ISPRS Journal of Photogrammetry and Remote Sensing 236 (2026) 99–119

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

TimeSenCLIP: A time series vision–language model for remote sensing

Pallavi Jain [a,b,f,*], Diego Marcos [b,f], Dino Ienco [b,c,e,f], Roberto Interdonato [b,d,e,f], Tristan Berchoux [a,e,f]

[a] Mediterranean Agronomic Institute of Montpellier - CIHEAM-IAMM, Montpellier, France
[b] Inria, Montpellier, France
[c] INRAE, Montpellier, France
[d] Cirad, Montpellier, France
[e] UMR TETIS, Montpellier, France
[f] University of Montpellier, Montpellier, France

ARTICLE INFO

Keywords:
VLMs
Time series
Multispectral
Remote sensing
Sentinel-2
Cross-view
Contrastive learning

ABSTRACT

Vision–language models (VLMs) have shown significant promise in remote sensing applications, particularly for land-use and land-cover (LULC) mapping via zero-shot classification and retrieval. However, current approaches face several key challenges, such as the dependence on caption-based supervision, which is often not available or very limited in terms of the covered semantics, and the fact of being adapted from generic VLM architectures that are suitable for very high resolution images. Consequently, these models tend to prioritize spatial context over spectral and temporal information, limiting their effectiveness for medium-resolution remote sensing imagery.

In this work, we present TimeSenCLIP, a lightweight VLM for remote sensing time series, using a cross-view temporal contrastive framework to align multispectral Sentinel-2 time series with geo-tagged ground-level imagery, without requiring textual annotations. Unlike prior VLMs, TimeSenCLIP emphasizes temporal and spectral signals over spatial context, investigating whether single-pixel time series contain sufficient information for solving a variety of tasks.

Our approach is trained on the LUCAS and Sen4Map datasets and evaluated across four main mapping tasks: land cover, land use, habitat mapping and crop type classification. The CLIP text encoder can be used to probe the learned representations using semantically meaningful categories, enabling effective zero-shot generalization without task-specific text supervision. We further extend our evaluation to bioregions mapping and country-level image retrieval. Although coarse, these tasks are valuable for probing whether the model captures geographically meaningful representations, such as regional climate regimes, vegetation patterns, and land-use structures. TimeSenCLIP achieves consistently better performance than existing CLIP-based remote sensing models in both zero-shot classification and cross-modal retrieval. Notably, single-pixel multispectral time series variants remain highly competitive, particularly with extended temporal coverage, demonstrating that temporal–spectral dynamics can compensate to a substantial degree for the reduced spatial footprint.

While larger spatial patches still offer advantages for tasks where spatial patterns are inherently informative, such as ecosystem type classification, the results suggest that single-pixel multispectral time series can provide effective remote sensing vision–language pipelines, supporting scalable and efficient modeling in scenarios where large spatial tiles or extensive textual annotations are impractical. Code is available at https://github.com/pallavijain-pj/TimeSenCLIP.

1. Introduction

Remote sensing technology is central to large-scale environmental monitoring, providing critical insights into land cover change, ecosystem health, biodiversity assessment, and agricultural productivity (Li et al., 2014; Soubry et al., 2021). The use of machine learning for the enhanced analysis of satellite data has enabled increasingly automated and accurate classification systems. However, the challenge of scaling out across diverse ecosystems, sensing modalities, and geographical

* Corresponding author at: Mediterranean Agronomic Institute of Montpellier - CIHEAM-IAMM, Montpellier, France.
E-mail addresses: pallavi.jain@inria.fr (P. Jain), diego.marcos@inria.fr (D. Marcos), dino.ienco@inria.fr (D. Ienco), roberto.interdonato@inria.fr (R. Interdonato), berchoux@iamm.fr (T. Berchoux).

https://doi.org/10.1016/j.isprsjprs.2026.03.043
Received 14 August 2025; Received in revised form 20 February 2026; Accepted 26 March 2026

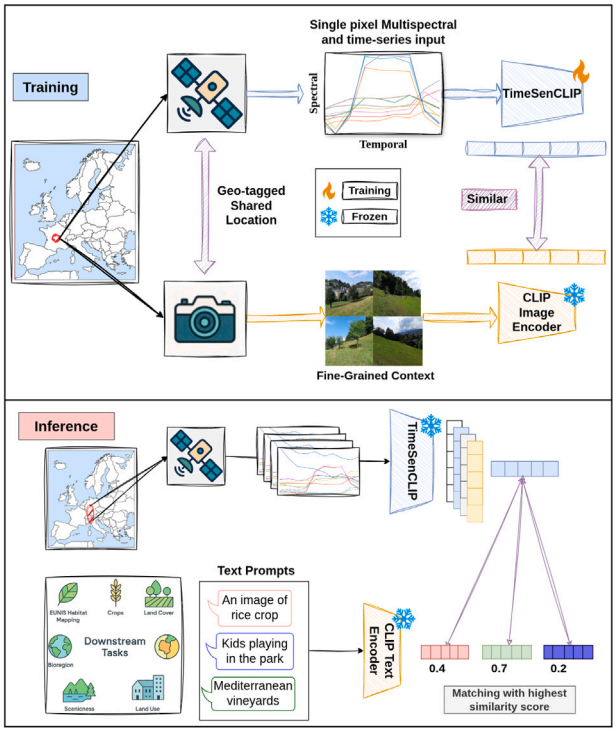


Fig. 1. TimeSenCLIP Pipeline Illustration: Satellite Sentinel-2 single pixel multispectral time series are aligned with geo-tagged ground-level images through cross-view learning, enabling the model to capture fine-grained ecological semantics without relying on large spatial context or text supervision.

regions remains largely unresolved, particularly when computational efficiency and generalization to unseen classes are required (Reichstein et al., 2019; Zhu et al., 2017; Qiu et al., 2024; Tan et al., 2024).

Recent advances in vision–language models (VLMs) have opened new possibilities in this domain. Models like CLIP (Radford et al., 2021), which learn aligned embeddings between images and natural language, have shown remarkable zero-shot generalization across a wide range of vision tasks such as object recognition, semantic segmentation, image captioning, and cross-modal retrieval (Tschannen et al., 2025). When adapted to remote sensing, VLMs allows for open-vocabulary classification and text-driven retrieval of overhead imagery data using natural language queries (e.g., "arid farmland with sparse vegetation" or "oil palm plantations near waterways"), significantly reducing the need for exhaustive, class-specific supervision (Dhakal et al., 2023; Jain et al., 2025; Mall et al., 2023; Vivanco Cepeda et al., 2023; Klemmer et al., 2025).

This capability is particularly appealing for ecosystem mapping, which extends beyond traditional land-use and land-cover (LULC) classification (e.g., "urban" or "cropland"), by combining diverse biophysical information to produce thematic maps that depict the spatial distribution, scale, and ecological linkages of natural systems within a given landscape (e.g., for mapping biodiversity, habitat, and ecological regions) (Keith et al., 2022; Lausch et al., 2016). Such distinctions are often fine-grained, context-dependent, and not well captured by fixed-class LULC taxonomies. Addressing this complexity typically requires either labor-intensive expert annotation or flexible models capable of adapting to unseen descriptions. VLMs bridge this gap by interpreting open-ended text prompts to identify nuanced classes without retraining (eg."Intensively managed broadleaf forest dominated by Fagus sylvatica" versus simply "forest") (Jain et al., 2025).

Despite these advantages, extending VLMs to remote sensing data faces several key challenges. **First**, many existing remote-sensing VLMs depend heavily on text-labeled training data or curated caption datasets. Such annotations are costly, exhibit vocabulary bias, and remain insufficient for fine ecosystem granularity, regional habitat terminology, or visually subtle classes. As a result, current text-supervised models often fail to generalize beyond the concepts in their rigid text training set. **Second**, most remote sensing VLMs are adapted from generic VLMs and adopt large spatial input images (e.g., 200 × 200 to 512 × 512 px) (Zhang et al., 2016; Zhu et al., 2017), rendering them more suitable for very high resolution imagery than medium resolution like Sentinel-2. While spatial context is often valuable, a high pixel count with medium resolution imagery results in a very large context that can be unreliable in fragmented or heterogeneous landscapes, particularly in ecological settings, where cloud contamination, mixed pixels, seasonal transitions (Koldasbayeva et al., 2024) or weak spatial autocorrelation reduce the discriminative power of spatial context. **Third**, a large body of remote sensing research shows that, at the medium spatial resolution such as Sentinel-2, many ecological, agricultural, and phenological classes are more distinctly characterized by multispectral and temporal signatures than by spatial patterns (Zhong et al., 2019; Rußwurm and Körner, 2018; Lausch et al., 2016; Pettorelli et al., 2016; Pesaresi et al., 2022). Yet, most VLM adaptations rely primarily on static RGB representations, leaving spectral and temporal information underexploited. **Finally**, the acquisition, computational and storage costs associated with high-resolution patches, multimodal fusion, and text-based supervision limit the practicality of VLMs for large-scale or long-term monitoring.

To address these limitations, we propose TimeSenCLIP, a lightweight model, that adapts the VLM paradigm to operate without unnecessary spatial context and without the need for additional text annotations. We specifically investigate whether the spectral–temporal signature of a single Sentinel-2 pixel contains sufficient semantic information to support accurate land-use, land-cover, and ecosystem classification. This choice directly challenges the assumption that large spatial neighborhoods are necessary, emphasizing the discriminative potential of spectral–temporal dynamics.

Rather than training with text labels or captions, TimeSenCLIP adopts a cross-view learning paradigm as shown in Fig. 1 (top), that aligns satellite image time series with geo-tagged ground-level photographs. This type approach can leverage any large collection of geotagged photos, such user generated ones on Flickr, used by Sat2Cap (Dhakal et al., 2023) and GRAFT (Mall et al., 2023), or the large scale LUCAS survey used by SenCLIP (Jain et al., 2025). This setup helps reduce spurious correlations that can arise from caption-based training, due to the limitations of building a caption dataset, and encourages the model to capture more fine-grained ecological signals, such as vegetation structure, land use intensity, or habitat composition. At inference time, satellite time series embeddings produced by TimeSenCLIP can be compared with textual descriptions of the classes of interest, as seen in Fig. 1 (bottom). By leveraging the rich language understanding of CLIP through ground-level images, TimeSenCLIP can go beyond generic text prompts based on class names or conventional aerial- or satellite-view textual prompts (e.g., *"a centered satellite image of a residential building"* or *"an aerial view of a plantation"*), enabling generalization to nuanced descriptions of land-use and ecosystem type from a ground-level perspective.

Unlike prior work, TimeSenCLIP explicitly incorporates Sentinel-2's multispectral and temporal dimensions, demonstrating their usefulness for tasks such as crop mapping, bioregion delineation, and habitat mapping. Because temporal–spectral data already provides the necessary signal to solve these tasks, we use a single-pixel time series as input to keep the model lightweight and computationally efficient. This potentially entails a mismatch between the objects depicted by each modality, with ground-level images capturing details of objects in the scene and a large spatial context, albeit in a single temporal moment, while Sentinel-2 time series capture the dynamics of the average reflectance of a 100 m^2 surface. However, our results suggest that both these different views of a location can be aligned meaningfully and are able to capture the essential land-use characteristics.

Our main contributions are:

1. We introduce TimeSenCLIP, a cross-view learning approach aligning satellite spectral–temporal observations with ground-level semantics using ground-level photos and a text-aligned vision model, eliminating dependence on caption-based supervision.
2. We investigate the role of temporal modeling in text-aligned satellite representations, disentangling the contributions of temporal, spectral, and spatial information and showing that temporal–spectral dynamics are critical for fine-grained ecological and land-use understanding.
3. We provide a comprehensive assessment of our framework across diverse tasks, including LULC classification, habitat mapping, bioregion mapping, crop type identification, and scenicness estimation, showing consistent zero-shot performance while maintaining computational efficiency through single-pixel temporal modeling.

Together, these contributions highlight the practical potential of temporally and spectrally informed vision–language models for scalable, fine-grained zero-shot mapping in remote sensing.

2. Related works

Our approach builds upon recent advances in vision–language alignment, remote sensing foundation models, and multivariate time series learning, but departs from prior work by leveraging satellite time series and ground level imagery alignment, without requiring paired-text supervision.

2.1. VLMs in remote sensing

VLMs such as CLIP (Radford et al., 2021) have significantly advanced zero-shot performance by projecting images and text into a shared embedding space, enabling flexible and scalable classification across diverse domains. Inspired by these advances, recent efforts have adapted VLMs to the remote sensing domain. Notable examples include GeoRSCLIP (Zhang et al., 2023), SkyCLIP (Wang et al., 2024), and RemoteCLIP (Liu et al., 2023), which leverage satellite imagery and natural language to perform open-vocabulary land cover classification. These models demonstrate strong generalization by pairing large-scale remote sensing imagery with curated textual descriptions. Beyond classification, models like GeoPixel (Shabbir et al., 2025) extend vision–language alignment to the pixel and patch level, enabling fine-grained captioning and grounding over high-resolution Earth observation data. However, such models often assume that semantic understanding can only emerge from larger spatial contexts. This is reflected in the use of large image patches (e.g., 224 × 224 pixels), which implicitly rely on local spatial continuity for semantic inference. While effective in many structured environments, this assumption becomes problematic in ecologically diverse or fragmented landscapes, such as mixed-crop, hedgerow mosaics, agroforestry systems, seasonal floodplain wetlands, or shrub-steppe regions, where spatial coherence is weak, discontinuous, or even misleading. A second limitation lies in their reliance on paired textual description supervision; training typically depends on manually curated image–text datasets (e.g., RS5M (Zhang et al., 2023)), which are costly to scale and often suffer from vocabulary bias, especially in underrepresented ecological contexts.

