LoCA: Enabling Multi-Modality in Geospatial Location Encoders

Anonymous authors
Paper under double-blind review

Abstract

Location encoders map geographic coordinates to high-dimensional descriptors through contrastive alignment with paired Earth Observation imagery. We prove that this construction is structurally limited: any Lipschitz location encoder produces an output of intrinsic dimension at most 2, regardless of architecture or bandwidth, which forces the contrastive loss to strictly exceed its asymptotic optimum. Empirically, joint training also collapses the encoder's embedding covariance to extremely low rank. Although dimension-limited, current location encoders sidestep this spectral collapse by training against a frozen image encoder, which we reinterpret as a covariance regularizer that supplies a high-rank target for the location encoder's output to fold onto. However, extending this approach to multimodal inputs is fundamentally challenging, as location encoders should simultaneously map a single location to multiple, potentially incompatible, descriptors. We identify two requirements a regularizer must satisfy for location encoding in multimodal regimes: *co-alignment* across modalities, since adapters alone cannot align independently pretrained encoders; and *high effective rank*, since the location encoder inherits its target's spectrum. We propose **LoCA** (Location Contrastive Alignment), a two-stage pipeline that satisfies both. *Stage 1* jointly trains Sentinel-1 and Sentinel-2 encoders with contrastive learning, producing a co-aligned, high-rank multimodal target. *Stage 2* freezes the image encoders and trains a location and a tabular metadata encoder as projection adapters with multi-view contrastive learning. Against five state-of-the-art encoders on 17 tasks ranging from land cover and wildfire forecasting to geophysical, socio-economic, biodiversity and climate applications, LoCA achieves the most task wins, is the only model to outperform every other on a majority of tasks, and wins 68.8% of all per-task comparisons.

1 Introduction

The rich and heterogeneous information available through Earth Observation (EO) instruments creates massive opportunities for understanding the complex processes governing our planet. Each EO sensor captures distinct aspects of the Earth's surface, from spectral reflectance to atmospheric dynamics, yet no single modality offers a complete representation. Effectively distilling and fusing these diverse sources holds the potential for a universal descriptor of any location on the planet. Such a *location encoder* could propel research across applications with inherent multimodal dependencies, including wildfire forecasting (Prapas et al., 2023), species distribution modeling (Sastry et al., 2024), and land cover classification (Sumbul et al., 2021).

A large body of geospatial applications rely on the persistent geophysical and semantic properties of a location. Recent work on EO foundation models (Tseng et al., 2025; Jakubik et al., 2025; Brown et al., 2025; Wang et al., 2025) aims at encoding such diverse information applicable to arbitrary downstream tasks. However, many such models are trained as functions of imagery: an observation image is required at inference, while the same location admits many vectors depending on which image is supplied. Coordinate-indexed embedding data products published from these models (Brown et al., 2025; Wang et al., 2025) are snapshots of the imagery at release time, not functions on coordinates, and cannot return a vector outside the published grid or the release-time observation window. The desired vector that describes the inherent characteristics of a location needs a function defined on all coordinates, easily queryable from latitude and longitude alone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

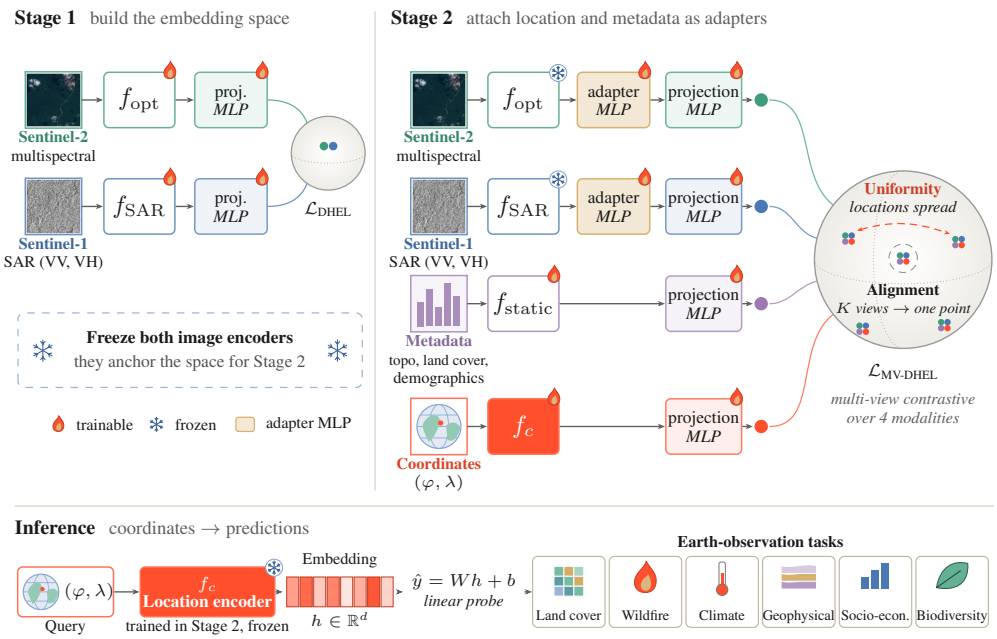

Figure 1: **LoCA pipeline.** ***Stage 1*:** jointly train Sentinel-1 SAR and Sentinel-2 optical encoders with DHEL to produce a co-aligned, high-rank multimodal target. ***Stage 2*:** freeze the image encoders and fit a location and a metadata encoder as adapters against the four-view system. **Inference:** a coordinate query passes through the encoder alone to produce an embedding for downstream tasks.

Contrastive learning is the natural fit for building such a function. The InfoNCE objective (van den Oord et al., 2018; Chen et al., 2020) reduces to categorical cross-entropy over a positive and $K-1$ negatives, the same loss shape as a linear classifier on a frozen representation, and produces globally separable semantic embeddings (Wang & Isola, 2020). Contrastive location encoders, such as SatCLIP (Klemmer et al., 2025), GeoCLIP (Cepeda et al., 2023) and CSP (Mai et al., 2023a), align location embeddings to frozen imagery embeddings with a two-view InfoNCE objective. However, we show that *a location encoder is structurally incompatible with the geometry of contrastive learning*.

First, we note that a Lipschitz map cannot increase intrinsic dimension, so any location encoder (f_c), including the spherical-harmonic and Siren architectures used by SatCLIP, sends S^k onto a set of intrinsic dimension at most k. With $k=2$ for geographic coordinates and embedding dimension $d-1=255$, the location embedding lies on a 2-dimensional surface inside a 255-dimensional sphere, regardless of bandwidth or parameterization. Second, we prove that this is the wrong substrate for contrastive learning: the asymptotic optimum of the InfoNCE objective is the uniform distribution on S^{d-1} (Wang & Isola, 2020), which such a surface cannot represent. Consequently, the contrastive loss strictly exceeds its unconstrained minimum at every aligned Lipschitz pair (Theorem 1), uniformly over SatCLIP-type encoder families (Corollary 1). Empirically, we further observe that joint training exhibits a spectral collapse: alignment drags the image encoder onto the same surface, and the empirical covariance of the **location encoder collapses to low rank**. A 255-dimensional embedding sphere ends up used along only a few linearly independent directions, severely limiting downstream concept separability.

Despite these geometric limitations, existing location encoders remain effective. We argue this is because, in practice, they freeze a pretrained unimodal encoder and thereby avoid the empirical spectral collapse. SatCLIP (Klemmer et al., 2025) and GeoCLIP (Cepeda et al., 2023) keep a pretrained image encoder fixed and only fit the location side, which changes the problem from *learning a representation jointly with another encoder* to *fitting an inverse map from coordinates to a fixed point cloud on the embedding sphere*. The geometric image of f_c remains 2-dimensional, but its empirical sample and spectrum now track those of the pretrained target. The frozen image encoder thus acts as a covariance regularizer, turning f_c into an adapter onto a high-rank cloud the pretrained model already produced.

However, *this paradigm does not extend to multiple sensors*. Two independently pretrained satellite image encoders, a Sentinel-1 SAR model and a Sentinel-2 optical model, **are not co-aligned** (Zavras et al., 2025): nothing in their separate objectives places paired inputs at the same point, so no single location function can fit both. Trainable MLP adapters absorb these per-modality transforms but cannot manufacture co-alignment that isn't already there. Multimodal location encoders therefore need a target that is co-aligned by construction. Motivated by the above limitations, we introduce **LoCA** (Location Contrastive Alignment), a two-stage pipeline determined by two key requirements for the targets: *co-alignment* across modalities, so that the location function fits one constraint per coordinate rather than one per modality; and *high covariance rank*, so that the spectrum the location encoder inherits is not itself collapsed. *Stage 1* enforces co-alignment by jointly training a Sentinel-1 and a Sentinel-2 satellite image encoders with DHEL(Koromilas et al., 2024), the contrastive objective that attains the highest effective rank. *Stage 2* freezes both image encoders, attaches a small projection adapter to each, and trains a location encoder f_c and a metadata encoder f_m (MLP on tabular data) against the four-encoder pipeline with MV-DHEL (Koromilas et al., 2025), the multi-view extension of DHEL. In this way, LoCA produces multimodal location embeddings of high variability across the 2 available intrinsic dimensions. We extensively benchmark LoCA comparing it against state-of-the-art baselines on a diverse benchmark suite covering all relevant semantic factors of variability inherent in the EO domain as defined in Kondylatos et al. (2025). Our contributions can be summarized as follows:

- **C1:** We **prove** that location encoders constrain contrastive learning away from its optimum, and empirically show that **image-location contrastive training from scratch collapses**.
- **C2:** We recast existing contrastive location encoders as adapters over pretrained image representations and argue that extending this formulation to multiple modalities requires (i) **cross-modal co-alignment** and (ii) **high effective rank** of the target space.
- **C3:** We **extend location encoders to utilize multimodal satellite representations and metadata** by introducing **LoCA**, a training pipeline achieving co-aligned, high-rank features by first jointly training Sentinel-1 and Sentinel-2 encoders and then adapting location and metadata encoders.
- **C4:** LoCA achieves the **best performance on more tasks** than state-of-the-art competitors and **beats each examined location encoders** on head-to-head comparisons.

Overall, our work provides a **principled framework for location encoders**: it identifies their failure modes, derives the requirements for extending them to multiple modalities, and uses them to enable multimodality through LoCA, laying the groundwork for future advances in location encoding.

2 Related Work

Geospatial Foundation Models (GFMs) & Embeddings. Several GFMs use Masked Image Modelling (MIM) with various inductive biases: multi-scale or scale-aware encodings Reed et al. (2023); Tang et al. (2023), spectral-aware masking Irvin et al. (2023); Zhu et al. (2023), and sensor-agnostic backbones that absorb heterogeneous EO inputs Tseng et al. (2023); Bountos et al. (2025); Xiong et al. (2024); Wang et al. (2025); Herzog et al. (2026); Forgaard et al. (2026). A second family blends MIM with contrastive or reconstruction objectives across satellite modalities Tseng et al. (2025); Guo et al. (2024); Fuller et al. (2023); Astruc et al. (2024), and a third uses generative or diffusion-based formulations Jakubik et al. (2025); Khanna et al. (2024); Jia et al. (2025). Building on this momentum, recent works release precomputed embedding products at fixed grids: AlphaEarth Brown et al. (2025) fuses optical, radar, LiDAR, and climate reanalysis into 64-dimensional annual embeddings at $10\,\mathrm{m}$, Copernicus-FM Wang et al. (2025) provides $0.25°$ multi-sensor averages, and TESSERA Feng et al. (2026) releases annual Barlow-Twins-pretrained embeddings with *int8* quantization. These products are observation-based rather than location-based: each cell is a snapshot of imagery tied to a specific acquisition window, not a function of coordinates. Recent work positions embeddings generated via vision models as one instantiation of an "Earth embedding", an AI-native representation of a given point in geographic space (and time) Klemmer et al. (2026).

Location Encoders. A second branch of models skips the image encoding step and produces embeddings directly as a function of coordinates. Early designs train MLPs to predict object presence from coordinates Mac Aodha et al. (2019), followed by geodesic-preserving multi-scale spherical encodings Mai et al. (2023b) and the spherical-harmonic / sinusoidal-MLP architecture Rußwurm

et al. (2024) that has since become standard. Contrastive learning is now the dominant regime for self-supervised training: GeoCLIP Cepeda et al. (2023) aligns coordinates with CLIP image embeddings, and SatCLIP Klemmer et al. (2025) aligns them with embeddings stemming from frozen Sentinel-2 encoders to produce general-purpose location embeddings. Multimodal extensions either add new contrastive losses per modality Liu et al. (2026); Shatwell et al. (2025) or align all pairwise combinations of image, location, and time through separate InfoNCE terms Shatwell et al. (2026). These designs share two structural costs: heterogeneous loss terms produce conflicting gradients, and the number of terms grows quadratically with the number of modalities. We address both by generalizing the SatCLIP framework to arbitrary modalities through a single multi-view objective.

Multi-View Contrastive Learning. Multi-view CL has been studied through aggregations of pairwise losses across heads, scales, classes, or feature levels (Wang et al., 2024; Li et al., 2023; Shah et al., 2023; Xu et al., 2022), by treating each view combination as an independent task (Tian et al., 2020), through alternative aggregations (Shidani et al., 2024), and via optimal transport over similarity tensorsPiran et al. (2024). MV-DHEL (Koromilas et al., 2025) is the multi-view objective whose optimum scales with the number of views without dimensional collapse, making it suitable to train our multimodal location encoders.

3 Revisiting the inner workings of Location Encoders

In this section we explore the modeling limitations in Location Encoders. In sections 3.1 and 3.2 we formulate the task of Location Encoding and provide theoritical bounds on the optima of contrastive learning with location encoders and the dimensionality of the produced representations. In section 3.3 we analyse the inner workings of existing location encoders, and reformulate them through the lens of a *coordinate-indexed lookup table*.

3.1 Setup and Notation

Let the location domain be the k-sphere $S^k \subset \mathbb{R}^{k+1}$ with an atomless probability measure μ (for SatCLIP, $k = 2$ and μ is the geographic distribution of training locations on Earth). The image domain $\mathcal{I}$ is an arbitrary measurable space. Pairs (c, I) are drawn from a joint distribution p with each image tied to a unique location, $p(c \mid I) = \delta_{c(I)}$. Two encoders $f_c : S^k \to S^{d-1}$ and $f_I : \mathcal{I} \to S^{d-1}$ share the embedding sphere of dimension $d-1$, with $d-1 > k$ (for SatCLIP, $d-1 = 255$). They minimize the population NT-Xent loss (Wang & Isola, 2020; Koromilas et al., 2024)

$$\mathcal{L}_\tau(f_c, f_I) = -\mathbb{E}_{(c,I)\sim p}\left[\frac{f_c(c)^\top f_I(I)}{\tau}\right] + \tfrac{1}{2}\,\mathbb{E}_c \log \mathbb{E}_{I'}\left[e^{f_c(c)^\top f_I(I')/\tau}\right] + \tfrac{1}{2}\,\mathbb{E}_I \log \mathbb{E}_{c'}\left[e^{f_c(c')^\top f_I(I)/\tau}\right], \tag{1}$$

the population (infinite-batch) limit of the symmetric CLIP loss (Radford et al., 2021). The first term rewards same-pair embeddings being close, while the next two push different-pair embeddings apart.

We use the *pushforward measure* $f_{\#}\mu$ on S^{d-1}, the distribution of $f(X)$ when $X \sim \mu$, so $f_{c\#}\mu$ is the distribution of location embeddings produced by f_c. We will use the ***Hausdorff dimension*** $\dim_H E$, which is related to an intrinsic-dimension notion on smooth manifolds and extends to irregular sets (definition in Section B.1). We write p-a.e. for "p-almost every," meaning except on a set of p-measure zero. Let $\mathcal{L}_\tau^{\text{free}}$ denote the infimum of $\mathcal{L}_\tau$ over all measurable encoder pairs. Among aligned pairs (satisfying $f_c(c) = f_I(I)$ for p-a.e. (c, I)), this infimum is attained when additionally $f_{c\#}\mu = \sigma$, the uniform measure on S^{d-1} (Wang & Isola, 2020). We denote V^* for the value at this aligned-uniform optimum: $V^* \geq \mathcal{L}_\tau^{\text{free}}$ by definition, with the reverse inequality not generally true because of the modality gap (Liang et al., 2022; Cai et al., 2026). We prove a strict gap to V^*, which suffices for a strict gap to $\mathcal{L}_\tau^{\text{free}}$.

3.2 Location encoders constrain contrastive learning away from its optimum

Theorem 1 (Dimensional gap for Lipschitz encoders)**.** *Let $k \geq 1$, $d > 2$ with $k < d-1$. Let $f_c : S^k \to S^{d-1}$ be Lipschitz and $f_I : \mathcal{I} \to S^{d-1}$ measurable. Set $\nu := f_{c\#}\mu$. Then*

(i) $\dim_H f_c(S^k) \leq k$*. The location encoder maps S^k onto a set of intrinsic dimension at most k, regardless of how f_c is parameterized or how high the embedding dimension d is.*

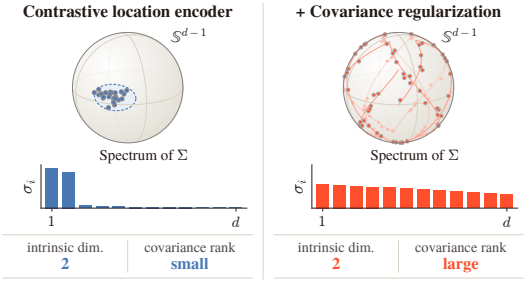

Figure 2: **Intrinsic dimension is bounded; the covariance is not.** Plain contrastive learning collapses the embedding to a 2-d patch with low variability; covariance regularization makes the 2-d manifold wind densely through S^{d-1}, raising the covariance rank without changing the intrinsic dimension.

Table 1: Effective rank Roy & Vetterli (2007) of the location-embedding marginal and its ratio to the total representation dimensionality D. Higher = more directions used. "Anchor" indicates the pretrained target whose spectrum the encoder starts from. Blue indicates training on Sentinel-2, Purple on RGB and Green on both Sentinel-1&2.

