4 for details.

Metrics. On OSV-5M we follow Dufour et al. (2025) and report GeoScore (Astruc et al., 2024), the mean haversine distance (km), and classification accuracy at country, region and city level. On Im2GPS3k and YFCC4k we report the radius-based accuracy Vo et al. (2017), denoting the fraction of predictions within 1, 25, 200, 750 and 2500 kilometers of the ground truth.

5.2 Main Results

Comparison to generative methods. In Table 2, all DIG-XL models surpass every prior generative method on every metric. The gains grow at finer scales, with region accuracy improving by 11–15 points and city accuracy roughly doubling.

Table 2: **Geolocation performance on OSV-5M** (Astruc et al., 2024) with official training and test splits. Dist. is mean distance. **Bold** indicates best and *italics* second best.

	Method	GeoS. ↑	Dist. ↓ (km)	Accuracy (%) ↑ Ctry.	Reg.	City
Deterministic	SC 0-shot (Haas et al., 2023)	2273	2854	38.4	20.8	14.8
	Regression (Astruc et al., 2024)	3028	1481	56.5	16.3	0.7
	ISNs (Muller-Budack et al., 2018)	3331	2308	66.8	39.4	4.2
	Hybrid (Astruc et al., 2024)	3361	1814	68.0	39.4	5.9
	SC Retr. (Haas et al., 2023)	3597	1386	73.4	45.8	19.9
	HierLoc (Gadi et al., 2026)	3963	861	82.9	55.0	*23.3*
Generative	Uniform (Dufour et al., 2025)	131	10052	2.4	0.1	0.0
	vMFMix (Izbicki et al., 2019)	1746	5662	34.2	11.1	0.3
	vMF (Dufour et al., 2025)	2776	2439	52.7	17.2	0.6
	FM $\mathbb{R}^3$ (Dufour et al., 2025)	3688	1149	74.9	40.0	4.2
	Diff $\mathbb{R}^3$ (Dufour et al., 2025)	3762	1123	75.9	40.9	3.6
	RFM $\mathcal{S}_2$ (Dufour et al., 2025)	3767	1069	76.2	44.2	5.4
	LocDiff (Wang et al., 2025)	–	–	77.0	46.3	11.0
	DIG-XL (ours)					
	GeoCLIP	4095	*851*	*83.4*	*60.1*	**24.6**
	SatCLIP	*4125*	912	*83.4*	**61.0**	22.4
	TTE	**4129**	**849**	**83.5**	57.4	18.4

In Table 3, we compare on Im2GPS3k and YFCC4k. On YFCC4k, each generative baseline trains on the 48M-image YFCC subset, whereas our models train on MP16-Pro (4.1M) or on YFCC-44M. Trained on $\sim$10$\times$ less data, our MP16-Pro GeoCLIP model already matches or exceeds $\text{RFM}_{10\text{M}}$ at 200–2500 km, and surpasses it by 7.7 points at 1 km. Trained on YFCC-44M at a matched 1M steps, our GeoCLIP model outperforms all generative method at every threshold. Notably, scaling to 10M steps widens the margin over $\text{RFM}_{10\text{M}}$ to 32.1 points at 1 km.

On Im2GPS3k, Diff $\mathbb{R}^3$, FM $\mathbb{R}^3$ and $\text{RFM}_{10\text{M}}$ train on YFCC-48M without filtering against the test set, which may inflate their scores. Our YFCC-44M GeoCLIP model nevertheless exceeds them at every threshold, and scaling to 10M steps further widens the gap. While LocDiff and LocDiff-H train on MP16 and use 16 augmented forward passes, our models use a single forward pass trained on the smaller MP16-Pro. Despite this, our GeoCLIP model is ahead at 1 and 25 km and within 3.2 points at 200–2500 km. Finally, the TTE and SatCLIP models match or exceed GeoCLIP at 200–2500 km but fall well behind at 1 and 25 km. We return to this trade-off in Sec. 6.

Comparison to deterministic methods. In Table 2, our GeoCLIP model surpasses the strongest deterministic method, HierLoc, on every metric. The SatCLIP and TTE models trail HierLoc at city level while exceeding it on GeoScore, country and region accuracy, and TTE also on mean error. Retrieval-based methods are inherently limited on OSV-5M, whose test images are separated from the training set by a spatial buffer, so no gallery entry is at the true location. Our generative model predicting in continuous space does not have this ceiling.

On Im2GPS3k and YFCC4k (Table 3), most deterministic methods also train on MP16 (Table 17). On Im2GPS3k, our MP16-Pro GeoCLIP model is second best at 1 and 2500 km and within 3.5–5.3 points of the best method elsewhere. On YFCC4k the gap to deterministic methods is larger at fine scales and closes at coarse ones. We attribute this to the geographic overlap between MP16 and the benchmarks, which retrieval exploits directly (App. A.3.2). We leverage this by composing DIG with GeoRanker, a strong retrieve-and-rerank method, which outperforms every deterministic method on both benchmarks, including GeoRanker itself. Scaling the training data has a similar effect. Our YFCC-44M GeoCLIP model trained for 10M steps surpasses all deterministic methods on YFCC4k and at 1 and 25 km on Im2GPS3k.

Table 3: **Geolocation performance on Im2GPS3k and YFCC4k.** Accuracy (%) within distance thresholds, with best results in **bold**. We train different DIG-XL variants on MP16-Pro and YFCC-44M. 10M denotes training for 10M steps; ‡ denotes reimplementation with public weights.

		Im2GPS3k					YFCC4k				
	Method	1 km	25 km	200 km	750 km	2500 km	1 km	25 km	200 km	750 km	2500 km
Deterministic	PlaNet (Weyand et al., 2016)	8.5	24.8	34.3	48.4	64.6	5.6	14.3	22.2	36.4	55.8
	CPlaNet (Seo et al., 2018)	10.2	26.5	34.6	48.6	64.6	7.9	14.8	21.9	36.4	55.5
	Translocator (Pramanick et al., 2022)	11.8	31.1	46.7	58.9	80.1	8.4	18.6	27.0	41.1	60.4
	GeoDecoder (Clark et al., 2023)	12.8	33.5	45.9	61.0	76.1	10.3	24.4	33.9	50.0	68.7
	PIGEOTTO Haas et al. (2024)	11.3	36.7	53.8	72.4	85.3	10.4	23.7	40.6	62.2	77.7
	GeoCLIP (Vivanco Cepeda et al., 2023)	14.1	34.5	50.7	69.7	83.8	9.6	19.3	32.6	55.0	74.7
	HierLoc (Gadi et al., 2026)	11.3	43.8	58.4	74.1	85.1	8.4	30.2	43.3	61.7	75.8
	Img2Loc (Zhou et al., 2024)	17.1	45.1	57.9	72.9	84.7	14.1	29.6	41.4	59.3	76.9
	G3 (Jia et al., 2024)	16.7	40.9	55.6	71.2	84.7	24.0	35.9	47.0	64.3	78.2
	GeoRanker (Jia et al., 2025)	18.8	45.1	61.5	76.3	89.3	32.9	43.5	54.3	69.8	82.5
Generative	vMFMix (Izbicki et al., 2019)	–	–	–	–	–	–	0.4	8.8	20.9	41.0
	vMF (Dufour et al., 2025)	–	–	–	–	–	–	4.8	15.0	30.9	53.4
	Diff $\mathbb{R}^3$ (Dufour et al., 2025)	0.0‡	16.2‡	46.8‡	63.1‡	79.1‡	0.0‡	11.1	37.7	54.7	71.9
	FM $\mathbb{R}^3$ (Dufour et al., 2025)	2.1‡	29.1‡	43.8‡	60.8‡	78.6‡	2.0‡	22.1	35.0	53.2	73.1
	RFM $\mathcal{S}_2$ (Dufour et al., 2025)	–	–	–	–	–	–	23.7	36.4	54.5	73.6
	RFM_{10M} $\mathcal{S}_2$ (Dufour et al., 2025)	8.2‡	37.2‡	51.3‡	66.8‡	81.7‡	9.8‡	33.5	45.3	61.1	77.7
	LocDiff (Wang et al., 2025)	10.9	34.0	53.3	72.5	85.2	–	33.3	46.7	65.2	79.9
	LocDiff-H (Wang et al., 2025)	15.3	36.5	56.4	75.2	87.4	–	–	–	–	–
	***DIG-XL (ours)**, trained on MP16-Pro*										
	TTE	3.9	34.5	56.3	73.0	85.5	1.8	23.4	47.7	65.0	77.6
	SatCLIP	2.4	36.4	55.5	72.0	84.8	1.7	26.2	47.7	64.3	78.9
	GeoCLIP	17.3	41.6	56.2	72.0	86.3	17.5	31.6	45.0	63.4	78.8
	GeoCLIP + GeoRanker	20.1	47.6	**64.1**	**78.1**	**89.6**	34.0	45.5	56.8	71.9	83.8
	***DIG-XL (ours)**, trained on YFCC-44M*										
	TTE	2.2	36.0	59.6	74.9	87.3	0.9	30.1	55.3	70.8	82.7
	SatCLIP	3.4	39.3	60.3	75.4	87.4	2.7	36.6	58.4	72.0	82.7
	GeoCLIP	20.2	45.4	58.4	73.6	86.8	30.1	47.5	58.3	71.1	82.9
	GeoCLIP_{10M}	**21.9**	**48.6**	60.8	75.3	87.6	**41.9**	**59.7**	**67.9**	**77.1**	**86.1**