Our work challenges these assumptions by exploring whether rich semantics can emerge from single pixel represented solely by their spectral–temporal signatures. We demonstrate that, without relying on large spatial context or curated text prompts, small inputs encode sufficient information to support ecosystem retrieval and alignment.

2.2. Cross-view ground–satellite alignment

Cross-view supervision has emerged as a compelling strategy for learning geospatial representations by aligning satellite information and ground-level imagery. Models such as Sat2Cap (Dhakal et al., 2023), SenCLIP (Jain et al., 2025), and GRAFT (Mall et al., 2023) leverage contrastive or generative objectives to bridge the domain gap between views. SenCLIP, in particular, introduces an attention pooling mechanism across directional ground-level images, processed via a frozen CLIP encoder and aligns them with Sentinel-2 inputs, inspiring the ground encoder design in our model. Similarly, GAIR (Liu et al., 2025) applies hierarchical fusion of satellite and ground imagery to learn coherent geospatial embeddings for downstream tasks.

In contrast to these approaches, we propose to align spectral–temporal pixel from Sentinel-2 with ground-level features through direct contrastive learning, avoiding any reliance on spatial priors, or language-based supervision. This enables learning from extremely small spatial inputs (as little as single pixel), while retaining semantic richness from high-resolution ground imagery. Our model thereby offers a supervision efficient alternative for geospatial representation learning, applicable across ecological, agricultural, and other land-use domains.

2.3. Multispectral and temporal remote sensing models

Temporal modeling is a cornerstone of remote sensing research, particularly for applications such as ecological and agricultural applications. Convolution-based architectures like TempCNN (Pelletier et al., 2019) and hybrid CNN–attention models such as L-TAE (Sainte Fare Garnot and Landrieu, 2020) and ConvTran (Foumani et al., 2024) have demonstrated strong performance on time series satellite data. However, their effectiveness often depends on extensive preprocessing and regularly sampled time series.

Recent advances in spectral–temporal VLMs for remote sensing have further explored text alignment using such models. Notable examples include Llama3-MS-CLIP, which extends RGB inputs to multispectral patches (Marimo et al., 2025); GeoLLAVA, which treats time series as video–language pairs (Elgendy et al., 2024); and EarthDial, which employs a two-stage RGB-to-multispectral and temporal fine-tuning scheme (Soni et al., 2024). While these models represent important progress towards truly spatio-spectral VLMs for Earth observation, they remain limited by their reliance on RGB priors, fixed temporal windows, or the need for extensive supervision. In contrast, our transformer-based framework directly processes raw spectral–temporal cubes, jointly modeling multispectral (10 Sentinel-2 bands) and multi-temporal information spatial upsampling, or multi-stage pretraining.

2.4. Ecological representation and remote sensing foundations

Recent work has advanced ecological representation learning by aligning remote sensing imagery with external ecological knowledge, including it in the form of text. For example, EcoWiKiRS (Zermatten et al., 2025) and WildSAT (Daroya et al., 2024) leverage species occurrence data and habitat preferences, while TaxaBind (Sastry et al., 2025) aligns a variety of ecologically relevant modalities, including satellite imagery. Other efforts focus on multi-sensor remote sensing data fusion for ecosystem modeling (Wang et al., 2025).

Our approach contributes an alternative, without requiring class labels, taxonomies, or curated prompts, by aligning Sentinel-2 spectral–temporal pixel with image-level supervision from ground-level imagery. It avoids spatial priors or dense annotations (e.g., maps), and instead learns directly from raw satellite image time series pixels through cross-view supervision.

This setup enables semantic representations to emerge from spectral–temporal structure alone, supporting a broader shift toward scalable, data-driven ecosystem modeling in remote sensing VLMs.

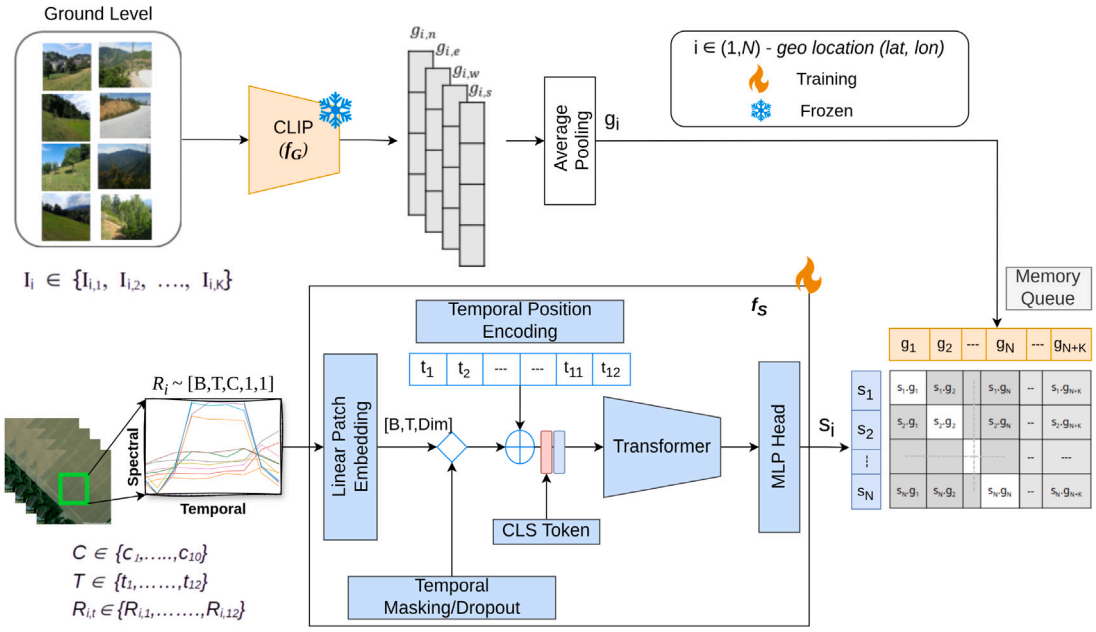


Fig. 2. TimeSenCLIP Model Training: Spectral–temporal patches from Sentinel-2 are aligned with ground-level CLIP features using contrastive learning. The satellite encoder learns from minimal spatial input via a transformer, while a memory queue enables efficient negative sampling.

3. Method

Our approach, TimeSenCLIP (Temporal VLM for Sentinel-2 imagery), illustrated in Fig. 2, builds on the prior satellite–ground alignment works of SenCLIP (Jain et al., 2025). SenCLIP leverages CLIP-based encoders for both ground-level and satellite modalities, making it suitable only for RGB imagery.

With TimeSenCLIP, we use a frozen CLIP image encoder only for the ground-level modality and train a multivariate time series model from scratch for the satellite modality, bridging spectral–temporal satellite observations with semantically rich ground-level representations.

Overview. During training, our approach makes use of two main modules:

- Ground-level encoder: The frozen CLIP image encoder provides semantically rich embeddings that act as a text-aligned proxy target. Multiple directional images are aggregated via attention pooling to focus on the most informative perspectives. This module is only used for training.
- Satellite encoder: A transformer-based architecture captures long-range spectral–temporal dependencies, accommodating monthly, quarterly, or annual time series. Via contrastive training, this model learns to provide a text-aligned representation of the time series. The resulting model is the one used at inference in order to perform zero-shot classification of Sentinel-2 time series.

Our design is guided by two main objectives: (1) to leverage the CLIP encoder as a semantic anchor, transferring knowledge from natural image–language domains into remote sensing; and (2) to capture spectral–temporal dependencies in Sentinel-2 data using a transformer-based encoder capable of handling monthly, quarterly and annual time series.

3.1. CLIP overview

CLIP (Contrastive Language–Image Pretraining) (Radford et al., 2021) is a large-scale vision–language model trained on hundreds of millions of image–text pairs. It learns to align images and their corresponding textual descriptions in a shared embedding space using a contrastive objective, such that matching image–text pairs have high cosine similarity and non-matching pairs have low similarity.

Formally, for sample i, let $\mathbf{v}_i$ and $\mathbf{t}_i$ denote the L2-normalized image and text embeddings from image encoder f_{img} and text encoder f_{text}:

$$\mathbf{v}_i = \frac{f_{\text{img}}(I_i)}{\|f_{\text{img}}(I_i)\|_2}, \quad \mathbf{t}_i = \frac{f_{\text{text}}(T_i)}{\|f_{\text{text}}(T_i)\|_2}. \tag{1}$$

The cosine similarity between embeddings is then the dot product $\mathbf{v}_i \cdot \mathbf{t}_j$.

Maximizing similarity for matched pairs and minimizing it for non-matched pairs produces a semantically and visually rich latent space that generalizes across natural and human-made scenes.

This is done using the InfoNCE loss, which is formulated as:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_j^-\}) = -\log \frac{\exp(\mathbf{x}\cdot\mathbf{x}^+/\tau)}{\exp(\mathbf{x}\cdot\mathbf{x}^+/\tau) + \sum_{j=1}^{J}\exp(\mathbf{x}\cdot\mathbf{x}_j^-/\tau)}, \tag{2}$$

where $\mathbf{x}$ is the anchor sample, $\mathbf{x}^+$ a positive sample, possibly from a different modality, and $\{\mathbf{x}_j\}$ represents a set of negative samples, and τ is a learnable temperature controlling the concentration of the similarity distribution. In CLIP, where both an image and a text models are learned jointly, the total loss would be $\mathcal{L}_{\text{CLIP}} = \sum_i \mathcal{L}_{\text{InfoNCE}}(\mathbf{v}_i, \mathbf{t}_i^+, \{\mathbf{t}_j^-\}_i) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{t}_i, \mathbf{v}_i^+, \{\mathbf{v}_j^-\}_i)$

3.2. Model architecture

During training, TimeSenCLIP makes use of an image encoder and a time-series encoder: a frozen CLIP image encoder f_G for the ground-level images and a trainable, transformer-based, satellite time-series encoder f_S. After training, only f_S is retained, and it can be used together with the CLIP text encoder f_T in order to infer the alignment between a satellite time series and any textual prompt.

Ground-level encoder (CLIP). For each geographic location, multiple ground-level images can be available: $\{I_1, \ldots, I_K\}$. These K images are individually processed by the pretrained CLIP image encoder to produce the corresponding feature embeddings $\{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$. These embeddings are then aggregated via an average pooling to yield a single ground-level descriptor.

Satellite encoder. For the **Sentinel-2 time series**, we design a transformer-based model $f_S(\cdot)$ that takes as input a spectral–temporal cube $\mathbf{R} \in \mathbb{R}^{T\times C\times H\times W}$, where T denotes the number of temporal frames, C is the number of spectral bands, and $H \times W$ is the spatial extent. In the single-pixel setting, $(H = 1, W = 1)$, whereas larger spatial dimensions can be used to incorporate additional spatial context. Each time slice $\mathbf{R}_t \in \mathbb{R}^{C\times H\times W}$ is flattened and projected with a linear layer:

$$\mathbf{r}_t = \text{Linear}\left(\text{Flatten}(\mathbf{R}_t)\right), \quad t = 1, \ldots, T \tag{3}$$

To improve robustness to missing or irregularly sampled temporal data, we apply temporal augmentations during training, including random quarterly masking and median pooling of time steps, simulating partially observed sequences (discussed in Section 3.3).

Learnable temporal position embeddings $\mathbf{p}_t$ are added to each time step, and a learnable *class token* $\mathbf{p}_{\text{cls}}$ is prepended to summarize the entire spectral–temporal trajectory:

$$\mathbf{Z} = \left[\mathbf{p}_{\text{cls}}; \mathbf{r}_1 + \mathbf{p}_1; \ldots; \mathbf{r}_T + \mathbf{p}_T\right] \tag{4}$$

The tokenised sequence is then processed by the transformer-based encoder:

$$\hat{\mathbf{s}} = f_S(\mathbf{Z}). \tag{5}$$

The output corresponding to the class token $\mathbf{p}_{\text{cls}}$ is extracted as the initial satellite representation and further processed by an MLP.

The choice of a vanilla transformer backbone is motivated by its ability to model long-range dependencies through global self-attention, which is particularly beneficial for spectral–temporal data. Unlike CNN-based architectures that rely on fixed local receptive fields, the transformer captures dynamic inter-band relationships and cross-seasonal temporal patterns within a unified representation space, improving flexibility across different sensors and temporal resolutions.

