

UNIGEOCLIP: Unified Geospatial Contrastive Learning

Guillaume Astruc^{1,3,4,*}, Eduard Trulls², Jan Hosang², Loic Landrieu^{1,4}, Paul-Edouard Sarlin²
¹ LASTIG, Univ Gustave Eiffel, IGN, ENSG, France ² Google, Switzerland ³ CNES, France ⁴
 LIGM, CNRS, Univ Gustave Eiffel, ENPC, Institut Polytechnique de Paris, Marne-la-Vallée, France
 guillaume.astruc@enpc.fr

Abstract

The growing availability of co-located geospatial data spanning aerial imagery, street-level views, elevation models, text, and geographic coordinates offers a unique opportunity for multimodal representation learning. We introduce UNIGEOCLIP, a massively multimodal contrastive framework to jointly align five complementary geospatial modalities in a single unified embedding space. Unlike prior approaches that fuse modalities or rely on a central pivot representation, our method performs all-to-all contrastive alignment, enabling seamless comparison, retrieval, and reasoning across arbitrary combinations of modalities. We further propose a scaled latitude-longitude encoder that improves spatial representation by capturing multi-scale geographic structure. Extensive experiments across downstream geospatial tasks demonstrate that UNIGEOCLIP consistently outperforms single-modality contrastive models and coordinate-only baselines, highlighting the benefits of holistic multimodal geospatial alignment. A reference implementation is available at gastruc.github.io/unigeoclip.

1. Introduction

Expressive and robust geospatial embeddings that capture both fine-grained semantic content and large-scale spatial structure are critical for automating downstream geospatial tasks such as urban land-use classification [16], monitoring land cover [14], and large-scale socio-economic inference [33]. Existing work largely falls into three successful paradigms. *Embedding fields* map geographic coordinates to latent vectors to enable localized interpolation [5, 18]. *Multimodal vision models* fuse multiple sensor observations into a single representation [3, 4, 36]. Finally, *contrastive approaches* align heterogeneous geospatial data sources in a shared embedding space, most notably GeoCLIP [38] and SatCLIP [15]. Despite their success, these paradigms exhibit limitations for general-purpose geospatial reasoning. Embedding fields provide static “snapshots” of a region and struggle

*Work done during an internship at Google.

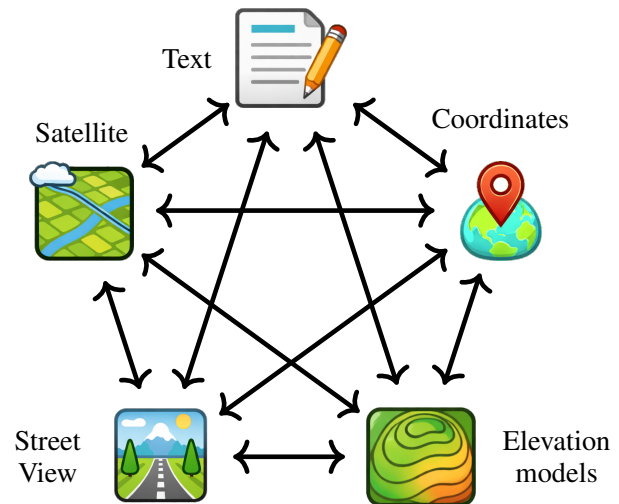


Figure 1. **Unified contrastive learning of geospatial data.** We jointly train encoders for five modalities (text, aerial imagery, street-level imagery, elevation, and geographic coordinates), which simultaneously are contrasted across all modality pairs. This yields a single unified embedding space that represents heterogeneous geospatial information.

to model dynamic. Multimodal fusion models collapse all available modalities into a single representation, preventing cross-modal comparison or retrieval. Existing contrastive approaches typically align geographic coordinates with a single Earth observation modality, most often top-down imagery, and largely ignore textual information despite its central role in modern vision-language models.

In this paper, we propose a *contrastive and massively multimodal* framework that jointly aligns five complementary geospatial modalities, enabling seamless translation, comparison, and retrieval across modalities. Specifically, we contrast high-resolution aerial imagery, geometry-dense Digital Surface Models (DSMs), street-level imagery, rich text descriptions, and raw geographic coordinates, embedded using a novel spatial encoder. Unlike prior multimodal contrastive frameworks in vision, such as UniBind [20] or ImageBind [10], which rely on a central pivot modality (typically im-



Figure 2. **Sample from our multimodal geospatial dataset.** Each location is represented through five complementary modalities: aerial imagery, street-level imagery, a Digital Surface Model (DSM), geographic coordinates, and an automatically generated text description.¹ All modalities are spatially co-registered and jointly contrasted during training.

ages), our approach adopts a fully holistic formulation: all modalities are directly contrasted with one another. This all-to-all alignment strategy yields a unified embedding space that supports robust reasoning under arbitrary availability of modalities. Observing that embeddings derived from raw positional encodings are often limited in expressiveness [28] and can become a bottleneck when contrasted with richer modalities, we propose a learned multi-scale geographic embedding that substantially increases representational capacity.

In summary, we make the following contributions:

- **UNIGEOCLIP: Unified Geospatial Contrastive Learning.** We introduce the first purely contrastive framework that jointly aligns an unprecedented set of georeferenced modalities: street-view imagery, aerial imagery, DSMs, text, and geographic coordinates.
- **Scaled Latitude–Longitude Encoder.** We propose a novel coordinate encoder that outperforms standard Fourier-feature and MLP-based embeddings by explicitly modeling multi-scale spatial dependencies.
- **Strong performance in geospatial tasks.** We demonstrate consistent improvements over single-modality contrastive models and coordinate-only baselines across a diverse suite of geospatial probing and downstream tasks.