Sampling in a semantic space. Figure 4 shows two test queries. The Cloud Gate sculpture in Chicago (top) is highly localizable (Astruc et al., 2024; Izbicki et al., 2019), and the cosine similarity between the sampled latent and location embeddings across the globe resolves to a sharp peak. The generated embeddings retrieve images of the sculpture itself or of the Chicago skyline from within 0.4 km, showing that they carry semantics and position aligned with the query. The forest track near Baie-Saint-Paul, Québec (bottom) instead offers few distinctive cues, and its boreal conifer forest also characterizes southern Scandinavia. The similarity map is correspondingly diffuse, concentrating on both regions, and one trajectory commits to the wrong continent. Yet its retrieval is a similar forest road, demonstrating that the error is semantically intuitive. We attribute this behavior to diffusing in a representation space where distant but semantically similar places lie close together. In App. A.2.3, we exploit this structure by drawing multiple samples and predicting from the densest mode, which improves accuracy.

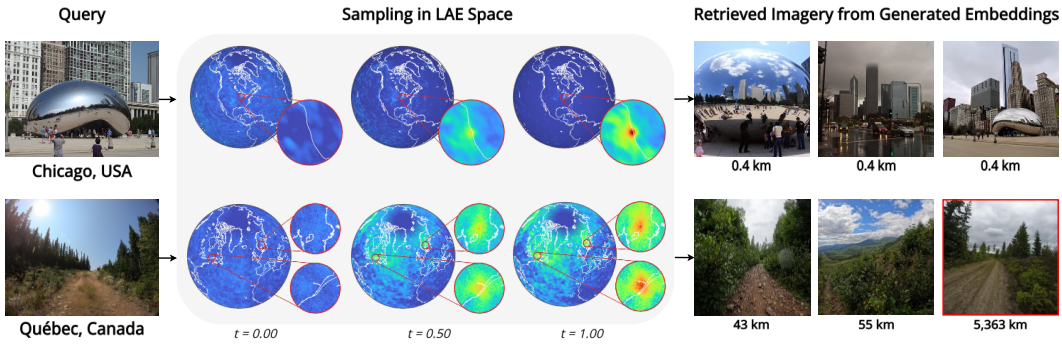


Figure 4: **Sampling in the LAE space.** For each query *(left)*, we show three sampling trajectories as the cosine similarity between the latents and location embeddings across the globe *(middle)*, and display the top images retrieved with the generated embeddings, including their error in km *(right)*.

5.3 DIG Design Choices

Diffusion ablations. Table 4 ablates the diffusion recipe with DIG-B on OSV-5M. *(a) Prediction and loss space.* Score matching (ϵ-pred, ϵ-loss) performs poorly, while flow matching (v-pred, v-loss) improves but still lags at fine scales (16.4% at city level). Predicting and regressing the clean embedding (x-pred, x-loss) is best on every metric. *(b) Schedule.* Dufour et al. (2025) use a skewed sigmoid that concentrates training on low-noise timesteps near the data. While this skewed sigmoid beats a linear schedule on GeoScore and mean error in our LAE diffusion space, it trails at city level (17.9 vs. 18.7). A symmetric sigmoid, which emphasizes both ends of the path equally, is best on all metrics. *(c) REPA.* Similar to Singh et al. (2026), we find that representation alignment is complementary to diffusing in the same representation space. REPA adds 38 GeoScore and 2.9 points of city accuracy. *(d) Conditioning.* Concatenating StreetCLIP and DINOv3 features outperforms either alone. We hypothesize this is due to StreetCLIP carrying geolocation-specific discriminative features while DINOv3 supplying general visual features, both beneficial for the diffusion model.

Table 4: **DIG ablations.** We train DIG-B models on OSV-5M using GeoCLIP LAEs; the highlighted row in each panel is our chosen configuration.

Method	GeoS. ↑	Dist. (km) ↓	Acc. (%) ↑ Ctry.	Reg.	City
ϵ-pred, ϵ-loss	3847	1007	79.3	47.2	11.6
v-pred, v-loss	3968	920	81.5	53.1	16.4
x-pred, ϵ-loss	3795	1053	78.4	45.4	10.5
x-pred, v-loss	3957	947	80.8	53.6	17.4
x-pred, x-loss	**4037**	**878**	**82.6**	**56.4**	**19.8**

(a) $\boldsymbol{x}$ **marks the prediction and loss.**

Schedule	GeoS. ↑	Dist. (km) ↓	Acc. (%) ↑ Ctry.	Reg.	City
linear	3967	951	81.1	54.3	18.7
skew-sigmoid	3986	923	82.2	54.3	17.9
sigmoid	**4037**	**878**	**82.6**	**56.4**	**19.8**

(b) **Diffusion schedule choice.**

Method	GeoS. ↑	Dist. (km) ↓	Acc. (%) ↑ Ctry.	Reg.	City
w/o REPA	3999	903	81.8	54.0	16.9
w/ REPA	**4037**	**878**	**82.6**	**56.4**	**19.8**

(c) **REPA improves performance.**

Encoder	GeoS. ↑	Dist. (km) ↓	Acc. (%) ↑ Ctry.	Reg.	City
DINOv3	3824	1206	77.7	52.1	17.5
StreetCLIP	3953	975	81.0	53.8	18.2
Both	**4037**	**878**	**82.6**	**56.4**	**19.8**

(d) **Combining vision representations.**

Importance of pretrained location representations. Table 5 varies the diffusion space, using DIG-B models on OSV-5M. Diffusing directly in coordinate space performs worst. SH spaces perform far better, with city-level performance increasing with spectral resolution (L), demonstrating the value of multi-scale positional information (Wang et al., 2025). Yet, SatCLIP, which learns a 256-dimensional projection from the SH ($L = 40$) encoding, improves over it on every metric, as do TTE and GeoCLIP likewise. Despite their lower dimensionality, learned representations that add semantic structure make for a better diffusion space.

Classifier-free guidance. Consistent with Dufour et al. (2025), guidance improves geolocalization performance (Figure 5). Without CFG ($s = 1$), GeoScore drops and mean error rises. Gains saturate by $s \approx 3$, so we use $s = 3$ throughout.

Table 5: **Importance of LAE latent spaces.** We train DIG-B models evaluated at OSV-5M.

Diffusion Space	GeoS. ↑	Dist. (km) ↓ Mean	Median	Classification Acc. (%) ↑ Country	Region	City
coordinates	3425	1604	199.0	68.7	38.8	6.0
SH ($L = 10$)	4042	**825**	116.9	81.0	47.9	9.1
SH ($L = 20$)	4051	856	101.9	81.5	50.6	10.7
SH ($L = 40$)	4023	979	83.4	81.1	53.8	13.9
LAE (ours)						
SatCLIP	**4096**	875	**65.6**	**82.9**	**57.6**	*18.4*
TTE	*4089*	*847*	82.0	82.5	54.1	15.4
GeoCLIP	4037	877	*70.9*	*82.6*	*56.4*	**19.8**

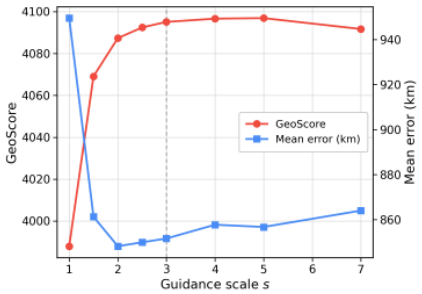


Figure 5: **Impact of CFG.** DIG-XL with GeoCLIP LAE evaluated on OSV-5M.

6 DISCUSSION

Location encoder trade-offs. GeoCLIP is strongest at fine scales and SatCLIP and TTE at coarse ones, but the latter trail at 1–25 km, especially on Flickr benchmarks. We see two likely reasons:

(1) Alignment modality. SatCLIP and TTE are aligned to Sentinel-2 satellite imagery (Klemmer et al., 2025), whereas GeoCLIP is aligned to ground-level photographs (Larson et al., 2017). DIG depends on the *semantics* these encoders carry (Table 5), so those semantics should ideally match the query imagery. On OSV-5M street views, the encoder gap is smaller as overhead appearance corresponds reasonably well to street-level semantics. On the Flickr benchmarks it is large, since a satellite view of Paris says little about a café menu or a museum hall photographed there, making the these encoders contribute little beyond position.