Method	**Anchor**	**Erank**	**% of D**
SatCLIP†	No	8	3.1
SatCLIP	MoCo R50	89	34.9
Copernicus-FM	DINOv2	158	20.6
LoCA (ours)	Stage 1	158	62

†image encoder trained from scratch (no MoCo init).

(ii) *$\nu \perp \sigma$, in the sense that ν and σ live on disjoint Borel subsets of S^{d-1}. The location-embedding distribution and the uniform target on the sphere have zero overlap.*

(iii) *If (f_c, f_I) is aligned, $\mathcal{L}_\tau(f_c, f_I) > \mathcal{L}_\tau^{\text{free}}$. The contrastive loss strictly exceeds its unconstrained minimum at every aligned Lipschitz pair.*

The corresponding proof is provided in Section B.3. From a high level, we prove that no aligned pair with a Lipschitz location encoder attains the optimum of contrastive learning. Part (i) states that a Lipschitz map **cannot increase intrinsic dimension**. Part (ii) shows that the location embeddings and the uniform target (the optimum of contrastive learning) live on disjoint pieces of the sphere with zero overlap. The consequence is that (iii) under alignment, the strict mismatch between ν and σ forces the loss strictly above its unconstrained minimum.

The above findings have direct consequences for existing location encoders. SatCLIP composes a spherical-harmonic feature map $\text{SH}_L : S^2 \to \mathbb{R}^{(L+1)^2}$ of degree L, a Siren MLP, and a normalization onto S^{d-1} (Klemmer et al., 2025; Sitzmann et al., 2020). We write $\mathcal{F}_L^{\text{SC}}$ for the resulting bandwidth-L family under a mild non-degeneracy condition (definitions in Section B.4). Each piece of the composition is Lipschitz with a constant independent of the trainable parameters, so every $f_c \in \mathcal{F}_L^{\text{SC}}$ is Lipschitz with a constant uniform across the family (Lemma 2), and Theorem 1 applies pointwise.

Corollary 1 (SatCLIP strict positivity over aligned pairs)**.** *Fix $d \geq 4$. For every finite $L \in \mathbb{N}$ and every admissible parameter set, the contrastive loss of any aligned SatCLIP pair is strictly larger than the unconstrained free optimum: even when minimizing jointly over the entire family $\mathcal{F}_L^{\text{SC}}$ and over all aligned image encoders, the infimum stays strictly above $\mathcal{L}_\tau^{\text{free}}$.*

These results extend from a single Lipschitz encoder to the infimum over the entire bandwidth-L class, so no parameter setting reaches the free optimum. See Section B.5 for the proof.

3.3 Geometry, variability collapse, and the role of regularization

Theorem 1(i) states that the image of a location encoder has intrinsic dimension at most k inside S^{d-1}, regardless of architecture or bandwidth. This is intuitively expected: standard encoders compress high-dimensional inputs into a lower-dimensional code; a location encoder does the opposite, decoding a low-dimensional coordinate into a higher-dimensional vector. As a result, *location encoders constitute the wrong substrate for contrastive learning* as they can never attain the point configurations that the training objective optimizes for. Theorem 1 constrains the geometry of the embedding, not its spectrum: a k-dimensional surface can still have full-rank covariance (Figure 2). Under Equation (1), alignment forces the essential range of f_I to lie in $f_c(S^k)$ (Section B.3), so both encoders inherit the location side's intrinsic dimension. Empirically, joint training also settles on a surface of low covariance rank (Table 1): a 255d embedding space is used along only a small number of linearly independent directions.

A natural fix is to break the symmetry of joint training by freezing one side, creating a setup in which contrastive objectives can work. Freezing f_I regularizes f_c: instead of co-collapsing with another encoder, f_c is fitted against a fixed, externally produced point cloud $\{y_i := f_I(I_i)\}_{i \leq N}$ on S^{d-1}. The geometric image $f_c(S^k)$ remains k-dimensional by Theorem 1, but the empirical sample $\{f_c(c_i)\}_{i \leq N}$ now tracks $\{y_i\}$, and its spectrum tracks the spectrum of the target cloud. The function thus memorizes high-rank labels by oscillating along its k-dimensional surface. Geometry and spectrum decouple (Figure 2) since the manifold stays low-dimensional while the empirical covariance reaches whatever rank the target supplies.

This is *why* existing contrastive location encoders work. SatCLIP (Klemmer et al., 2025) and GeoCLIP (Cepeda et al., 2023) are not performing contrastive learning in the usual sense of jointly learning a representation with another encoder; they instead use the contrastive objective as a *proxy for fitting a coordinate-indexed lookup table* whose values come from a pretrained image encoder acting as a regularizer that provides f_c a high-rank target to fold onto. This adjustment to the image modality is also effective in other domains through the Locked-image Tuning (Zhai et al., 2022).

4 Adapting Location Encoders to Multimodal Embeddings

In Section 3.3 the location encoder is reframed as an adapter that indexes a pretrained label cloud. We now ask: *what must a label cloud satisfy for this view to extend to multiple sensors?* We identify two requirements in Section 4.1 and Section 4.2 and introduce LoCA that satisfies both in Section 4.3.

4.1 A location encoder cannot align to two targets

Extending the *coordinate-indexed lookup table* paradigm to multiple modalities is far from trivial, since the targets it supplies must be mutually consistent across modalities. Formally, suppose we freeze a Sentinel-1 encoder f_{I_1} and a Sentinel-2 encoder f_{I_2}, both pretrained independently, and demand a single f_c to align with both. Alignment requires $f_c(c_i) = f_{I_1}(I_i^{S_1}) =: y_i^{S_1}$ and $f_c(c_i) = f_{I_2}(I_i^{S_2}) =: y_i^{S_2}$ on every paired sample, which forces $y_i^{S_1} = y_i^{S_2}$. Independently trained encoders do not satisfy this: their embedding spaces are determined up to an arbitrary orthogonal transformation by their respective objectives and initializations, and there is no mechanism by which paired inputs would land at the same point. The location encoder is presented with two incompatible label sets and cannot fit either. Inserting trainable projection adapters between f_c and each frozen target relaxes this rigidity by absorbing per-modality orthogonal transformations. However, adapters cannot manufacture co-alignment that is absent from the joint distribution of the targets themselves: no choice of adapters can pull $y_i^{S_1}$ and $y_i^{S_2}$ to the same point unless the underlying f_{I_1}, f_{I_2} already agree up to such a transformation. This is a structural limitation and extends to any setup combining M not co-trained encoders. Adapting to multiple modalities therefore requires a target that is ***co-aligned by construction***: paired-sample labels for distinct modalities must coincide on S^{d-1}, so that f_c is asked to satisfy one constraint per coordinate rather than M.

4.2 Regularizer embeddings should have high rank

Through the regularizer view of Section 3.3 the number of linearly independent directions in the representation space of f_c is bounded to that of the target. Alignment forces $f_c(c_i) \approx y_i$, so a low-rank target propagates its collapse to the location encoder. This rules out a natural choice of target *i.e.*, multimodal GFMs, primarily relying on MIM objectives (Section 2). MIM objectives concentrate variance on a small number of directions, a pathology documented for self-supervised methods broadly (Jing et al., 2022) and for GFMs specifically (Hackel et al., 2026; Robinson & Corley, 2026; Rao et al., 2026). Using such a model as a pretrained anchor would serve f_c a low-rank label cloud and reproduce the very collapse the adapter is supposed to avoid. We call this the ***rank requirement***. This can be met in two ways: a) guide the representation space with external, high-rank anchors *e.g.* Copernicus-FM distills from Dinov2, or b) utilize optimization processes that enforce a high-rank. The first introduces unnecessary axes of variation potentially reflecting exogenous features (*e.g.* from RGB natural imagery) which may limit its utility to domain-relevant applications (see section 5.1), while the latter creates a label cloud for the target modalities themselves.

Table 2: **Downstream evaluation.** Mean over 10 seeds (3 for SatBird); metric $\times 100$, higher is better. **Green**: best across both probes; **blue**: best within probe (MLP, Linear) when not best overall. Environmental tasks predict relative ordering of climate variables. Bottom two rows: head-to-head comparisons (Table 11). N/A: embeddings not retrievable (Section E).

Task ↓ Model →	Observation-based FMs				Location encoders								Metric
	AlphaEarth		Copernicus FM		SatCLIP		GeoCLIP		GT-Loc		LoCA (Ours)		
	MLP	Linear	MLP	Linear	MLP	Linear	MLP	Linear	MLP	Linear	MLP	Linear	
Biodiversity & Ecology													
Countries	N/A	N/A	80.9	88.3	95.5	96.5	91.5	79.8	94.7	94.4	**96.4**	**97.0**	$F1_w$
Biome	N/A	N/A	84.0	82.4	89.8	86.6	86.9	75.6	88.1	81.9	**91.0**	**87.0**	$F1_w$
Ecoregions	N/A	N/A	59.9	75.1	79.7	84.4	60.6	42.6	69.6	75.9	**80.2**	**85.0**	$F1_w$
iNaturalist	49.1	59.4	40.4	61.8	**56.4**	69.8	44.8	58.4	49.3	59.3	55.0	**71.4**	*Acc.*
SatBird (USA summer)	**67.5**	56.5	59.6	57.9	60.7	58.2	63.3	47.8	63.3	52.7	60.8	**61.1**	*top-k Acc.*
Land cover													
reBEN	42.8	43.3	57.4	**55.2**	57.3	54.9	56.0	44.1	56.4	52.4	**58.3**	54.7	$F1_\mu$
Wildfire forecasting													
Wildfire forecasting	**80.5**	70.1	74.9	71.0	79.1	75.8	79.8	73.2	79.6	75.5	80.2	**76.2**	*AUPRC*
Geophysical & Socio-economic													
Median income	28.1	11.0	32.9	26.5	50.2	35.2	45.0	12.1	45.9	31.9	**51.9**	**39.4**	R^2
Population	80.2	69.6	77.0	67.9	**82.1**	67.9	78.1	66.0	78.5	**69.7**	81.8	68.2	R^2
Cali. housing	54.9	31.1	43.3	39.2	63.3	35.3	77.3	48.4	**78.3**	**66.1**	65.5	36.4	R^2
Air temperature	**98.2**	42.8	90.6	81.0	93.9	80.1	94.3	32.9	94.2	60.4	90.3	**81.4**	R^2
Elevation	**97.9**	87.4	94.1	**90.6**	89.9	64.9	83.0	56.0	83.9	62.5	89.4	64.4	R^2
Environmental	**LightGBM ranking**												
Temperature	N/A		**79.8**		59.6		72.0		74.4		77.9		*Ranking acc.*
Precipitation	N/A		**77.7**		70.6		72.1		72.5		**77.7**		*Ranking acc.*
MSLP	N/A		73.2		**83.2**		77.8		79.1		76.3		*Ranking acc.*
Wind speed	N/A		**75.6**		71.6		73.8		71.4		**75.6**		*Ranking acc.*
Relative humidity	N/A		53.5		**65.2**		47.8		43.5		47.5		*Ranking acc.*
# Best (out of 17)	4		3		3		0		1		**8**		
# Top-2 (out of 17)	4		6		8		3		3		**11**		
LoCA wins vs. model	**5/9**		**11/17**		**12/17**		**12/17**		**13/17**		–		
Pairwise win rate (%)	57.8		35.1		55.8		36.4		45.5		**68.8**		

4.3 LoCA: Location Contrastive Alignment

We introduce LoCA (Location Contrastive Alignment), a two-stage pipeline that addresses all aforementioned limitations. The full pipeline is illustrated in Figure 1 and detailed below.

***Stage 1*: Jointly Contrastive Multimodal Pretraining.** The first stage involves training information-rich Sentinel-1 and Sentinel-2 encoders jointly with the DHEL loss, which enforces the representation space to utilize a big fraction of the available dimensionality (Koromilas et al., 2024). Furthermore, joint contrastive training enforces co-alignment as a property of the loss optimum, since paired inputs are pulled to the same embedding by construction. These two properties result in a co-aligned, high-rank georeferenced label cloud $\{y_i^{\text{multi}}\}$, which can be effectively used as a target.

***Stage 2*: Multi-View Contrastive Adapter Fitting.** In this stage we freeze both image encoders, letting the location encoder adapt to a fixed multimodal embedding space. We attach a trainable MLP projection to each uni-modal encoder, and introduce a location encoder f_c and a metadata encoder f_m (a standard MLP that processes all 14 variables as a single feature vector). Notably, Theorem 1 applies to f_m as well with $k = 14$, restricting its usage as a *Stage 1* anchor. We train this four-encoder system with the MV-DHEL (Koromilas et al., 2025) loss. MV-DHEL is a multi-view extension of DHEL whose optimum scales with the number of views without dimensional collapse. Because both image encoders are frozen, their embeddings remain co-aligned and the target space retains its high effective rank. As a result, the empirical spectrum of f_c at training reaches the projection-head dimension, even though $f_c(S^2)$ is still a 2-dimensional manifold by Theorem 1.

5 Empirical Evaluation & Discussion

Experimental setup. As detailed in Section C we extend the S2-100K Klemmer et al. (2025) (SatCLIP train set) to align co-located Sentinel-1 (S1) imagery with the original Sentinel-2 (S2) acquisitions and augment them with 14 metadata variables, including elevation, land cover, population density and others. We utilize ResNet-50 backbones for all LoCA image encoders. The location encoder is defined with a degree-$L = 40$ spherical harmonic feature map followed by a 2-hidden-layer

Table 3: Isolating the contributions of (i) the Stage 1 and (ii) the metadata encoder. We report the change in mean linear probe score per task family with respect to the full model. See also Table 13.

Removed component	Biodiv. & Ecology	Land cover	Wildfire	Geophysical	Environmental
Stage 1	−56.9	−53.6	−26.8	−61.3	−21.1
Metadata encoder	−0.1	+0.5	+0.3	−3.5	−1.3

SIREN, while the metadata encoder is a 3-layer MLP. All modalities share a 2-layer projection head. We train both stages for 100 epochs each. Full implementation details are provided in Section D.

Benchmark. We curate a diverse benchmark by covering all Semantic Factors (SF) of variability inherent to the EO domain, as defined in (Kondylatos et al., 2025), excluding the spatial-arrangement SF, which is not relevant to general-purpose location embeddings. Ground sampling distances range from a few meters to multiple kilometers per pixel; target granularities span fine-grained species identification to coarse biome classification; and geographic and thematic domains include Mediterranean wildfire forecasting and bird encounter rate prediction across the USA. Our tasks cover five categories of location-dependent attributes: (i) land cover from reBEN (Clasen et al., 2025); (ii) wildfire danger from Mesogeos (Kondylatos et al., 2023); (iii) long-term climate from five ranking tasks over temporal means of variables from SeasFire (Karasante et al., 2025); (iv) geophysical and socioeconomic structure and (v) biodiversity, both from the SatCLIP evaluation (Klemmer et al., 2025), the latter complemented by species distribution from SatBird (Teng et al., 2023).

We evaluate downstream performance on frozen embeddings in two ways. (i) For the environmental tasks, we train a LambdaRank objective (Burges et al., 2007) as implemented in LightGBM (Ke et al., 2017) to test whether the representations contain enough information to recover the relative ordering of long-term mean climate values. (ii) For all remaining downstream tasks, we train both a linear probe and an MLP, with identical architecture and hyperparameters across tasks. We compare against two families of baselines: location encoders that are themselves coordinate functions, *i.e.* SatCLIP (Klemmer et al., 2025), GeoCLIP (Cepeda et al., 2023), and GT-Loc (Shatwell et al., 2025); and precomputed embedding fields released by geospatial foundation models, namely AlphaEarth (Brown et al., 2025) and Copernicus-FM (Wang et al., 2025). Further details are provided in Section E.

5.1 Downstream performance of location encoders

Across the 17 tasks in Table 2, LoCA ranks first on 8 and among the top-2 models on 11. In pairwise comparisons, LoCA outperforms each competing model on the majority of tasks winning in 68.8% of the cases (see Table 11). Under linear probing, LoCA exhibits the best performance on 8 out of 12 tasks, demonstrating that *a wide spectrum of location-dependent information is readily available in its embeddings*, consistent with the motivation of Section 4 and the findings of Table 1. LoCA excels in ***Land cover*** and ***Biodiversity & Ecology*** tasks, leading on reBEN, Countries, iNaturalist, Biome, and Ecoregions. AlphaEarth leads on SatBird and ***Wildfire forecasting***, with LoCA a close second on the latter (-0.3%). ***Geophysical & Socio-economic*** results are mixed: LoCA leads on Median Income, SatCLIP on Population (with LoCA following closely), and AlphaEarth on Air Temperature and Elevation. GT-Loc and GeoCLIP stand out on California housing, despite each achieving top-2 performance in only 3 tasks overall. This may reflect their pretraining dataset (MP-16 (Larson et al., 2017)), a global Flickr image dataset, with substantial coverage of California. Its images of streets, houses, and indoor scenes may provide useful priors for house-price estimation. For ***Environmental*** tasks, SatCLIP leads on MSLP and Relative Humidity, while Copernicus-FM leads on Temperature, followed by LoCA, and ties with LoCA on Precipitation and Wind Speed. However, Copernicus-FM requires multimodal imagery for its embeddings, whereas LoCA uses coordinates alone. Among coordinate-based encoders, LoCA is best on Temperature, Precipitation, and Wind speed.

5.2 LoCA increases embeddings variability

In this section we empirically verify that LoCA satisfies the two requirements identified in Section 4.
Co-alignment. We measure alignment using cosine similarity ($\cos$) and L_{align}, the expected distance between pairs, from (Wang & Isola, 2020). The two imagery encoders trained jointly in *Stage 1* exhibit $\cos = 0.94$ and $L_{\text{align}} = 0.12$, while pairing two independently pretrained encoders for the

Table 4: Mean F1 on 10 single-label land-cover classes for 4 supervised encoders, before/after LoCA's *Stage 2*. "S1+S2": trained on both modalities; "All": concatenated embeddings.