2. Related Work

Multimodal Geospatial Models Recent developments in geospatial representation learning have transitioned from specialized, task-specific encoders to broader foundation models

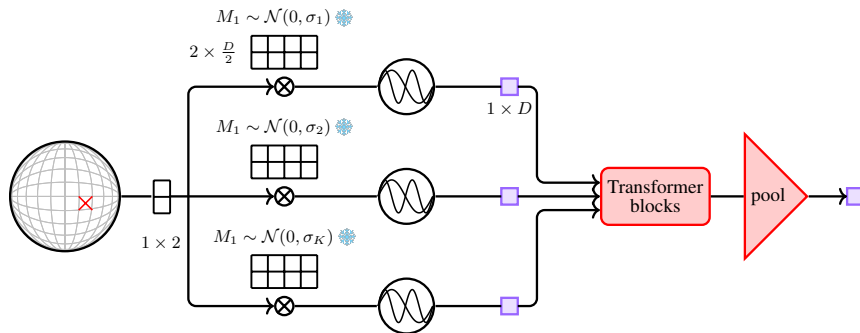
designed to capture both fine-grained semantic content and large-scale spatial structures [16]. Significant focus has been placed on *embedding fields* [5, 18], which map geographic coordinates directly to latent features to allow localized spatial interpolation. Although these fields excel at specific downstream tasks, they essentially act as static snapshots—frozen in time and strictly bound to the geographic distribution of their training sets. In parallel, multimodal vision models such as *AnySat* [4], *Panopticon* [39], and *Galileo* [36] have begun to explore the fusion of multiple satellite sensors (*e.g.*, SAR, multispectral). However, these models combine all sensor observations into a single representation, preventing cross-modal comparison or retrieval. Moreover, these frameworks remain largely focused on bird’s-eye-view data modalities, ignoring the rich ground-level perspective provided by street-level imagery or textual data.

Unified Multimodal Binding and Contrastive Alignment

The “binding” paradigm seeks to align disparate data streams into a single latent space to support arbitrary modality availability. *ImageBind* [10] popularized the use of a central pivot modality (typically vision) to align sensors, a strategy recently adapted for ecological data in *TaxaBind* [29]. In the geospatial domain, contrastive approaches have historically focused on the image-location relationship. Seminal works like *SatCLIP* [15] and *GeoCLIP* [38] established the standard for pairing imagery with geographic coordinates. While these have been refined through improved retrieval functions [6], temporal dynamics [31], and localized attention [19], they typically align only two modalities at a time. Crucially, existing contrastive frameworks largely ignore textual information, despite its central role in modern vision-language

¹Analytical use of proprietary data sources was done with special permission from Google. Faces and license plates are anonymized by blurring.

Figure 3. **Multi-Scale Coordinate Encoder.** Latitude–longitude coordinates are first projected using multiple random Fourier feature matrices with increasing spectral bandwidths. Each frequency projection is treated as a token and processed through self-attention to enable inter-scale interaction. The resulting tokens are averaged to produce a unified D -dimensional geographic embedding.



models. Unlike *UniBind* [20], which still relies on a pivot, our approach adopts a fully holistic all-to-all formulation, ensuring that text, DSM, and visual sensors are all primary citizens in the embedding space.

Geolocation and Cross-Modal Retrieval The boundaries of global-scale positioning have been pushed by leveraging the relationship between ground-level and overhead perspectives [12, 13]. Recent geolocation models like *PIGEON* [11], *OpenStreetView* [2] and *Plonk* [8] focus on predicting geolocation from street-level image only. Specialized retrieval frameworks, such as *CityLoc* [21] and *Text2Loc* [40], have shown success in urban understanding; however, they are often limited to narrow pairings (e.g., text-to-image or image-to-GPS). *ScalingGeoloc* [18] aligns Street-View images to aerial image and cell code prototypes. By integrating five modalities simultaneously, our work addresses the limitations of these specialized models, enabling a more robust geospatial reasoning framework that can infer socio-economic variables [33] or monitor land cover [14] by cross-referencing ground-level, top-down, and elevation data within a single, unified manifold.

3. Method

We consider a multimodal sample x characterized by a set of M modalities $x = \{x^1, x^2, \dots, x^M\}$, including street-level imagery (‘SV’), aerial imagery (‘sat’), elevation models (‘DSM’), textual descriptions (‘txt’), and geographic coordinates (‘GPS’). Our objective is to jointly train modality-specific encoders $\{\phi^m\}_{m \in \mathcal{M}}$ to output representations that are directly comparable across modalities.

3.1. Architecture

Each modality is embedded with a dedicated encoder.

Embedding Earth Observation and Text. Street-level and aerial images are encoded with modality-specific image encoders ϕ^{SV} and ϕ^{sat} , respectively. Both are instantiated from the image encoder of SigLIP-2 [35]. Text inputs are embedded using the SigLIP-2 text encoder ϕ^{txt} . For terrain information, we train a Digital Surface Model (DSM) en-

coder ϕ^{DSM} from scratch. This encoder is implemented as a Vision Transformer with register tokens, and use the class token of the last layer as the modality embedding.