Cross-modal alignment with contrastive learning. To align satellite spectral–temporal embeddings with ground-level visual embeddings, we adopt an InfoNCE-based (Oord et al., 2018) contrastive learning framework. Let $\mathbf{s}$ denote a satellite embedding and $\mathbf{g}$ its corresponding ground-level embedding. Both embeddings are L2-normalized:

$$\mathbf{s} = \frac{\hat{\mathbf{s}}}{\|\hat{\mathbf{s}}\|_2}, \quad \mathbf{g} = \frac{\hat{\mathbf{g}}}{\|\hat{\mathbf{g}}\|_2}. \tag{6}$$

The positive pair $(\mathbf{z}_S, \mathbf{z}_G)$ represents the satellite and ground images from the same location, while negative samples are drawn from a large MoCo-style (He et al., 2020) memory bank $\mathcal{Q}$ that stores previously seen ground embeddings. This design allows the model to scale efficiently without relying on extremely large batch sizes. The final loss we use is thus:

$$\mathcal{L}_{\text{TimeSenCLIP}}(\{\mathbf{s}_i\}, \{\mathbf{g}_i\}, \mathcal{Q}) = \sum_i \mathcal{L}_{\text{InfoNCE}}(\mathbf{s}_i, \mathbf{g}_i^+, \{\mathbf{g}_j^-\}_i \cup \mathcal{Q}). \tag{7}$$

Here, $\mathbf{g}_i^+$ denotes the ground image corresponding to the same location i as the satellite time series $\mathbf{s}_i$, whereas $\mathbf{g}_j^-$ represents a negative samples, any other locations within the batch and $\mathcal{Q}$ is a memory bank storing ground embeddings from previously seen locations.

Minimizing this loss encourages the satellite embedding to be close to its corresponding ground embedding while pushing it away from negatives, effectively aligning cross-modal representations in the CLIP latent space. This approach provides both robust alignment and computational scalability for large spectral–temporal datasets.

3.3. Temporal augmentation strategy

To account for missing or irregular observations in satellite time series due to cloud cover, seasonal occlusions, or sensor issues, we apply stochastic temporal masking during training. These augmentations are applied on-the-fly to the linear patch embeddings with a 50% probability per batch, ensuring diverse perturbations across epochs.

We randomly applied one of several temporal augmentations with equal probability, simulating partial or aggregated temporal observations:

- Median Pooling: Collapses all temporal frames into a single median vector, simulating temporally aggregated data (e.g., annual composites). This encourages robustness to varying temporal granularity while preserving spectral–temporal patterns.
- Random Quarter Mask: Randomly masks a contiguous subset of temporal frames (e.g., quarterly) to simulate realistic seasonal gaps caused by cloud cover or missing acquisitions.
- Random Temporal Mask: Randomly masks between 1 and 11 temporal frames, ensuring at least one frame is retained. This teaches the model to handle variable-length sequences and temporal inconsistencies.

These three **training-only strategies** allow the model to learn robust spectral–temporal representations under incomplete or aggregated inputs. At inference, the model can flexibly operate on **monthly, quarterly, or single-aggregated sequences** (discussed in Section 5.2.1), enabling evaluation under different temporal resolutions without retraining.

3.4. Prompt design

For zero-shot evaluation, we construct three types of textual prompts: (i) **class names**, (ii) **generic** templated prompts of the form "a centered satellite image of {class name}", and (iii) **class descriptive** natural language prompts. Five descriptive variations per class were generated using GPT-4, following the style of SenCLIP (Jain et al., 2025).

Using multiple prompt types allows us to evaluate the model's ability to capture fine-grained semantics: class names provide minimal, direct information; template prompts add context to help differentiate visually similar categories; and class descriptive prompts offer alternative phrasings that account for natural variation in how a class may be described. Because TimeSenCLIP is trained with ground-level images, it captures broader contextual information about locations. This enables the model to better leverage descriptive prompts, often yielding higher zero-shot performance than using only class names or templates. Moreover, as ground-level image representations rather than textual captions are used during training, the model is not limited to the usual remote sensing terminology and can generalize to natural language descriptions better adapted to characterize outdoor scenes.

We performed an exploration study to assess the contribution of each prompt type independently, reporting performance using only class names, only template prompts, or only descriptive prompts. For descriptive prompts, late prompt ensembling is applied on the CLIP embeddings by averaging the embeddings of the five prompts per class (Menon and Vondrick, 2022; Roth et al., 2023) and then computing the similarity with the satellite embeddings, producing a single class prediction for each satellite image time series. Examples of descriptive prompts are provided in Table 1.

4. Dataset

4.1. Cross-view training dataset

Ground-level images. The LUCAS (Land Use/Cover Area frame Survey) is a European Union-wide in-situ data collection campaign designed to systematically monitor land use and land cover (LULC) across Europe. The survey, conducted in 2018 (d'Andrimont et al., 2020), encompasses approximately 235,000 georeferenced sampling points distributed across 28 (EU member states and United Kingdom) countries. Each site is documented with high-resolution ground-level images taken in four cardinal directions (north, east, south, and west), accompanied by detailed annotations of land use, land cover (LULC), and crop types.

Table 1
Examples of descriptive prompts, together with their associated category.

Category	Descriptive Prompts
Sparsely wooded grasslands (EUNIS)	Sparse canopy with open fields; scattered trees on grass; visible bare ground between patches.
Seasonally wet and wet grasslands (EUNIS)	Wet meadows with reflective surfaces; dark wet zones amid dry land; green patches with water.
Boreal (Bio-Region)	Dense pine and spruce forests; mossy forest floors in coniferous areas.
Mediterranean (Bio-Region)	Olive/vineyard rows on slopes; dry, dusty terrain with sparse summer vegetation.
Common Wheat (Crops)	Wheat fields near harvest; thriving wheat across Europe.
Olive Groves (Crops)	Mediterranean olives, oil and pickles.; traditional olive landscapes in Greece, Italy, Spain.
Scenicness	Wide motorway; busy highway; misty mountain lake; grassy bog surrounded by hills.

Table 2
Class distribution statistics for the evaluation datasets across six label types. The table lists the total number of classes and the eight most frequent classes by proportion.

Label Type	# Classes	Top 8 Classes (by proportion)	# Samples	Proportion (%)	Label Type	# Classes	Top 8 Classes (by proportion)	# Samples	Proportion (%)
Land Cover	8	Woodland	17,813	35.48	EUNIS Ecosystem	44	Arable land and market gardens	15,732	31.34
		Cropland	13,007	25.91			Broadleaved deciduous woodland	8228	16.39
		Grassland	10,954	21.82			Coniferous woodland	6496	12.94
		Artificial Land	3173	6.32			Mesic grasslands	6197	12.34
		Shrubland	2715	5.41			Mixed deciduous and coniferous woodland	2549	5.08
		Bare Land	1139	2.27			Buildings of cities, towns and villages	1778	3.54
		Wetlands	1009	2.01			Low density buildings	1533	3.05
		Water	395	0.79			Dry grasslands	1091	2.17
Land Use	16	Agriculture	22,309	44.44	Bio Region	8	Continental	15,642	31.16
		Forestry	15,081	30.04			Mediterranean	11,609	23.12
		Semi-natural and natural areas not in use	6806	13.56			Atlantic	10,287	20.49
		Residential	2065	4.11			Boreal	7320	14.58
		Transport, Communication Networks, Storage, Protection Works	1860	3.70			Alpine	3714	7.40
		Recreation, Leisure, Sport	596	1.19			Pannonian	1195	2.38
		Abandoned areas	403	0.80			Steppic	331	0.66
		Community services	346	0.69			Black Sea	107	0.21
Crops	40	Common wheat	2422	4.82	Countries	28	France	7219	14.38
		Maize	1716	3.42			Spain	6547	13.04
		Barley	1277	2.54			Italy	4195	8.36
		Rape and turnip rape	780	1.55			Germany	4016	8.00
		Sunflower	402	0.80			Sweden	4000	7.97
		Oats	364	0.73			Poland	3463	6.90
		Durum wheat	361	0.72			United Kingdom	2582	5.14
		Rye	338	0.67			Romania	2507	4.99

Sentinel-2 time series data. We employ the Sen4Map dataset (Sharma et al., 2024), which provides co-registered multi-spectral and multi-temporal Sentinel-2 observations aligned with the LUCAS (2018) in-situ survey points across Europe. Each sample represents a 64 × 64 pixel Sentinel-2 patch centered on a LUCAS location and includes 10 spectral bands at 10 m and 20 m spatial resolutions, aggregated into monthly median composites to ensure cloud-free annual temporal coverage, resulting in 12 time steps per year. We used the Sen4Map random split, comprising 140k training, 30k validation, and 50k test samples. We utilized train split for contrastive pre-training of the model and evaluated our model on test split.

Although the dataset provides 64 × 64 spatial patches, TimeSenCLIP operates primarily on single-pixel (1 × 1) inputs extracted from the center of each patch. This design choice minimizes computational overhead while ensuring that each location is represented solely by its temporal–spectral signature, rather than by larger spatial context. For ablation studies, we also experiment with 5 × 5 and 9 × 9 patches to assess the influence of spatial context.

4.2. Evaluation tasks and datasets

We evaluate TimeSenCLIP on multiple downstream tasks derived from the same geo-referenced dataset, encompassing land classification, habitat mapping, and perceptual quality prediction.

Land use/land cover and crop type. We use coarse-grained land use and land cover (LULC) labels derived from the LUCAS field survey,[1] which follows the CORINE classification system. These labels include land cover categories such as artificial surfaces, croplands, grasslands, wetlands and forest, and define a multi-class classification task, based on in-situ observations across Europe. The Level 0 taxonomy comprises 8 land cover and 16 land use classes, while Level 1 further distinguishes croplands into 38 specific crop types. Fig. A.1 in appendix showcases

[1] https://ec.europa.eu/eurostat/documents/205002/8072634/LUCAS2018-C3-Classification.pdf#page=10.09

the Level-0 land cover class distribution across the EU within evaluation data split.

Biogeographical zones. The biogeographical regions of Europe were obtained from the European Environmental Agency (EEA)[2] (European Environment Agency, 2016). This dataset divides Europe into 11 ecologically distinct zones (e.g., Alpine, Boreal, Mediterranean) based on climatic, topographic, and vegetational characteristics. Each LUCAS site was spatially assigned to its corresponding biogeographical zone. Among the 11 regions, Arctic, Anatolian, Macaronesian, and Outside Europe were excluded, as these zones were not represented in the LUCAS geotag coverage.

Habitat mapping. The European Nature Information System (EUNIS) habitat classification, provided by the EEA (European Environment Agency, 2019), offers a continent-wide ecosystem map at 100 m spatial resolution.[3] We employ Level 2 of the EUNIS hierarchy, comprising 44 terrestrial habitat types (e.g., Mixed deciduous and coniferous woodland, Arable land and market gardens). These fine-grained ecological categories enrich the dataset by capturing detailed habitat characteristics that complement broader land-use labels.

Scenicness prediction. Scenicness prediction evaluates the ability of TimeSenCLIP to transfer learned representations to subjective, human-centered perceptual tasks, extending the typical classification-focused applications of remote sensing vision–language models (VLMs) to regression tasks. This task is formulated as a visual regression problem aimed at predicting the perceived beauty of landscapes (Workman et al., 2017; Levering et al., 2021, 2024). We use the ScenicOrNot (SoN) dataset (Seresinhe et al., 2015), which contains geotagged ground-level images across the United Kingdom (UK) rated by human annotators on a scale from 1 to 10. These crowd-sourced scores serve as ground-truth labels for perceived aesthetic quality. Predicting scenicness is challenging due to subjective variability, sparse spatial coverage, and the temporal–spectral complexity of satellite observations. To address this, we adopt the scenicness-oriented textual prompts from (Levering et al., 2024) (e.g., “a busy highway” vs. “a mountain”) and apply late prompt ensembling to capture concept-level aesthetics. For evaluation, we use 2411 Sen4Map samples from the UK, aligned to the nearest SoN image (within 100 m), maintaining the temporal and spectral information of the satellite time series. This setup demonstrates that TimeSenCLIP can generalize from semantic alignment to regression-based perceptual tasks, broadening the scope of remote sensing VLM applications.

Together, these datasets enable comprehensive evaluation of the model’s capacity for landscape characterization, capturing both LULC patterns and broader ecological attributes across diverse European regions. Table 2 provides an overview of the evaluation dataset, detailing the number of classes per label type and highlighting the eight most frequent classes with their corresponding proportions.

5. Experimental setup

5.1. Implementation details

Ground-level encoder. We use the ViT-B/32 CLIP model (Radford et al., 2021), pretrained on natural images, as the ground-level encoder. Each LUCAS ground-level image is independently passed through the frozen CLIP encoder, producing a 512-dimensional embedding.

Satellite time series encoder. The satellite branch processes Sentinel-2 multispectral time series represented as tensors of shape $T \times C \times H \times W$, where T denotes the number of temporal observations, C the number of spectral bands, and $H \times W$ the spatial extent; in the single-pixel setting, ($H = 1$, $W = 1$), while larger spatial dimensions enable the incorporation of additional spatial context. All Sentinel-2 bands are linearly rescaled to the range [0, 1] using per-band min–max normalization, where the minimum and maximum statistics are computed over the entire dataset. Learnable temporal positional embeddings of size 512 are added to each frame. The resulting sequence with CLS token is then fed into a 6-layer Transformer encoder with 8 attention heads, a hidden dimension of 256, and a latent size of 512. The Transformer employs GELU activations and LayerNorm throughout. The final class token output is passed through a lightweight two-layer MLP projection head to obtain the satellite embedding.

Contrastive alignment. Both the CLIP-based ground embeddings and the satellite embeddings are projected into a shared 512-dimensional latent space prior to alignment. Both ground and satellite embeddings are L2-normalized and aligned via the contrastive objective described in Section 3. The momentum-based memory queue maintains $K = 2048$ negative samples and is updated at each iteration through an enqueue–dequeue mechanism, where newly computed embeddings replace the oldest to keep the queue size constant.