	Location		**S1**		**S2**		**S1+S2**		**All**	
	base	*Stage 2*	*base*	*Stage 2*	*base*	*Stage 2*	*base*	*Stage 2*	*base*	*Stage 2*
Mean F1	74.7	**76.6**	69.6	**75.9**	78.4	**80.9**	78.1	**80.8**	80.9	**84.3**

same modalities yields $L_{\text{align}} = 1.00$ ($\cos = 0.50$), confirming that their embedding spaces are not co-aligned. This misalignment translates to decreased downstream performance, as shown in Table 3.

High effective rank. Effective rank (Roy & Vetterli, 2007) measures how many linearly independent directions an embedding populates and is a standard proxy for linear-probe quality (**?**). Across 343,928 land locations on a $0.25°$ global grid (Table 1), LoCA populates 158 of its 255 dimensions (62%), three times the fraction of Copernicus-FM (21%) and $1.8\times$ that of SatCLIP (35%). Joint training from scratch instead collapses the covariance to an effective rank of 8 (SatCLIP†). The effective rank of GT-Loc and GeoCLIP is misleading for Earth Observation, since their CLIP-derived encoders inflate it with irrelevant language information, as discussed in Section 4.2 and validated in Table 2, while that of AlphaEarth is unavailable, as its embeddings were only partially retrievable.

5.3 Ablations on LoCA components

Impact of *Stage 1*. Section 4 identifies *Stage 1* as the construction that properly supplies the *Stage 2* target. This is validated in Table 3, where by replacing the jointly-trained S1 and S2 encoders with two encoders pretrained independently with MoCo violates the co-alignment requirement collapsing performance across the entire benchmark confirming the necessity of *Stage 1*. This validates that *SatCLIP cannot be trivially extended to multiple modalities* by freezing two independent encoders.
Impact of *Stage 2*. Encoders trained in a supervised fashion on the same classes converge to a similar embedding structure (Koromilas et al., 2026), satisfying the goal of *Stage 1*. This enables us to isolate the impact of *Stage 2*, by replacing *Stage 1* with three supervised encoders (one for each of S1, S2, and stacked S1+S2) trained on BigEarthNet (Sumbul et al., 2021). We freeze these encoders alongside a pretrained SatCLIP location encoder, and train only their adapters and shared projections under MV-DHEL (Koromilas et al., 2025) (see Section E.3.2). *Stage 2* improves every encoder's linear probing performance (see Table 4). Concatenating the adapted embeddings raises mean F1 from 80.9% to 84.3% further demonstrating the benefit of *Stage 2* beyond simple multimodal fusion.
Impact of the metadata encoder. We assess the contribution of the metadata encoder by removing it from *Stage 2* (see Table 3). The metadata encoder primarily benefits tasks aligned with its geophysical and slow-climate variables: removing it reduces mean scores by 3.5% and 1.3% for Geophysical & Socio-economic and Environmental tasks respectively. Its effect in other task groups is small, with a 0.1% drop in Biodiversity and minor gains in Land Cover and Wildfire tasks when omitted, despite the inclusion of coarse land cover fractions as metadata. We hypothesize that these coarse descriptors add little of the fine-grained signal required for lower hierarchy multi-label classification.

6 Conclusion

In this work, we established the theoretical foundations for multimodal location encoding. We proved fundamental limitations of jointly trained location encoders and identified two requirements for effective multimodal representations: co-alignment and high effective rank. Building on this theoretical ground, we introduced LoCA, a multimodal location encoder that satisfies both and can scale to any number of modalities. LoCA outperfoms competing models across a broad range of critical EO tasks , while substantially improving embedding space utilization. Our work opens promising research directions including further expanding geospatial location encoders with temporal information, meteorological variables, and a broader array of land cover metadata.

Limitations. LoCA has no temporal dimension, so seasonal and shorter-term dynamics are captured only indirectly. Moreover, *Stage 1* utilizes only S1 and S2. Expanding to other sensors and climate variables would add complementary structure, given that the framework can absorb any amount modalities without modification. Finally, coordinate-based benchmarks are sensitive to spatial autocorrelation. Blocked or leave-region-out evaluations are important future research directions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with creating and modifying scientific figures, brainstorming on engineering choices, identifying relevant literature for the paper and for the mathematics associated with the proofs, creating tables, and minimally polishing the writing to improve readability. For the mathematical proofs, we used AI for generic brainstorming and to identify possible errors and flaws, including through numerical tests. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, or interpreting results. We have reviewed all AI-assisted work: figures were drafted with AI assistance and revised by the authors; tables were verified against source results; all mathematical proofs were constructed by the authors; all text was written by the authors, and any AI-polished passages were revised by them; and the authors read the literature suggested by AI. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical results state their assumptions, including the Lipschitz condition, spherical coordinate domain, embedding sphere, and alignment condition; theorem statements and intuition are given in the main paper, with full proofs in Section B. We describe the dataset construction, downstream task splits, evaluation metrics, and downstream training hyperparameters in Sections C and E. Full implementation details for LoCA pretraining, including the image encoder architectures, optimizer settings, batch sizes, number of epochs, and projection dimensions, are provided in Section D. Section E also reports the embedding dimensions and the number of seeds used per downstream task: 10 seeds for all tasks except SatBird, which uses 3; the mean and standard deviation reported in Table 10 reflect variability from downstream predictor initialization and stochastic optimization under fixed embeddings and fixed data splits. Section D additionally specifies the compute resources used for pretraining, including GPU/CPU type, memory, and runtime. Our codebase builds on two publicly available repositories, referenced in Section D; at submission time we are not releasing an anonymized code repository or complete reproduction scripts, but plan to make the code publicly available in a future version of this work. We further plan to release our training set which is an extension of the S2-100K Klemmer et al. (2025) dataset.

REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

M. Bondarenko, R. Priyatikanto, N. Tejedor-Garavito, W. Zhang, T. McKeen, A. Cunningham, T. Woods, J. Hilton, D. Cihan, B. Nosatiuk, T. Brinkhoff, A. Tatem, and A. Sorichetta. The spatial distribution of population in 2015–2030 at a resolution of 30 arc-seconds, approximately 1 km at the equator: R2025A version v1, 2025. URL `https://hub.worldpop.org/geodata/listing?id=137`. Global mosaics; 1 km resolution; years used: 2021, 2022, and 2023. Global Demographic Data Project, funded by the Bill and Melinda Gates Foundation (INV-045237). Accessed: 2026-05-04.

Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. Fomo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27858–27868, 2025.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan,

Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Christopher J. Burges, Robert Ragno, and Quoc V. Le. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

Yichao Cai, Zhen Zhang, Yuhang Liu, and Javen Qinfeng Shi. The geometric mechanics of contrastive representation learning: Alignment potentials, entropic dispersion, and cross-modal divergence. *arXiv preprint arXiv:2601.19597*, 2026.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=I18BXotQ7j.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Kai Norman Clasen, Leonard W. Hackel, Tom Burgert, Gencer Sumbul, Begum Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *2025 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1264–1268, 2025. URL https://api.semanticscholar.org/CorpusID:271039888.

Copernicus Climate Change Service, Climate Data Store. Land cover classification gridded maps from 1992 to present derived from satellite observations, 2019. LCCS land-cover class maps at approximately 300 m resolution; years used: 2021, 2022, and 2023; downloaded from the Copernicus Data Space Ecosystem S3 endpoint. Accessed: 2026-05-04.

Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, Matt Hansen, Harvey Locke, Erle C. Ellis, Benjamin Jones, Charles Victor Barber, Randy Hayes, Cyril Kormos, Vance Martin, Eileen Crist, Wes Sechrest, Lori Price, Jonathan E. M. Baillie, Don Weeden, Kierán Suckling, Crystal Davis, Nigel Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, José C. Brito, Othman A. Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, Jonathan Timberlake, Heinz Klöser, Yara Shennan-Farpón, Roeland Kindt, Jens-Peter Barnekow Lillesø, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, and Muhammad Saleem. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6):534–545, 2017. doi: 10.1093/biosci/bix014.

European Space Agency. Copernicus DEM – global and european digital elevation model (GLO-30), 2022. URL https://doi.org/10.5270/ESA-c5d3d65. Accessed: 2026-25-09.

Kenneth J. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2nd edition, 2003.

Herbert Federer. *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag, New York, 1969.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2nd edition, 1999.

Theodor Forgaard, Jarle H Reksten, Anders U Waldeland, Valerio Marsocci, Nicolas Longépé, Michael Kampffmeyer, and Arnt-Børre Salberg. Thor: A versatile foundation model for earth observation climate and society applications. *arXiv preprint arXiv:2601.16011*, 2026.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27662–27673, 2024. doi: 10.1109/cvpr52733.2024.02613.

Leonard Hackel, Tom Burgert, and Begüm Demir. How much of a model do we need? Redundancy and slimmability in remote sensing foundation models. *arXiv preprint arXiv:2601.22841*, 2026.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34806–34817, 2026.

Josh Hooker, Gregory Duveiller, and Alessandro Cescatti. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Scientific Data*, 5(1):180246, 2018. doi: 10.1038/sdata.2018.246.

Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y Ng. Usat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199*, 2023.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7383–7394, 2025.

Junteng Jia and Austin R. Benson. Residual correlation in graph neural network regression. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 588–598. Association for Computing Machinery, 2020. doi: 10.1145/3394486.3403101.

Yuru Jia, Valerio Marsocci, Ziyang Gong, Xue Yang, Maarten Vergauwen, and Andrea Nascetti. Can generative geospatial diffusion models excel as discriminative geospatial foundation models? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8429–8440, 2025.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Ilektra Karasante, Lazaro Alonso, Ioannis Prapas, Akanksha Ahuja, Nuno Carvalhais, and Ioannis Papoutsis. Seasfire cube: A multivariate dataset for global wildfire modeling. *Scientific Data*, 12 (1):368, 2025. doi: 10.1038/s41597-025-04546-3.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, 1995.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

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.

Konstantin Klemmer, Esther Rolf, Marc Rußwurm, Gustau Camps-Valls, Mikolaj Czerkawski, Stefano Ermon, Alistair Francis, Nathan Jacobs, Hannah Kerner, Lester Mackey, Gengchen Mai, Oisin Mac Aodha, Markus Reichstein, Caleb Robinson, David Rolnick, Evan Shelhamer, Vincent Sitzmann, Devis Tuia, and Xiaoxiang Zhu. Earth embeddings: Toward artificial intelligence-centric representations of our planet. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–15, 2026. doi: 10.1109/MGRS.2026.3710416.

Spyros Kondylatos, Ioannis Prapas, Gustau Camps-Valls, and Ioannis Papoutsis. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the mediterranean. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=VH1vxapUTs.

Spyros Kondylatos, Nikolaos Ioannis Bountos, Dimitrios Michail, Xiao Xiang Zhu, Gustau Camps-Valls, and Ioannis Papoutsis. On the generalization of representation uncertainty in earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6552–6562, 2025.

Panagiotis Koromilas, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Bridging mini-batch and asymptotic analysis in contrastive learning: From InfoNCE to kernel-based losses. In *Proceedings of the International Conference on Machine Learning*, 2024.

Panagiotis Koromilas, Efthymios Georgiou, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis A Nicolaou, and Yannis Panagakis. A principled framework for multi-view contrastive learning. *arXiv preprint arXiv:2507.06979*, 2025.

Panagiotis Koromilas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Neural collapse by design: Learning class prototypes on the hypersphere. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VN8RbsdxYs.

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.

Jinxing Li, Beicheng Liang, Xiangwei Lu, Mu Li, Guangming Lu, and Yong Xu. From global to local: Multi-patch and multi-scale contrastive similarity learning for unsupervised defocus blur detection. *IEEE Transactions on Image Processing*, 32:1158–1169, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 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.

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, 2019. URL https://api.semanticscholar.org/CorpusID:186206739.

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.),

Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 23498–23515. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/mai23a.html.

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, 2023b. doi: 10.1016/j.isprsjprs.2023.06.016.

OpenStreetMap contributors. Openstreetmap planet dump, 2023. URL https://www.openstreetmap.org. Map data copyrighted OpenStreetMap contributors and available under the Open Database License.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. URL https://ideas.repec.org/a/eee/stapro/v33y1997i3p291-297.html.

Zoe Piran, Michal Klein, James Thornton, and Marco Cuturi. Contrasting multiple representations with the multi-marginal matching gap. In *International Conference on Machine Learning*, volume 235. PMLR, 2024.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Ioannis Prapas, Nikolaos-Ioannis Bountos, Spyros Kondylatos, Dimitrios Michail, Gustau Camps-Valls, and Ioannis Papoutsis. Televit: Teleconnection-driven transformers improve subseasonal to seasonal wildfire forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3754–3759, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Arjun Rao, Marc Rußwurm, Konstantin Klemmer, and Esther Rolf. Measuring the intrinsic dimension of earth representations. In *International Conference on Learning Representations*, 2026.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076. IEEE, 2023.

Caleb Robinson and Isaac Corley. Compressing Earth embeddings, 2026. URL https://geospatialml.com/posts/compressing-earth-embeddings/. Blog post, accessed 2026-04-21.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

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

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7121–7130. IEEE, 2024.

Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M de Melo, and Rama Chellappa. Multi-view action recognition using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3381–3391, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2025.

David G Shatwell, Sirnam Swetha, and Mubarak Shah. Tiger: A unified framework for time, images and geo-location retrieval. *arXiv preprint arXiv:2603.24749*, 2026. To appear at CVPR 2026.

Amitis Shidani, R Devon Hjelm, Jason Ramapuram, Russell Webb, Eeshan Gunesh Dhekane, and Dan Busbridge. Poly-view contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=iHcTLIor0m.

Vincent Sitzmann, Julien N.P̃. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, 2020.

Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep learning with geospatial data. *ACM Transactions on Spatial Algorithms and Systems*, 11(4):1–28, 2025. doi: 10.1145/3707459. Originally released as arXiv:2111.08872.

Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.

Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023.

Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75925–75950. Curran Associates, Inc., 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 60280–60300. PMLR, 2025.

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

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Lei Wang, Piotr Koniusz, Tom Gedeon, and Liang Zheng. Adaptive multi-head contrastive learning. In *European Conference on Computer Vision*, pp. 404–421. Springer, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9888–9899, 2025.

D. Woods, T. McKeen, A. Cunningham, R. Priyatikanto, A. Sorichetta, A. J. Tatem, and M. Bondarenko. Worldpop high resolution, harmonised annual global geospatial covariates: Distance to OSM major roads 2023, 2025. URL `https://hub.worldpop.org/geodata/listing?id=125`. Layer derived from OpenStreetMap highway data. Accessed: 2026-05-04.

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 preprint arXiv:2403.15356*, 2024.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16030–16039, 2022.

Angelos Zavras, Dimitrios Michail, Begüm Demir, and Ioannis Papoutsis. Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:270–287, 2025.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2022. doi: 10.1109/CVPR52688.2022.01759.

Lingxuan Zhu, Jiaji Wu, Biao Wang, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7):3728, 2023.

A Appendix

B Proofs

B.1 Preliminaries

We collect the metric and measure-theoretic notions used in the proof. Standard references are (Falconer, 2003, Chapter 2) for Hausdorff measure and dimension, and (Kechris, 1995, Chapter 17) for standard Borel spaces and Borel isomorphisms.

Diameter and Lipschitz maps. For a subset U of a metric space (X, ρ), the diameter of U is

$$\operatorname{diam} U \;:=\; \sup\{\rho(x, y) : x, y \in U\}. \tag{2}$$

Diameter is the largest distance between two points of U. A ball of radius r has diameter $2r$; a line segment of length L has diameter L. A map $f : (X, \rho_X) \to (Y, \rho_Y)$ is L-Lipschitz if $\rho_Y(f(x), f(x')) \le L\, \rho_X(x, x')$ for all $x, x' \in X$. As a consequence,

$$\operatorname{diam} f(U) \;\le\; L \operatorname{diam} U \quad \text{for every } U \subset X, \tag{3}$$

i.e. Lipschitz maps cannot increase diameter by more than the factor L.

Hausdorff measure. For $E \subset X$ and $s \ge 0$, the s-dimensional Hausdorff (outer) measure is

$$\mathcal{H}^s(E) \;:=\; \lim_{\delta \to 0^+} \inf\Big\{ \sum_i (\operatorname{diam} U_i)^s \;:\; E \subset \bigcup_i U_i,\ \operatorname{diam} U_i < \delta \Big\}. \tag{4}$$

The construction is: cover E by small sets, raise each diameter to the s-th power, sum the contributions, and take the tightest such cover. The exponent s controls how the sum scales when we dilate the cover. At $s = 1$, doubling each diameter doubles the sum, which mirrors how length scales. At $s = 2$, doubling each diameter quadruples the sum, mirroring area. At $s = 3$, the sum scales as volume. Up to constant factors that depend on the convention for diam, $\mathcal{H}^s$ recovers the standard length, area,

and volume in those dimensions (Falconer, 2003, Section 2.1). The limit $\delta \to 0$ forces the cover to consist of arbitrarily small sets, which is what makes the construction sensitive to fine geometric structure rather than coarse extent.

Hausdorff dimension. The function $s \mapsto \mathcal{H}^s(E)$ is non-increasing in s, and for any E there is a unique critical value at which it jumps from ∞ to 0:

$$\dim_H E \;:=\; \sup\{s \geq 0 : \mathcal{H}^s(E) = \infty\} \;=\; \inf\{s \geq 0 : \mathcal{H}^s(E) = 0\}. \tag{5}$$

For a smooth k-dimensional submanifold of $\mathbb{R}^n$, $\dim_H = k$ and $\mathcal{H}^k$ recovers the standard k-dimensional surface measure. For irregular sets such as Cantor sets or fractal curves, $\dim_H$ takes non-integer values. The notion is purely metric, so it requires no smoothness, no manifold structure, and no embedding of X into Euclidean space.

Relation to intrinsic dimension in ML. The term "intrinsic dimension" in machine learning is a loose umbrella for several quantities aimed at capturing the effective degrees of freedom of a dataset, including Hausdorff dimension, Minkowski (box-counting) dimension, correlation dimension, and the dimension of an underlying smooth manifold. For smooth manifolds and well-behaved data supports, these notions coincide (Falconer, 2003). Empirical intrinsic dimension estimators such as TwoNN and MLE-based methods target correlation-style dimensions that, on a smooth support, agree with $\dim_H$ (Ansuini et al., 2019; Pope et al., 2021). We work with $\dim_H$ because the Lipschitz scaling property has a clean classical statement in this setting.