Embedding GPS Coordinates. We propose a novel learned coordinate encoder ϕ^{GPS} for geographic locations $x^{\text{GPS}} = (x^{\text{lon}}, x^{\text{lat}})$. To mitigate distortions induced by spherical geometry, we first apply the *Equal Earth Projection* [30], mapping latitude–longitude coordinates to a planar representation. Inspired by GeoCLIP [34, 38], we adopt Random Fourier Features to encode spatial information. We define a random spectral projection matrix $\mathbf{M} \in \mathbb{R}^{\frac{D}{2} \times 2}$ with entries sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. This matrix is fixed and not learned during training. The encoding is obtained by concatenating the sine and cosine components:

$$\gamma_{\mathbf{M}}(x^{\text{GPS}}) = [\cos(2\pi\mathbf{M}x^{\text{GPS}}), \sin(2\pi\mathbf{M}x^{\text{GPS}})]^{\top}. \quad (1)$$

To capture multi-scale spatial structure, we introduce a scalable multi-frequency fusion mechanism. We perform projections using a set of K matrices $\{\mathbf{M}_k\}_{k=1}^K$, sampled with increasing spectral variances $\{\sigma_k\}_{k=1}^K$. Each projected embedding is treated as a token. Unlike prior approaches such as GeoCLIP [38], which process each scale independently using separate MLPs and aggregate them by averaging, we explicitly allow interactions across spatial scales. Specifically, the K tokens are processed by B self-attention blocks with register tokens, enabling information exchange between frequencies. The final GPS embedding is obtained by averaging the output tokens, yielding a single D -dimensional representation.

3.2. Supervision

We consider a batch $\{x_1, \dots, x_B\}$ of multimodal samples, where each sample $x_i = \{x_i^1, x_i^2, \dots, x_i^M\}$ corresponds to a geographic location observed through M co-located modalities. Each modality is encoded by a modality-specific encoder ϕ^m into a shared D -dimensional embedding $f_i^m = \phi^m(x_i^m)$. Our objective is to learn embeddings that are *spatially consistent*: representations associated to the same location are close in the embedding space, while those from different locations are far apart.

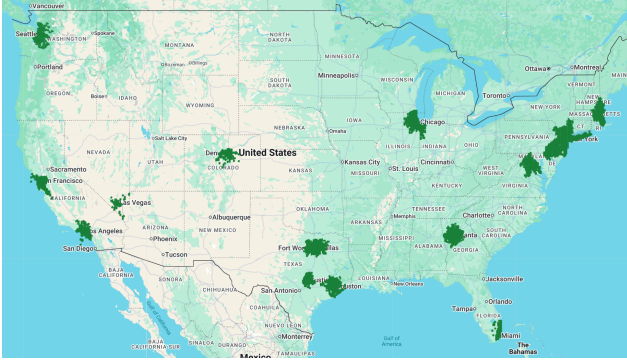


Figure 4. **Geographic Coverage.** Spatial distribution of sampled locations across the continental United States. Green regions indicate areas containing samples after spatial filtering and farthest-point sampling.

Multi-Way Contrastive Objective. We supervise the encoders using a multi-way contrastive objective that jointly aligns all modalities. Specifically, we minimize the average InfoNCE loss [23] over all ordered modality pairs $(m, n) \in \mathcal{M}^2$:

$$\mathcal{L} = \frac{1}{M^2} \sum_{(m,n) \in \mathcal{M}^2} \mathcal{L}_{m \rightarrow n}, \quad (2)$$

$$\mathcal{L}_{m \rightarrow n} = -\frac{1}{B} \sum_{i=1}^B \log \left(\frac{\exp(\langle f_i^m, f_i^n \rangle / \tau)}{\sum_{j=1}^B \exp(\langle f_i^m, f_j^n \rangle / \tau)} \right), \quad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is a temperature parameter. By exhaustively contrasting all modality pairs, this objective enforces a fully shared latent space across modalities, enabling robust cross-modal retrieval and reasoning without relying on a privileged pivot modality.

4. Experiments

We first describe the dataset used to train UNIGEO-CLIP (Sec. 4.1), then present quantitative evaluations on cross-modal retrieval (Sec. 4.2) and additional downstream tasks (Sec. 4.3), followed by an ablation study (Sec. 4.4).

4.1. Dataset

To train and evaluate UNIGEOCLIP, we assemble a large-scale multimodal dataset providing a dense, multi-perspective representation of urban environments across the continental United States.

Spatial Extent and Sampling. The dataset spans the continental USA, restricted to the largest metropolitan centers, which contain the richest multimodal data. To ensure uniform spatial coverage, we partition the territory using S2 cells [17]. We consider cells at level $L = 16$, which roughly corresponds to an area of 150×150 m. Our full spatial

coverage is composed of $\sim 800k$ S2 cells. Within each cell, we apply Farthest Point Sampling [9] to select up to 120 street-level panoramas, enforcing a minimum separation of 40 m between samples. Cells containing fewer than five valid observations are discarded. This strategy yields a spatially balanced dataset while avoiding excessive clustering in dense metropolitan areas.

Temporal Split. To evaluate robustness under temporal distribution shift, we adopt a strict temporal split. We run evaluations using data from year 2023, while data from years 2017–2024 (excluding 2023) is used for training, following the same evaluation setup as Lindenberger *et al.* [18]. This protocol prevents temporal leakage and assesses the ability of the learned representations to generalize across evolving urban landscapes.