Training configuration. Models are trained using the AdamW optimizer with an initial learning rate of 10^{-4}, weight decay of 1×10^{-6}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. A cosine annealing schedule with 10 warm-up epochs is used over a total of 200 epochs, with a batch size of 1024. All experiments are conducted on a single NVIDIA TITAN X GPU. During training, positive pairs are formed between satellite and corresponding ground embeddings, while negatives are drawn from both the in-batch and the MoCo-style memory queue of size 2048.

5.2. Evaluation setup

The overall evaluation framework is illustrated in appendix Fig. A.2. During inference, only the trained satellite encoder is employed. The evaluation encompasses three components that jointly assess the versatility and generalization of the proposed TimeSenCLIP framework across diverse remote sensing tasks: (1) **Zero-shot classification**, which measures the model’s ability to identify land cover, land use, crop type, and habitat classes purely from semantic text descriptions, without any task-specific fine-tuning; (2) **Cross-modal retrieval**, which evaluates the quality of representation alignment by computing similarity between embeddings from the satellite time-series and ground-level domains, enabling both Satellite-to-Ground (S2G) and Ground-to-Satellite (G2S) retrieval; and (3) **Scenicness prediction**, which leverages the learned multimodal embeddings to estimate perceptual “scenicness” scores, demonstrating the model’s capacity to generalize beyond categorical classification to subjective, aesthetic evaluation. More details on each evaluation tasks are in Appendix A.1.

In addition, we describe how temporal aggregations are applied to summarize multi-temporal inputs, outline the baseline models used for performance comparison, and specify the evaluation metrics adopted for each evaluation task.

5.2.1. Temporal aggregation

In addition, during inference, we evaluate the model under three temporal aggregation strategies. The monthly setting uses all 12 individual time steps. The quarterly setting reduces the sequence to 4 time steps, each computed as the median of a 3-month window. The annual (Single) setting further compresses the sequence to a single representation by taking the median across all 12 months.

5.2.2. Baselines

We compare TimeSenCLIP against a range of CLIP-based VLMs, including CLIP (Radford et al., 2021), GeoRSCLIP (Zhang et al.,

[2] https://www.eea.europa.eu/en/analysis/maps-and-charts/biogeographical-regions-in-europe-2

[3] https://sdi.eea.europa.eu/catalogue/srv/api/records/7c0cf3f2-ab54-4cd0-a635-b322df7197f6

Table 3
Top-1 zero-shot classification accuracy (%) for CLIP-based baselines and **TimeSenCLIP** across five geospatial tasks: Land Cover, Land Use, Habitat, Crops, and Bioregion. Results are reported using *class-name (C)*, *generic (G)*, and *description-level (D)* textual prompts. All baselines operate on 64 × 64 pixel Sentinel-2 image patches. **TimeSenCLIP-P1** uses single-pixel inputs (1 × 1) whereas **TimeSenCLIP-P64** leverages SenCLIP-pretrained spatial embeddings combined with a temporal transformer and is trained and evaluated using 64 × 64 pixel inputs. TimeSenCLIP-P1 is the only model evaluated on a 12-timestep temporal sequence; other baselines use a single timestamp. Both RGB and multispectral (MS) variants are compared. Values are averaged over five random seeds, with the standard deviation reported in the overall column. Underline indicates the best performance within each pixel and temporal configuration, while **bold** denotes the best overall performance across all temporal and spatial settings.

T	Bands	Model	Land Cover			Land Use			Habitat			Crops			Bioregion			Overall		
			C	G	D	C	G	D	C	G	D	C	G	D	C	G	D	C	G	D
64 × 64 Pix																				
1	RGB	CLIP	30.26	29.40	33.82	40.20	46.35	38.38	11.71	14.89	20.61	3.03	2.52	4.10	16.64	17.55	16.38	20.37±.22	22.14±.21	22.66±.20
	RGB	GeoRSCLIP	35.18	34.66	42.79	35.58	47.16	46.87	13.98	18.15	25.19	3.12	3.53	3.87	18.88	20.19	17.89	21.35±.10	24.74±.22	27.32±.15
	RGB	RemoteCLIP	28.45	38.08	37.27	39.22	43.05	43.08	13.19	13.13	16.73	2.47	4.76	2.46	18.44	19.88	17.03	20.35±.14	23.78±.10	23.31±.11
	RGB	SkyCLIP	22.41	26.46	35.13	3.78	4.39	26.22	2.14	2.56	3.79	2.09	2.11	3.13	9.14	16.30	15.13	7.91±.06	10.36±.10	16.68±.10
	RGB	SenCLIP	38.13	37.86	42.06	32.04	32.35	39.23	27.03	23.28	22.70	3.29	3.58	4.57	28.28	32.50	20.87	25.75±.10	25.91±.10	25.89±.14
	MS	Llama3-MS-CLIP	45.39	42.32	26.49	44.02	50.67	49.63	10.79	19.57	20.98	2.84	2.65	1.07	18.41	11.55	33.21	24.29±.12	25.35±.11	26.28±.13
	RGB	TimeSenCLIP-P64	53.18	53.96	55.20	56.61	**61.21**	61.01	24.16	24.43	**31.80**	7.28	7.43	13.90	32.15	34.34	33.86	34.68±.16	36.27±.15	39.15±.24
Single Pix																				
1	RGB	TimeSenCLIP-P1	52.33	53.59	53.64	**56.64**	60.31	61.56	26.57	23.84	25.98	4.76	4.84	5.36	21.36	22.25	15.90	32.33±.16	32.97±.14	32.49±.12
	MS	TimeSenCLIP-P1	55.00	56.27	57.23	50.14	59.17	61.16	**33.92**	**29.60**	29.42	9.17	8.72	13.39	28.08	28.61	26.09	35.26±.11	36.47±.13	37.46±.15
12	MS	TimeSenCLIP-P1	**62.40**	**61.35**	**66.49**	54.45	60.67	**64.59**	24.06	25.01	30.90	**29.33**	**28.42**	**40.36**	**49.87**	**45.65**	**34.42**	**42.02 ± .15**	**44.2 ± .16**	**47.35 ± .18**

2023), RemoteCLIP (Liu et al., 2023), SkyCLIP (Wang et al., 2024), SenCLIP (Jain et al., 2025), and Llama3-MS-CLIP (Marimo et al., 2025). All baselines rely on spatial representations and operate on full 64 × 64 RGB image patches, with the exception of Llama3-MS-CLIP, which uses multispectral Sentinel-2 inputs that are resized to 224 × 224 to match the standard CLIP input resolution.

For a fair comparison, TimeSenCLIP is evaluated under multiple configurations: (i) **TimeSenCLIP-P1-RGB**, which models single-pixel (1 × 1) RGB time series and isolates the effect of temporal reasoning without spatial context; (ii) **TimeSenCLIP-P64-RGB**, which builds on SenCLIP-pretrained spatial embeddings extracted from 64 × 64 RGB patches and augments them with our temporal encoding module; and (iii) **TimeSenCLIP-P1-MS**, our main model, which leverages multispectral single-pixel (1 × 1) time series to fully exploit both temporal and spectral information.

Each model is trained separately. This setup enables a direct comparison between spatially grounded VLMs and our temporally and spectrally enhanced representations, disentangling the contributions of pure temporal modeling (P1-RGB), temporally enriched spatial embeddings (P64-RGB), and temporally–spectrally informed multispectral inputs (P1-MS).

5.2.3. Evaluation metrics

To comprehensively assess the performance of our proposed framework, we employ task-specific evaluation metrics tailored to each downstream objective. The three evaluation settings are quantitatively evaluated as follows:

- Zero-shot classification: We report the *Top-1 accuracy*, which measures the proportion of test samples for which the predicted label (i.e., the class with the highest similarity score between the image and text embeddings) exactly matches the ground-truth class. This metric reflects the model's direct classification capability without any fine-tuning or task-specific adaptation.
- Cross-Modal retrieval: Retrieval performance is quantified using *Recall@1*, which indicates the percentage of queries where the top-ranked retrieved image corresponds to the same class as the query. This metric evaluates the discriminative quality of the learned embeddings and their alignment across modalities (e.g., Satellite-to-Ground and Ground-to-Satellite retrieval).
- Scenicness regression: The predicted scenicness scores are evaluated using both the *Pearson correlation coefficient (R)* and *Kendall's tau (τ)*. Pearson's R measures the linear correlation between predicted and ground-truth scores, while Kendall's τ assesses their rank correlation.

6. Results and discussion

This section is organized into four parts. First, Section 6.1 reports quantitative results for zero-shot classification, cross-modal retrieval, and scenicness regression. Second, Section 6.2 analyzes the effects of architectural design choices, spatial context, augmentation strategies, and computational efficiency through ablation studies. Third, Section 6.3 provides qualitative insights via per-pixel prediction maps and retrieval examples. Finally, Section 6.4 contextualizes these findings and discusses their broader technical and ecological implications.

6.1. Quantitative results

We evaluate TimeSenCLIP on three scenarios: zero-shot classification, cross-modal retrieval, and scenicness regression. These tasks collectively probe the model's ability to generalize across domains, modalities, and temporal resolutions without task-specific fine-tuning. All experiments compare TimeSenCLIP against multiple CLIP-based baselines, enabling a consistent assessment of temporal, spectral, and spatial contributions.

6.1.1. Zero-shot classification performance

For zero-shot classification, we report Top-1 accuracy using three types of textual prompts: *class-name*, *generic*, and *description-level*, as introduced in Section 3.4. Evaluations are conducted across five geospatial tasks: Land Cover, Land Use, Habitat, Crops, and Bioregion, under varying temporal lengths (T), spatial input resolutions (*Pix*), and spectral modalities (RGB and multispectral). Table 3 summarizes the zero-shot classification performance across all baseline models and TimeSenCLIP variants across these settings. Each experiment was repeated with five random seeds; column values show mean performance, while the final column reports the average standard deviation.

Standard CLIP models show limited transfer to Sentinel-2 imagery due to the lack of temporal cues and the domain gap between natural and multispectral satellite images. GeoRSCLIP, and SenCLIP, improve upon standard CLIP but still struggle to capture intra-annual variations, leading to inconsistent performance across tasks and prompt types. Llama3-MS-CLIP further improves over RGB-based baselines for Land Cover and Land Use, suggesting that multispectral inputs help capture broad material and surface properties. However, its performance degrades on more complex tasks such as Crops, Habitat, and Bioregion.

TimeSenCLIP consistently outperforms all prior CLIP-based approaches across tasks, temporal settings, and prompt formulations. Notably,

Table 4
Cross-Modal retrieval performance across land cover, land use, habitat, crops, bioregion, and country tasks. Results are reported as **Ground-to-Satellite (G2S)** and **Satellite-to-Ground (S2G)** Recall@1 retrieval. Experiments are grouped by temporal sequence length (T), with RGB and multispectral (MS) variants compared within each temporal setting. Underline indicates the best performance within each pixel and temporal configuration, while **bold** denotes the best overall performance across all temporal and spatial settings.

T	Bands	Model	Land Cover		Land Use		Habitat		Crops		Bioregion		Country		Overall	
			G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G
64 × 64 Pix																
1	RGB	CLIP	0.438	0.141	0.513	0.152	0.360	0.269	0.157	0.127	0.361	0.231	0.130	0.047	0.327	0.161
	RGB	GeoRSCLIP	0.439	0.313	0.628	0.274	0.386	0.241	0.165	0.087	0.425	0.263	0.116	0.077	0.360	0.209
	RGB	RemoteCLIP	0.394	0.406	0.539	0.496	0.368	0.188	0.060	0.078	0.283	0.200	0.099	0.025	0.291	0.232
	RGB	SkyCLIP	0.114	0.305	0.257	0.255	0.064	0.121	0.025	0.030	0.221	0.242	0.075	0.049	0.126	0.167
	RGB	SenCLIP	0.483	0.387	0.589	0.424	0.342	0.263	0.103	0.077	0.473	0.486	0.183	0.197	0.362	0.306
	MS	Llama3-MS-CLIP	0.267	0.269	0.301	0.263	0.317	0.135	0.050	0.064	0.276	0.196	0.098	0.078	0.218	0.167
	RGB	TimeSenCLIP-P64	0.539	0.563	0.628	0.635	0.404	0.417	0.132	0.188	0.531	0.601	**0.261**	0.355	0.416	0.460
Single Pix																
1	RGB	TimeSenCLIP-P1	0.507	0.534	0.594	0.601	0.349	0.340	0.125	0.095	0.295	0.300	0.114	0.113	0.331	0.331
	MS	TimeSenCLIP-P1	0.534	0.574	0.573	0.626	0.343	0.321	0.174	0.180	0.338	0.342	0.150	0.146	0.352	0.365
12	MS	TimeSenCLIP-P1	**0.599**	**0.660**	**0.646**	**0.682**	**0.407**	**0.433**	**0.244**	**0.411**	**0.526**	**0.634**	0.255	**0.392**	**0.446**	**0.535**

TimeSenCLIP-P1-RGB, trained solely on single-pixel RGB time series, surpasses all existing CLIP baselines, demonstrating that temporal modeling alone can capture discriminative phenological patterns even without spatial context. This effect is especially pronounced for temporally driven tasks such as Crops and Bioregion.