(2) Spectral resolution. GeoCLIP's random Fourier features encode much higher spatial frequencies than the band-limited spherical harmonics of SatCLIP or TTE's soft site assignment, making its space more locally discriminative, accounting for its fine-grained accuracy. We see the same trend in Table 5, where a higher Legendre degree L, and thus finer spectral resolution, improves city-level accuracy. Conversely, the global bases and semantic tokens of SatCLIP and TTE capture structure at continental scale, which explains their higher accuracy at coarse scales. Leveraging an ensemble of location encoders or ones aligned to multiple modalities, is a natural future direction.

Sub-kilometer accuracy. Our decoders target reconstruction accuracy at the kilometer-level, matching the scale of global geolocalization. Yet many tasks, such as trajectory geolocalization (Gao et al., 2026), operate below 1 km. We probe reconstruction at this scale in Table 6, finding raw decoding to be poor. Analytic refinement resolves this for GeoCLIP and TTE, while SatCLIP plateaus regardless of refinement. Matching location encoders to higher-resolution geolocalization tasks remains open.

Table 6: **Sub-km reconstruction.**

Decoder	<0.01 km	<0.1 km	<0.5 km	GFLOPs
SatCLIP	0.10	9.31	89.0	0.076
GeoCLIP	0.01	1.23	26.1	0.076
TTE	0.09	8.70	85.6	0.076
+ 5 refinement steps				
SatCLIP	90.2	99.4	100.0	0.161
GeoCLIP	99.7	99.8	99.9	0.737
TTE	100.0	100.0	100.0	0.315

7 CONCLUSION

In this work, we proposed DIG, a diffusion model for image geolocalization that leverages pretrained location encoders by diffusing in their representation space. Geolocating a photograph requires understanding both where a place is and its semantic context, which location encoders capture together in a single embedding. Instead of predicting the image's coordinates directly, we propose to generate their location embeddings with DIG, a conditional latent diffusion model. To use these generating representations for prediction, we introduce a Location Autoencoder (LAE) that decodes them back to original coordinates. Using this recipe, DIG sets a new state of the art on OSV-5M, Im2GPS3k and YFCC4k, and we demonstrate that its performance emerges from using location encoders that learn semantics beyond their positional encoding. Notably, DIG can be composed with retrieve-and-rerank systems to improve their performance at every distance threshold. Location encoders were built to describe the Earth; we show they are also the space in which to search it.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models and conceptual frameworks, to design and provide feedback on our research methodology and experiments, to implement methods, and to clean and reformat datasets. We have not used generative AI tools to propose or refine hypotheses, to interpret results, to support qualitative or thematic data analysis, or to assist with translation. Generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to refine the writing of sections of this paper and to help identify related literature. We have reviewed all AI-assisted work. All AI-generated code was reviewed and verified for correctness by the authors, and all AI-suggested references were checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

In this work, we study how pretrained representations can be leveraged within generative models for geolocalization. All experiments use established public datasets and benchmarks, namely MP16-Pro, OSV-5M, YFCC4K, Im2GPS3k, and YFCC-100M, each of which we cite from its original source and use in accordance with its research license. We did not collect any new data, and no human subjects were involved in this research. We note that these datasets contain user-uploaded photographs depicting people or private settings. We use these images as released and do not attempt to identify individuals. More broadly, we recognize that improvements in geolocalization carry risk. Geolocalization can enable nonconsensual tracking and expose private residences, personal routines, or sensitive locations. Our models are intended for research and benchmarking, and we do not endorse their use to surveil or track individuals. Performance varies geographically, with potentially larger errors in regions underrepresented in training data. These disparities and privacy risks warrant careful consideration before deployment.

REFERENCES

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21967–21977. IEEE, 2024.

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=LFUaHUCO1a.

Opeyemi Bamigbade, John Sheppard, and Mark Scanlon. Computer vision for multimedia geolocation in human trafficking investigation: A systematic literature review. *arXiv preprint arXiv:2402.15448*, 2024.

Daniel Cher, Hamza Iqbal, Eric Xing, Brian Wei, and Nathan Jacobs. Tessellating the earth. In *European Conference on Computer Vision*, pp. 443–459. Springer, 2026a.

Daniel Cher, Eric Xing, Kexing Li, Brian Wei, Isaac Corley, and Nathan Jacobs. Chronosphere: Space-time tessellation of local climate experts, 2026b. URL https://arxiv.org/abs/2609.21872.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23182–23190. IEEE, 2023.

Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International conference on machine learning*, pp. 6320–6342. PMLR, 2023.

Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, Hannah Kerner, and Nathan Jacobs. Sat2cap: Mapping fine-grained textual descriptions from satellite images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 533–542. IEEE, 2024.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24680–24689. IEEE, 2025.

Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23016–23026, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Hari Krishna Gadi, Daniel Matos, Hongyi Luo, Lu Liu, Yongliang Wang, Yanfeng Zhang, and Liqiu Meng. Hierloc: Hyperbolic entity embeddings for hierarchical visual geolocation. *arXiv preprint arXiv:2601.23064*, 2026.

Tianyi Gao, Jiayu Lin, Danielle Beaulieu, and Nathan Jacobs. Trajectory-aware cross-view geolocalization with sequential observations. In *European Conference on Computer Vision*, 2026.

Minghui Guo, Shengqiong Wu, and Hao Fei. V-rae: Rethinking video latent spaces for generation. *arXiv preprint arXiv:2608.13556*, 2026.

Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275*, 2023.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12893–12902. IEEE, 2024.

James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Mike Izbicki, Evangelos E. Papalexakis, and Vassilis J. Tsotras. Exploiting the earth's spherical geometry to geolocate images. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet (eds.), *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, volume 11907 of *Lecture Notes in Computer Science*, pp. 3–19, Cham, 2019. Springer International Publishing. ISBN 978-3-030-46147-8. doi: 10.1007/978-3-030-46147-8_1.

Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. In *European Conference on Computer Vision (ECCV)*, 2026.

Furong Jia, Ling Dai, Wenjin Deng, Fan Zhang, Chen Hu, Daxin Jiang, and Yu Liu. Spotagent: Grounding visual geo-localization in large vision-language models through agentic reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 2062–2073, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817948. URL https://doi.org/10.1145/3770855.3817948.

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

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Yixuan Li. Georanker: Distance-aware ranking for worldwide image geolocalization. *arXiv preprint arXiv:2505.13731*, 2025.

Pengyue Jia, Derong Xu, Yingyi Zhang, Xiaopeng Li, Wenlin Zhang, Yi Wen, Yuanshao Zhu, and Xiangyu Zhao. Georouter: Dynamic paradigm routing for worldwide image geolocalization. *arXiv preprint arXiv:2603.24376*, 2026b.

Modi Jin, Yiming Zhang, Boyuan Sun, Dingwen Zhang, Ming-Ming Cheng, and Qibin Hou. Geoagent: Learning to geolocate everywhere with reinforced geographic characteristics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41352–41364, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025.

Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24 (1):93–96, 2017.

Tung-Duong Le-Duc, Hoang-Quoc Nguyen-Son, and Minh-Son Dao. Geosearch: Augmenting worldwide geolocalization with web-scale reverse image search and image matching. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3913–3918, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809936. URL https://doi.org/10.1145/3805712.3809936.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

Ling Li, Yu Ye, Yao Zhou, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *International Conference on Machine Learning (ICML)*, 2024.

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

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Yong Li, Furong Jia, Dacheng Yin, Kang Rong, Fengyun Rao, Jing Lyu, and Fan Zhang. Reverse: Reinforcing evidence verification and search for agentic image geo-localization. *arXiv preprint arXiv:2605.26861*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9595–9605. IEEE, 2019.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *The Eighth International Conference on Learning Representations*. openreview, 2020.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–579, 2018.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

OpenAI, Josh Achiam, et al. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pp. 196–215. Springer, 2022.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *International Conference on Learning Representations*, volume 2024, pp. 1746–1759, 2024.

Srikumar Sastry, Dan Cher, Brian Wei, Aayush Dhakal, Subash Khanal, Dev Gupta, and Nathan Jacobs. TerraDiT: Point-conditioned diffusion transformer for satellite image synthesis. *arXiv preprint arXiv:2603.02172*, 2026.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *European Conference on Computer Vision*, pp. 544–560. Springer, 2018.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, volume 2026, pp. 154506–154537, 2026.