Modalities. As shown in Fig. 2, we collect five complementary data modalities for each geographic location:

- **Aerial Imagery.** High-resolution overhead imagery is resampled to 60 cm/pixel resolution and cropped into 256×256 tiles centered at each location.
- **Street-Level Imagery.** Each panorama contributes four perspective crops. Panoramic imagery is stitched and rendered using a pinhole camera model, from which 224×224 crops are sampled with randomized roll, pitch, yaw, and field-of-view to encourage viewpoint robustness.
- **Digital Surface Models (DSM).** Elevation data provides dense geometric structure aligned with the visual modalities. DSM patches are extracted at resolution 60 cm/pixel and spatially co-registered with the aerial imagery.
- **Text Descriptions.** Each location is associated with an automatically generated textual description derived from large-scale georeferenced data. These descriptions capture semantic attributes such as land use, built environment characteristics, and context such as local landmarks: see Fig. 2 for an example.
- **Geographic Coordinates.** Raw latitude and longitude corresponding to each sampled location.

Data Sources and Licensing. The modalities used in this work are collected from a combination of commercial and proprietary data sources under standard licensing agreements.

4.2. Cross-Modality Retrieval

We evaluate cross-modal alignment through a zero-shot geospatial retrieval task, with results reported in Tab. 1. Given a street-view query, the objective is to retrieve the geographically matching instance in another modality. Performance on this task quantifies the consistency of semantic

Table 1. **Cross-Modal Street View Retrieval.** We report Acc@100 m for cross-modal retrieval and specify the modalities contrasted during training. OOD denotes the out-of-domain evaluation setting. refers to geocells hashcodes. GeoCLIP is scaled up to the parameter count of our model and retrained on our data.

retrieval →	training modalities	Target			
				{, , ,	OOD
GeoCLIP [38]		-	24.6	-	4.5
ScalingGeoloc [18]		45.8	-	-	56.9
UNIGEOCLIP		-	41.2	-	24.8
UNIGEOCLIP		83.9	-	-	41.3
UNIGEOCLIP		76.7	46.5	75.6	32.3 29.0
UNIGEOCLIP		77.2	47.0	81.9	33.5 29.6
UNIGEOCLIP		88.2	69.4	91.0	41.2 41.2

and spatial representations across modalities.

Evaluation Protocol. Following the standard retrieval-based localization paradigm, we ℓ_2 -normalize all embeddings and compute cosine similarities between a *query modality* and a *database of georeferenced candidates*. We choose ground-level images as queries, and use modalities in {, , , } as targets. The predicted location corresponds to the candidate with the highest similarity score. To ensure a fair comparison with GeoCLIP, we fine-tuned it on our training set, scaled its coordinate encoder to match our parameter count, and utilized an identical training regime. We evaluate street-view queries against the following targets:

- **Aerial** . This corresponds to the classic cross-view retrieval task, matching ground-level imagery to overhead observations.
- **GPS Coordinates** . This setting evaluates direct image-to-location retrieval [12], where street-view embeddings are matched against encoded geographic coordinates.
- **Multimodal Ensembling** {, , , }. To assess complementarity across modalities, we compute separate similarity matrices for each available modality and aggregate them by simple averaging before selecting the global maximum. This measures the synergy between heterogeneous geospatial signals. In Table 1, this indicates ensembling the data modalities each model was trained with (excluding ground-level images).
- **Out-of-Distribution Aerial** (OOD). We evaluate cross-view retrieval on a geographically distinct region (Amsterdam) to assess spatial generalization beyond the training distribution (USA).
- **Geocells** . Following the protocol of Scaling Geoloc [18], we evaluate localization via spatial discretization. The study area is partitioned into discrete cells, each represented by encoding the centroid coordinates using our latitude–longitude encoder. Since our model is not explicitly trained on geocell tokens, this setting probes its ability to generalize to discretized spatial representations.

Table 2. **Satellite Image Encoder.** We evaluate the ability of our satellite image encoder to analyze aerial and satellite imagery of various image encoders on two geospatial benchmarks: solar panel detection and land-cover segmentation . Models are grouped into cross-modal contrastive frameworks (top), Earth observation foundation models (middle), and general-purpose vision foundation models (bottom). We underline the best performance among contrastive models and **bold** the best overall performance.

classif / semseg →	training modalities	model	m-pv4ger classif (OA)	m-chesapeake semseg (mIoU)
SatClip [15]		ViT-B	93.0	59.3
SigLip2 [35]		ViT-B	95.7	60.9
UNIGEOCLIP		ViT-B	96.9	65.9
UNIGEOCLIP		ViT-B	<u>97.0</u>	<u>66.3</u>
SenPaMAE [25]		ViT-B	87.1	46.9
DOFA [41]		ViT-L	97.4	59.2
AnySat [4]		ViT-B	92.8	61.7
Panopticon [39]		ViT-B*	96.4	60.8
Dinov2 [24]		ViT-B*	97.5	64.0
Dinov3 Web [32]		ViT-7B	98.3	76.5

Analysis. From the results in Tab. 1, we draw the following conclusions:

- **Stronger Cross-Modal Alignment.** UNIGEOCLIP consistently outperforms a retrained GeoCLIP model [38] for retrieval-based geolocation. This indicates improved cross-modal harmonization.
- **Impact of Multi-Contrastive Formulation.** Our models improve the more modalities we contrast. Note that contrasting with text requires a larger batch size to converge.
- **Complementarity of Modalities.** Multimodal ensembling consistently surpasses the best individual modality. Ground-level imagery captures fine-grained semantics, whereas aerial imagery and DSMs encode structural and spatial layout cues; their combination yields a more robust and discriminative retrieval capacity.
- **Generalization to Spatial Discretization.** Under the geocell protocol, GeoCLIP struggles to generalize. In contrast, our model achieves competitive localization accuracy despite never being trained for cell-based classification, indicating greater spatial flexibility.
- **Out of Distribution.** We show that our model can generalize to unseen areas with different statistics, by evaluating on a city in Europe (Amsterdam) using a model trained in the USA.