Spatial context remains beneficial when temporal coverage is limited. **TimeSenCLIP-P64-RGB**, which combines SenCLIP-pretrained spatial embeddings with a temporal transformer, achieves strong performance on spatially structured tasks such as Land Use and Land Cover. However, as temporal depth increases, single-pixel models often match or exceed their larger-patch counterparts, indicating that temporal evolution can outweigh spatial detail for many geospatial categories.

Incorporating multispectral inputs further enhances performance. **TimeSenCLIP-P1-MS** consistently improves over RGB variants, particularly for Land Cover and Crops, highlighting the complementary role of spectral information when combined with temporal modeling. Overall, these results confirm that temporal modeling is the dominant factor for zero-shot generalization in remote sensing, with spatial and spectral cues providing task-dependent gains.

Noticeably, Crops and Bioregion benefit most from temporal information due to strong seasonal signatures. Land Cover and Land Use gain moderate improvements from combining spatial and temporal cues. In contrast, Habitat classification shows limited sensitivity to increased temporal depth, suggesting that habitat-level semantics are driven primarily by spectral composition and spatial structure rather than seasonal evolution.

6.1.2. Cross-modal retrieval performance

For cross-modal retrieval, we evaluate cross modality matching performance using Recall@1 in both Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G) directions. Experiments are conducted across six semantic tasks: Land Cover, Land Use, Habitat, Crops, Bioregion, and Country, while varying temporal length (T), spatial input resolution (*Pix*), and spectral modality (RGB versus multispectral). The results compare all CLIP-based baselines and TimeSenCLIP variants under consistent retrieval protocols.

The cross-modal retrieval results in Table 4 reveal the same fundamental trends observed for the zero-shot classification scenario. Standard CLIP and SkyCLIP achieve only moderate performance and exhibit a strong degradation in Satellite-to-Ground (S2G) retrieval, reflecting limited robustness to viewpoint, scale, and acquisition-time discrepancies. GeoRSCLIP and SenCLIP improve retrieval for structurally stable categories such as Land Cover and Land Use, yet remain sensitive to seasonal variability, leading to weaker performance on temporally dynamic tasks such as Crops. Multispectral adaptation in Llama3-MS-CLIP yields only marginal retrieval gains.

In contrast, TimeSenCLIP delivers the strongest and most stable cross-modal alignment across tasks and retrieval directions, with performance improving consistently with temporal depth and peaking on phenology-driven categories, highlighting temporal dynamics as a view-invariant signal for matching ground and satellite observations. Although larger spatial patches remain beneficial for coarse or regionally defined tasks, temporally rich single-pixel multispectral sequences often achieve comparable or superior performance, indicating that temporal–spectral information can compensate for reduced spatial context. Moreover, TimeSenCLIP maintains balanced accuracy between Ground-to-Satellite and Satellite-to-Ground retrieval, unlike conventional CLIP-based models that exhibit pronounced asymmetry, demonstrating that temporal encoding produces stable, view-agnostic representations that support consistent semantic alignment across perspectives.

6.1.3. Scenicness regression with prompt ensembling

We evaluate scenicness prediction using prompt-ensembling strategies and report performance with Pearson's R and Kendall's τ. Table 5 compares early and late prompt ensembling approaches for scenicness prediction following (Levering et al., 2024), reporting performance on Sentinel-2 imagery using both correlation metrics. More details on setup are in Appendix A.1

Table 5 shows that with single pixel temporal depth is critical for perceptual landscape estimation. With a single timestamp (T=1), performance is limited (R=0.318, τ=0.205 under late ensembling), indicating that spectral information alone is insufficient. However, increasing temporal coverage yields consistent improvements, with quarterly aggregation enhancing performance and monthly temporal input (T=12) achieving the best results under single pixel (R=0.527, τ=0.363).

CLIP continues to serve as a competitive non-temporal baseline, but its static representations (R = 0.517, τ = 0.362) fall short of TimeSenCLIP-P64-RGB, demonstrating that temporal cues are valuable even for tasks commonly considered visually "static". Under single-temporal input, TimeSenCLIP-P64-RGB achieves the strongest performance (R = 0.660, τ = 0.450), and also shows competitive performance with CLIP on ground-level imagery (R = 0.660, τ = 0.465). These results highlight two key aspects of our design: (1) temporal encoding substantially enhances the SenCLIP spatial encoder, equipping it to capture seasonal cycles and long-term landscape appearance; and (2) large spatial context (P64) remains important for modeling holistic perceptual attributes such as scenicness, where multi-scale texture, landform geometry, and spatial composition strongly influence aesthetic ratings.

Additional detailed results are provided in Appendix A.2.

Table 5
Scenicness estimation performance of TimeSenCLIP using multispectral single-pixel inputs compared to CLIP and SenCLIP. We report Early and Late ensembling, where Late columns report the average over late ensembling strategies (2, 5, 8, 10 prompts).

Model	T	Pix	Early		Late	
			R	τ	R	τ
CLIP (Ground Images) (Levering et al., 2024)	1	–	0.544	0.390	0.660	0.465
CLIP (Satellite)	1	64	0.510	0.358	0.519	0.363
SenCLIP	1	64	0.396	0.272	0.421	0.283
TimeSenCLIP-RGB	1	64	**0.532**	**0.385**	**0.660**	**0.450**
TimeSenCLIP-MS	1	1	0.336	0.223	0.318	0.205
TimeSenCLIP-MS	4	1	0.438	0.309	0.451	0.312
TimeSenCLIP-MS	12	1	**0.518**	**0.361**	**0.527**	**0.363**

Table 6
Top-1 Zero-shot accuracy across classical time-series model architecture and the proposed transformer model on tasks using Class (C), Generic (G), and Description (D)-level textual prompts. Results report the mean accuracy over five runs with different random seeds and standard deviation reported as average in overall column. Best performance is shown in **bold**.

Model	Land Cover			Land Use			Habitat			Crops			Bioregion			Overall		
	C	G	D	C	G	D	C	G	D	C	G	D	C	G	D	C	G	D
Supervised Upper Bound	78.65	–	–	79.30	–	–	60.52	–	–	64.60	–	–	85.39	–	–	73.70	–	–
CNN1D	60.60	60.83	64.80	50.70	58.60	61.32	25.23	25.23	29.65	22.61	23.86	34.99	36.08	41.30	33.65	39.04±.19	41.96±.22	44.88±.15
ConvTran	60.29	60.26	65.06	52.11	59.91	62.49	26.59	25.87	29.81	23.77	25.07	38.70	37.55	42.69	33.38	40.06±.24	42.76±.24	45.89±.22
MLP	60.91	60.74	65.80	51.73	61.19	64.26	26.46	26.38	30.39	28.50	27.64	39.97	38.42	43.50	33.19	41.20±.24	43.89±.22	46.72±.14
TempCNN	**64.30**	**63.43**	**66.53**	53.82	59.30	**65.41**	**35.70**	**33.01**	**33.37**	23.48	24.66	34.33	32.42	36.86	27.41	41.94±.24	43.45±.22	45.41±.11
Transformer (ours)	62.40	61.35	66.49	**54.45**	**60.67**	64.59	24.06	25.01	30.90	**29.33**	**28.42**	**40.36**	**49.87**	**45.65**	**34.42**	**42.02 ± .15**	**44.22 ±.16**	**47.35 ± .18**

6.2. Ablation study

6.2.1. Conventional temporal baseline comparison

For a fair comparison with TimeSenCLIP's transformer-based encoder, we train classic temporal models CNN1D (Kiranyaz et al., 2015), MLP (Rumelhart et al., 1986), ConvTran, and TempCNN, following the same training protocol as TimeSenCLIP and using identical multispectral single-pixel time series inputs.

These architectures are largely adopted for the classification of satellite image time series data and provide complementary inductive biases: MLPs capture per-timestep spectral relationships, CNN1D models learn local temporal patterns, TempCNN extends this with deeper temporal convolutions, and ConvTran introduces hybrid convolution–transformer reasoning. By restricting all baselines to purely temporal–spectral signals and removing spatial context, we ensure that performance differences reflect the models' temporal modeling capabilities rather than advantages from spatial structure.

Table 6 reports the zero-shot evaluation results, with each experiment repeated over five random seeds. Reported values correspond to the mean performance for each task, while the final column summarizes the average standard deviation across runs. The results indicate that classic time series architectures achieve competitive zero-shot performance, but their effectiveness varies notably across model families and task types. MLP and TempCNN generally provide stronger results than CNN1D and ConvTran, particularly on tasks where spectral or phenological cues play a larger role. TempCNN achieves the highest scores on couple of benchmarks, especially for Land Cover and Habitat, reflecting the strength of convolutional temporal filters for modeling smooth spectral–temporal trends. MLP remains consistently stable across tasks, offering strong performance despite its lightweight structure. CNN1D and ConvTran deliver more modest accuracy overall, showing limitations in capturing the complexity of multispectral sequences in a zero-shot setting.

TimeSenCLIP's transformer variant performs competitively across all tasks and often best performing models, particularly for Crops and Bioregions, where its learned representations capture class distinctions that benefit from subtle spectral–temporal differences. Although differences between architectures are sometimes limited, the transformer-based approach demonstrates balanced behavior across all prompt types Class, Generic, and Description, resulting in the highest mean Overall accuracy under all types of prompts. This suggests that the combination of our training strategy and transformer design produces embeddings that generalize well across diverse ecosystems and LULC categories without reliance on spatial context or tailored text supervision.

The supervised upper bound provides the highest overall accuracy, but the performance gap between supervised and zero-shot models varies substantially across tasks. Smaller gaps for Land Cover and Land Use suggest that these broad categories are reasonably well represented in the CLIP text encoder, enabling effective semantic alignment even without task-specific training. In contrast, larger discrepancies appear for Habitat Crops, and Bioregions. This can be attributed not only to limitations in the pretrained text encoder, which offers less coverage of specialized ecological terminology, but also to the intrinsic difficulty of these tasks: they involve a much larger number of classes (e.g., 44 habitat types and 39 crop types) in contrast to land cover and land use (e.g., 8 land cover types and 16 land use types), making semantic separation in the embedding space more challenging. Additionally, many classes in these domains exhibit weak or overlapping phenological and ecological signatures (e.g., similar crop growth cycles, or habitats without distinct seasonal patterns), reducing the discriminative cues available for zero-shot inference.

6.2.2. Impact of spatial context

We investigate how spatial context, controlled via patch size, affects zero-shot classification performance. Specifically, we compare single-pixel (P1), 5 × 5 (P5), 9 × 9 (P9), 16 × 16 (P16), and 64 × 64 (P64) patches, where larger patches provide broader spatial context to the temporal encoder. A separate model is trained for each patch size and then evaluated on corresponding size and with different temporal settings. Fig. 3 illustrates the impact of patch size on Top-1 accuracy across several geospatial classification tasks.

The results reveal that single-pixel (P1) time series are highly competitive, often matching or exceeding larger patches. This demonstrates that temporal and spectral dynamics alone capture much of the discriminative information required for zero-shot classification, with minimal

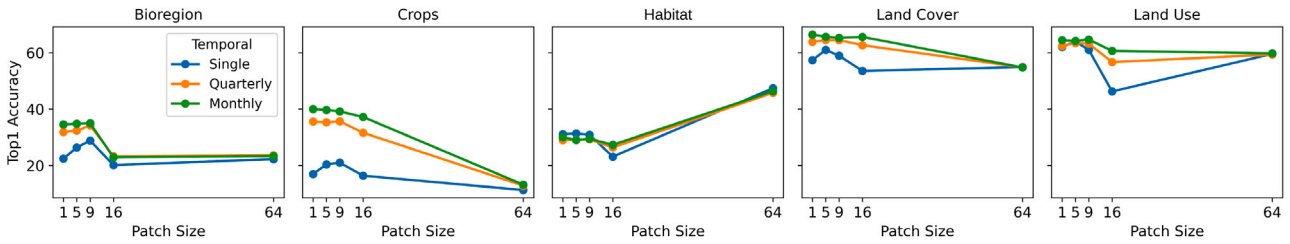


Fig. 3. Impact of spatial information on zero-shot classification Top-1 accuracy (%) across different mapping tasks. Single-pixel (P1) time series are often competitive with larger patches.

Table 7
Effect of multispectral (MS) and temporal (TS) dropout applied during training (over 12 temporal and 10 spectral) on zero-shot classification performance (Top-1 accuracy, %). The table reports Top-1 accuracy averaged across different textual prompt types and evaluated using both single-timestamp and monthly temporal inputs. (✓) denotes that dropout is enabled during training, whereas (✗) indicates that no dropout is applied.

T	TS Drop	MS Drop	Land Cover	Land Use	Habitat	Bioregion	Crops	Average
1	✗	✗	28.69	25.86	9.52	18.68	3.23	17.20
	✗	✓	23.76	27.88	8.39	18.18	2.93	16.23
	✓	✗	**56.32**	**57.83**	**31.29**	**25.30**	**12.61**	**36.67**
	✓	✓	53.52	56.15	28.03	25.00	7.81	34.10
12	✗	✗	61.72	59.76	23.43	38.71	31.82	43.09
	✗	✓	63.12	58.77	25.15	40.47	31.14	43.73
	✓	✗	**63.63**	**60.03**	25.54	**40.24**	**32.23**	**44.34**
	✓	✓	62.42	58.17	**26.68**	39.42	31.64	43.67

spatial context being sufficient for robust geospatial representations. Small spatial neighborhoods (P5) provide modest gains for some tasks; for example, Land Cover zero-shot accuracy increases from 56.08% (P1) to 58.14%, indicating that limited spatial information can complement temporal and spectral cues. Larger patches (P16–P64) mainly benefit coarse or spatially homogeneous tasks such as Habitat or Land Cover/Use, whereas fine-grained categories like Crops or Bioregion can degrade due to mixed-signal pixels.