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.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Golsa Tahmasebzadeh, Sherzod Hakimov, Ralph Ewerth, and Eric Müller-Budack. Multimodal geolocation estimation of news photos. In *European Conference on Information Retrieval*, pp. 204–220. Springer, 2023.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2640–2649. IEEE, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhangyu Wang, Zeping Liu, Jielu Zhang, Zhongliang Zhou, Qian Cao, Nemin Wu, Lan Mu, Yang Song, Yiqun Xie, Ni Lao, et al. Locdiff: Identifying locations on earth by diffusing in the hilbert space. *Advances in Neural Information Processing Systems*, 38:620–647, 2025.

Brian Wei, Srikumar Sastry, Daniel Cher, Eric Xing, and Nathan Jacobs. TerraDiT-Ω: Unified spatial control for satellite image synthesis with any geospatial primitive. In *European Conference on Computer Vision*, 2026.

Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, pp. 37–55. Springer, 2016.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Ryo Yokota, Yoshinori Hawai, Ken'ichi Tsuchiya, Daisuke Imoto, Manato Hirabayashi, Norimitsu Akiba, Hidetoshi Kakuda, Kosuke Tanabe, Masakatsu Honma, and Kenji Kurosawa. A revisited visual-based geolocalization framework for forensic investigation support tools. *Forensic science international: digital investigation*, 35:301088, 2020.

Bo Yu, Fengze Yang, Yiming Liu, Chao Wang, Xuewen Luo, Taozhe Li, Ruimin Ke, Xiaofan Zhou, and Chenxi Liu. Locatability-guided adaptive reasoning for image geo-localization with vision-language models. *arXiv preprint arXiv:2603.13628*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval*, pp. 2749–2754, 2024.

A Appendix

A.1 Location Autoencoders

A.1.1 Location Encoders

We evaluate spherical harmonic (SH) encodings (Rußwurm et al., 2024) at three truncation levels, $L \in \{10, 20, 40\}$, and three pretrained location encoders: SatCLIP (Klemmer et al., 2025), TTE (Cher et al., 2026a), and GeoCLIP (Vivanco Cepeda et al., 2023). The pretrained encoders use contrastive learning to align geographic coordinates with features of images captured at those locations.

Spherical harmonics. Spherical harmonics provide a Fourier-like basis on the sphere. Following Rußwurm et al. (2024), we encode latitude φ and longitude λ by evaluating the real spherical harmonics of degrees $0, \ldots, L-1$:

$$\mathrm{PE}_L(\varphi, \lambda) = \left[Y_l^m(\varphi, \lambda) \right]_{0 \le l < L,\ |m| \le l} \in \mathbb{R}^{L^2}, \tag{8}$$

where

$$Y_l^m(\varphi, \lambda) = N_l^m P_l^{|m|}(\sin \varphi) \begin{cases} \cos(m\lambda), & m > 0, \\ 1, & m = 0, \\ \sin(|m|\lambda), & m < 0. \end{cases} \tag{9}$$

Here, $P_l^{|m|}$ is an associated Legendre polynomial, and N_l^m normalizes each basis function to unit norm on the sphere. Larger L includes higher-frequency harmonics and supports finer spatial detail. The encoding has L^2 dimensions: 100, 400, and 1,600 for $L = 10$, 20, and 40, respectively. We use these fixed features directly as the diffusion latent.

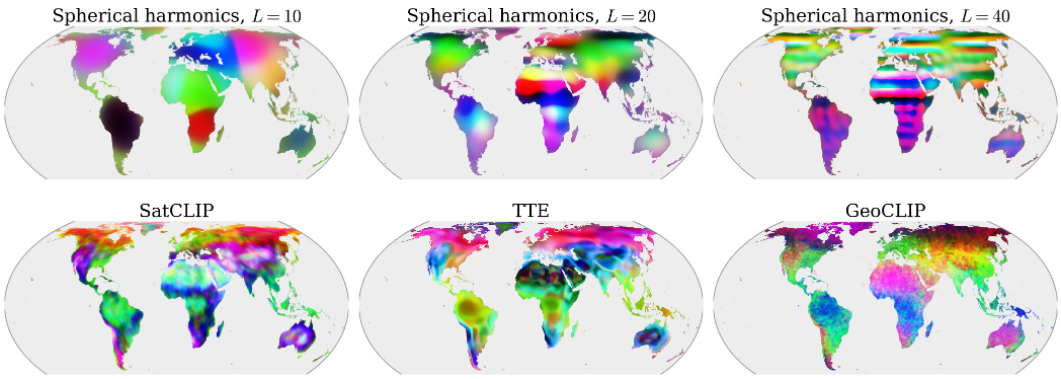


Figure 6: **ICA of location embeddings over land.** Top: spherical harmonics with L=10, 20, 40. Bottom: learned encoders SatCLIP, TTE, and GeoCLIP. Each panel maps its first three independent components to RGB.

Comparing representations. We visualize each encoder with independent component analysis (ICA) in Figure 6, following Dhakal et al. (2025): embeddings are evaluated over land, standardized, and reduced to three independent components shown as RGB. Colors are arbitrary per panel, so only spatial structure is comparable. For SH, increasing L raises the spatial frequency, from continent-scale blobs at L=10 to regular banding along parallels at L=40. SatCLIP's SirenNet maps the same L=40 basis to smooth regions that follow geographic structure rather than parallels, and has been shown to better capture location semantics (Klemmer et al., 2025). TTE's map is smooth, since the soft assignment interpolates between neighboring sites. GeoCLIP's multi-scale RFF positional encoding produces a much higher-frequency map than the other encoders.

A.1.2 LAE Implementation Details

Decoder architecture. All decoders share one design. The encoder's raw output (d=512 for GeoCLIP and TTE, d=256 for SatCLIP, and d=L^2 for SH) is layer-normalized and projected to width

w. It is then processed by 8 pre-norm residual MLP blocks. Each block applies LayerNorm, a linear layer from w to $4w$, GELU, and a linear layer from $4w$ back to w. The last layer of each block is zero-initialized so that every block starts as the identity. A final LayerNorm and linear layer map the features to $\mathbb{R}^3$, and the output is normalized onto the unit sphere. We use w=768.

Decoder training. Decoders are trained only on synthetic pairs. Each step draws a fresh batch of 4096 locations uniformly by area over land, using rejection sampling against a 0.05° raster of the 1:10m Natural Earth land and minor-island polygons[1]. The frozen encoder embeds these locations, and the decoder minimizes the mean great-circle distance in kilometers between decoded and true locations. We use Adam (Kingma & Ba, 2014) (β_1=0.9, β_2=0.999) with a constant learning rate of 3×10^{-4} after a 1k-step linear warm-up, and clip gradients at norm 1.0. Training runs on a single H100 GPU. We evaluate an exponential moving average of the weights with decay 0.999. Locations are resampled at every step, so no location is repeated and no decoder is fit to any benchmark's coordinate distribution. We measure reconstruction on a fixed set of 1M land locations drawn the same way with a separate seed. Table 7 lists the training length for each encoder.

Table 7: **Official decoder training.** All decoders have depth 8 and use batch size 4096, with locations drawn uniformly over land. Density is locations seen per km^2 of land (1.50×10^8 km^2).

Encoder	Input dim	Params	Steps	Locations seen	Density (/km^2)
GeoCLIP	512	38.2M	6.1M	2.50×10^{10}	167
TTE	512	38.2M	1.5M	6.14×10^{9}	41
SatCLIP	256	38.0M	400k	1.64×10^{9}	11
SH (L=10)	100	37.9M	200k	8.19×10^{8}	5.5
SH (L=20)	400	38.1M	200k	8.19×10^{8}	5.5
SH (L=40)	1600	39.0M	200k	8.19×10^{8}	5.5

Levenberg–Marquardt refinement. We refine each decoded location $\hat{\ell}_0 = D_\psi(x)$ by minimizing the encoder residual $r(\ell) = x - E(\ell)$ with Equation 2. Updates are taken in a local tangent frame at $\hat{\ell}_k$, with unit east and north vectors u_e and u_n and steps measured in kilometers. The Jacobian $J_k \in \mathbb{R}^{d\times2}$ is computed exactly with two forward-mode automatic differentiation passes, one along u_e and one along u_n. Solving the damped normal equations gives a step $\delta_k = (\delta_e, \delta_n) \in \mathbb{R}^2$, and $\oplus$ follows the great circle in that direction,

$$\hat{\ell} \oplus \delta = \cos\Big(\frac{\|\delta\|}{R}\Big)\,\hat{\ell} + \sin\Big(\frac{\|\delta\|}{R}\Big)\,\frac{\delta_e u_e + \delta_n u_n}{\|\delta\|}, \tag{10}$$

where $R = 6371$ km is the Earth's radius. A step is accepted only if it lowers $\|r\|$, and a rejected step leaves the location unchanged. Each location keeps its own damping λ_k, starting at $\lambda_0 = 10^{-3}$. It is divided by 3 after an accepted step and multiplied by 10 after a rejected one. Each step costs two forward-mode passes, one encoder evaluation for the acceptance check, and a batched 2×2 solve, and the procedure is deterministic. We use five steps for GeoCLIP, one for SatCLIP and TTE.