4.3. Encoder Downstream Evaluation

We evaluate the capacity of individual encoders trained with our framework to generalize to downstream tasks. Encoders are frozen and assessed via linear probing.

Satellite Image Encoder. We evaluate the aerial encoder on two complementary geospatial tasks: (i) *m-pv4ger* [16, 22] is a photovoltaic panel detection benchmark based on high-resolution aerial imagery, requiring the identification

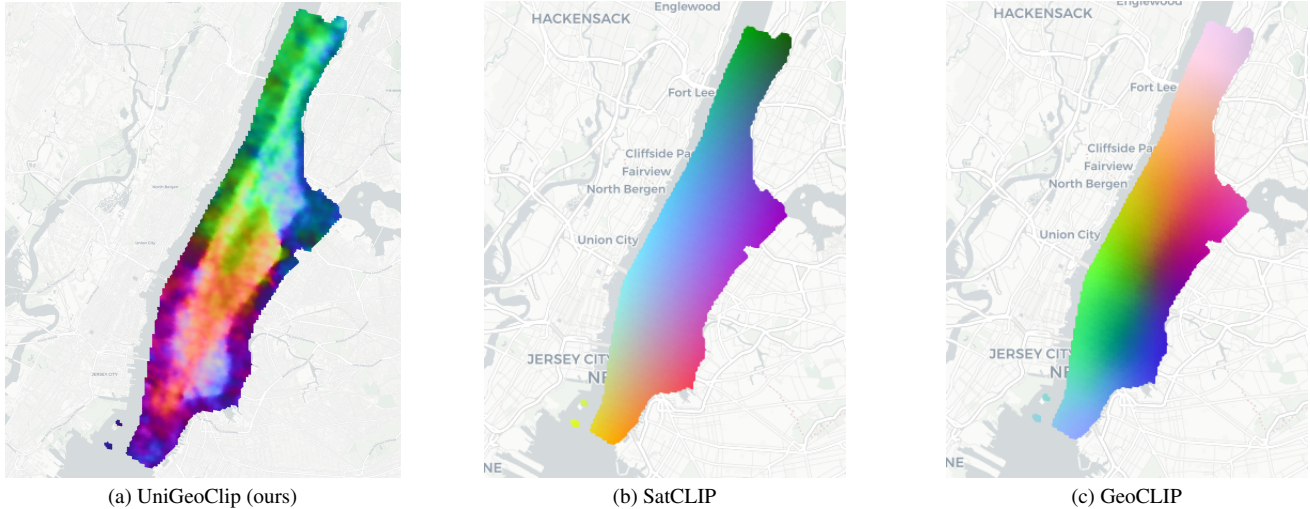


Figure 5. **PCA of Coordinate Embeddings.** Embeddings computed over a dense grid in Manhattan, NYC are projected using PCA, with the top three principal components mapped to RGB. UniGeoCLIP produces spatial patterns that reflect underlying urban structure (e.g., Central Park and surrounding neighborhoods), indicating semantically informed representations. In contrast, SatCLIP and GeoCLIP exhibit smoother, predominantly position-driven gradients.

Table 3. **Evaluation of the Location Encoder.** Performance on 27 downstream regression tasks spanning health, socio-economic, and environmental indicators. Models are grouped into contrastive approaches (top) and pre-computed embedding fields (bottom). Variants marked with † are retrained on our dataset. ★ denotes models specifically trained for socio-economic prediction. We **bold** the best overall performance and underline the best performance among contrastive approaches.

Regression 🌍 → 🎯	target 🎯			mean R^2
	health	social	environmental	
SigLIP 2 [35]	34.5	48.5	72.3	40.2
SatClip [15]	25.7	34.4	70.7	30.1
† SatClip [15]	32.3	48.5	43.7	36.7
GeoClip [38]	45.2	65.7	84.2	49.8
† GeoClip [38]	47.3	67.9	77.9	51.6
UNIGEOCLIP	53.1	69.9	81.1	<u>57.0</u>
AlphaEarth [5]	23.1	29.9	83.3	29.0
★ PDFM [1]	73.9	82.6	82.3	74.5

of small structures within large scenes; (ii) *m-chesapeake* [16, 26] is a land-cover semantic segmentation dataset requiring dense pixel-wise classification. For segmentation, we extract patch embeddings from the final transformer layer and project them to class logits of size $P \times P \times C$, where P denotes the patch size and C the number of classes. Together, these tasks assess both global semantic discrimination and fine-grained spatial reasoning.

As shown in Tab. 2, **UniGeoCLIP** achieves the strongest performance among all contrastive frameworks. It surpasses both *SigLIP2* and a fine-tuned version of *SatCLIP* [15] on m-pv4ger (97.0% OA) and m-chesapeake (66.3 mIoU), highlighting the representational advantages of multimodal geospatial pre-training.