Lipschitz scaling of Hausdorff measure. Combining Equation (3) with Equation (4): applying the Lipschitz diameter bound to each set in any δ-cover of A yields a $(L\delta)$-cover of $f(A)$, with diameters scaled by at most L. Each term $(\operatorname{diam} U_i)^s$ in the sum becomes at most $L^s(\operatorname{diam} U_i)^s$ in the image cover. Summing and taking infimum,

$$\mathcal{H}^s(f(A)) \;\leq\; L^s\,\mathcal{H}^s(A) \quad \text{for every } s \geq 0. \tag{6}$$

Taking $s > \dim_H A$ yields $\mathcal{H}^s(A) = 0$ by Equation (5), hence $\mathcal{H}^s(f(A)) = 0$, hence $\dim_H f(A) \leq s$. Letting $s \downarrow \dim_H A$ gives (Falconer, 2003, Proposition 2.2(b)):

$$\dim_H f(A) \;\leq\; \dim_H A. \tag{7}$$

Pushforward and mutual singularity. Given a measure μ on X and a measurable $f : X \to Y$, the pushforward $f_{\#}\mu$ is the measure on Y defined by $(f_{\#}\mu)(B) := \mu(f^{-1}(B))$ for Borel $B \subset Y$. Two measures ν_1, ν_2 on Y are mutually singular, written $\nu_1 \perp \nu_2$, if there is a Borel partition $Y = B_1 \sqcup B_2$ with $\nu_1(B_2) = 0$ and $\nu_2(B_1) = 0$. Mutual singularity is the measure-theoretic statement that ν_1 and ν_2 live on disjoint pieces of Y.

Standard Borel spaces and Borel isomorphism. A standard Borel space is a measurable space isomorphic, as a measurable space, to a Polish space with its Borel σ-algebra. Both S^k and S^{d-1} are standard Borel. By (Kechris, 1995, Theorem 17.41), any two atomless standard Borel probability spaces are isomorphic as measure spaces: there exists a Borel bijection $T : X \to Y$ such that $T_{\#}\mu_X = \mu_Y$ and T^{-1} is also Borel measurable.

B.2 Aligned-pair characterization of the free optimum

The asymptotic InfoNCE characterization of (Wang & Isola, 2020; Koromilas et al., 2024) is established for single-encoder setups with matching marginals. The two-encoder cross-modal objective Equation (1) is more delicate: the multimodal landscape can favor population-level marginal mismatch (the modality gap of (Liang et al., 2022)), so a clean cross-marginal lower bound on $\mathcal{L}_\tau$ in terms of the single-marginal Wang-Isola minimum does not hold without further structure. The lemma below reduces to the single-encoder case by restricting to *aligned* encoder pairs, which is the regime in which a location encoder serves its intended purpose: producing embeddings that match the image encoder on paired samples.

Definition 1 (Aligned encoder pair)**.** *A measurable pair (f_c, f_I) is* aligned *if $f_c(c) = f_I(I)$ for p-a.e. (c, I).*

Under the Dirac structure $p(c \mid I) = \delta_{c(I)}$, alignment is equivalent to $f_I(I) = f_c(c(I))$ for p_I-a.e. I, and forces $\nu_c := f_{c\#}\mu = f_{I\#}p_I =: \nu_I$ since the marginal of p on S^k is μ.

Lemma 1 (Wang-Isola characterization restricted to aligned pairs)**.** *For any aligned pair (f_c, f_I) with common marginal $\nu := \nu_c = \nu_I$,*

$$\mathcal{L}_\tau(f_c, f_I) \;=\; -\frac{1}{\tau} \;+\; \mathbb{E}_{\boldsymbol{u}\sim\nu} \log \mathbb{E}_{\boldsymbol{v}\sim\nu}[K(\boldsymbol{u}, \boldsymbol{v})], \tag{8}$$

where $K(\boldsymbol{u}, \boldsymbol{v}) := e^{\boldsymbol{u}^\top \boldsymbol{v}/\tau}$. The right-hand side is the asymptotic InfoNCE loss of (Wang & Isola, 2020, Theorem 1) for a single encoder with marginal ν, uniquely minimized over $\nu \in \mathcal{P}(S^{d-1})$ at $\nu = \sigma$, with minimum value $V^* := -1/\tau + \mathbb{E}_{\boldsymbol{u}\sim\sigma} \log \mathbb{E}_{\boldsymbol{v}\sim\sigma}[K(\boldsymbol{u}, \boldsymbol{v})]$.

Proof. The proof has two parts. We first show that under alignment with matched marginals, the symmetric multimodal loss collapses to a single-encoder expression in ν. We then identify σ as the unique minimizer of this expression, which is established in the relaxed measure-level analysis of Wang & Isola (2020, Appendix A.2).

Collapse to single-encoder form. Write the loss as $\mathcal{L}_\tau(f_c, f_I) = A + \frac{1}{2}(U_1 + U_2)$ following the population NT-Xent expansion in Equation (1), where

$$A \;:=\; -\frac{1}{\tau}\, \mathbb{E}_{(c,I)\sim p}\big[f_c(c)^\top f_I(I)\big]$$

is the alignment term and U_1, U_2 are the two uniformity terms (each without a $1/2$ factor; the $1/2$ sits in front of the sum).

We handle the alignment term first. Under Definition 1, $f_c(c) = f_I(I)$ for p-a.e. pairs (c, I). Both encoders take values on S^{d-1}, so $f_c(c)^\top f_I(I) = \|f_c(c)\|^2 = 1$ for p-a.e. pairs, giving $A = -1/\tau$. Alignment saturates this term at its lowest possible value.

For the uniformity terms, observe that under matching marginals $\nu_c = \nu_I = \nu$, the inner kernel expectations agree across modalities. For any test vector $\boldsymbol{u} \in S^{d-1}$,

$$\mathbb{E}_{I'\sim p_I}\big[e^{f_I(I')^\top \boldsymbol{u}/\tau}\big] \;=\; \mathbb{E}_{\boldsymbol{w}\sim\nu_I}\big[e^{\boldsymbol{w}^\top \boldsymbol{u}/\tau}\big] \;=\; \mathbb{E}_{\boldsymbol{w}\sim\nu_c}\big[e^{\boldsymbol{w}^\top \boldsymbol{u}/\tau}\big] \;=\; \mathbb{E}_{c'\sim\mu}\big[e^{f_c(c')^\top \boldsymbol{u}/\tau}\big],$$

where the first equality is a change of variables, the second uses $\nu_I = \nu_c = \nu$, and the third is again change of variables. So both uniformity terms reduce to the same expression in ν:

$$U_1 \;=\; \mathbb{E}_{\boldsymbol{u}\sim\nu} \log \mathbb{E}_{\boldsymbol{v}\sim\nu}[K(\boldsymbol{u}, \boldsymbol{v})] \;=\; U_2. \tag{9}$$

Their average is the same value. Combining $A = -1/\tau$ with Equation (9),

$$\mathcal{L}_\tau(f_c, f_I) \;=\; -\frac{1}{\tau} + \mathbb{E}_{\boldsymbol{u}\sim\nu} \log \mathbb{E}_{\boldsymbol{v}\sim\nu}[K(\boldsymbol{u}, \boldsymbol{v})],$$

which is Equation (8).

The right-hand side is uniquely minimized at σ. Write $h_\nu(u) := \mathbb{E}_{v\sim\nu}[K(u, v)]$ and $W(\nu) := \mathbb{E}_{u\sim\nu} \log h_\nu(u)$, so that the right-hand side of Equation (8) equals $-1/\tau + W(\nu)$. The minimization of W over all of $\mathcal{P}(\mathbb{S}^{d-1})$ is exactly the relaxed problem studied by Wang & Isola (2020, Appendix A.2, Eq. (13)) in the proof of their Theorem 1, where it is shown that σ is its unique minimizer. The argument has two steps. First, a first-variation argument shows that any minimizer $\nu^\star$ must have $h_{\nu^\star}$ constant $\nu^\star$-almost surely. On this set of measures, Jensen's inequality holds with equality, so $W(\nu) = \log \iint K \,\mathrm{d}(\nu \otimes \nu) = \frac{1}{\tau} + \log \iint e^{-\|u-v\|^2/(2\tau)} \,\mathrm{d}(\nu \otimes \nu)$. Second, the Gaussian energy on the right is uniquely minimized over $\mathcal{P}(\mathbb{S}^{d-1})$ by σ (Wang & Isola, 2020, Proposition 1), and σ itself satisfies the constancy condition, since h_σ is constant by rotation invariance. Hence σ is the unique minimizer of W over $\mathcal{P}(\mathbb{S}^{d-1})$, and

$$\min_{\nu\in\mathcal{P}(\mathbb{S}^{d-1})} \Big(-\tfrac{1}{\tau} + W(\nu) \Big) = -\tfrac{1}{\tau} + \mathbb{E}_{u\sim\sigma} \log \mathbb{E}_{v\sim\sigma}[K(u, v)] = V^*.$$

We stress that the conditional phrasing of Wang & Isola (2020, Theorem 1, result 2) ("if perfectly uniform encoders exist") concerns only whether σ is attained by an encoder, not the uniqueness of σ among all Borel probability measures.

□

Corollary 2 (Aligned-pair optimum and its relation to the free infimum)**.** *$V^* \geq \mathcal{L}_\tau^{\text{free}}$, with V^* attained at any aligned pair (f_c, f_I) with $f_{c\#}\mu = \sigma$. For any aligned pair with $f_{c\#}\mu \neq \sigma$, $\mathcal{L}_\tau(f_c, f_I) > V^*$.*

Proof. Lemma 1 gives that aligned pairs achieve V^* iff their common marginal is σ, with strict inequality $\mathcal{L}_\tau(f_c, f_I) > V^*$ when the common marginal differs from σ. The Borel-isomorphism construction in Section B.3 (Step 3) realizes an aligned pair with $f_{c\#}\mu = \sigma$, attaining V^*. By definition $\mathcal{L}_\tau^{\text{free}} = \inf_{f_c, f_I} \mathcal{L}_\tau(f_c, f_I) \leq V^*$, since the construction is a measurable pair. □

Remark on non-aligned pairs. The aligned-pair optimum is characterized in the corollary and the inequality $V^* \geq \mathcal{L}_\tau^{\text{free}}$ is given. The reverse inequality $V^* \leq \mathcal{L}_\tau^{\text{free}}$ is not established, and there is no general reason to expect it to hold: for non-aligned pairs, the symmetric multimodal landscape can support stable equilibria with $\nu_c \neq \nu_I$ at values strictly less than V^* (Cai et al., 2026; Liang et al., 2022), so $\mathcal{L}_\tau^{\text{free}} < V^*$ is possible in principle. This does not affect the strict-gap result of Theorem 1(iii), which establishes $\mathcal{L}_\tau(\text{aligned Lipschitz}) > V^* \geq \mathcal{L}_\tau^{\text{free}}$ and uses only the side $V^* \geq \mathcal{L}_\tau^{\text{free}}$. Restricting to aligned pairs is appropriate for our setting: a location encoder whose embeddings do not align with the image encoder on paired data fails its purpose as a learned coordinate. The dimension and singular-marginal conclusions of Theorem 1 (parts (i) and (ii)) hold for any Lipschitz f_c regardless of alignment; only the strict per-pair loss gap (part (iii)) requires the alignment hypothesis.

B.3 Proof of Theorem 1

The proof is delivered in three steps. In Step 1 the dimension of the image is bounded by the domain dimension. In Step 2 the dimension bound is turned into mutual singularity of measures via the codimension gap. And finally, in Step 3 Lemma 1 is invoked, which collapses the symmetric multimodal NT-Xent to single-encoder Wang-Isola for aligned pairs, to convert mutual singularity into a strict per-pair inequality (under alignment).

Step 1: dimension of the image. By hypothesis, $f_c : S^k \to S^{d-1}$ is Lipschitz, and S^k has Hausdorff dimension k as a smooth k-manifold. Applying the Lipschitz dimension bound Equation (7) from Section B.1 with $A = S^k$,

$$\dim_H f_c(S^k) \;\leq\; \dim_H S^k \;=\; k, \tag{10}$$

proving (i). □

Step 2: singular marginal. We set $A := f_c(S^k) \subset S^{d-1}$. Since S^k is compact and f_c is Lipschitz hence continuous, A is compact, hence Borel. By Step 1, $\dim_H A \leq k$. The setup imposes $k < d-1$, so by the definition of Hausdorff dimension,

$$\mathcal{H}^{d-1}(A) \;=\; 0. \tag{11}$$

The uniform measure σ on S^{d-1} is the normalized $(d-1)$-Hausdorff measure: for every Borel $B \subseteq S^{d-1}$,

$$\sigma(B) \;=\; \frac{\mathcal{H}^{d-1}(B)}{\mathcal{H}^{d-1}(S^{d-1})}. \tag{12}$$

This identification follows because Hausdorff measure is metric and hence invariant under isometries of the ambient space; rotations $R \in O(d)$ act by isometry on $\mathbb{R}^d$ and preserve S^{d-1}, so $\mathcal{H}^{d-1}|_{S^{d-1}}$ is rotation-invariant. The sphere S^{d-1} is a smooth compact $(d-1)$-manifold embedded in $\mathbb{R}^d$, hence $\mathcal{H}^{d-1}(S^{d-1})$ is finite and positive (it equals the standard surface area $2\pi^{d/2}/\Gamma(d/2)$); see (Federer, 1969, § 3.2.46) or (Falconer, 2003, Ch. 3). Normalizing, $\mathcal{H}^{d-1}/\mathcal{H}^{d-1}(S^{d-1})$ is a rotation-invariant Borel probability measure on S^{d-1}. The orthogonal group $O(d)$ acts transitively on S^{d-1}, and there is a unique rotation-invariant Borel probability measure on a homogeneous space of a compact group (Folland, 1999, Ch. 11), which we call σ. Hence the two coincide. Combined with Equation (11),

$$\sigma(A) \;=\; 0. \tag{13}$$

On the other hand, A contains the image of S^k under f_c, so

$$\nu(A) \;=\; \mu\big(f_c^{-1}(A)\big) \;=\; \mu(S^k) \;=\; 1, \qquad \nu(S^{d-1} \setminus A) \;=\; 0. \tag{14}$$

Equations Equation (13) and Equation (14) exhibit a Borel partition $S^{d-1} = A \sqcup (S^{d-1} \setminus A)$ such that $\nu(S^{d-1} \setminus A) = 0$ and $\sigma(A) = 0$, which is the definition of mutual singularity from Section B.1. Therefore $\nu \perp \sigma$, proving (ii). □

Step 3: per-pair strict gap (under alignment).

What we want to show: any aligned pair (f_c, f_I) where f_c is Lipschitz strictly exceeds the free infimum, $\mathcal{L}_\tau(f_c, f_I) > \mathcal{L}_\tau^{\text{free}}$. The argument has two parts. We first construct an explicit pair attaining the value V^*, which sets a target value to compare against. We then use Step 2 plus the uniqueness of the minimizer in Lemma 1 to show every Lipschitz aligned pair lies strictly above this target.

An explicit pair attains V^*. The aligned-uniform value V^* is the value of the loss when the location encoder distributes its outputs uniformly on S^{d-1} and the image encoder agrees with it on paired data. We construct such a pair to confirm V^* is actually attained somewhere.

The location and embedding domains have very different intrinsic dimensions (k versus $d-1$), so we cannot construct a smooth bijection between them. We can, however, construct a measure-theoretic bijection. Both (S^k, μ) and (S^{d-1}, σ) are atomless standard Borel probability spaces, and any two such spaces are isomorphic as measure spaces (Section B.1; this is a classical theorem in descriptive set theory). We therefore have a measure-preserving Borel isomorphism $T : S^k \to S^{d-1}$, meaning T is a Borel bijection and $T_\# \mu = \sigma$.

We use T to define an aligned pair. Let

$$f_c^\circ(c) \;:=\; T(c), \qquad f_I^\circ(I) \;:=\; T(c(I)). \tag{15}$$

The Dirac structure $p(c \mid I) = \delta_{c(I)}$ gives $c = c(I)$ for p-a.e. (c, I), so $f_c^\circ(c) = T(c) = T(c(I)) = f_I^\circ(I)$ for p-a.e. pairs; the pair is aligned in the sense of Definition 1. The pushforward $f_{c\#}^\circ \mu$ equals $T_\# \mu$ by definition, which equals σ by the measure-preserving property. So (f_c°, f_I°) is aligned with common marginal σ, and Corollary 2 gives $\mathcal{L}_\tau(f_c^\circ, f_I^\circ) = V^*$. Since this is a measurable pair, $\mathcal{L}_\tau^{\text{free}} \le V^*$.

The map T cannot be continuous when $k < d-1$: a continuous bijection from the compact space S^k onto the Hausdorff space S^{d-1} would be a homeomorphism, which invariance of dimension rules out. Continuity alone, however, does not exclude the value V^*, since non-injective space-filling maps are continuous. What excludes it is the regularity of f_c: Lipschitz maps cannot increase Hausdorff dimension (Equation (7)), so by Step 2 the aligned-uniform value V^* is attained outside any Lipschitz class. More generally, α-Hölder maps satisfy $\dim_H f(A) \le \dim_H A/\alpha$ (Falconer, 2003, Ch. 2), so the same conclusion holds for any α-Hölder f_c with $\alpha > k/(d-1)$.

Lipschitz aligned pairs are strictly above V^*. Suppose f_c is Lipschitz and (f_c, f_I) is aligned. We want $\mathcal{L}_\tau(f_c, f_I) > V^*$.

By Step 2 applied to this f_c, the pushforward $\nu_c := f_{c\#}\mu$ is mutually singular with σ. In particular $\nu_c \neq \sigma$, since σ is not mutually singular with itself ($\sigma(S^{d-1}) = 1$ rules out $\sigma \perp \sigma$).