A.1.3 Additional LAE Experiments

Analytic refinement. We vary the number of LM steps and measure the effect on reconstruction (Table 9) and on geolocalization (Table 8). For reconstruction, a single step raises the GeoCLIP decoder from 67.91% to 99.10% within 1 km, and five steps reach 99.89%. The SatCLIP and TTE decoders start at about 99.6% and reach 100.00% after one step. Cost grows linearly with the number of steps. Decoding with five GeoCLIP steps costs 0.74 GFLOPs per location, under 1% of the roughly 84 GFLOPs needed to sample one latent with DIG-XL (50 steps with CFG). We therefore use five steps for GeoCLIP and one step for SatCLIP and TTE.

Table 8 decodes the same generated samples with 0 to 5 steps. On every benchmark, refinement changes accuracy at 25 km and above by at most 0.3 points, so its effect is limited to the 1 km threshold. On OSV-5M, fewer than 0.3% of predictions fall within 1 km for any encoder or step

[1] https://www.naturalearthdata.com

Table 8: **Analytic refinement on geolocalization.** XL models trained on each benchmark's training set (IM2GPS3K on MP16-Pro). Every row decodes the same generated samples. We bold improvements over vanilla > 0.5.

	GeoCLIP					SatCLIP					TTE				
Decoder	1 km	25 km	200 km	750 km	2500 km	1 km	25 km	200 km	750 km	2500 km	1 km	25 km	200 km	750 km	2500 km
OSV-5M/OSV-5M															
vanilla	0.21	33.3	69.4	83.2	91.9	0.13	30.0	72.8	84.2	91.2	0.10	22.5	71.7	84.9	92.1
+1	0.26	33.4	69.4	83.2	91.9	0.12	30.3	72.8	84.2	91.2	0.09	22.7	71.8	84.9	92.1
+2	0.27	33.4	69.4	83.2	91.9	0.12	30.3	72.8	84.2	91.2	0.09	22.7	71.8	84.9	92.1
+3	0.28	33.4	69.4	83.2	91.9	0.12	30.3	72.8	84.2	91.2	0.09	22.7	71.8	84.9	92.1
+4	0.27	33.4	69.4	83.2	91.9	0.12	30.3	72.8	84.2	91.2	0.09	22.7	71.8	84.9	92.1
+5	0.27	33.4	69.4	83.2	91.9	0.12	30.3	72.8	84.2	91.2	0.09	22.7	71.8	84.9	92.1
MP16-Pro/IM2GPS3K															
vanilla	5.6	41.6	56.2	72.1	86.4	2.7	36.3	55.5	72.0	84.8	2.9	34.4	56.3	73.0	85.5
+1	**12.3**	41.6	56.2	72.0	86.4	2.4	36.4	55.5	72.0	84.8	**3.9**	34.5	56.3	73.0	85.5
+2	**15.7**	41.7	56.2	72.0	86.3	2.4	36.4	55.5	72.0	84.8	**3.9**	34.5	56.3	73.0	85.5
+3	**16.8**	41.7	56.2	72.0	86.3	2.4	36.4	55.5	72.0	84.8	**3.9**	34.5	56.3	73.0	85.5
+4	**17.2**	41.6	56.2	72.0	86.3	2.4	36.4	55.5	72.0	84.8	**3.9**	34.5	56.3	73.0	85.5
+5	**17.3**	41.6	56.2	72.0	86.3	2.4	36.4	55.5	72.0	84.8	**3.9**	34.5	56.3	73.0	85.5
YFCC/YFCC4K															
vanilla	7.8	47.4	58.3	71.1	82.9	2.7	36.5	58.4	72.0	82.7	0.9	30.0	55.3	70.8	82.7
+1	**22.6**	47.4	58.3	71.1	82.9	2.7	36.6	58.4	72.0	82.7	0.9	30.1	55.3	70.8	82.7
+2	**27.6**	47.5	58.3	71.1	82.9	2.7	36.6	58.4	72.0	82.7	0.9	30.1	55.3	70.8	82.7
+3	**29.1**	47.5	58.3	71.1	82.9	2.7	36.6	58.4	72.0	82.7	0.9	30.1	55.3	70.8	82.7
+4	**29.9**	47.5	58.3	71.1	82.9	2.7	36.6	58.4	72.0	82.7	0.9	30.1	55.3	70.8	82.7
+5	**30.1**	47.5	58.3	71.1	82.9	2.7	36.6	58.4	72.0	82.7	0.8	30.1	55.3	70.8	82.7

Table 9: **Analytic refinement across location encoders.** Fraction of 1M uniform land coordinates localized within 1 km, and decoder cost, as a function of LM refinement steps. The highlighted row is our chosen configuration.

	GeoCLIP		SatCLIP		TTE	
Decoder	<1 km	GFLOPs	<1 km	GFLOPs	<1 km	GFLOPs
Base	67.91	0.076	99.65	0.076	99.59	0.076
+1	99.10	0.208	100.00	0.093	100.00	0.124
+2	99.67	0.341	100.00	0.110	100.00	0.172
+3	99.79	0.473	100.00	0.127	100.00	0.220
+4	99.85	0.605	100.00	0.144	100.00	0.267
+5	99.89	0.737	100.00	0.161	100.00	0.315

count. On the Flickr benchmarks, GeoCLIP gains the most, rising from 5.6% to 17.3% at 1 km on Im2GPS3k and from 7.8% to 30.1% on YFCC4k. Here many samples land close to the true location, and the base decoder's kilometer-scale error decides whether they fall within 1 km. The first two steps give most of the gain (10.1 of 11.7 points on Im2GPS3k and 19.8 of 22.3 on YFCC4k). The SatCLIP and TTE decoders are already accurate below 1 km, so refinement changes their results by at most one point. The step count therefore trades a small cost for accuracy at the finest scale, with diminishing returns after two steps.

Comparing to retrieval. Since DIG generates location embeddings, an alternative to decoding is to retrieve the location whose embedding is most similar among a gallery covering the Earth. We follow GeoCLIP (Vivanco Cepeda et al., 2023), whose gallery is 100k coordinates sampled from the training set, and also test a gallery of 100k coordinates sampled uniformly over land. Table 10 decodes the same generated embeddings all three ways. Decoding is best on every benchmark, and the gap is largest on OSV-5M, where the test locations are kept at a distance from every training location.

Table 10: **Decoding versus gallery retrieval.** We compare reading out generated embeddings using our LAE decoders with retrieving over training-set or uniform-land location galleries. Reported metrics use DIG-XL GeoCLIP model (OSV-5M uses the OSV-5M model; YFCC4k the YFCC-44M model; IM2GPS3K the MP16-Pro model) Best in **bold**.

	OSV-5M					YFCC4k					IM2GPS3K				
Readout	GeoS.	Mean	Ctry	Reg.	City	1 km	25 km	200 km	750 km	2500 km	1 km	25 km	200 km	750 km	2500 km
Decoder	**4095**	**852**	**83.4**	**60.1**	**24.6**	**30.1**	**47.5**	**58.3**	**71.1**	**82.9**	**17.3**	**41.6**	**56.2**	**72.0**	**86.3**
Training set gallery	4040	878	82.2	56.2	19.9	28.4	46.7	58.0	71.0	**82.9**	**17.3**	41.4	**56.2**	71.7	**86.3**
Uniform-land gallery	4067	861	82.3	55.9	17.5	0.0	22.8	55.4	70.1	82.5	0.0	17.4	54.0	71.0	85.9

Table 11: Diffusion model sizes. Widths and depths follow DiT (Peebles & Xie, 2023); the MLP hidden size is $4w$. Parameter counts are for a 512-dimensional location embedding and a 2048-dimensional image condition, and exclude the REPA head (2–3M parameters).

Model	Width w	Depth D	MLP hidden	Params
B	768	12	3072	82M
L	1024	24	4096	284M
XL	1152	28	4608	417M

A.2 DIG Additional Details

A.2.1 DIG Implementation Details

This section gives the details for DIG as trained for the main results. Notation follows Sec. 4.