On semantic segmentation, our model also outperforms specialized Earth observation models such as *AnySat* and *Panopticon*, despite never been trained for dense tasks. Although large-scale general-purpose models such as *DINOv3 Web* [32] achieve higher absolute performance, they rely on substantially larger architectures (e.g., ViT-7B) and web-scale pre-training.

Overall, these results suggest that the multi-modal contrastive objective effectively distills structural geographic information into the aerial encoder, enabling strong downstream performance without task-specific supervision or increased model scale.

Spatial Coordinate Encoder. We evaluate the representation power of our novel coordinate encoder on the 27 downstream regression tasks introduced by Sun *et al.* [33], where geographic coordinates are mapped to health, socio-economic, and environmental indicators. We limit this evaluation to locations that overlap with our training set. This yields 1,447 training locations and 179 test locations.

We then perform linear probing on our frozen coordinate embeddings and compare against several baselines: (i) other contrastive models, both off-the-shelf and retrained on our dataset, (ii) the text encoder *SigLIP 2* [35], with which the geographic information is provided via a detailed textual prompt that includes the country, state, county, city, and geographic coordinates, (iii) pre-computed embedding fields inferred by *AlphaEarth* [5] from Earth observations by *PDFM* [1] from rich auxiliary socio-economic and environmental covariates.

Table 3 shows that our encoder achieves a mean R^2 of 57.0, substantially outperforming existing embedding fields from *AlphaEarth* (29.0) and contrastive baselines like *Sat-*

CLIP (30.1) and *GeoCLIP* (49.8). Importantly, even when compared to our retrained versions of these baselines under identical data conditions, our coordinate encoder exhibits stronger spatial generalization.

While *PDFM* remains state-of-the-art on this regression-focused benchmark, it leverages a substantially broader set of auxiliary spatial signals—including thematic maps, search trends, and environmental covariates. In contrast, our approach relies solely on multimodal alignment of fundamental geospatial inputs. These results indicate that contrastive multi-modal training induces rich spatial representations that correlate strongly with socio-economic and environmental indicators.

Retraining SatCLIP and GeoCLIP on our dataset significantly improves performance in health and social regression tasks. This gain is likely due to the spatial alignment between our training locations and the benchmark, allowing the models to capture localized socio-economic nuances. Conversely, environmental performance remains stagnant or decreases; these tasks likely benefit more from the vast geographic diversity of the original models’ broader training sets, which capture large-scale ecological patterns that our metropolitan-focused data may lack.



We visualize in Fig. 5 the representations produced by our coordinate encoder by applying Principal Component Analysis (PCA) to embeddings computed over a dense grid of locations in Manhattan, New York. The resulting projection reveals a continuous and semantically structured spatial representation. Rather than discretely hashing geographic regions, the encoder learns a smooth manifold characterized by coherent clusters and gradual transitions that reflect underlying urban structure, such as Central Park.

Compared to alternative spatial encodings, our embeddings exhibit sharper boundaries and more distinctive spatial organization, indicating improved representational fidelity and multi-scale geographic modeling.

Evaluation of the DSM Encoder. We consider the task of pixelwise land-cover semantic segmentation of DSM images from the **MDAS** dataset [14], which contains 1,702 high-resolution images totaling 24G annotated pixels. Unlike aerial imagery, DSM data currently lacks widely adopted large-scale foundation models. We therefore compare our DSM encoder against two standard baselines trained from scratch: a U-Net and a Vision Transformer (ViT). In contrast, our DSM encoder is evaluated under a linear probing protocol with frozen weights.

As shown in Tab. 4, our model outperforms both baselines by a substantial margin. By aligning elevation data with semantic text and visual imagery during pre-training, the encoder acquires structurally informed representations that standard architectures trained solely with semantic supervision fail to capture. These results indicate that cross-modal contrastive pre-training serves as an effective initialization

Table 4. **Evaluation of the DSM Encoder on MDAS.** We evaluate the accuracy of land-cover prediction from DSM-only input. Our linearly-probed DSM encoder significantly outperforms U-Net and ViT-B models trained from scratch.

 ↔ 	accuracy
U-Net [27]	45.5
ViT-B [7]	39.1
UNIGEOCLIP	72.0

strategy for DSM understanding, compensating for the absence of domain-specific pre-trained models and yielding a more robust backbone for elevation-driven tasks.

4.4. Discussion and Ablation Study

We present an ablation study to analyze the impact of key design choices and an analysis to better understand the behavior of our multimodal alignment framework.

Which Modalities Contribute Most? To quantify the contribution of each modality to the shared embedding space, we evaluate retrieval performance under different training configurations. As shown in Tab. 1, performance improves consistently as additional modalities are incorporated. A model trained only on aerial imagery and coordinates already achieves competitive zero-shot localization. However, adding Street View significantly enhances spatial precision, and the full multimodal configuration (including text and DSM) reaches a peak Accuracy@100 m of 69.4%. These results confirm that each modality contributes complementary spatial information.

Can We Align Heterogeneous Modalities? Our framework contrasts modalities that are fundamentally different in nature: text, imagery, elevation data, and raw geographic coordinates. Although these modalities vary in structure and dimensionality, they share a common latent factor: the geographic content of a location. We thus wonder whether such heterogeneous signals are really embedded into a shared space without sacrificing representational expressivity. To investigate this, we visualize in Fig. 6 the learned latent space using t-SNE [37].

Embeddings of distinct modalities form compact clusters corresponding to individual geographic locations. The consistent co-localization of all five modalities within these clusters indicates that the model learns a modality-invariant geographic representation, rather than segregating modalities into disjoint subspaces. Within each cluster, embeddings of aerial and DSM are generally closer to each other, as are those of location and street view.