Lemma 1 applies to the aligned pair and gives

$$\mathcal{L}_\tau(f_c, f_I) \;=\; -\frac{1}{\tau} + \mathbb{E}_{\boldsymbol{u}\sim\nu_c} \log \mathbb{E}_{\boldsymbol{v}\sim\nu_c}[K(\boldsymbol{u}, \boldsymbol{v})]. \tag{16}$$

The right-hand side is the asymptotic InfoNCE loss for a single encoder with marginal ν_c. By Lemma 1, this expression is uniquely minimized over $\nu \in \mathcal{P}(S^{d-1})$ at $\nu = \sigma$. Since $\nu_c \neq \sigma$, we get a strict inequality:

$$\mathbb{E}_{\boldsymbol{u}\sim\nu_c} \log \mathbb{E}_{\boldsymbol{v}\sim\nu_c}[K(\boldsymbol{u}, \boldsymbol{v})] \;>\; \mathbb{E}_{\boldsymbol{u}\sim\sigma} \log \mathbb{E}_{\boldsymbol{v}\sim\sigma}[K(\boldsymbol{u}, \boldsymbol{v})]. \tag{17}$$

Adding $-1/\tau$ to both sides,

$$\mathcal{L}_\tau(f_c, f_I) \;>\; -\frac{1}{\tau} + \mathbb{E}_{\boldsymbol{u}\sim\sigma} \log \mathbb{E}_{\boldsymbol{v}\sim\sigma}[K(\boldsymbol{u}, \boldsymbol{v})] \;=\; V^* \;\ge\; \mathcal{L}_\tau^{\text{free}}, \tag{18}$$

proving (iii). □

Remark (image encoder inherits the dimension bound under alignment). Theorem 1 parts (i) and (ii) bound $f_c(S^k)$ alone. Under alignment of (f_c, f_I), the dimension ceiling extends to the image

side. By Definition 1 and the Dirac structure $p(c \mid I) = \delta_{c(I)}$, alignment gives $f_I(I) = f_c(c(I))$ for p_I-a.e. I. Hence the essential range of f_I under p_I is contained in $f_c(S^k)$:

$$\operatorname{supp}(\nu_I) \;\subseteq\; f_c(S^k), \qquad \dim_H \operatorname{supp}(\nu_I) \;\le\; k, \qquad \nu_I \;=\; \nu_c \;\perp\; \sigma, \tag{19}$$

where $\nu_I := f_{I\#}p_I$. The dimension ceiling and singular marginal carry from the location side to the image side at any aligned pair, regardless of which encoder pair achieves the partial optimum. The conclusion is about the pushforward measure ν_I, not about the literal image set $f_I(\mathcal{I})$, since f_I may take arbitrary values off a p_I-null set without affecting the loss.

B.4 SatCLIP function class and uniform Lipschitz constant

Function class. The SatCLIP location encoder is parameterized by a bandwidth $L \in \mathbb{N}$ and a Siren network. For each L, $\mathrm{SH}_L : S^2 \to \mathbb{R}^{(L+1)^2}$ stacks the real-valued spherical harmonic basis up to degree L, and $g_{\boldsymbol{\phi}} : \mathbb{R}^{(L+1)^2} \to \mathbb{R}^d$ is a Siren network with K hidden layers, frequency hyperparameter ω_0, and parameter vector $\boldsymbol{\phi}$ (Sitzmann et al., 2020). Write $P(L)$ for the Siren parameter count and $\Phi_L \subset \mathbb{R}^{P(L)}$ for a compact set of admissible parameters. Fix a margin $\varepsilon > 0$ and assume the non-degeneracy condition

$$\inf_{\boldsymbol{\phi}\in\Phi_L,\, c\in S^2} \|g_{\boldsymbol{\phi}}(\mathrm{SH}_L(c))\| \ge \varepsilon, \tag{20}$$

which keeps the post-Siren normalization onto S^{d-1} well-defined. The bandwidth-L SatCLIP function class is

$$\mathcal{F}_L^{\mathrm{SC}} \;:=\; \Big\{\, c \mapsto \frac{g_{\boldsymbol{\phi}}(\mathrm{SH}_L(c))}{\|g_{\boldsymbol{\phi}}(\mathrm{SH}_L(c))\|} \;:\; \boldsymbol{\phi} \in \Phi_L \,\Big\}. \tag{21}$$

Uniform Lipschitz constant.

Lemma 2 (SatCLIP encoders are uniformly Lipschitz)**.** *There exists $M = M(\Phi_L, L, \varepsilon) < \infty$ such that every $f_c \in \mathcal{F}_L^{\mathrm{SC}}$ satisfies $\mathrm{Lip}(f_c) \le M$.*

Proof. A SatCLIP encoder is a composition of three pieces, $f_c = N \circ g_{\boldsymbol{\phi}} \circ \mathrm{SH}_L$, where $N(\boldsymbol{u}) := \boldsymbol{u}/\|\boldsymbol{u}\|$ is the normalization onto the unit sphere. Lipschitz constants compose multiplicatively, so we bound each piece separately and multiply.

Spherical-harmonic features. Each coordinate of SH_L is a polynomial in the ambient $\mathbb{R}^3$ coordinates, restricted to S^2. Polynomials are smooth, and a smooth function on a compact set is Lipschitz with constant equal to the maximum gradient norm. We write L_{SH} for this constant; it depends only on the bandwidth L (higher L means polynomials of higher degree, with larger gradients), not on any of the encoder's trainable parameters.

Siren MLP. The Siren $g_{\boldsymbol{\phi}}$ alternates affine maps $\boldsymbol{u} \mapsto \boldsymbol{W}_k\boldsymbol{u} + \boldsymbol{b}_k$ with the activation $\sin(\omega_0 \cdot)$ applied componentwise. Each affine map is $\|\boldsymbol{W}_k\|_{\mathrm{op}}$-Lipschitz (the bias does not affect Lipschitz constants). The activation is ω_0-Lipschitz, since $|\sin'| = |\cos| \le 1$ implies $|\sin(\omega_0 a) - \sin(\omega_0 b)| \le \omega_0 |a - b|$. With K activation layers and $K + 1$ affine layers, multiplying through the chain gives

$$\mathrm{Lip}(g_{\boldsymbol{\phi}}) \;\le\; \omega_0^K \prod_{k=0}^{K} \|\boldsymbol{W}_k\|_{\mathrm{op}}. \tag{22}$$

The right-hand side is a continuous function of the parameter vector $\boldsymbol{\phi}$, and we have assumed Φ_L is compact. A continuous function on a compact set attains its maximum, so the right-hand side is bounded uniformly across $\boldsymbol{\phi} \in \Phi_L$ by some finite constant $M_g(\Phi_L)$. The compactness assumption on Φ_L is what gives us a single Lipschitz constant for the Siren that works for the entire family at once.

Normalization. The map $N : \boldsymbol{u} \mapsto \boldsymbol{u}/\|\boldsymbol{u}\|$ is smooth on $\mathbb{R}^d \setminus \{\boldsymbol{0}\}$, with Jacobian operator norm $1/\|\boldsymbol{u}\|$ at each point. On the annular region $\{\boldsymbol{u} : \|\boldsymbol{u}\| \ge \varepsilon\}$ this gives a pointwise bound of $1/\varepsilon$, but a pointwise gradient bound only translates to a Lipschitz constant on a *convex* domain (via the mean value theorem applied along straight-line paths). The annular region $\{\boldsymbol{u} : \|\boldsymbol{u}\| \ge \varepsilon\}$ is not convex: it is the complement of the open ball of radius ε around the origin, and the straight line between two points on opposite sides of the ball passes through it, leaving the region.

We get around this by routing through a path that stays inside the annular region. The argument is elementary: parameterize a smooth path $\gamma : [0, 1] \to \{\boldsymbol{u} : \|\boldsymbol{u}\| \geq \min(\|\boldsymbol{u}\|, \|\boldsymbol{v}\|)\}$ from $\boldsymbol{u}$ to $\boldsymbol{v}$ that avoids the inner ball (for instance, a piecewise-linear path that detours along the sphere of radius $\min(\|\boldsymbol{u}\|, \|\boldsymbol{v}\|)$ when the straight line between the endpoints would dip too close to the origin), apply the pointwise Jacobian bound along this path, and observe that the path length is at most twice the straight-line distance $\|\boldsymbol{u} - \boldsymbol{v}\|$. This gives the global bound

$$\Big\| \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} - \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \Big\| \;\leq\; \frac{2\,\|\boldsymbol{u} - \boldsymbol{v}\|}{\min(\|\boldsymbol{u}\|, \|\boldsymbol{v}\|)} \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \neq \mathbf{0}. \tag{23}$$

The factor of 2 accounts for the detour around the origin; on a convex domain it would be 1. On the set where $\min(\|\boldsymbol{u}\|, \|\boldsymbol{v}\|) \geq \varepsilon$, the bound becomes $(2/\varepsilon)\|\boldsymbol{u} - \boldsymbol{v}\|$, so N is $(2/\varepsilon)$-Lipschitz on this set.

The non-degeneracy condition Equation (20) is precisely what ensures the input to N stays in this annular region: $\|g_{\boldsymbol{\phi}}(\mathrm{SH}_L(c))\| \geq \varepsilon$ for every $\boldsymbol{\phi} \in \Phi_L$ and every $c \in S^2$.

Composition. Putting the three pieces together,

$$\mathrm{Lip}(f_c) \;\leq\; \mathrm{Lip}(N) \cdot \mathrm{Lip}(g_{\boldsymbol{\phi}}) \cdot \mathrm{Lip}(\mathrm{SH}_L) \;\leq\; \frac{2}{\varepsilon} \cdot M_g(\Phi_L) \cdot L_{\mathrm{SH}}. \tag{24}$$

The right-hand side does not depend on $\boldsymbol{\phi}$, so it serves as a uniform constant $M = M(\Phi_L, L, \varepsilon)$ across the entire family. □

B.5 Proof of Corollary 1

The strategy is to reduce the SatCLIP infimum to a minimization problem over the location encoder alone, then use compactness plus continuity to show this minimization is attained at some specific $f_c^\star \in \mathcal{F}_L^{\mathrm{SC}}$, and finally apply the strict inequality from Corollary 2 to that minimizer.

By Lemma 1, for any aligned pair (f_c, f_I) with $f_c \in \mathcal{F}_L^{\mathrm{SC}}$, the loss depends on f_c only through its pushforward $\nu_c := f_{c\#}\mu$:

$$\mathcal{L}_\tau(f_c, f_I) \;=\; -\frac{1}{\tau} \;+\; W(\nu_c), \qquad W(\nu) := \mathbb{E}_{\boldsymbol{u}\sim\nu} \log \mathbb{E}_{\boldsymbol{v}\sim\nu}[K(\boldsymbol{u}, \boldsymbol{v})]. \tag{25}$$

The image encoder drops out: any aligned f_I gives the same value for fixed f_c. We write $\mathcal{L}^*(f_c) := -1/\tau + W(\nu_c)$ for this common value. The SatCLIP infimum over aligned pairs is therefore $\inf_{f_c \in \mathcal{F}_L^{\mathrm{SC}}} \mathcal{L}^*(f_c)$, and we show this infimum is strictly above $\mathcal{L}_\tau^{\mathrm{free}}$.

The proof has three steps. Step A shows the family $\mathcal{F}_L^{\mathrm{SC}}$ is compact in a suitable topology, so any continuous function on it attains its infimum. Step B shows $\mathcal{L}^*$ is continuous in that topology, so the infimum is attained at a specific element $f_c^\star$. Step C applies Corollary 2 to $f_c^\star$ to get the strict inequality.

A. Outer compactness. We show $\mathcal{F}_L^{\mathrm{SC}}$ is compact in $C(S^2; \mathbb{R}^d)$ under the uniform topology, using Arzelà-Ascoli plus a closedness check.

By Lemma 2, every $f_c \in \mathcal{F}_L^{\mathrm{SC}}$ has Lipschitz constant at most M, so the family is equicontinuous. Every f_c takes values in S^{d-1}, so the family is pointwise bounded. Arzelà-Ascoli then gives precompactness in $C(S^2; \mathbb{R}^d)$ under the uniform topology.

For closedness, we use that $\mathcal{F}_L^{\mathrm{SC}}$ is the image of the compact parameter set Φ_L under the parameterization map $\boldsymbol{\phi} \mapsto g_{\boldsymbol{\phi}}(\mathrm{SH}_L(\cdot))/\|g_{\boldsymbol{\phi}}(\mathrm{SH}_L(\cdot))\|$. This map is continuous: the joint map $\Phi_L \times S^2 \ni (\boldsymbol{\phi}, c) \mapsto g_{\boldsymbol{\phi}}(\mathrm{SH}_L(c))$ is continuous (Siren is smooth in its parameters and inputs), and the non-degeneracy condition Equation (20) keeps the denominator bounded away from zero, so the normalization stays continuous. The continuous image of a compact set is compact, so $\mathcal{F}_L^{\mathrm{SC}}$ is closed. Combined with precompactness, $\mathcal{F}_L^{\mathrm{SC}}$ is compact in $C(S^2; \mathbb{R}^d)$.

B. Continuity of $\mathcal{L}^*$. By Equation (25), $\mathcal{L}^*$ is continuous in f_c if and only if the composition $f_c \mapsto \nu_c \mapsto W(\nu_c)$ is continuous. We check the two pieces.

Uniform convergence of encoders implies weak convergence of pushforwards. Suppose $f_c^{(n)} \to f_c$ uniformly on S^2. We construct a coupling that controls how far $\nu_c^{(n)}$ is from ν_c. The synchronous

coupling, where the same input c feeds both encoders, gives a joint distribution on $S^{d-1} \times S^{d-1}$ with marginals $(\nu_c^{(n)}, \nu_c)$. Its expected coordinate distance is $\mathbb{E}_{c\sim\mu}\|f_c^{(n)}(c) - f_c(c)\|$, which is at most $\|f_c^{(n)} - f_c\|_\infty$. By the Kantorovich-Rubinstein duality, the W_1 Wasserstein distance is the infimum of expected distance over all couplings, so

$$W_1(\nu_c^{(n)}, \nu_c) \;\le\; \mathbb{E}_{c\sim\mu}\,\|f_c^{(n)}(c) - f_c(c)\| \;\le\; \|f_c^{(n)} - f_c\|_\infty. \tag{26}$$

Uniform convergence of $f_c^{(n)} \to f_c$ therefore drives $W_1(\nu_c^{(n)}, \nu_c) \to 0$. On a compact metric space such as S^{d-1}, W_1 convergence implies weak convergence (this is standard; see any reference on Wasserstein distances), so $\nu_c^{(n)} \rightharpoonup \nu_c$.

Weak convergence of ν implies convergence of $W(\nu)$. We need $W(\nu) = \int \log h_\nu(\boldsymbol{u})\, d\nu(\boldsymbol{u})$ to be continuous in ν under weak convergence, where $h_\nu(\boldsymbol{u}) := \mathbb{E}_{\boldsymbol{v}\sim\nu}[K(\boldsymbol{u}, \boldsymbol{v})]$ is the inner kernel expectation.

The kernel $K(\boldsymbol{u}, \boldsymbol{v}) = e^{\boldsymbol{u}^\top \boldsymbol{v}/\tau}$ is bounded continuous on $S^{d-1} \times S^{d-1}$, with values in $[e^{-1/\tau}, e^{1/\tau}]$. So $h_\nu(\boldsymbol{u})$ is bounded between $e^{-1/\tau}$ and $e^{1/\tau}$ for any ν and $\boldsymbol{u}$, and $\log h_\nu(\boldsymbol{u})$ is bounded between $-1/\tau$ and $1/\tau$. The Portmanteau theorem says that, applied to the bounded continuous test function $\boldsymbol{v} \mapsto K(\boldsymbol{u}, \boldsymbol{v})$, weak convergence $\nu_n \rightharpoonup \nu$ implies $h_{\nu_n}(\boldsymbol{u}) \to h_\nu(\boldsymbol{u})$ pointwise in $\boldsymbol{u}$, and joint continuity of K in $(\boldsymbol{u}, \boldsymbol{v})$ upgrades this to joint continuity of $h_\nu(\boldsymbol{u})$ in $(\nu, \boldsymbol{u})$. Composing with $\log$, the integrand $\log h_\nu(\boldsymbol{u})$ is bounded and jointly continuous in $(\nu, \boldsymbol{u})$.

A bounded jointly continuous integrand integrated against ν gives a weakly continuous functional. Concretely, $W(\nu_n) = \int \log h_{\nu_n}(\boldsymbol{u})\, d\nu_n(\boldsymbol{u}) \to \int \log h_\nu(\boldsymbol{u})\, d\nu(\boldsymbol{u}) = W(\nu)$ along any weakly convergent sequence $\nu_n \rightharpoonup \nu$. So W is weakly continuous on $\mathcal{P}(S^{d-1})$.

Composing the two pieces. Uniform convergence $f_c^{(n)} \to f_c$ implies weak convergence $\nu_c^{(n)} \rightharpoonup \nu_c$ implies $W(\nu_c^{(n)}) \to W(\nu_c)$, hence $\mathcal{L}^*(f_c^{(n)}) \to \mathcal{L}^*(f_c)$. So $\mathcal{L}^*$ is continuous on $\mathcal{F}_L^{\mathrm{SC}}$.

C. Conclusion. By Step A, $\mathcal{F}_L^{\mathrm{SC}}$ is compact. By Step B, $\mathcal{L}^*$ is continuous on $\mathcal{F}_L^{\mathrm{SC}}$. A continuous function on a compact set attains its infimum, so $\inf_{f_c\in\mathcal{F}_L^{\mathrm{SC}}} \mathcal{L}^*(f_c) = \mathcal{L}^*(f_c^\star)$ for some $f_c^\star \in \mathcal{F}_L^{\mathrm{SC}}$.