Latent standardization. The diffusion operates on a standardized embedding: with $\mu, \sigma \in \mathbb{R}^d$ the per-dimension mean and standard deviation of $E(\ell)$ over the training coordinates, the model sees $x = (E(\ell) - \mu)/\sigma$, and samples are de-standardised before decoding. This matters for GeoCLIP, whose raw embeddings have per-dimension standard deviation ≈ 0.014, two orders of magnitude below the unit-variance noise the path assumes.

Time sampling. Training times are not uniform. We draw $u \sim \mathcal{U}(0, 1)$ and warp it through the sigmoid schedule of Dufour et al. (2025),

$$t(u) = \frac{\sigma(b) - \sigma\big(a + (b - a)\, u\big)}{\sigma(b) - \sigma(a)}, \qquad a = -5,\ b = 5, \tag{11}$$

which places more mass near both ends of the path ($t \approx 0$ and $t \approx 1$) than in the middle. The same warp defines the sampling grid: the N Euler steps are equally spaced in u, hence non-uniform in t, with t clamped to $\leq 1 - 10^{-3}$ so the velocity conversion of Sec. 4 stays finite. Dufour et al. (2025) use the asymmetric setting $(a, b) = (-7, 3)$; the symmetric one above performed better in our schedule ablation.

Representation alignment. The similarity in $\mathcal{L}_{\text{REPA}}$ has a value term and a direction term,

$$\text{sim}(a, x) = -\,\|a - x\|_2^2 \;+\; \lambda_{\cos}\, \frac{a^\top x}{\|a\|\,\|x\|}, \qquad \lambda_{\cos} = 0.5, \tag{12}$$

so that $\mathcal{L}_{\text{REPA}} = \mathbb{E}\big[\|h_\phi(f_k) - x\|^2 + \lambda_{\cos}(1 - \cos(h_\phi(f_k), x))\big]$ up to a constant. REPA as published uses the cosine alone; we add the squared error so that $h_\phi(f_k)$ matches x in magnitude as well as direction, i.e. it is an x-prediction at depth k. The head reads block $k = 8$ and is $\text{LN} \to \text{Linear}(w \to 2048) \to \text{SiLU} \to \text{Linear}(2048 \to d)$. We use $\lambda = 0.5$ and discard the head at inference.

Conditioning vector. A 288-channel sinusoidal embedding of the timestep (two-layer SiLU MLP to width w) is summed with a linear projection of the image features $c \in \mathbb{R}^{2048}$ (StreetCLIP $\|$ DINOv3 CLS tokens, precomputed); the null condition $\varnothing$ is the projection's bias. Each block maps the sum to AdaLN scale, shift and gate via $\text{SiLU} \to \text{Linear}(w \to 3w)$.

Table 12: Training configuration of the XL model.

Latent d	512 (GeoCLIP), standardized per dimension
Width w / blocks / MLP expansion	1152 / 28 / 4
Parameters	420M
Alignment	$k = 8$, $\lambda = 0.5$, $\lambda_{\cos} = 0.5$
Time schedule	Eq. 11, $(a, b) = (-5, 5)$
Condition dropout p_{drop}	0.1
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.05
Learning rate	8×10^{-4}; linear warm-up 1k steps; cosine decay to 0
Batch size / steps	1024 / 1M
Gradient clipping	global norm 1.0
Precision	bf16 autocast
EMA	decay 0.9999; used for all evaluation

Table 13: **Generative metrics of the unconditional distribution.** Precision, recall, density and coverage Naeem et al. (2020) of samples drawn from $p(\mathbf{y} \mid \varnothing)$ against the test-set locations of OSV-5M (210k) and YFCC4k (4.5k), following Dufour et al. (2025). Best in **bold**, second *italicized*.

	OSV-5M				YFCC4k			
Method	precision↑	recall↑	density↑	coverage↑	precision↑	recall↑	density↑	coverage↑
Uniform	0.29	0.98	0.21	0.21	0.59	0.99	0.38	0.22
vMF	0.598	**0.982**	0.499	0.446	0.667	**0.993**	0.542	0.599
vMFMix	0.513	*0.980*	0.422	0.358	0.626	*0.988*	0.474	0.498
Diff $\mathbb{R}^3$	0.822	0.916	0.752	0.568	0.938	0.959	0.959	0.837
FM $\mathbb{R}^3$	0.845	0.907	0.799	*0.575*	*0.953*	0.959	1.037	*0.920*
RFM $\mathcal{S}_2$	0.841	0.896	0.797	**0.590**	**0.957**	0.952	*1.060*	**0.926**
LocDiff	–	–	0.795	0.572	–	–	**1.072**	0.915
Ours (GeoCLIP)	**0.953**	0.792	**0.932**	0.351	0.946	0.934	1.032	0.879
Ours (SatCLIP)	*0.951*	0.779	*0.925*	0.378	0.946	0.930	1.021	0.879
Ours (TTE)	0.950	0.794	0.920	0.383	0.945	0.932	1.019	0.875

A.2.2 Generative Comparison.

Table 13 evaluates unconditional generation using the precision, recall, density, and coverage metrics from PLONK. On OSV-5M, DIG achieves the highest precision and density but trails the coordinate-space baselines in recall and coverage. A random sample of OSV-5M training locations obtains 0.950 precision, 0.791 recall, 0.923 density, and 0.385 coverage, closely matching all three DIG variants. This suggests that the lower recall and coverage arise from the shift between the training and test distributions rather than mode dropping. For example, the United States accounts for 24.9% of the training set but only 11.4% of the test set, and the country distributions have a total variation distance of 0.46. Matching the training distribution therefore undersamples regions overrepresented in the test set. The baselines trade precision and density for higher recall and coverage by producing a broader distribution. On YFCC4k, the metrics saturate, and all methods are within noise of the ground-truth ceiling.

A.2.3 Leveraging the Posterior for Prediction

Every prediction in the main tables is a single draw from the DIG's conditional distribution. A single draw can land in a minor mode or in the tail of the correct one, and it says nothing about how certain the model is. Drawing several samples exposes that structure. Where the samples agree the model is confident, and where they split between regions the split itself is informative. Here we draw 16 samples per test image with different seeds and test rules for combining them into one prediction.

Aggregation rules. *Single sample* uses the first draw and matches the main-paper protocol. *Center* is the spherical mean of the 16 decoded coordinates. The remaining rules first cluster the samples, joining any two within 200 km of each other, and keep only the cluster with the most samples. *Cluster mean* is the mean coordinate of that cluster. *Cluster medoid* is the sample closest to the

Table 14: **Aggregating 16 posterior samples.** Each test image is sampled 16 times with different seeds, and predictions are combined by the rule in each row. OSV-5M uses the OSV-5M models; YFCC4k the YFCC-44M models; IM2GPS3K the MP16-Pro models. Best rule per encoder in **bold**.

	OSV-5M					YFCC4k					IM2GPS3K				
Rule	GeoS.	Mean	Ctry	Reg.	City	1 km	25 km	200 km	750 km	2500 km	1 km	25 km	200 km	750 km	2500 km
GeoCLIP															
Single sample	4095	851	83.4	60.1	24.6	30.1	47.5	58.3	71.1	82.9	17.3	41.6	56.2	72.0	86.3
Center	4081	**767**	82.6	58.7	24.2	26.1	41.5	54.8	69.3	82.1	16.4	38.5	55.1	72.2	86.3
Cluster mean	**4142**	790	**84.5**	61.1	25.4	28.5	48.5	**61.9**	**72.6**	**83.3**	17.5	42.0	**58.1**	**73.3**	**86.7**
Cluster latent mean	4141	790	**84.5**	**61.2**	**25.5**	30.3	49.0	61.8	**72.6**	**83.3**	17.9	42.5	**58.1**	**73.3**	**86.7**
Cluster medoid	4141	790	84.4	61.1	25.3	**30.8**	**49.5**	61.6	**72.6**	**83.3**	**18.1**	**42.8**	58.0	73.2	**86.7**
SatCLIP															
Single sample	4125	912	83.4	61.0	22.4	2.7	36.6	58.4	72.0	82.7	2.4	36.4	55.5	**72.0**	**84.8**
Center	4103	**848**	82.7	60.2	22.3	2.5	32.7	55.3	70.5	83.1	**3.1**	35.9	55.0	71.0	84.7
Cluster mean	**4148**	860	**83.9**	**61.4**	22.8	**2.8**	37.3	**61.0**	**73.2**	**84.0**	**3.1**	**36.6**	**55.8**	71.5	84.7
Cluster latent mean	**4148**	860	**83.9**	**61.4**	**22.9**	**2.8**	37.2	60.9	**73.2**	**84.0**	**3.1**	**36.6**	**55.8**	71.6	84.7
Cluster medoid	**4148**	860	**83.9**	**61.4**	22.8	2.6	**37.6**	60.7	73.1	**84.0**	2.6	**36.6**	55.7	71.5	84.7
TTE															
Single sample	4129	849	83.5	57.4	18.4	0.9	30.1	55.3	70.8	82.7	**3.9**	34.5	56.3	73.0	**85.5**
Center	4112	**783**	82.9	56.8	18.2	0.9	27.4	52.8	70.0	82.5	3.6	32.8	54.8	72.3	84.7
Cluster mean	**4157**	794	**84.1**	**58.1**	**18.7**	1.0	30.9	**58.8**	72.7	**83.5**	3.8	33.9	56.6	73.0	85.3
Cluster latent mean	**4157**	794	**84.1**	**58.1**	**18.7**	1.0	31.0	58.7	72.8	**83.5**	3.8	34.3	56.5	73.0	85.3
Cluster medoid	4156	794	**84.1**	**58.1**	**18.7**	**1.1**	**31.6**	58.4	**72.9**	83.4	3.8	**34.8**	**56.7**	**73.1**	85.4

Table 15: **DIG as a prior for GeoRanker.** Accuracy (%) within distance thresholds and median error (km). GeoRanker and DIG-XL (GeoCLIP) are trained on MP16-Pro.