The temporal depth consistently dominates over spatial size. Increasing the number of timestamps improves zero-shot accuracy across all patch sizes, and long P1 time series can match or surpass larger patch variants, especially for phenology-driven tasks. These results highlight that minimal spatial input combined with sufficient temporal coverage provides a strong, generalizable geospatial representation while keeping computational costs low.

6.2.3. Impact of dropout strategies

To assess how TimeSenCLIP handles incomplete or noisy observations, common in real-world remote sensing, we study two dropout-based regularization strategies on $T = 12$ temporal setting, applied only during training:

- Temporal Drop (TS Drop): for each batch, with a probability of 50%, all timesteps of the time series are either randomly masked, quarterly masked or collapsed into a single median value as discussed in Section 3.3. This simulates missing temporal observations or coarsely aggregated products.
- Multispectral Dropout (MS Drop): for each batch, non-RGB spectral bands are randomly masked with a probability of 50%, encouraging the model to rely on robust spectral cues.

During evaluation, we assess the robustness of models trained with dropout under different temporal aggregations: Single and Monthly, as described in Section 5.2.1.

Zero-shot classification results in Table 7 indicate that temporal dropout is highly effective when temporal coverage is limited. In the single-temporal setting, TS Drop substantially improves performance, increasing the average Top-1 accuracy from 17.2% to 36.7%. Without temporal augmentation, the model tends to overfit to the fixed temporal signature of a single acquisition. TS Drop forces the model to learn more stable, temporally robust cues, demonstrating that temporal regularization is particularly valuable for sparse time series.

In contrast, MS Drop provides minimal gains in some settings but often reduces performance. Removing non-RGB bands discards spectral features critical for distinguishing vegetation and ecosystem types, limiting its usefulness for zero-shot classification. Combining TS Drop with MS Drop does not produce additional improvements, suggesting that the primary source of robustness comes from temporal augmentation rather than spectral masking.

These findings highlight temporal dropout as a simple yet powerful strategy for improving TimeSenCLIP's generalization under limited temporal observations. As temporal depth increases, the model naturally develops temporal robustness, and the relative benefit of TS Drop diminishes.

Table 8
Computational efficiency of CLIP-Based Model and TimeSenCLIP models on Batch Size 1024.

Model	FLOPs (GMac)	Parameters (M)	Throughput (samples/s)	Peak Memory (MB)
CLIP	4.46	151.28	565.33	3045.13
TimeSenCLIP 1 × 1	0.105	8.17	19 530.50	584.60
TimeSenCLIP 5 × 5	0.106	8.29	19 424.24	596.32
TimeSenCLIP 9 × 9	0.110	8.58	19 428.41	624.08
TimeSenCLIP 16 × 16	0.121	9.48	17 798.35	709.12

6.2.4. Computational efficiency and inference performance

Table 8 summarizes the computational efficiency of the proposed TimeSenCLIP model in comparison to CLIP-based baselines. All CLIP-based models behave identically in terms of FLOPs, parameters, and throughput, as they share the same image encoder. TimeSenCLIP, which processes 12 temporal frames with 10 spectral bands, achieves a 97%–98% reduction in FLOPs and 94% fewer parameters compared to CLIP when operating on a single pixel (1 × 1), while delivering substantially higher throughput (over 33 × faster) and requiring significantly less GPU memory. Even when processing larger spatial patches such as 5 × 5, 9 × 9, or 16 × 16, TimeSenCLIP maintains low computational cost, with only modest increases in FLOPs, memory, and throughput. These results demonstrate that TimeSenCLIP is highly efficient and well-suited for scalable remote sensing tasks, particularly in resource-constrained or real-time settings. As TimeSenCLIP-P64 uses SenCLIP for embeddings, so it was not included in this comparison.

6.3. Qualitative analysis

6.3.1. Per-pixel prediction maps

To evaluate the capability of TimeSenCLIP to perform per-pixel zero-shot classification using natural-language class descriptions, we derive prediction maps for LULC and crop type categories. We report such maps in Figs. 4 and 5, together with reference ground-truth information to facilitate qualitative comparison. Because dense semantic

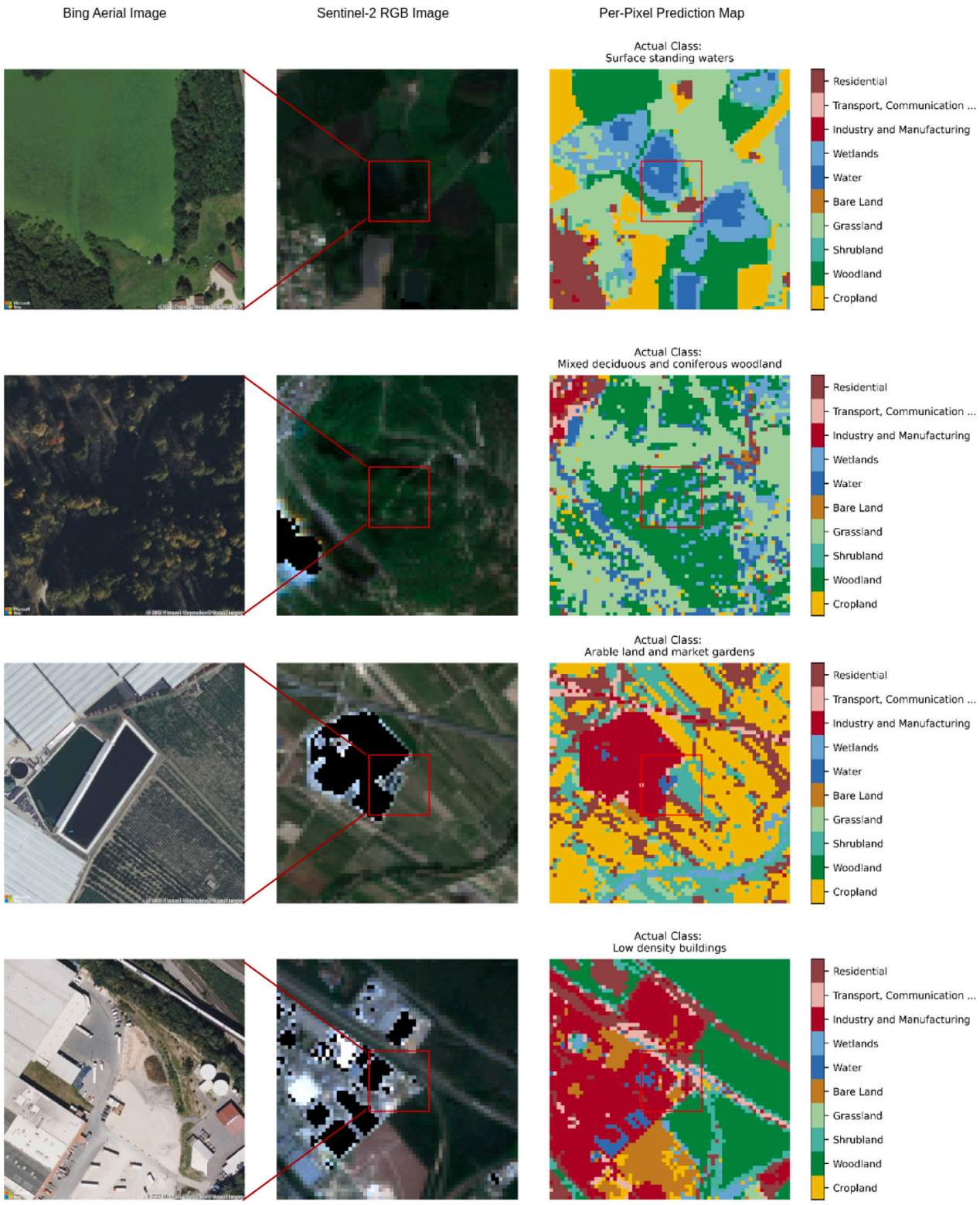


Fig. 4. Per-pixel land-cover prediction map generated from text-prompt-based LULC class descriptions using monthly Sentinel-2 time series data (64 × 64-pixel tile). The accompanying Bing aerial image provides a high-resolution visualization of the same geographic area for reference only and is centered on the corresponding Sentinel-2 tile. As the two datasets are acquired at different times and resolutions, discrepancies in appearance and feature alignment are expected. The color bar indicates the set of LULC classes predicted within the tile. The actual class for this geo-tagged location corresponds to the habitat ground-truth label as they provide broader class context for larger spatial than LUCAS LULC labels.

Fig. 5. Per-pixel crop type prediction map generated using text-prompt-based crop class descriptions on monthly Sentinel-2 time series (64 × 64-pixel tile). The Bing aerial image provides a high-resolution visual reference of the same region, centered on the Sentinel-2 tile. Because the Bing and Sentinel-2 images were acquired at different times, visual appearance may differ; the Bing image is included solely as contextual reference. The color bar indicates the set of crop classes predicted within this tile. The actual class for this geo-tagged location corresponds to the LUCAS ground-truth label. To improve interpretability, fine-grained crop classes were color-coded according to their higher-level agronomic groups.

segmentation annotations are not available in LUCAS, only the class label at the center pixel is provided; this central label is therefore shown as the ground-truth reference rather than a full pixel-wise segmentation map. Each extract spans 640 $m \times 640m$ (corresponding to a 64 × 64 Sentinel-2 pixel tile) and represents a monthly multispectral composite. Classifications are produced at the pixel level using the single-pixel TimeSenCLIP model, which analyzes each pixel time series independently and performs zero-shot classification using descriptive text prompts (Section 3.4). Each pixel is assigned to the class whose textual description best matches its spectral–temporal signature, following the same zero-shot procedure described in Sections 6.1.1 and 6.2.1, but applied densely across the 64 × 64 tile.

The maps exhibit spatially coherent and semantically meaningful patterns, illustrating the model's ability to associate high-level language-based class concepts with pixel-scale temporal behavior. At the same time, localized salt-and-pepper noise appears as isolated pixel-level fluctuations between neighboring classes. This effect is expected because inference is performed pixel-wise and no spatial smoothing post-processing is used. Nevertheless, broader spatial structures remain consistent, indicating that spectral–temporal representations capture meaningful land-cover organization despite the absence of explicit spatial modeling.

To facilitate visual interpretation, we include a high-resolution Bing aerial image for each location as a reference. Although the Bing imagery is not perfectly aligned with Sentinel-2 due to differences in acquisition date, spatial resolution, and viewing geometry, it still provides useful contextual cues regarding settlement structure, vegetation distribution, and general land-use patterns. The LUCAS point label associated with each location is also shown to indicate the ground-truth class at the sample center, helping assess the semantic correctness of the predicted patterns.

For LULC zero-shot pixel classification, we merged land-use and land-cover taxonomies and resolved overlapping categories by unifying semantically equivalent classes. For example, the land-cover class Cropland and the land-use class Agriculture were consolidated into a single category, Cropland. As shown in Fig. 4, the model correctly identifies a wide range of LULC categories and, in ambiguous cases, assigns semantically consistent or closely related classes, including built-up classes such as residential and industrial areas, as well as natural categories such as grassland and shrubland.

For zero-shot crop pixel classification, we used the full set of fine-grained crop class descriptions. After generating prediction maps at the fine-grained level, classes were aggregated into higher-level agronomic groups for visualization. For example, Common wheat, Maize, and Barley were grouped under Cereals, while Olive groves and Vineyards were grouped as Permanent Mediterranean. As shown in Fig. 5, although certain crop categories are over-predicted, likely reflecting the difficulty of zero-shot discrimination between spectrally similar crops; the model consistently identifies the correct crop class at the central LUCAS location and effectively distinguishes cropland from artificial surfaces or non-cropland areas.

6.3.2. Text-to-image retrieval

Fig. 6 presents the quantitative results of text-to-image retrieval, where free-form textual descriptions are used to search for the most semantically relevant satellite image patches. Each query corresponds to a specific land cover, land use, habitat class, biogeographic region, or crop type.

Our model, TimeSenCLIP, demonstrates strong performance in matching descriptive text prompts to appropriate satellite imagery, despite the fact that these descriptions are often abstract, high-level, or grounded in human perception (e.g.,*"traditional olive landscapes in Greece, Italy, Spain"*.).

For each text query, we display the Top-5 retrieved satellite patches. These examples showcase the model's ability to generalize across tasks and semantic levels, from broad biogeographic zones to fine-grained crop types. Retrieved visually and semantically consistent matches, indicating that the learned representation aligns textual semantics with satellite spectral–temporal features effectively.

This experiment demonstrates the practical utility of TimeSenCLIP in real-world scenarios, where users can retrieve relevant satellite imagery using natural language descriptions without relying on predefined class labels. It also underscores the model's strong zero-shot generalization capabilities, as the textual prompts and categories used during retrieval were not part of the training process. Instead, the model learns to associate semantic meaning through supervision from ground-level imagery, enabling it to match unseen descriptions to appropriate satellite scenes.