We apply Corollary 2 to this $f_c^\star$. By Step 2 of Theorem 1 applied to $f_c^\star$ (which is Lipschitz because it lives in $\mathcal{F}_L^{\mathrm{SC}}$), the pushforward $\nu_c^\star := f_{c\#}^\star\mu$ is mutually singular with σ, hence $\nu_c^\star \neq \sigma$. Lemma 1, applied to any aligned pair built from $f_c^\star$, gives $W(\nu_c^\star) > W(\sigma)$ strictly. Therefore

$$\mathcal{L}^*(f_c^\star) \;=\; -1/\tau + W(\nu_c^\star) \;>\; -1/\tau + W(\sigma) \;=\; V^* \;\ge\; \mathcal{L}_\tau^{\mathrm{free}}.$$

Combining with the reduction at the start of the proof,

$$\inf_{\substack{f_c\in\mathcal{F}_L^{\mathrm{SC}},\, f_I \text{ measurable} \\ (f_c, f_I) \text{ aligned}}} \mathcal{L}_\tau(f_c, f_I) \;=\; \inf_{f_c\in\mathcal{F}_L^{\mathrm{SC}}} \mathcal{L}^*(f_c) \;=\; \mathcal{L}^*(f_c^\star) \;>\; \mathcal{L}_\tau^{\mathrm{free}}, \tag{27}$$

so the SatCLIP infimum strictly exceeds $\mathcal{L}_\tau^{\mathrm{free}}$. □

C Data acquisition

C.1 Metadata

To provide rich spatial context for our geospatial analysis, we compiled a global dataset of static environmental and anthropogenic features at approximately 1 km spatial resolution. This dataset supports location-aware modeling by capturing terrain characteristics, infrastructure patterns, and land-use composition.

We use a regular latitude-longitude grid (EPSG:4326) with 0.01 degree resolution in both dimensions, corresponding to roughly 1 km at the equator. The grid spans the full globe: 180W to 180E longitude (36,000 grid points) and 90S to 90N latitude (18,000 grid points). All source datasets were resampled or interpolated to this common grid to guarantee spatial coherence across variables.

Digital Elevation Model (DEM): Elevation was obtained from Copernicus DEM GLO-30 (COP-DEM GLO-30)European Space Agency (2022) through the Microsoft Planetary Computer STAC API. Native 30 m tiles were resampled to 1 km using area-weighted averaging.

From elevation, we derived:

- Slope, computed as $\sqrt{(\partial z/\partial x)^2 + (\partial z/\partial y)^2}$ (radians).
- Aspect, computed as $\arctan 2(-\partial z/\partial x, \partial z/\partial y)$ (degrees).
- Curvature, computed from the Laplacian of the elevation field.

Road network proximity: Road distance data was obtained from WorldPop's global highway-distance productWoods et al. (2025) derived from OpenStreetMap (OSM) (2023 release)OpenStreetMap contributors (2023). Country-level rasters at 100 m were mosaicked into a global raster, then resampled to the target 1 km grid using bilinear interpolation. The resulting variable `roads_dist_km` is the Euclidean distance (km) from each cell to the nearest OSM road.

Population density: We assembled annual WorldPop global population-density products for 2021, 2022, and 2023 at 1 kmBondarenko et al. (2025), and stored them as a temporal stack. Values were converted from persons/m^2 to persons/km^2.

Land-cover fractions: Land cover came from the Copernicus Climate Change Service (C3S) Global Land Cover datasetCopernicus Climate Change Service, Climate Data Store (2019) (LCCS classes, native resolution about 300 m), downloaded for 2021–2023 from the Copernicus Data Space Ecosystem S3 endpoint. We defined eight land-cover groups:

- Agriculture (LCCS 10–40)
- Forest (LCCS 50–100)
- Grassland (LCCS 110, 130)
- Shrubland (LCCS 120–122)
- Sparse vegetation (LCCS 140–153, 200–202, 220)
- Wetland (LCCS 160–180)
- Settlement (LCCS 190)
- Water bodies (LCCS 210)

For each group, we generated a binary mask at native resolution and aggregated it to 1 km by computing the fraction of subpixels per 1 km cell (3x3 coarsening).

Merging and alignement: All datasets were processed independently and then merged in three steps:

1. Build a base archive with COPDEM elevation, slope, aspect, curvature, and OSM road distance.
2. Integrate multi-year population density, aligned to exact target coordinates.
3. Append the eight land-cover fraction variables, aligned with nearest-neighbor interpolation to preserve categorical consistency.

Variables in the final archive: The final Zarr archive contains the following variables :

Table 5: Variables included in the global static-feature archive.

Variable	Dimensions	Unit	Description
`elevation`	(lat, lon)	m	Height above reference ellipsoid
`slope`	(lat, lon)	rad	Magnitude of terrain gradient
`aspect`	(lat, lon)	deg	Direction of steepest ascent
`curvature`	(lat, lon)	–	Local topographic curvature
`roads_dist_km`	(lat, lon)	km	Distance to nearest OSM road
`pop_density`	(lat, lon, year)	persons/km^2	WorldPop annual estimates
`lc_agriculture`	(lat, lon, year)	[0,1]	Fractional agriculture cover
`lc_forest`	(lat, lon, year)	[0,1]	Fractional forest cover
`lc_grassland`	(lat, lon, year)	[0,1]	Fractional grassland cover
`lc_shrubland`	(lat, lon, year)	[0,1]	Fractional shrubland cover
`lc_sparse_vegetation`	(lat, lon, year)	[0,1]	Fractional sparse-vegetation cover
`lc_wetland`	(lat, lon, year)	[0,1]	Fractional wetland cover
`lc_settlement`	(lat, lon, year)	[0,1]	Fractional settlement cover
`lc_water_bodies`	(lat, lon, year)	[0,1]	Fractional water-body cover

For LoCA pretraining, temporal metadata variables are resolved to a single year before being passed to the metadata encoder, yielding 14 scalar inputs per coordinate: elevation, slope, aspect, curvature, road distance, population density, and eight land-cover fractions. We use the mean of the years 2021-2023. Coordinates include `lat`, `lon`, and `year`; all data are provided in EPSG:4326.

C.2 Sentinel-2/Sentinel-1 paired image dataset

We constructed a paired optical-radar dataset, containing 100K image pairs, by sampling Sentinel-2 (S2) patches and matching each patch with the closest Sentinel-1 (S1) GRD acquisition over the same geographical location. The goal was to produce spatially aligned S2/S1 pairs.

S2 candidate scenes were drawn from the same GeoParquet index used in the SatCLIP(Klemmer et al., 2025) pipeline. Following the same setup, candidate rows were filtered to valid cloud-cover values before sampling.

For each attempt, a base timestamp is sampled from the parquet index, and a local candidate pool is formed within a temporal window of ± 14 days. From this pool, one S2 L2A item is selected, and a random 256×256 crop is then extracted (with configurable edge margin).

Each S2 patch is quality-checked and rejected if either of the following exceeds 10%:

- all-band missing fraction (pixels that are NaN in all bands),
- all-band zero fraction (pixels that are zero in all bands).

Given the accepted S2 patch location, we query Sentinel-1 GRD items intersecting that location in a window of ± 6 days around the S2 acquisition time.

From returned S1 items, we keep only scenes containing both `VV` and `VH` assets, then sort by absolute temporal distance to the S2 timestamp and try candidates in that order.

The stacked S1 patch (VV, VH) is then subject to the same empty/zero quality thresholds as S2.

For each valid pair index `idx`, the pipeline writes:

- `patch_{idx}_s2.tif`: S2 patch (12 bands, `uint16`),
- `patch_{idx}_s1.tif`: S1 patch (2 bands: VV/VH, `float32`).

D LoCA Implementation Details

Pretraining Dataset. We construct our pretraining dataset by aligning three data sources: (i) a global metadata dataset of 648 M points, each described by 14 static environmental and anthropogenic features, (ii) the S2-100K Sentinel-2 image dataset comprising 100,000 multispectral patches (256×256 px at 10 m/px, 12 spectral bands, nearly uniformly distributed across global land mass), and (iii) co-registered Sentinel-1 GRD SAR image patches (2-channel VV+VH, 224×224 px).

Each S2 image patch covers a circular footprint of radius $\sim$1.28 km ($= 256/2 \times 10$ m). The global metadata grid has $\sim$1 km spacing between points, where each metadata point covers a radius of $\sim$500 m. We compute circle-circle overlap fractions between every image footprint and nearby metadata grid cells via binary search on the regular grid. A metadata point is assigned to an image if at least 75% of its coverage circle falls within the image footprint; on average $\sim$5 metadata points are matched per image at this threshold. Sentinel-1 images are selected to match the Sentinel-2 images, so each row in the final parquet file represents a (metadata point, S2 image, S1 image) triplet. A consolidated HDF5 archive is created for each image modality, converting TIF files into a format that substantially reduces data-loading time during training.

As shown in Table 5, the 14 metadata features comprise 8 land-cover fractions (agriculture, forest, grassland, settlement, shrubland, sparse vegetation, water bodies, wetland), population density, distance to nearest road, and 4 terrain variables (elevation, slope, aspect, curvature). Missing values are handled with per-variable fill rules: land-cover and roads $\rightarrow -1$; population density, elevation, curvature, slope $\rightarrow 0$; aspect $\rightarrow 361$ (an out-of-range sentinel, since the valid range is $[-180, 180]$). Standard scaling (z-score normalisation) with precomputed global mean and standard deviation is applied to all metadata features. The full dataset is used for contrastive pretraining (no train/test split); downstream tasks use their own independent splits.

Two-Stage Training Pipeline. *Stage 1 (S1+S2 end-to-end pretraining, 2 modalities):* Both image encoder backbones—the S2 ResNet-50 and the S1 ResNet-50—together with their projection heads, are trained end-to-end with the contrastive loss; all backbone weights are trainable. Optimization hyperparameters for both stages are listed in Table 6.

Stage 2 (full 4-modality alignment): The S1 and S2 backbones are frozen and their weights are transferred from stage 1; their embeddings are precomputed once and cached to disk, so the frozen backbones are never executed during stage-2 training and no different augmentations are applied to the same image across iterations. The metadata encoder and location encoder are introduced as fresh, randomly initialised modules. Only the metadata encoder, location encoder, and their MLP heads are trained.

Training sampling in stage 2: We load the (metadata point, S2 image, S1 image) triplets by grouping metadata points by their associated (S2, S1) image pair. Each training sample includes the coordinates of the S2 image, the S2 image itself, the matched S1 image, and one randomly selected metadata point from the grouped set. In the next iteration of the same image group, the sampled metadata features may differ.

Stage transition: Projection heads from stage 1 are discarded at the start of stage 2; only backbone weights are loaded (with `strict=False`). A fresh optimiser, scheduler, and epoch counter are initialised for stage 2.

Encoder Architectures. *S2 Image Encoder:* We use the SatCLIP pretrained ResNet-50 visual backbone (`microsoft/SatCLIP-ResNet50-L40`). The final fully-connected layer is replaced with an identity mapping for feature extraction, yielding a backbone output dimension of 2048. When frozen (stage 2), a 2-layer trainable MLP head (adapter) maps $2048 \rightarrow 256$. In both stages, a 2-layer projection head MLP maps $256 \rightarrow 128$.

S1 Image Encoder: We use the torchgeo ResNet-50 pretrained on Sentinel-1 GRD via MoCo v2 (`ResNet50_Weights.SENTINEL1_GRD_MOCO`) with 2 input channels (VV + VH). The architecture follows the same pattern as the S2 encoder (backbone $\rightarrow$ optional trainable head (adapter) $\rightarrow$ projection head), with backbone output dimension 2048.

Location Encoder: We adopt the SatCLIP location encoder architecture. Input coordinates (longitude, latitude) are first mapped to a high-dimensional representation via a spherical harmonics positional encoding with Legendre polynomial degree $L = 40$, producing an embedding of dimension $L \times L = 1600$. This representation is then processed by a Sinusoidal Representation Network (SIREN) Sitzmann et al. (2020) consisting of 2 hidden layers of width 256 and an output layer of dimension 256. In a SIREN, each layer computes $\mathbf{h}_{i+1} = \sin(\omega_0 \cdot (\mathbf{W}_i \mathbf{h}_i + \mathbf{b}_i))$, where ω_0 is a frequency scaling factor ($\omega_0 = 30$ for the first layer, $\omega_0 = 1$ for subsequent layers) and the weights $\mathbf{W}_i$ are initialised following the scheme of Sitzmann et al. (2020): $\mathbf{W}_0 \sim \mathcal{U}(-1/n, 1/n)$ and $\mathbf{W}_i \sim \mathcal{U}\Big(-\sqrt{6/n}/\omega_0,\ \sqrt{6/n}/\omega_0\Big)$ for $i \geq 1$, where n is the layer's fan-in. The location encoder

is randomly initialised (not loaded from the pretrained Microsoft checkpoint). A 2-layer projection head MLP maps $256 \to 128$.

Metadata Encoder: The encoder is a 3-layer MLP ($14 \to 256 \to 256 \to 256$) with BatchNorm, ReLU, and dropout ($p = 0.1$), followed by a 2-layer pretraining head ($256 \to 128 \to 128$).

Ensemble: All active modality encoders are wrapped in an `nn.ModuleDict` (`EnsembleEncoders`). All projection heads follow the structure [Linear $\to$ BatchNorm $\to$ ReLU $\to$ Linear].

Contrastive Loss. We use MV-DHEL (Multi-View Decoupled Hyperspherical Energy Loss) Koromilas et al. (2025). With 2 modalities (stage 1), MV-DHEL reduces analytically to DHEL; with 4 modalities (stage 2), it uses the full MV-DHEL formulation with multiplier 4. The temperature is $\tau = 0.2$ in Stage 1 and $\tau = 0.3$ in Stage 2. All embeddings are ℓ_2-normalised before loss computation. Alignment is computed across all modality pairs simultaneously; uniformity is computed per-view via within-view similarity matrices, so that alignment and uniformity optimisation are fully decoupled.

Table 6: Hyperparameters for the two-stage training pipeline. In both stages we use SGD with a linear warmup of 10 epochs followed by cosine annealing over the remaining epochs.

Hyperparameter	**Stage 1**	**Stage 2**
Modalities	S2, S1	static, location, S2, S1
Optimiser	SGD	SGD
Learning rate	0.2	0.0016
Momentum	0.9	0.9
Weight decay	1×10^{-6}	5×10^{-6}
Scheduler	warmup-cosine	warmup-cosine
Warmup epochs	10	10
Batch size	256	1024
Epochs	100	100
Temperature τ	0.2	0.3
Loss	DHEL	MV-DHEL
Num workers	15	12

Image Preprocessing. *S2 images* are scaled by $1/10000$. A zero-filled band is inserted at position 10 (B10 placeholder) to expand from 12 to 13 channels, matching the SatCLIP pretrained backbone input. Augmentations comprise `RandomCrop(256)`, `RandomHorizontalFlip`, `RandomVerticalFlip`, and `GaussianBlur` (kernel size 3). A coordinate jitter of $\sim$1 km radius is applied to image centroid coordinates. *S1 images* are used as raw 2-channel float32 tensors (VV + VH) with no additional augmentation or normalisation.

Inference. For downstream evaluation, we extract 256-dimensional location encoder embeddings from the stage-2 final checkpoint.

Pretraining Evaluation Metrics. Alignment is measured as the mean squared ℓ_2 distance between positive pairs across all modality combinations Wang & Isola (2020). Uniformity is estimated via a Monte-Carlo estimator with 200,000 random pairs using a Gaussian kernel ($t = 2$), computed per-view in chunks of 50,000 for memory efficiency. Representation quality is assessed by the numerical rank of the embedding covariance matrix (tolerance 10^{-5}) and the effective rank based on the entropy of normalised singular values Garrido et al. (2023). All metrics are computed on held-out data every 10 epochs.

Computational Setup. Stage 1 is run on a single NVIDIA H100 GPU node with 64GB HBM2 RAM on EuroHPC with 20 CPU cores, completing in approximately 14 hours. Stage 2 is run on a single NVIDIA A100-SXM4-40GB GPU with 10 CPU cores, completing in approximately 5 hours. The total end-to-end training time is approximately 19.5 hours. We use PyTorch with float32 matrix-multiplication precision set to `high` and deterministic operations via `CUBLAS_WORKSPACE_CONFIG=:4096:8`. Reproducibility is ensured via `torch.manual_seed(42)`. Training is logged with Weights & Biases in offline mode (synced post-job). Checkpoints are saved every 10 epochs, including optimiser and scheduler states for resumability.

Code. Our implementation builds on the SatCLIP codebase Klemmer et al. (2025) for location and S2 image encoders, torchgeo Stewart et al. (2025) for S1 image encoders, and the Multi-View CL framework Koromilas et al. (2025) for MV-DHEL losses.

E Downstream tasks evaluation

We evaluate the utility of learned location representations on a downstream benchmark spanning regression, classification, and ranking tasks. The benchmark includes the original SatCLIP evaluation tasks for direct comparison with prior work, and extends them with additional environmental applications covering biodiversity modeling, land-cover characterization, long-term climate ordering, and wildfire danger forecasting. We compare against SatCLIP (Klemmer et al., 2025), GeoCLIP (Cepeda et al., 2023), GT-Loc (Shatwell et al., 2025), and geospatial foundation-model embeddings from AlphaEarth Foundations (AEF) (Brown et al., 2025) and Copernicus-FM (Wang et al., 2025).

E.1 Datasets Overview

All datasets are formulated as location-conditioned prediction tasks: each sample is associated with a latitude–longitude coordinate pair, which is encoded by the location encoder and passed to a downstream predictor. Table 7 summarizes the dataset sources, sample sizes, splits, metrics, and geographic coverage.

SatCLIP original benchmarks. For Population (Klemmer et al., 2025), Elevation (Klemmer et al., 2025), and Air Temperature (Hooker et al., 2018), we use the publicly released SatCLIP downstream datasets. The original SatCLIP Countries (Klemmer et al., 2025), Ecoregions (Dinerstein et al., 2017) and Biomes (Dinerstein et al., 2017) datasets use the same 100,000 approximately uniformly sampled global locations and include an ocean class (Klemmer et al., 2025). In our inspection, approximately 70% of these locations fell over ocean, making the ocean class dominant and causing overall accuracy to reflect ocean classification more strongly than performance across land classes. We therefore construct revised, land-only versions of these three datasets by sampling locations uniformly over the sphere and retaining only those that intersect a global land mask. For California Housing (Pace & Barry, 1997), we use the standard `scikit-learn` version of the dataset. For Median Income (Jia & Benson, 2020), we construct a county-level dataset from USDA Economic Research Service 2022 median household income estimates and match each county to a representative geographic coordinate using U.S. Census county boundaries. For iNaturalist (Van Horn et al., 2018), we follow the SatCLIP setup by pairing each observation's location with pretrained InceptionV3 (Mac Aodha et al., 2019) image features and the corresponding species label.