Should We Scale the Location Encoder? The location encoder operates on low-dimensional inputs consisting of latitude and longitude. This raises the question of whether increasing architectural capacity is justified for mapping two

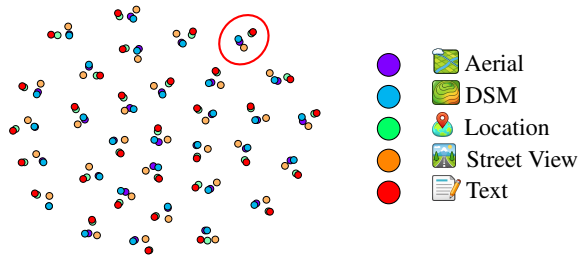


Figure 6. **Location Embedding Visualization.** T-SNE projection of embeddings for 48 distinct locations. Each cluster \circ corresponds to a single geographic location and contains embeddings from all modalities.

Table 5. **Ablation of the number of blocks of the location encoder.** This ablation is conducted with lower batch size.

Retrieval	Target \odot				
			{,,}		
Depth 0	58.1	55.0	10.2	13.6	4.4
Depth 4	77.2	73.1	40.7	30.0	25.4
Depth 8	77.6	73.1	44.0	27.8	27.6
Depth 12	79.1	74.4	47.0	29.2	29.3

scalar values into a rich embedding space. To examine the relationship between model capacity and representational power, we vary the number of self-attention blocks in the GPS encoder and evaluate performance across multiple retrieval tasks. As reported in Tab. 5, increasing the depth of the encoder yields consistent gains across all cross-modal retrieval settings, including aerial and out-of-distribution (OOD) evaluation.

At *Depth 0*, the encoder reduces to fixed Random Fourier Features without transformer layers. In this regime, performance is limited: coordinate retrieval reaches only 10.2, and cell-code generalization drops to 4.4. As depth increases, retrieval accuracy improves steadily. With 12 self-attention blocks, the coordinate retrieval rises to 47.0, and multimodal ensembling reaches 79.1. These results suggest that, despite the low dimensionality of its input, the coordinate encoder benefits substantially from increased capacity. The transformer layers enable non-linear interactions across spatial frequencies, allowing the model to capture complex geographic structure beyond simple spectral projections.

Comparison with Alternative Location Encoding Paradigms. To evaluate the impact of our multi-scale design, we replace our coordinate encoder with two existing alternatives scaled to match our parameter count: *GeoCLIP* [38], using Random Fourier Features [34], and *SatCLIP* [15], using SIREN [28]. All models are trained using identical modalities and supervision, for a fair comparison.

Despite parity in capacity, our approach significantly outperforms the scaled baselines on coordinate regression tasks (Tab. 3) and also improves downstream aerial representation

Table 6. **Comparison of different location encoders.** Our model is trained on same data modalities as the baselines.

location encoder architecture	location retrieval	geocell retrieval	aerial semseg	socio-economic variables
SirenNet (SatCLIP)	-	-	59.3	35.5
RFF encoder (GeoCLIP)	24.6	4.5	-	52.6
Our location encoder	41.2	24.8	65.9	56.7

quality (Tab. 2). In particular, our joint aerial–coordinate pre-training consistently exceeds the performance of the scaled *SatCLIP* baseline. Our multi-scale frequency fusion mechanism leverages depth to model complex geographic dependencies through cross-scale interactions that shallow MLP- or RFF-based encoders fail to capture.

Generalization Beyond the Training Region. To assess the generalization capacity of our learned representations, we perform zero-shot geolocation in an out-of-distribution (OOD) setting. We curate an evaluation dataset from Amsterdam, a region excluded from pre-training and exhibiting a substantial domain shift relative to the training data (U.S. cities). The dataset contains aligned aerial imagery, GPS coordinates, and street-level observations.

Localization is performed via cross-modal retrieval: Street View queries are matched against an aerial image gallery, and the latitude–longitude of the top-retrieved aerial image is used as the predicted location. As reported in Tab. 1, the model trained solely with aerial and Street View contrast achieves the strongest OOD performance. This is expected, as incorporating additional modalities may encourage specialization to the geographic distribution of the training data. Nevertheless, the full multimodal model maintains competitive performance, demonstrating that the learned embeddings retain strong generalization beyond the training region. Importantly, performance trends in this OOD evaluation remain consistent with those observed in the primary in-distribution experiments. This stability across geographic domains highlights the robustness of our multimodal alignment strategy.

5. Conclusion

We introduced UNIGEOCLIP, a unified multimodal framework for global geographic representation learning. By adopting a multimodal, all-to-all contrastive objective, our method is able to align five heterogeneous modalities (aerial imagery, Street View, coordinates, DSM, and text) within a shared embedding space. This holistic alignment improves zero-shot geolocation, cross-modal retrieval, and downstream encoder performance. We also propose a transformer-based coordinate encoder that outperforms traditional spatial encoder baselines. Overall, UNIGEOCLIP provides a scalable foundation for multimodal geospatial reasoning and large-scale Earth observation applications.

Acknowledgments

The authors thank Simon Lynen and Filip Saina for inspiring discussions and valuable feedback on the manuscript. We thank Mimi Sun for helping with the evaluation of PDFM and Matthew Pereira for helping with the text summaries.