	Im2GPS3k						**YFCC4k**					
Method	1 km	25 km	200 km	750 km	2500 km	Med.	1 km	25 km	200 km	750 km	2500 km	Med.
GeoRanker	18.6	45.0	61.5	76.2	89.2	63.5	32.6	43.3	53.8	70.0	82.6	121.2
GeoRanker + prior	18.9	46.5	62.9	76.9	89.4	47.7	33.4	45.4	55.8	71.4	83.4	79.9
GeoRanker + prior + samples	19.6	47.4	63.5	77.9	**89.6**	38.0	33.6	**45.5**	56.4	**72.2**	**84.1**	75.4
GeoRanker + prior + samples + snap	**20.1**	**47.6**	**64.1**	**78.1**	**89.6**	**36.0**	**34.0**	**45.5**	**56.8**	71.9	83.8	**73.7**

cluster mean. *Cluster latent mean* averages the generated location embeddings of that cluster and decodes the result.

Results. Table 14 shows the same pattern for all three encoders. Center lowers the mean error on OSV-5M but hurts accuracy at nearly every threshold, since averaging across modes pulls the prediction between them. Restricting to the largest cluster fixes this, and the three cluster rules improve on the single sample on OSV-5M and YFCC4k at almost every metric, with smaller and less consistent gains on Im2GPS3k. Cluster mean is best at 200 km for every encoder, gaining 2.6 to 3.6 points on YFCC4k. Cluster medoid is best at 25 km for every encoder, raising GeoCLIP from 47.5 to 49.5 on YFCC4k and from 41.6 to 42.8 on Im2GPS3k, and from 30.1 to 30.8 at 1 km on YFCC4k. On OSV-5M the cluster rules add 23 to 47 GeoScore and 0.5 to 1.1 points of country accuracy. Sixteen samples therefore give a consistent improvement at 16 times the sampling cost, and cluster medoid, the rule that returns an actual sample, is the safest choice.

A.2.4 DIG WITH GEORANKER

GeoRanker. GeoRanker (Jia et al., 2025) is a learned reranker over a small set of candidate locations. For each query image, the candidates are retrieved MP16-Pro (Jia et al., 2024) images with their coordinates and place names, plus GPS guesses from GPT-4V (OpenAI et al., 2024). A fine-tuned Qwen2-VL-7B (Wang et al., 2024) scores each candidate against the query with a scalar reward trained to be high for candidates near the true location, and the highest-reward candidate is used for coordinate prediction. We use the released weights, prompts and candidate lists (12 retrievals and 3 guesses) without modification and reproduce their reported accuracies in Table 15.

Injecting DIG. We draw $K = 16$ samples $\hat{\ell}_1, \ldots, \hat{\ell}_K$ per image from DIG-XL (GeoCLIP) trained on MP16-Pro and use them both as candidates and as a prior over the candidates. As candidates, each sample is added to the pool as a GPS-only guess and scored by the frozen reranker. As a prior, we add to every candidate's reward the fraction of samples within R km of it,

$$s(c) = r(q, c) + \lambda \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\big[\, d_{\mathcal{S}^2}(c, \hat{\ell}_k) \le R \,\big], \qquad \hat{\ell} = \arg\max_{c \in \mathcal{C}} s(c), \tag{13}$$

where $\mathcal{C}$ is the candidate pool and $d_{\mathcal{S}^2}$ the great-circle distance. We use $\lambda = 1$ and $R = 50$ km. We also try snapping each sample to its nearest MP16-Pro training point and add that image with its place name as a retrieval-style candidate. We find that snapping improves performance, likely because of the high geographic overlap between the training and test sets (Table 16). GeoRanker is never fine-tuned, and since rewards are pointwise every candidate is scored once, so all rows of Table 15 share one set of rewards.

Results. Table 15 reports Im2GPS3k and YFCC4k. Adding DIG as a prior improves GeoRanker at every threshold on both benchmarks and cuts the median error from 63.5 to 47.7 km and from 121.2 to 79.9 km. Adding the samples as candidates improves it further, and snapping adds a small gain at fine thresholds. The full system gains 2.6 and 2.2 points at 25 km, 2.6 and 3.0 points at 200 km, and reduces the median error by 43% and 39%.

A.3 Datasets

A.3.1 Metrics

Distance error. All metrics start from the great-circle distance d in km between the predicted and the ground-truth coordinate on a sphere of radius 6371 km. We report the mean and the median of d over the test set. The mean is dominated by the few predictions on the wrong continent, and the median by the typical case.

GeoScore. Following OSV-5M (Astruc et al., 2024), each prediction receives $5000 \cdot \exp(-d/1492.7)$ and GeoScore is the average over the test set. It ranges from 0 to 5000 and rewards fine-grained accuracy while remaining bounded for large errors.

Accuracy within a radius. Accuracy at r km is the percentage of test images with $d < r$. On Im2GPS3k and YFCC4k we use the standard radii of 1, 25, 200, 750 and 2500 km (Hays & Efros, 2008), corresponding to street, city, region, country and continent level.

Administrative accuracy. On OSV-5M we follow Dufour et al. (2025) and reverse-geocode the predicted coordinate offline to a country, region and city with GeoNames, and count a prediction as correct when its label matches the ground-truth label at that level.

A.3.2 Dataset Distributions

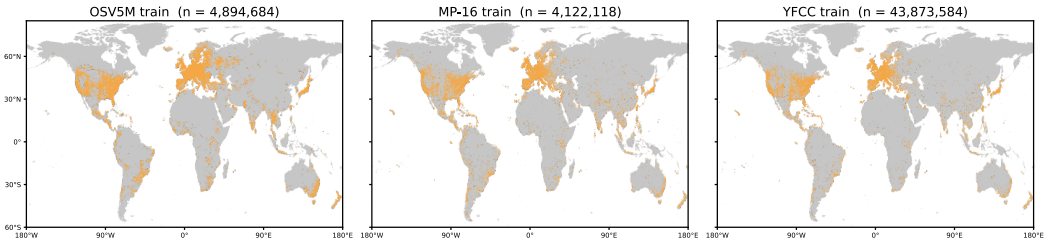


Figure 7: Geographic distribution of training sets.

Table 16: **Train–test geographic overlap.** For every test location, the great-circle distance to its nearest training location. "Within" columns give the share of test locations with a training location inside that radius.

	Distance (km)		Within (%)	
Training set	Median	Mean	1 km	25 km
Test: OSV-5M				
OSV-5M	4.32	15.7	0.1	86.4
Test: Im2GPS3k				
MP16-Pro	0.081	1.03	86.5	99.0
YFCC-44M	0.022	0.82	92.0	99.1
Test: YFCC4k				
MP16-Pro	0.004	0.69	93.7	99.7
YFCC-44M	0.000	0.19	97.2	99.8

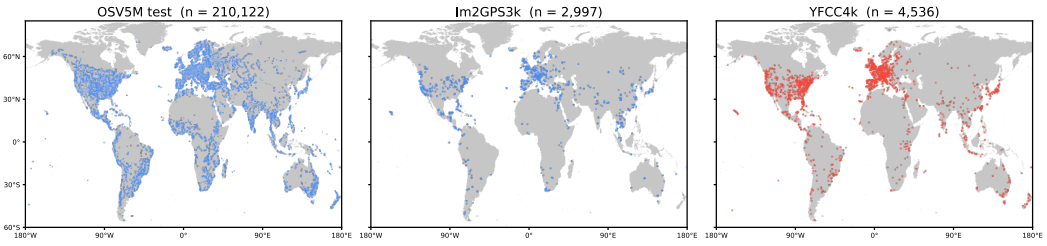


Figure 8: Geographic distribution of test sets.