6.4. Discussion

Temporal dynamics play an important role in learned feature representations from remote sensing data, providing the strongest discriminative signal across zero-shot classification and cross-view retrieval, particularly for phenology-driven categories such as Crops and Bioregions where static spatial appearance is not informative enough. When temporal frequency is dense, single-pixel time series often match or exceed larger spatial patches, indicating that temporal information can partially compensate for limited spatial context and challenging the assumption that broader spatial extent is always required for geospatial understanding. Nevertheless, the relative contribution of temporal and spatial cues is task-dependent: Crops and Bioregions are primarily influenced by seasonal dynamics, Land Cover and Land Use benefit from spatial clues, and Habitat shows limited sensitivity to temporal depth, reflecting its reliance on fine spatial structure or additional semantic priors.

Ablation studies provide ecological insight, showing that time series dynamics alone capture phenologically distinct ecosystem signatures, with ecosystem identity expressed through temporal–spectral patterns rather than spatial structure. Higher accuracy for bioregions than finer crop or habitat classes reflects the greater temporal stability of large-scale ecological systems, while robustness to temporal dropout suggests ecosystems differ in resilience to seasonal gaps. Phenology thus serves as a proxy for ecosystem function, enabling ecological mapping with limited spatial context (Lausch et al., 2016).

Multispectral inputs enhance these temporal representations, improving the separability of visually similar classes with distinct spectral–temporal dynamics, especially for Land Cover and Crops. Textual descriptive prompts also provide complementary guidance, injecting domain knowledge not readily observable from imagery alone and reinforcing the importance of prompt design in zero-shot remote sensing models.

Finally, these performance gains are achieved with substantial computational efficiency. TimeSenCLIP's temporal transformer processes multi-spectral time series for single pixels or small patches with dramatically lower FLOPs, memory, and inference time compared to standard CLIP-based architectures, while maintaining or exceeding accuracy. This efficiency enables scalable, high-throughput analysis of large satellite datasets, highlighting a favorable trade-off between representational richness and computational cost.

7. Conclusion

This work presents TimeSenCLIP, a temporally and spectrally-aware multimodal framework for zero-shot ecological classification and retrieval. By leveraging fine-grained spectral–temporal inputs and regularization strategies, TimeSenCLIP demonstrates consistent and substantial improvements over existing baselines across diverse remote sensing tasks, including Land Cover, Land Use, habitat mapping, Bioregions, Crop types, and Scenicness prediction. Our findings highlight the importance of taking into account temporal dynamics particularly at a monthly resolution to capture seasonal and phenological patterns that are critical for fine-grained geospatial understanding.

Ablation studies confirm that temporal dropout enhances robustness to missing or inconsistent observations, while spatial context offers task-specific benefits. Notably, TimeSenCLIP maintains strong performance even when constrained to single pixel inputs, underscoring the power of spectral–temporal learning via contrastive learning. For tasks such as Crop type classification and Scenicness prediction, temporal ensembling and prompt diversification further improve performance, illustrating the value of aligning learned representations with temporal variations.

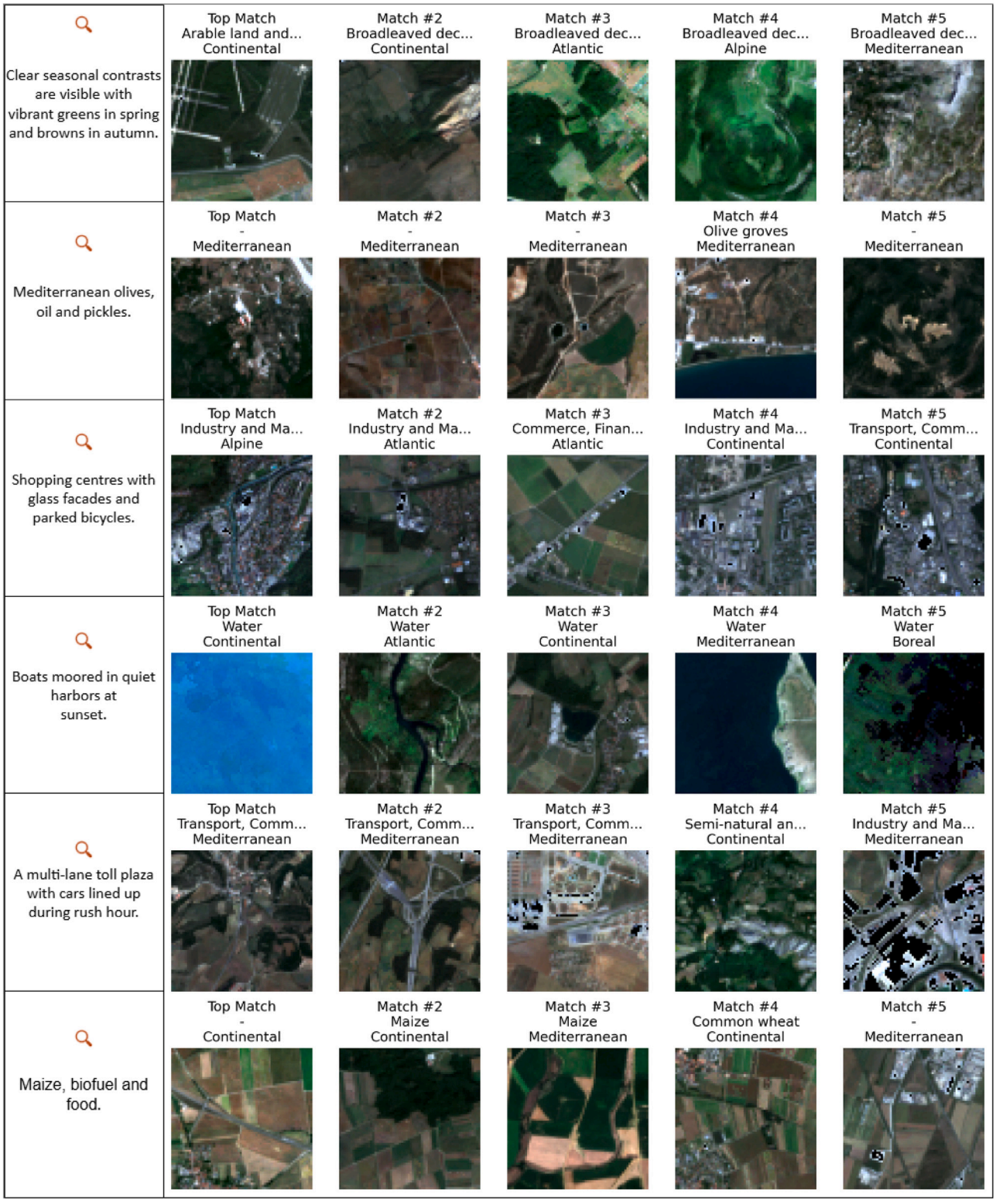


Fig. 6. Text-to-image retrieval performance using TimeSenCLIP with descriptive prompts. While the model processes single pixel monthly temporal composites during operation, these visualizations show representative RGB 64 × 64 patches for interpretability. Results demonstrate the model's ability to match textual descriptions with relevant satellite imagery across diverse land cover, land use, bioregions, habitat and crops classes. Each row presents the Top-5 retrieved time-series with their corresponding class labels and bioregion. Class labels are drawn from LULC, Habitat and Crops classes, when a crop label is unavailable, it is denoted with "-" .

Limitations and future work. Although our model shows strong performance on ecological tasks in Europe, it has only been evaluated using Sentinel-2 and LUCAS data from the EU. This limits the understanding of its generalizability to other sensors and regions with different climates and phenologies. Extending the framework to manage multi-sensor information (e.g. Sentinel-1 and Sentinel-2) and scale up globally by combining multiple sources of ground-level photographs are key directions to improve our approach. Another inherent limitation resides in the CLIP representation itself. Being based on the alignment between Internet photographs and their captions, many characteristics that are only observable across time may not be captured by the CLIP image embeddings, and cannot, thus, be transferred to TimeSenCLIP. This could be achieved by complementing the ground-level image representations with textual geocoded descriptions designed to fill these gaps.

CRediT authorship contribution statement

Pallavi Jain: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Diego Marcos:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Dino Ienco:** Writing – review & editing, Supervision, Resources. **Roberto Interdonato:** Writing – review & editing, Supervision, Resources. **Tristan Berchoux:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of AI-Assisted Writing

During the preparation of this work the author(s) used LLMs in order to assist with sentence structuring, and grammar improvements. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the ‘Giving Rural Actors Novel Data and Re-Usable Tools to Lead Public Action in Rural Areas’ (GRANULAR) project, which has received funding from the European Union’s Horizon Europe Research and Innovation Programme under Grant Agreement No. 101061068. This work was also supported in part by the ANR project OBTEA (ANR-22-CPJ1-0054-01) and I-SITE Excellence Program of the University of Montpellier, under the 3Investissements France 2030.

Appendix A

A.1. Evaluation setup

The overall evaluation framework is illustrated in Fig. A.2. As shown in Section 5.2, we assess the proposed models across three settings that is zero-shot mapping, cross-modal retrieval, and scenicness regression. We describe the detailed evaluation protocol for each task in this section.

A.1.1. Zero-shot evaluation setup

As illustrated in Fig. A.2, zero-shot classification is performed by aligning satellite time-series data with textual prompts in a shared latent space. We utilize the frozen CLIP text encoder to transform textual prompts, ranging from simple class names and generic templates to descriptive, LLM-generated descriptions into textual embeddings. Simultaneously, the proposed TimeSenCLIP encoder processes the satellite time-series into embeddings. Both textual and time-series embeddings are L_2 normalized to ensure a consistent scale. The model identifies the correct class by calculating the cosine similarity between the satellite embedding and each candidate text embedding. The class corresponding to the highest similarity score is selected as the final prediction. This approach eliminates the need for task-specific labels or fine-tuning, allowing the model to generalize to novel categories based purely on their semantic descriptions.

A.1.2. Cross-modal retrieval setup

As illustrated in Fig. A.2, cross-modal retrieval is conducted by aligning ground-level images with corresponding satellite single-pixel time-series in a joint embedding space. The ground-level images are processed through a frozen CLIP vision encoder, while the satellite time-series are encoded using the proposed TimeSenCLIP model. Both representations are L_2 normalized. Retrieval is performed by calculating the cosine similarity between a query embedding from one domain and a gallery of candidate embeddings from the other. For instance, in ground-to-satellite retrieval, the satellite time-series with the highest similarity to the ground-level query is selected as the top match. This capability demonstrates that TimeSenCLIP successfully captures the physical and phenological characteristics of a location, enabling accurate matching across vastly different perspectives without the need for shared metadata or geographical coordinates.

A.1.3. Scenicness regression prompt ensembling setup

Scenicness regression is performed using frozen TimeSenCLIP time-series embeddings together with frozen CLIP text embeddings as done with zero-shot evaluation. The evaluation follows the late prompt-ensembling strategy of Levering et al. (2021). In this setting (Fig. A.2), late ensembling means that the scenicness prediction is built step-by-step for each individual voter before combining the results. First, the image (or time-series data) is converted into image embeddings, while every text description from a voter is converted into text embeddings. The model then compares the image embedding with each text embedding to produce logits per text, which measure how strongly the image matches each description.

Next, these logits are grouped per voter. For each voter, the logits are passed through a softmax function, which turns them into normalized activation values. These activations indicate how much each description contributes to that voter’s scenicness judgment. The activations are then multiplied by the user-provided rating scores, producing a voter-specific scenicness estimate for the image. Finally, the individual scenicness estimates from all voters are averaged together using a regressor to obtain the final scenicness prediction for the image.

This approach differs from early ensembling, where all text descriptions from all voters are merged into a single pool before computing logits and softmax. Early ensembling therefore produces one shared prediction immediately, while late ensembling keeps each voter’s contribution separate until the final averaging step, allowing the model to better reflect differences in human perception of scenicness.

A.2. Scenicness regression analysis

Table A.1 reports the complete scenicness evaluation, including late-ensembling and early strategies, baseline CLIP variants, and TimeSenCLIP models with both single-pixel (P1-MS) and large-patch (P64-RGB) inputs. Baseline models operate on single-temporal inputs only, whereas TimeSenCLIP incorporates temporal encoding and is evaluated

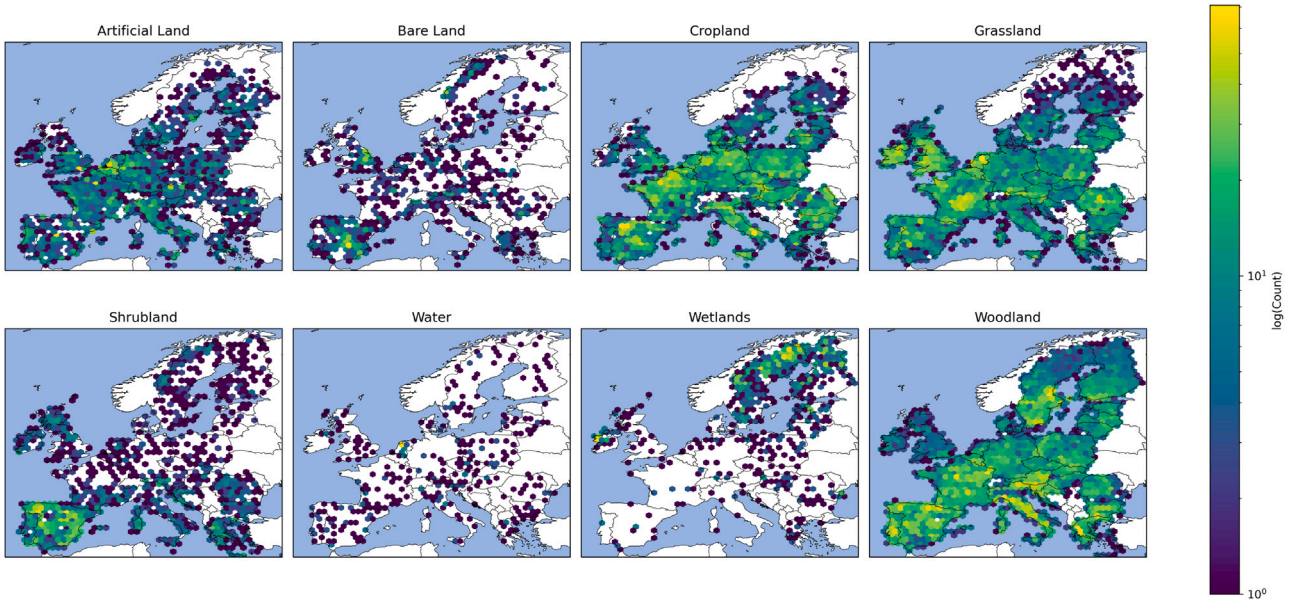


Fig. A.1. Land cover class distribution across the EU in the evaluation dataset.