Additional downstream tasks. We extend the benchmark with four environmental datasets: SatBird (Teng et al., 2023), SeasFire Cube (Karasante et al., 2025), BigEarthNet V2 (reBEN) (Clasen et al., 2025), and Mesogeos (Kondylatos et al., 2023). SatBird evaluates biodiversity modeling: we use the USA-summer subset and predict bird species encounter rates from hotspot coordinates, using the official split. To evaluate long-term climate structure, we sample 10,000 global land locations from SeasFire Cube and compute the 2012–2021 temporal mean for five climate variables: 2 m temperature, total precipitation, mean sea-level pressure, 10 m wind speed, and relative humidity. We evaluate whether their relative ordering can be recovered from frozen location embeddings. For each variable separately, locations are organized into groups of five. Within each group, the locations are assigned ordinal relevance labels according to their climate mean. A LightGBM model is trained using the frozen embeddings as features. We report pairwise ordering accuracy, defined as the fraction of within-group location pairs for which the predicted score ordering agrees with the ordering of the true climate means. We use reBEN to assess land-cover classification by associating each Sentinel-2 patch with its geographic coordinate pair and using the predefined split. Finally, for wildfire danger forecasting, we use the ML-ready dataset released with Mesogeos and follow its original temporal split; each sample is represented by the latitude–longitude coordinate pair of the corresponding grid-cell centroid, and performance is evaluated with metrics appropriate for imbalanced wildfire events, namely F1-score and AUPRC.

Table 7: Summary of benchmark evaluation tasks using the location encoder. For iNaturalist, 10% of the official training split is used for validation, and the official validation split is used as the test set.

Task name	Source	Number of samples	Dataset split	Evaluation metric	Geographic coverage
Regression					
Air temperature (Hooker et al., 2018)	SatCLIP	3076	60-20-20	R^2	Global
Median income (Jia & Benson, 2020)		3219	50-20-30	R^2	Cont. USA
Cali housing (Pace & Barry, 1997)		20640	50-20-30	R^2	Calif., USA
Elevation (Klemmer et al., 2025)		100000	30-10-60	R^2	Global
Population (Klemmer et al., 2025)		74512	30-10-60	R^2	Global
SatBird – USA summer (Teng et al., 2023)	SatBird	122593	Predefined	Top-k Acc.	USA
Climate ranking					
Temperature (Karasante et al., 2025)	SeasFire	10000	Spatial	Pairwise Acc.	Global
Precipitation (Karasante et al., 2025)		10000	Spatial	Pairwise Acc.	Global
Mean sea-level pressure (Karasante et al., 2025)		10000	Spatial	Pairwise Acc.	Global
10 m wind speed (Karasante et al., 2025)		10000	Spatial	Pairwise Acc.	Global
Relative humidity (Karasante et al., 2025)		10000	Spatial	Pairwise Acc.	Global
Classification					
Countries (Klemmer et al., 2025)	SatCLIP	100000	30-10-60	weighted F1	Global
iNaturalist (Van Horn et al., 2018)		460406	Predefined	Accuracy	Global
Biome (Dinerstein et al., 2017)		100000	30-10-60	weighted F1	Global
Ecoregions (Dinerstein et al., 2017)		100000	30-10-60	weighted F1	Global
reBEN (Clasen et al., 2025)	BigEarthNet	549488	Predefined	micro-F1	Europe
Wildfire forecasting (Kondylatos et al., 2023)	Mesogeos	16242	Predefined	F1 / AUPRC	Mediterranean region

E.2 Training Details

Embedding extraction. For each downstream task, we first extract a fixed representation for every geographic coordinate pair using each evaluated encoder. We benchmark SatCLIP (Klemmer et al., 2025), GeoCLIP (Cepeda et al., 2023), GT-Loc (Shatwell et al., 2025), AlphaEarth Foundations

Table 8: Dimensionality of the fixed representations extracted from each location encoder or geospatial foundation-model embedding product.

Encoder	Embedding dimension
SatCLIP (Klemmer et al., 2025)	256
GeoCLIP (Cepeda et al., 2023)	512
GT-LOC (Shatwell et al., 2025)	512
AlphaEarth Foundations (AEF) (Brown et al., 2025)	64
Copernicus-FM (Wang et al., 2025)	768
LoCA (ours)	256

Table 9: Downstream training hyperparameters used for all linear and MLP downstream probes unless otherwise specified. The separate climate-ranking configuration is described in the text.

Hyperparameter	Value
Epochs	100
Batch size	1024
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-5}
MLP hidden dimension	128
MLP number of layers	2
Dropout	0.5
Evaluation models	Linear probe, MLP
Seeds	0, 1, 7, 11, 42, 100, 1234, 2021, 8657, 41674

(AEF) (Brown et al., 2025), and Copernicus-FM (Wang et al., 2025). The dimensionality of the extracted representation for each encoder is reported in Table 8. These embeddings are used as fixed input features for lightweight downstream predictors; the encoders themselves are not fine-tuned on any downstream task. This protocol isolates the quality of the learned geographic representations and ensures that performance differences are attributable to the embeddings rather than task-specific adaptation of the encoders.

Downstream predictors. For the regression and classification tasks and for each encoder, we train two downstream models: a linear probe and a multilayer perceptron (MLP). The MLP consists of two layers with hidden dimension 128 and dropout rate 0.5. Unless otherwise specified, downstream models are trained for 100 epochs using AdamW with learning rate 10^{-3}, weight decay 10^{-5}, and batch size 1024. We repeat all other experiments over 10 random seeds and report the mean performance across runs. The only exception is SatBird, which is trained for 20 epochs and evaluated over 3 random seeds due to its substantially higher computational cost. The five SeasFire climate tasks use a separate ranking protocol. For each climate variable, we sample groups of five locations from the corresponding train, validation, and test sets and assign ordinal labels according to the long-term climate mean within each group. We construct 4,000 training groups, 1,000 validation groups, and 1,000 test groups. A LightGBM ranker is trained on the frozen embeddings with at most 1,000 estimators, learning rate 0.05, 31 leaves, a minimum of 20 child samples, and early stopping with patience 50 using validation NDCG. The complete ranking procedure is repeated over 10 seeds. The full set of downstream hyperparameters is summarized in Table 9.

Task-specific inputs. Most downstream models use only the extracted location embedding as input, apart from iNaturalist (Van Horn et al., 2018) and wildfire danger forecasting (Kondylatos et al., 2023). For iNaturalist , we follow the SatCLIP evaluation protocol and concatenate the location embedding with pretrained InceptionV3 image features released by (Mac Aodha et al., 2019) before training the species classifier. For wildfire danger forecasting , we additionally include a cyclic encoding of the day of year to capture seasonality:

$$\sin\left(2\pi\frac{d}{365}\right), \qquad \cos\left(2\pi\frac{d}{365}\right),$$

where d denotes the day of year. These two temporal features are concatenated with the location embedding before training the downstream predictor.

Dataset splits. We use the dataset splits summarized in Table 7. For the original SatCLIP downstream tasks, we follow the splits used in the SatCLIP benchmark (Klemmer et al., 2025). For iNaturalist, we use the same training and test sets as SatCLIP and reserve 10% of the training data for validation, replicating SatCLIP's approach. For datasets with predefined splits, including SatBird (Teng et al., 2023), reBEN (Clasen et al., 2025), and Mesogeos (Kondylatos et al., 2023), we keep the official partitions. For the SeasFire-derived climate-variable tasks (Karasante et al., 2025), we use a spatially blocked train/validation/test split. Geographic coordinates are represented on the three-dimensional unit sphere and clustered spatially, after which complete geographic clusters are assigned to individual partitions.

Evaluation metrics. We use the evaluation metric associated with each downstream task, as summarized in Table 7. For Countries, Biomes, and Ecoregions, we report weighted F1; for iNaturalist, we report accuracy, and for scalar regression tasks, we report R^2. For SatBird, we report top-k retrieval accuracy because the task predicts a vector of bird species encounter rates at each hotspot; ranking metrics evaluate whether the model recovers the species most likely to be observed, which is more informative for species-distribution use cases than only measuring pointwise regression error over all species. This follows the SatBird benchmark, which reports MAE, MSE, Top-10, Top-30, and adaptive Top-k metrics, and notes that MSE can be misleading for rarely observed species because near-zero predictions can achieve low error (Teng et al., 2023). For reBEN, we report micro-F1 because the dataset is based on BigEarthNet-style multi-label land-cover classification, where each image patch may contain multiple labels and classes are imbalanced (Clasen et al., 2025). For wildfire danger forecasting, we report F1 and AUPRC, following Mesogeos (Kondylatos et al., 2023), since wildfire events are rare and the task requires evaluating the precision–recall trade-off rather than overall accuracy. For the SeasFire climate-ranking tasks, we report pairwise ordering accuracy. Within each group of five locations, every unordered pair is compared, and a prediction is correct when the ordering induced by the ranker's scores agrees with the ordering of the true long-term climate means. Ties receive half credit.

Foundation-model embeddings. For the geospatial foundation-model baselines, we extract fixed embeddings from the released spatial embedding products. For Copernicus-FM (Wang et al., 2025), we use the publicly released Copernicus-Embed-025deg grid. This product provides a global embedding map at 0.25° resolution with shape $721 \times 1440 \times 768$, where each grid cell contains a 768-dimensional embedding obtained by averaging Copernicus-FM representations across modalities. For each downstream latitude–longitude coordinate pair, we map the pair to the corresponding 0.25° grid cell and retrieve the corresponding embedding.

For AlphaEarth Foundations (AEF) (Brown et al., 2025), we use the annual embedding field for 2023 from the released Google Earth Engine dataset. AEF provides annual, analysis-ready embedding fields over Earth's terrestrial surface, with 64 embedding bands per pixel. We sample the 2023 AEF image at each downstream coordinate using Google Earth Engine. Since some coordinates may fall on masked pixels or outside the valid AEF coverage, we first attempt direct point sampling and then apply a nearest-valid-pixel fallback: for missing points, we search within increasing buffer radius up to 10 km and use the closest valid embedding if one is found. Points for which no valid AEF embedding is available within this radius are treated as missing.

For the additional environmental tasks, we were unable to retrieve AEF embeddings reliably, as a large fraction of task coordinates returned missing values under direct sampling. We also attempted

nearest-valid-pixel retrieval for these datasets, but this procedure led to repeated computational failures at the required scale. We therefore exclude AEF from the additional environmental-task comparisons and report it only for downstream tasks where embeddings could be retrieved reliably. Because AEF is released over land rather than as a complete global land–ocean embedding field, we also do not evaluate AEF on tasks that require embeddings for ocean locations.We therefore compute aggregate win counts only over available method-task pairs and do not treat missing AEF entries as failures.

Statistical significance Table 10 reports mean $\pm$ one standard deviation over downstream predictor seeds. For all tasks except SatBird, we use 10 seeds; for SatBird, we use 3 seeds because of its substantially higher computational cost. For the regression and classification tasks, the embeddings and dataset splits are fixed across seeds, so the reported variability reflects downstream predictor initialization and stochastic optimization rather than variability in the pretrained encoders or data splits. For the SeasFire climate-ranking tasks, the spatial train/validation/test split is fixed, while the ranking groups are resampled across seeds and the ranker's random state also changes.

E.3 Additional Results

E.3.1 Comparison among state-of-the-art encoders

Table 11 gives a pairwise view of the results in Table 2. For each model and task we use its best probe (the LightGBM score for climate ranking), and each pair is compared only on the tasks both models were evaluated on: 9 for pairs involving AlphaEarth, 17 otherwise. Ties at the displayed precision count for neither model, and the win rate pools a model's wins over all of its pairwise comparisons.

LoCA is the only encoder that beats every other model on a majority of shared tasks, and it has the highest win rate (68.8%). Its advantage is therefore broad. It is never outperformed on land-cover or ecological classification, and its losses concentrate on geophysical and climate targets. Its most frequent loss, air temperature, comes from the MLP probe: under the linear probe LoCA obtains the best air-temperature score of all models (Table 2). This is consistent with its high effective rank (Table 1), since location information is linearly accessible in LoCA's embeddings, whereas other encoders need a non-linear probe to reach it. The most controlled comparison is with SatCLIP, which shares LoCA's location-encoder architecture: LoCA wins 12 of these 17 tasks, which isolates the benefit of its co-aligned, high-rank multimodal target.

Beyond LoCA, the ranking follows how each encoder's target is formed. SatCLIP, which like LoCA aligns to EO imagery, is second by majority wins (4/5). The encoders anchored to models trained on natural images rank lower: GT-Loc and GeoCLIP win majorities only against each other, and Copernicus-FM wins no majority despite its 768-dimensional, imagery-derived embeddings. This supports the argument of Section 4.2 that rank obtained from an external anchor partly reflects axes of variation that are not specific to EO, and does not translate into downstream performance. AlphaEarth is LoCA's closest competitor, since LoCA wins 5 of 9 tasks.

E.3.2 Supervised co-aligned targets

Table 12 provides the per-class results for the supervised BigEarthNet experiment summarized in the main paper. Rows marked "(ours)" use the same frozen base encoders as the corresponding baseline row, but train the adapters and shared projection head with MV-DHEL.

The MV-DHEL adapter stage improves the mean F1 for every modality family. The location encoder improves from 74.7 to 76.6, S1 from 69.6 to 75.9, S2 from 78.4 to 80.9, S1+S2 from 78.1 to 80.8, and the full four-modality system from 80.9 to 84.3. The largest gain is for S1, the weakest standalone image modality, suggesting that co-aligned multimodal supervision is especially helpful when one view is individually less discriminative. The full system, All (ours), obtains the best score in every class, supporting the claim that Stage 2 can exploit an already co-aligned supervised target even without Stage 1.

E.3.3 Ablating LoCA's Multi-View Components.

The purpose of this ablation is to separate the two distinct roles played by LoCA's additional views: joint Stage 1 pretraining should construct a coherent Sentinel-1/Sentinel-2 target for Stage 2, while

Table 10: Extended results for Table 2: mean ± std over 10 seeds (3 in SatBird), metric × 100, higher is better. Both halves share one highlighting convention across all six models: **best across both probes** in green, **best per probe** in blue when not best overall, second per probe in **bold**. N/A indicates the embeddings could not be retrieved (Section E).

Task ↓ Model →	AlphaEarth		Copernicus FM		SatCLIP		Metric
	MLP	Linear	MLP	Linear	MLP	Linear	
Land cover							
reBEN	42.8 ± 0.43	43.3 ± 0.71	**57.4 ± 0.37**	**55.2 ± 0.28**	57.3 ± 0.54	**54.9 ± 0.59**	$F1_\mu$
Wildfire forecasting							
Wildfire forecasting	**80.5 ± 0.59**	70.1 ± 0.43	74.9 ± 0.16	71.0 ± 0.27	79.1 ± 0.47	**75.8 ± 0.31**	*AUPRC*
Geophysical & Socio-economic							
Air temperature	**98.2 ± 0.02**	42.8 ± 2.15	90.6 ± 0.17	**81.0 ± 2.83**	93.9 ± 0.28	80.1 ± 0.85	R^2
Median income	28.1 ± 0.69	11.0 ± 0.54	32.9 ± 1.10	26.5 ± 1.32	**50.2 ± 0.64**	**35.2 ± 1.06**	R^2
Cali. housing	54.9 ± 0.59	31.1 ± 0.14	43.3 ± 0.13	39.2 ± 0.35	63.3 ± 0.36	35.3 ± 0.21	R^2
Elevation	**97.9 ± 0.03**	**87.4 ± 0.05**	**94.1 ± 0.04**	**90.6 ± 0.18**	89.9 ± 0.10	64.9 ± 0.09	R^2
Population	80.2 ± 0.08	**69.6 ± 0.06**	77.0 ± 0.21	67.9 ± 0.13	**82.1 ± 0.14**	67.9 ± 0.09	R^2
Biodiversity & Ecology							
SatBird (USA summer)	**67.5 ± 0.05**	56.5 ± 0.04	59.6 ± 0.03	57.9 ± 0.05	60.7 ± 0.07	**58.2 ± 0.06**	*top-k Acc.*
Countries	N/A	N/A	80.9 ± 1.08	88.3 ± 0.17	**95.5 ± 0.08**	**96.5 ± 0.07**	$F1_w$
iNaturalist	49.1 ± 0.96	59.4 ± 0.20	40.4 ± 1.96	61.8 ± 0.49	**56.4 ± 0.34**	**69.8 ± 0.24**	*Acc.*
Biome	N/A	N/A	84.0 ± 0.42	82.4 ± 0.15	**89.8 ± 0.09**	**86.6 ± 0.10**	$F1_w$
Ecoregions	N/A	N/A	59.9 ± 1.14	75.1 ± 0.06	**79.7 ± 0.16**	**84.4 ± 0.14**	$F1_w$
Environmental							
Temperature	N/A		**79.8 ± 0.56**		59.6 ± 2.20		*Ranking acc.*
Precipitation	N/A		**77.7 ± 0.33**		70.6 ± 2.60		*Ranking acc.*
MSLP	N/A		73.2 ± 0.60		**83.2 ± 1.57**		*Ranking acc.*
Wind speed	N/A		**75.6 ± 0.46**		71.6 ± 2.18		*Ranking acc.*
Relative humidity	N/A		**53.5 ± 0.92**		**65.2 ± 1.62**		*Ranking acc.*