Geographic bias. Figure 7 shows the geographic distribution of each training set. North America and Europe account for 70% of OSV-5M, 77% of MP16-Pro, and 81% of YFCC-44M. Africa accounts for 4.0% of OSV-5M and less than 2% of either Flickr dataset. We assign continents using Natural Earth admin-0 polygons. The test distributions differ substantially (Figure 8). YFCC4k closely follows its training distribution, with 77% of locations in North America and Europe and 1.7% in Africa. Im2GPS3k remains concentrated in North America and Europe at 61%. The OSV-5M test set is more balanced across countries. North America and Europe account for 39%, while Africa and South America account for 17% and 12%, respectively. OSV-5M therefore evaluates a larger geographic distribution shift, while the Flickr benchmarks remain closer to their training distributions.

Train–test spatial overlap. Table 16 reports the distance from each test location to its nearest training location. The Flickr benchmarks have substantial spatial overlap with their training sets. The median distance is 81 m for Im2GPS3k/MP16-Pro and tens of metres for YFCC4k, with 87–97% of test locations falling within 1 km of a training location. This overlap follows from their shared Flickr provenance, where photos often cluster around the same landmarks and viewpoints. Retrieval from a training gallery and classification over cells anchored at training coordinates benefit directly from this overlap. Sub-kilometre accuracy on these benchmarks therefore depends in part on training-set coverage. OSV-5M has much less overlap. Only 0.1% of its test locations fall within 1 km of a training location, and the median nearest-training distance is 4.3 km. Copying training coordinates is therefore insufficient for street-level accuracy. Together with the geographic rebalancing described above, this difference helps explain the stronger relative performance of retrieval methods on Im2GPS3k and YFCC4k.

A.3.3 Dataset Filtering

YFCC-44M. We build YFCC-44M from the public YFCC100M metadata dump Thomee et al. (2016), distributed by Multimedia Commons as a SQLite file, following the filtering of Dufour et al. (2025); Wang et al. (2025). We keep photos, discarding videos, that have coordinates with a Flickr

Table 17: Reported training sources of baselines on IM2GPS3k and YFCC4k.

	IM2GPS3k		YFCC4k	
Method	Source	# Images	Source	# Images
PlaNet (Weyand et al., 2016)	Web	126M	Web	126M
CPlaNet (Seo et al., 2018)	Flickr	30.3M	Flickr	30.3M
Translocator (Pramanick et al., 2022)	MP-16	4.7M	MP-16	4.7M
GeoDecoder (Clark et al., 2023)	MP-16	4.7M	MP-16	4.7M
PIGEOTTO (Haas et al., 2024)	Flickr/Wikipedia	4M	Flickr/Wikipedia	4M
GeoCLIP (Vivanco Cepeda et al., 2023)	MP-16	4.7M	MP-16	4.7M
HierLoc (Gadi et al., 2026)	MP-16	4.7M	MP-16	4.7M
Img2Loc (Zhou et al., 2024)	training-free	–	training-free	–
G3 (Jia et al., 2024)	MP16-Pro	4.1M	MP16-Pro	4.1M
GeoRanker (Jia et al., 2025)	MP16-Pro	4.1M	MP16-Pro	4.1M
vMFMix (Izbicki et al., 2019)	–	–	YFCC	48M
vMF (Dufour et al., 2025)	–	–	YFCC	48M
Diff $\mathbb{R}^3$ (Dufour et al., 2025)	YFCC	48M	YFCC	48M
FM $\mathbb{R}^3$ (Dufour et al., 2025)	YFCC	48M	YFCC	48M
RFM $\mathcal{S}_2$ (Dufour et al., 2025)	–	–	–	–
$\text{RFM}_{10\text{M}}$ $\mathcal{S}_2$ (Dufour et al., 2025)	YFCC	48M	YFCC	48M
LocDiff (Wang et al., 2025)	MP-16	4.7M	YFCC	48M
LocDiff-H (Wang et al., 2025)	MP-16	4.7M	–	–

geotag accuracy of at least 12 on Flickr's 1–16 scale, where 16 is street level. This leaves 43.90M photos. We remove the YFCC4k test set by Flickr photo ID. We download the remaining images from the Multimedia Commons S3 mirror and discard those that are unavailable. Finally, we remove by photo ID the Im2GPS3k test photos that also appear in YFCC100M, leaving 43,873,584 training images.

A.4 Additional Qualitative Results

Figure 9 visualizes DIG's predictive distribution by decoding Monte Carlo samples into a density over the globe. The posterior adapts its shape to the image. For the Vietnamese roadside, it places several modes on visually consistent tropical regions, including the correct coast; for the Japanese street, whose vegetation and architecture are distinctive, it collapses to a single sharp mode; for the Kazakh steppe, it spreads over the plausible region, keeping considerable mass near the ground truth while also entertaining a lookalike mode in Turkey.

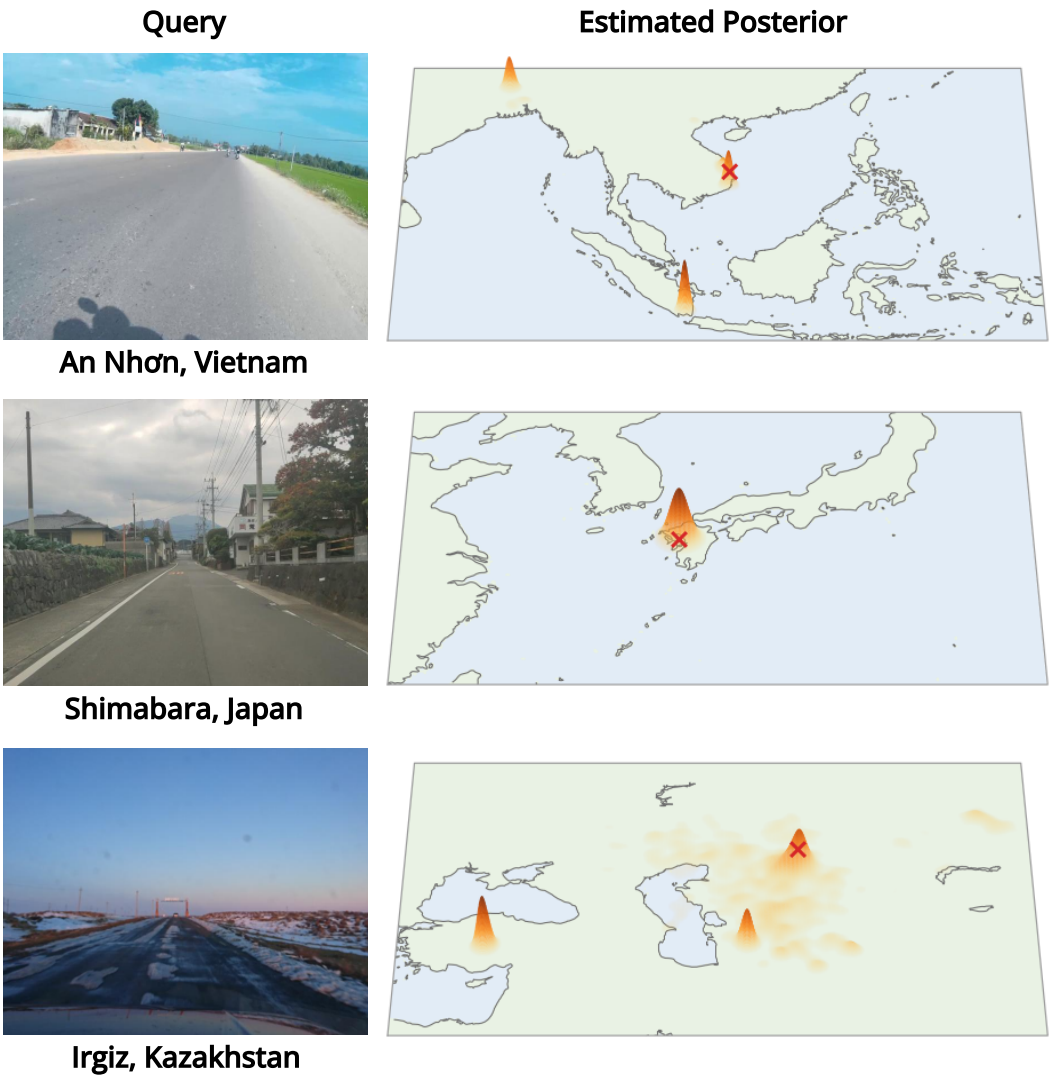


Figure 9: **Estimated posterior distribution.** We use DIG-XL (GeoCLIP) trained on OSV-5M.