References

- [1] Mohit Agarwal, Mimi Sun, Chaitanya Kamath, Arbaaz Muslim, Prithul Sarker, Joydeep Paul, Hector Yee, Marcin Sieniek, Kim Jablonski, Swapnil Vispute, et al. General Geospatial Inference with a Population Dynamics Foundation Model. *arXiv:2411.07207*, 2024. **6**
- [2] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronsson, 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 *CVPR*, 2024. **3**
- [3] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-Supervised Modality Fusion for Earth Observation. In *ECCV*, 2024. **1**
- [4] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth Observation Model for Any Resolutions, Scales, and Modalities. In *CVPR*, 2025. **1, 2, 5**
- [5] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv:2507.22291*, 2025. **1, 2, 6**
- [6] Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. RANGE: Retrieval Augmented Neural Fields for Multi-Resolution Geo-Embeddings. In *CVPR*, 2025. **2**
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. **7**
- [8] Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation. In *CVPR*, 2025. **3**
- [9] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *Transactions on image processing*, 1997. **4**
- [10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *CVPR*, 2023. **1, 2**
- [11] Lukas Haas, Silas Alberti, and Michal Skreta. PIGEON: Predicting Image Geolocations. In *CVPR*, 2024. **3**
- [12] James Hays and Alexei A Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008. **3, 5**
- [13] James Hays and Alexei A Efros. Large-Scale Image Geolocalization. *Multimodal location estimation of videos and images*, 2015. **3**
- [14] Jingliang Hu, Rong Liu, Danfeng Hong, Andrés Camero, Jing Yao, Mathias Schneider, Franz Kurz, Karl Segl, and Xiao Xiang Zhu. MDAS: A New Multimodal Benchmark Dataset for Remote Sensing. *Earth System Science Data*, 15 (1):113–131, 2023. **1, 3, 7**
- [15] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. In *AAAI*, 2025. **1, 2, 5, 6, 8**
- [16] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward Foundation Models for Earth Monitoring. *NeurIPS*, 2023. **1, 2, 5, 6**
- [17] Dan Larkin-York, Google Inc., Koordinates Limited, Mike Playle, and Tiago Brito. S2 Geometry Library. <https://github.com/google/s2geometry>, 2015. [Online; accessed 13-May-2025]. **4**
- [18] Philipp Lindenberger, Paul-Edouard Sarlin, Jan Hosang, Matteo Balice, Marc Pollefeys, Simon Lynen, and Eduard Trulls. Scaling Image Geo-Localization to Continent Level. *NeurIPS*, 2025. **1, 2, 3, 4, 5**
- [19] Zeping Liu, Fan Zhang, Junfeng Jiao, Ni Lao, and Gengchen Mai. GAIR: Improving Multimodal Geo-Foundation Model with Geo-Aligned Implicit Representations. *arXiv:2503.16683*, 2025. **2**
- [20] Yuanhui Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. UniBind: LLM-Augmented Unified and Balanced Representation Space to Bind Them All. In *CVPR*, 2024. **1, 3**
- [21] Qi Ma, Runyi Yang, Bin Ren, Nicu Sebe, Ender Konukoglu, Luc Van Gool, and Danda Pani Paudel. CityLoc: 6DoF Pose Distributional Localization for Text Descriptions in Large-Scale Scenes with Gaussian Representation. *arXiv:2501.08982*, 2025. **3**
- [22] Kevin Mayer, Benjamin Rausch, Marie-Louise Arlt, Gunther Gust, Zhecheng Wang, Dirk Neumann, and Ram Rajagopal. 3d-pv-locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3d. *Applied Energy*, 310:118469, 2022. **5**
- [23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*, 2018. **4**
- [24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024. **5**
- [25] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor Parameter Aware Masked Autoencoder for Multi-Satellite Self-Supervised Pretraining. In *GCPR*, 2024. **5**

- [26] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019. 6
- [27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 7
- [28] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks. In *ICLR*, 2024. 2, 8
- [29] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. TaxaBind: A Unified Embedding Space for Ecological Applications. In *WACV*, 2025. 2
- [30] Bojan Šavrič, Tom Patterson, and Bernhard Jenny. The equal earth map projection. *International Journal of Geographical Information Science*, 33(3):454–465, 2019. 3
- [31] David G Shatwell, Ishan Rajendrakumar Dave, Sirnam Swetha, and Mubarak Shah. GT-Loc: Unifying When and Where in Images Through a Joint Embedding Space. In *ICCV*, 2025. 2
- [32] 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:2508.10104*, 2025. 5, 6
- [33] Mimi Sun, Chaitanya Kamath, Mohit Agarwal, Arbaaz Muslim, Hector Yee, David Schottlander, Shailesh Bavadekar, Niv Efron, Shravya Shetty, and Gautam Prasad. Community search signatures as foundation features for human-centered geospatial modeling. *arXiv:2410.22721*, 2024. 1, 3, 6
- [34] 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. *NeurIPS*, 2020. 3, 8
- [35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv:2502.14786*, 2025. 3, 5, 6
- [36] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favien Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. In *ICML*, 2025. 1, 2
- [37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [38] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In *NeurIPS*, 2023. 1, 2, 3, 5, 6, 8
- [39] Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing Any-Sensor Foundation Models for Earth Observation. In *CVPR Workshops*, 2025. 2, 5
- [40] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2Loc: 3D Point Cloud Localization from Natural Language. In *CVPR*, 2024. 3
- [41] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. *arXiv:2403.15356*, 2024. 5