Task ↓ Model →	GeoCLIP		GT-Loc		LoCA (Ours)		Metric
	MLP	Linear	MLP	Linear	MLP	Linear	
Land cover							
reBEN	56.0 ± 0.18	44.1 ± 0.05	56.4 ± 0.35	52.4 ± 0.04	**58.3 ± 0.35**	54.7 ± 0.22	$F1_\mu$
Wildfire forecasting							
Wildfire forecasting	79.8 ± 0.25	73.2 ± 0.23	79.6 ± 0.60	75.5 ± 0.05	**80.2 ± 0.43**	**76.2 ± 0.08**	*AUPRC*
Geophysical & Socio-economic							
Air temperature	**94.3 ± 0.07**	32.9 ± 0.20	94.2 ± 0.13	60.4 ± 0.41	90.3 ± 0.44	**81.4 ± 0.68**	R^2
Median income	45.0 ± 0.44	12.1 ± 0.12	45.9 ± 0.60	31.9 ± 0.13	**51.9 ± 0.50**	**39.4 ± 0.67**	R^2
Cali. housing	**77.3 ± 0.36**	**48.4 ± 0.11**	**78.3 ± 0.21**	**66.1 ± 0.05**	65.5 ± 0.65	36.4 ± 0.12	R^2
Elevation	83.0 ± 0.27	56.0 ± 0.01	83.9 ± 0.18	62.5 ± 0.01	89.4 ± 0.11	64.4 ± 0.01	R^2
Population	78.1 ± 0.15	66.0 ± 0.02	78.5 ± 0.18	**69.7 ± 0.01**	**81.8 ± 0.11**	68.2 ± 0.02	R^2
Biodiversity & Ecology							
SatBird (USA summer)	**63.3 ± 0.06**	47.8 ± 0.11	**63.3 ± 0.01**	52.7 ± 0.01	60.8 ± 0.02	**61.1 ± 0.06**	*top-k Acc.*
Countries	91.5 ± 0.20	79.8 ± 0.02	94.7 ± 0.06	94.4 ± 0.02	**96.4 ± 0.03**	**97.0 ± 0.02**	$F1_w$
iNaturalist	44.8 ± 1.33	58.4 ± 0.24	49.3 ± 1.23	59.3 ± 0.25	**55.0 ± 0.34**	**71.4 ± 0.19**	*Acc.*
Biome	86.9 ± 0.21	75.6 ± 0.03	88.1 ± 0.11	81.9 ± 0.02	**91.0 ± 0.07**	**87.0 ± 0.04**	$F1_w$
Ecoregions	60.6 ± 0.66	42.6 ± 0.03	69.6 ± 0.30	75.9 ± 0.03	**80.2 ± 0.13**	**85.0 ± 0.13**	$F1_w$
Environmental							
Temperature	72.0 ± 0.73		74.4 ± 0.92		**77.9 ± 0.90**		*Ranking acc.*
Precipitation	72.1 ± 0.49		72.5 ± 0.63		**77.7 ± 1.02**		*Ranking acc.*
MSLP	77.8 ± 0.71		**79.1 ± 0.69**		76.3 ± 1.07		*Ranking acc.*
Wind speed	73.8 ± 0.72		71.4 ± 1.63		**75.6 ± 1.49**		*Ranking acc.*
Relative humidity	47.8 ± 0.85		43.5 ± 1.38		47.5 ± 1.41		*Ranking acc.*

Table 11: Head-to-head comparison over the 17 tasks of Table 2 (best probe per model). Cell (i, j): tasks on which the row model outperforms the column model, out of those both were evaluated on; **green** marks a majority. Ties at displayed precision count for neither model. LoCA is the only encoder that beats every other encoder on a majority of shared tasks.

Row beats column →	LoCA	SatCLIP	AlphaEarth	GT-Loc	GeoCLIP	Copernicus FM	Majority wins	Win rate
LoCA (Ours)	–	**12/17**	**5/9**	**13/17**	**12/17**	**11/17**	**5/5**	68.8%
SatCLIP	5/17	–	**5/9**	**11/17**	**10/17**	**12/17**	4/5	55.8%
AlphaEarth	4/9	4/9	–	**6/9**	**6/9**	**6/9**	3/5	57.8%
GT-Loc	4/17	6/17	3/9	–	**12/17**	**10/17**	2/5	45.5%
GeoCLIP	5/17	7/17	3/9	4/17	–	**9/17**	1/5	36.4%
Copernicus FM	4/17	5/17	3/9	7/17	8/17	–	0/5	35.1%

Table 12: Per-class F1 on BigEarthNet land-cover classification (10 classes), comparing pure location encoders (rows 1–2) against supervised ResNet-18 image encoders with and without the MV-DHEL multimodal training. "Mean F1" is the unweighted mean across the 10 class-wise F1 scores; remaining columns report class-wise F1. **Best per column** is highlighted in **green** and **second best** in **bold**.

Model	Mean F1	Arable Land	Mixed Forest	Coniferous Forest	Transitional Woodland	Broad-leaved Forest	Urban Fabric	Inland Waters	Complex Cultivation	Marine Waters	Pastures
Location	74.7	71.0	77.1	80.4	67.5	73.8	65.8	72.8	69.6	89.3	79.6
Location (ours)	76.6	71.4	79.8	83.8	68.5	77.6	63.0	73.0	70.7	**97.8**	80.6
S1	69.6	72.9	70.0	73.1	63.8	64.0	75.9	78.4	52.5	95.2	50.1
S1 (ours)	75.9	76.1	74.6	76.7	67.0	70.2	**78.0**	84.7	64.8	95.6	70.9
S2	78.4	78.5	78.4	84.5	68.9	76.7	73.8	82.5	67.9	96.1	76.6
S2 (ours)	**80.9**	**81.7**	**81.7**	**86.9**	**71.4**	79.1	75.1	83.5	71.4	96.7	**81.2**
S1+S2	78.1	78.0	78.4	83.7	67.8	76.6	72.2	83.5	67.9	97.2	76.1
S1+S2 (ours)	80.8	**81.7**	**81.7**	86.4	70.9	78.9	74.3	83.6	**72.0**	97.3	**81.2**
All	**80.9**	81.0	80.9	85.8	70.6	**79.2**	76.9	**85.5**	71.3	**97.8**	79.8
All (ours)	**84.3**	**83.9**	**83.6**	**88.7**	**73.8**	**83.5**	**80.3**	**89.0**	**77.0**	**99.3**	**83.9**

the metadata encoder should add complementary non-image signal only on tasks where such signal is useful. Table 13 reports the complete 17-task version of the ablation.

The main failure mode is the removal of joint Stage 1 pretraining: when the Sentinel-1 and Sentinel-2 targets are replaced by independently pretrained MoCo encoders, Stage 2 no longer receives a co-aligned multi-view target, and performance collapses on nearly all regression tasks and is substantially worse on classification tasks. For example, reBEN drops from 54.7 to 1.1 micro-F1, air-temperature prediction from 81.4 to -5.8 R^2, and Countries from 97.0 to 4.4 weighted F1. The effect is particularly clear on the climate-ranking tasks: pairwise accuracy lies between 49.6% and 50.6% across all five variables, i.e., approximately chance-level ordering. This supports the role of *Stage 1* in constructing a coherent multimodal target before adapting the location encoder and re-enforces the fact that SatCLIP cannot be trivially extended to multiple modalities

Removing metadata gives a more selective effect, which is the expected behavior. The metadata view is not expected to help uniformly across all tasks: when Sentinel-1, Sentinel-2, or coarse geography already capture the relevant signal, adding metadata should have little effect. Metadata improves the full model over the metadata-free variant on 9 of the 17 tasks. The largest gains are observed for several geophysical and climate-ranking tasks, including air temperature (76.0 to 81.4), elevation (57.1 to 64.4), MSLP ranking (70.0% to 76.3%), and wind-speed ranking (71.2% to 75.6%).

Overall, the ablation indicates two qualitatively different effects. *Stage 1* is necessary for obtaining a useful multimodal location representation, with its removal causing broad performance collapse and chance-level climate ranking. The metadata view instead contributes task-dependent complementary information: it improves several targets, but does not provide a uniform gain across all downstream tasks.

F Extended Related Work

F.1 Geospatial Foundation Models (GFMs) & Embeddings.

Several GFMs have relied on Masked Image Modelling (MIM) objectives as their pretraining strategy. ScaleMAE Reed et al. (2023) and Cross-Scale MAE Tang et al. (2023) utilize MIM with GSD-aware

Table 13: Ablation isolating the contributions of (i) the joint *Stage-1* contrastive pretraining and (ii) the metadata encoder, under the linear probe. Values are metric $\times$ 100; higher is better. Highlighting follows Table 2. Without *Stage 1* the embeddings collapse: downstream probes drop to (or below) the trivial predictor and environmental ranking is at chance (50). All variants are evaluated on all 17 tasks.

Task ↓ Variant →	**LoCA** w/o *Stage 1*	**LoCA** w/o metadata	**LoCA** (full)	**Metric**
Biodiversity & Ecology				
Countries	4.4	**96.9**	**97.0**	$F1_w$
Biome	9.7	**85.8**	**87.0**	$F1_w$
Ecoregions	1.7	**85.4**	**85.0**	$F1_w$
iNaturalist	56.6	**71.6**	**71.4**	*Acc.*
SatBird (USA summer)	44.5	**61.2**	**61.1**	*top-k Acc.*
Land cover				
reBEN	1.1	**55.2**	**54.7**	$F1_\mu$
Wildfire forecasting				
Wildfire forecasting	49.5	**76.5**	**76.2**	*AUPRC*
Geophysical & Socio-economic				
Median income	-7.9	**36.5**	**39.4**	R^2
Population	-0.8	**66.0**	**68.2**	R^2
Cali. housing	-1.5	**36.8**	**36.4**	R^2
Air temperature	-5.8	**76.0**	**81.4**	R^2
Elevation	-0.7	**57.1**	**64.4**	R^2
Environmental	**LightGBM ranking**			
Temperature	49.9	**84.4**	**77.9**	*Ranking acc.*
Precipitation	49.6	**75.9**	**77.7**	*Ranking acc.*
MSLP	49.9	**70.0**	**76.3**	*Ranking acc.*
Wind speed	50.6	**71.2**	**75.6**	*Ranking acc.*
Relative humidity	**49.8**	47.1	**47.5**	*Ranking acc.*
# Best (out of 17)	1	7	**9**	

positional encoding and cross-prediction for multi-scale learning. USatMAE Irvin et al. (2023) and SpectralMAE Zhu et al. (2023) introduce spectral-aware masking for paired NAIP/Sentinel-2 and hyperspectral data respectively. Extending the MIM paradigm, Presto Tseng et al. (2023), FoMo-Net Bountos et al. (2025), DOFA Xiong et al. (2024), Copernicus-FM Wang et al. (2025), OlmoEarth Herzog et al. (2026), THOR Forgaard et al. (2026) modify the MIM framework to train sensor agnostic encoders. Finally, Galileo Tseng et al. (2025) blends MIM with contrastive objectives introducing a dual global/local contrastive loss for multi-scale feature learning. Following a different direction, SkySense Guo et al. (2024) aligns WorldView RGB, Sentinel-2, and Sentinel-1 through contrastive and cross-modal objectives. In a similar fashion, CROMA Fuller et al. (2023) and OmniSAT Astruc et al. (2024) fuse Sentinel-1/2 via contrastive and reconstruction losses, with CROMA limited to single timestamps and OmniSAT extending to time series and aerial data. Recent works have explored generative approaches for the creation of GFMs TerraMind Jakubik et al. (2025) is the first any-to-any generative multimodal EO model, pretrained autoregressively on dual-scale token- and pixel-level representations across nine modalities. DiffusionSat Khanna et al. (2024) conditions on geolocation metadata for temporal generation, super-resolution, and in-painting. SatDiFuser Jia et al. (2025) demonstrates the potential of diffusion-based approaches for the creation of GFMs. Driven by the increasing capabilities of GFMs, recent works have utilized them to store precomputed location embeddings, promising to replace raw observations altogether. AlphaEarth Brown et al. (2025) fuses optical, radar, LiDAR, and climate reanalysis data into 64-dimensional embeddings at 10m resolution, released annually. Copernicus-FM Wang et al. (2025) provides a global embedding map at $0.25°$ by averaging multi-sensor representations. TESSERA **?** applies a Barlow Twins pretraining framework, releasing annual embeddings with *int8*

quantization. However, the pretraining schemes underlying these models are observation-based rather than location-based, yielding degenerate representations that undermine their stated purpose.

F.2 Location Encoders.

Location encoders follow a different paradigm, aiming to learn location-specific representations beyond single observations, identifying the key characteristics and conditions that govern and connect geospatial locations. In Mac Aodha et al. (2019) the authors introduced an early location encoder training a fully connected residual network (FCNet) to estimate the probability of object categories at a given location and time. Sphere2Vec Mai et al. (2023b) proposed a multi-scale encoding scheme based on Double Fourier Sphere (DFS) features that preserves geodesic distances on the spherical surface. In SirenNet Rußwurm et al. (2024) the authors introduced spherical harmonic basis functions as positional encoding which they utilized in a sinusoidal representation networks. Contrastive learning has emerged as the dominant paradigm for training location encoders at scale. GeoCLIP Cepeda et al. (2023) aligns CLIP-based image embeddings with GPS coordinates. SatCLIP Klemmer et al. (2025) extends the contrastive paradigm to satellite imagery, training a location encoder against Sentinel-2, frozen image encoders to produce general-purpose location embeddings. Recent approaches have moved beyond unimodal location encoders.GAIR Liu et al. (2026) jointly aligns remote sensing imagery, street-view photographs, and geolocation through two contrastive losses: one aligning RS with street-view embeddings and another anchoring each visual modality to the location representation. This pairwise formulation, however, does not scale gracefully: each additional modality requires new loss terms that increasingly distort the shared embedding space. GT-Loc Shatwell et al. (2025) incorporates temporal context by mapping images, locations, and timestamps into a joint embedding space, pairing a frozen pretrained image encoder with learned location and time encoders optimized via a GeoCLIP-style loss for image–location alignment and a cross-entropy loss for image–time alignment. While effective for three modalities, the heterogeneous loss design again constrains extensibility. TIGeR Shatwell et al. (2026) addresses pairwise completeness by aligning image–location–time triplets through six InfoNCE losses covering all pairwise combinations. This is suboptimal for two reasons: a) the competing objectives produce conflicting gradients that hinder the optimization process, and b) the number of required losses grows quadratically with the number of modalities, making further extension impractical. We address these limitations by generalizing the SatCLIP framework to arbitrarily many modalities through a single multi-view contrastive objective.

F.3 Multi-View Contrastive Learning.

The effectiveness of leveraging multiple views in CL has been demonstrated by aggregating pairwise losses from views generated through various means, such as multiple projection heads Wang et al. (2024), image scales Li et al. (2023), classes Shah et al. (2023), and feature levels Xu et al. (2022). However, the way to efficiently utilize multiple views in Contrastive Learning has been indipendently studied. An early approach by Tian et al. (2020) framed multi-view learning as a collection of independent tasks, treating each combination of views separately and simply aggregating pairwise losses. Recently, Shidani et al. (2024) has incorporated another aggregation method, while utilizing optimal transport on a high similarity tensor was proposed by Piran et al. (2024). The MV-DHEL (Koromilas et al., 2025) objective was recently proposed and is able to properly utilize all view interactions resulting to consistent performance gains that scale with the number of views and embeddings that avoid dimensionality collapse. In this work we draw from the MV-DHEL benefits to train multimodal location encoders.

G Licenses and terms for existing assets

We use existing datasets, models, and code only for research purposes and credit the original creators through the citations in the main paper and appendix. Table 14 summarizes the licenses and terms of use for the external assets used for pretraining, metadata construction, baselines, and downstream evaluation. We do not redistribute third-party raw imagery, pretrained baseline checkpoints, or third-party benchmark datasets unless their license explicitly permits redistribution.

Table 14: Existing assets used in this work and their licenses or terms of use. "Use in this paper" describes how the asset is used; it does not imply that we redistribute the original asset.

Asset	Use in this paper	License / terms
SatCLIP S2-100K dataset	Starting point for coordinate sampling and Sentinel-2 pairing.	Research purposes only; license field not specified in the dataset card.
SatCLIP code and pretrained models	Baseline location encoder; architecture reference.	MIT License.
Copernicus DEM GLO-30	Elevation, slope, aspect, and curvature metadata.	Copernicus DEM free license.
WorldPop population density	Population-density metadata.	Creative Commons Attribution 4.0 International (CC BY 4.0).
WorldPop distance-to-roads product	Road-distance metadata.	Creative Commons Attribution 4.0 International (CC BY 4.0).
OpenStreetMap	Underlying source for road-distance data.	Open Data Commons Open Database License 1.0 (ODbL 1.0).
C3S / ESA CCI Land Cover	Land-cover fraction metadata.	Dataset page lists ESA CCI licence, CC-BY licence, and VITO licence.
AlphaEarth Foundations satellite embedding dataset	Observation-based embedding-field baseline.	Creative Commons Attribution 4.0 International (CC BY 4.0).
Copernicus-FM and Copernicus-Embed-025deg	Observation-based embedding-field baseline.	Code: Apache-2.0 with third-party MIT/CC-BY-NC-4.0 components; pretrained weights and datasets: CC BY 4.0; Copernicus-Embed-025deg: CC BY 4.0.
GeoCLIP	Baseline location encoder.	MIT License.
GT-Loc	Baseline location encoder.	Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).
SatBird	Biodiversity downstream benchmark.	Repository license: GPL-3.0.
BigEarthNet v2.0 / reBEN	Land-cover downstream benchmark; supervised encoder experiment.	Community Data License Agreement – Permissive, Version 1.0 (CDLA-Permissive-1.0).
SeasFire Cube	Long-term climate-ranking downstream tasks.	Creative Commons Attribution 4.0 International (CC BY 4.0).
Mesogeos	Wildfire danger forecasting downstream task.	Creative Commons Attribution 4.0 International (CC BY 4.0).
iNaturalist / iNat competition data	Species-prediction downstream task.	iNaturalist competition terms: non-commercial research and educational use; no redistribution of images.
RESOLVE Ecoregions 2017	Countries/biome/ecoregion-style ecological labels and evaluation resources.	Creative Commons Attribution 4.0 International (CC BY 4.0).
California Housing via scikit-learn	Downstream regression benchmark.	scikit-learn is BSD-3-Clause; original dataset is from StatLib.
USDA ERS median household income data	Median-income downstream regression benchmark.	U.S. Government public-domain information / USDA website terms.
U.S. Census TIGER/Line county boundaries	County-coordinate construction for median-income task.	U.S. Census Bureau public geographic data.