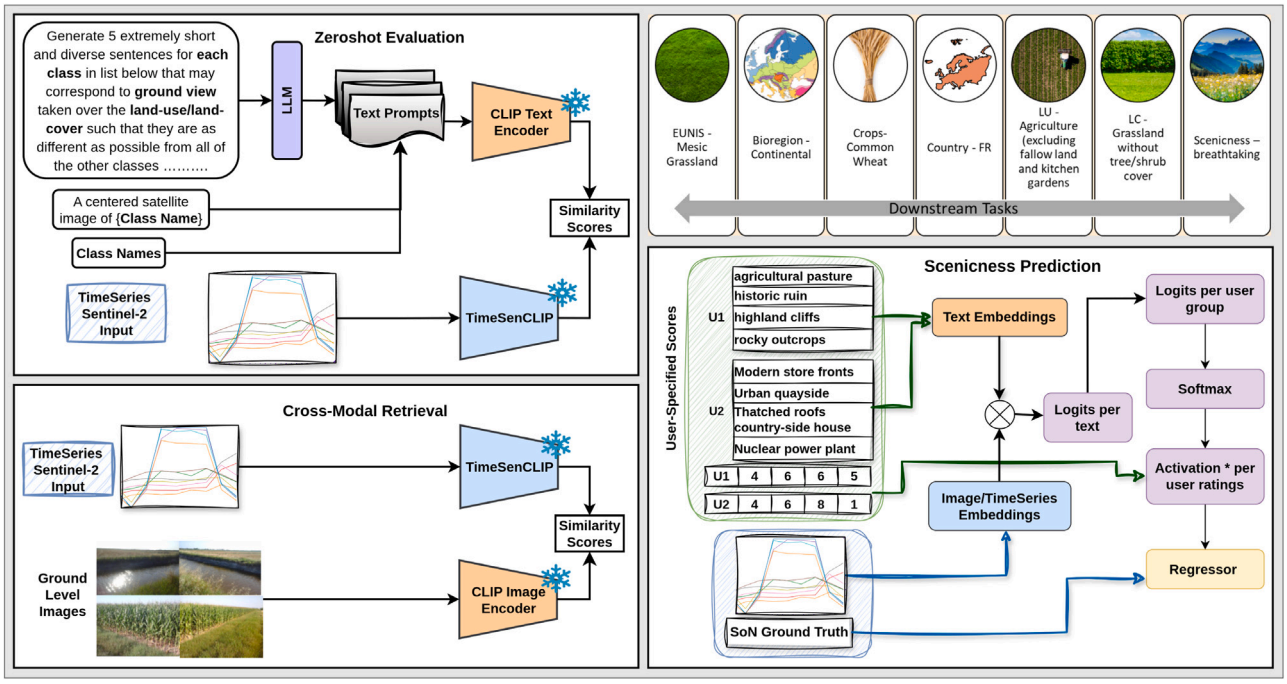


Fig. A.2. The infographic presents the comprehensive evaluation workflow of the TimeSenCLIP model. The pipeline integrates three key components: (1) **Zero-shot classification**, where the trained TimeSenCLIP (time series encoder) and CLIP text encoder (text prompt encoder) are used to perform inference across multiple downstream tasks. (2) **Cross-modal retrieval**, where satellite time-series and ground-level image embeddings are generated using the CLIP image encoder, and similarity scores are computed in both Satellite-to-Ground (S2G) and Ground-to-Satellite (G2S) directions to assess class-consistent retrieval performance; and (3) **Scenicness assessment**, which evaluates perceptual and aesthetic qualities of landscapes using the learned visual embeddings.

under single (T=1), quarterly (T=4), and monthly (T=12) temporal aggregation.

Under single-temporal input, TimeSenCLIP-P64-RGB achieves the strongest satellite-based performance (R=0.634, τ=0.437), surpassing CLIP, GeoRSCLIP, RemoteCLIP, and SenCLIP, and approaching CLIP performance on ground-level imagery (R=0.637, τ=0.450). This highlights two design insights: (1) temporal encoding substantially strengthens the SenCLIP spatial encoder by capturing seasonal and long-term landscape variation; and (2) large spatial context remains important for perceptual attributes such as scenicness, which depend on texture, landform geometry, and spatial composition.

While CLIP remains a strong non-temporal baseline (R=0.517, τ=0.362), it is consistently outperformed by TimeSenCLIP-P64-RGB, demonstrating the added value of temporal cues even for visually static perception tasks. In contrast, TimeSenCLIP-P1-MS underperforms in the single-temporal regime, indicating that temporal richness alone cannot compensate for minimal spatial context when only one timestamp is available.

Table A.1
Scenicness estimation on UK data points using Sentinel-2 satellite imagery. Performance is reported in terms of Pearson's R and Kendall's Tau (τ) across different temporal resolutions: Single, Quarterly (4 months), and Monthly (12 months). Multiple columns correspond to different prompt ensembling strategies as proposed by Levering et al.2024: **Early**, 2, 5, 8, and 10, followed by the average (**Avg**) across all strategies. The first row shows CLIP performance on ground-level images (SoN dataset) for reference. **Bold** text indicates the best overall performance across all temporal settings, while underline text indicates the best performance within each temporal resolution. This table demonstrates that TimeSenCLIP, using either multispectral single-pixel (P1-MS) or larger RGB patches (P64-RGB), can achieve competitive scenicness prediction performance compared to models trained on ground-level images.

T	Model	Early		2		5		8		10		Avg	
		R	τ	R	τ	R	τ	R	τ	R	τ	R	τ
Baselines (RGB and 64 Pix)													
1	CLIP (SON Images)(Levering et al., 2024)	0.544	0.390	0.692	0.485	0.654	0.447	0.673	0.479	0.623	0.450	0.637	0.450
1	CLIP	0.510	0.358	0.547	0.384	0.547	0.384	0.536	0.379	0.444	0.307	0.517	0.362
	GeoRSCLIP	0.478	0.346	0.520	0.378	0.512	0.369	0.466	0.336	0.306	0.212	0.456	0.328
	RemoteCLIP	0.435	0.295	0.477	0.321	0.473	0.318	0.478	0.325	0.397	0.260	0.452	0.304
	SkyCLIP	0.365	0.249	0.207	0.124	0.246	0.154	0.245	0.153	0.391	0.255	0.291	0.187
	SenCLIP	0.396	0.272	0.459	0.310	0.440	0.297	0.420	0.286	0.366	0.241	0.416	0.281
TimeSenCLIP (ours)													
1	P1-MS	0.336	0.223	0.345	0.231	0.346	0.231	0.308	0.200	0.274	0.167	0.322	0.210
	P64-RGB	**0.532**	**0.385**	0.681	0.461	0.672	0.460	0.671	0.461	0.614	0.420	0.634	0.437
4	P1-MS	0.438	0.309	0.479	0.324	0.468	0.325	0.445	0.313	0.413	0.288	0.449	0.312
	P64-RGB	0.514	0.374	0.682	0.461	0.672	0.460	0.671	0.460	0.617	0.420	0.631	0.435
12	P1-MS	0.518	0.361	0.558	0.371	0.548	0.373	0.526	0.366	0.478	0.343	0.526	0.363
	P64-RGB	0.518	0.378	**0.690**	**0.466**	**0.681**	**0.465**	**0.681**	**0.466**	**0.624**	**0.423**	**0.639**	**0.440**

Table A.2
Cross-Modal retrieval performance for Satellite-to-Ground (S2G) and Ground-to-Satellite (G2S) retrieval across temporal model architectures, reported using Recall@1. **Bold** values indicate the best-performing model for each task and retrieval direction.

Model	Land Cover		Land Use		Habitat		Crops		Bioregion		Country		Overall	
	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G	G2S	S2G
CNN1D	**0.632**	0.506	0.656	0.492	0.407	0.303	0.283	0.128	0.366	0.442	0.202	0.181	0.424	0.342
ConvTran	0.618	0.492	**0.684**	0.484	0.370	0.266	**0.297**	0.117	0.372	0.410	0.228	0.143	0.428	0.319
MLP	0.568	0.583	0.674	0.610	**0.413**	0.355	0.295	0.207	0.421	0.510	0.210	0.250	0.430	0.419
TempCNN	0.542	0.573	0.657	0.594	0.394	0.286	0.263	0.167	0.403	0.475	0.192	0.250	0.409	0.391
Transformer (ours)	0.598	**0.666**	0.642	**0.695**	0.378	**0.432**	0.283	**0.434**	**0.534**	**0.648**	**0.253**	**0.416**	**0.448**	**0.549**

Increasing temporal coverage reinforces these trends. Quarterly aggregation improves performance across configurations, and monthly aggregation yields the best overall results, with TimeSenCLIP-P64-RGB reaching R=0.639 and τ=0.440. Importantly, even the single-pixel TimeSenCLIP-P1-MS becomes competitive when enriched with full temporal information (R=0.526, τ=0.363), showing that detailed temporal structure can partially offset limited spatial context.

The temporal encoding markedly enhances scenicness prediction from overhead imagery. TimeSenCLIP-P64-RGB consistently outperforms satellite-only baselines, while single-pixel variants achieve competitive performance only when supported by dense temporal observations. These findings underscore the central role of temporal dynamics in modeling perceptual and aesthetic landscape qualities from multispectral satellite time series.

A.3. Conventional temporal baseline comparison for cross-modal retrieval

While TimeSenCLIP's transformer architecture shows only moderate gains over classical temporal models in zero-shot classification, its advantages are more pronounced in cross-modal retrieval. As shown in Table A.2, TimeSenCLIP generally outperforms CNN1D, ConvTran, MLP, and TempCNN for Satellite-to-Ground (S2G) retrieval, effectively capturing temporal patterns that aid cross-view alignment. For Ground-to-Satellite (G2S) retrieval, some conventional architectures, such as CNN1D or MLP, perform similarly or slightly better for finer-grained tasks like Land Cover, Land Use, Habitat, and Crops, likely because local ground images are easier to match to the broader satellite context. In contrast, for coarser categories such as Bioregion and Country, the transformer's ability to integrate long-term temporal sequences provides a stable signal, improving G2S performance and highlighting the advantages of temporal modeling for large-scale, temporally coherent classes. In general, TimeSenCLIP often reverses the typical S2G/G2S performance gap, outperforming baselines in S2G retrieval and occasionally exceeding G2S accuracy, suggesting its learned embeddings form a stable semantic anchor that supports accurate cross-view localization even under drastic viewpoint differences.

Intuitively, TimeSenCLIP bridges the gap between local ground observations and satellite imagery: it encodes distinctive spectral–temporal patterns while aligning them with rich semantic signals from geo-tagged ground images. This combination enables robust retrieval across diverse land-cover, land-use, ecological, and crop categories, without relying on large spatial patches or extra paired image–text supervision. Its transformer backbone and cross-view training strategy produce embeddings that are both semantically meaningful and generalizable, effectively balancing spatial and temporal cues according to task-specific requirements.

A.4. Ablation

A.4.1. Impact of spatial context

We report the impact of spatial context for cross-modal retrieval in Fig. A.3. Results in both Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G) directions broadly mirror the behavior observed in zero-shot classification, indicating that the balance between spatial and temporal context transfers consistently across evaluation settings.

Single-pixel (P1) time series already achieve strong retrieval accuracy across most tasks, highlighting the discriminative strength of temporal signatures for cross-view alignment. Incorporating limited spatial context through small patches (e.g., 5 × 5 to 16 × 16) yields task-dependent improvements, particularly for categories with structured spatial organization such as land cover and land use. Expanding the spatial extent beyond this range provides minimal additional

Table A.3
Effect of Multispectral (MS) and Temporal (TS) Dropout on retrieval performance (Recall@1). The table reports Recall averaged across S2G and G2S. (✓) indicates dropout applied; (✗) indicates no dropout.

T	TS Drop	MS Drop	Land Cover	Land Use	Habitat	Bioregion	Crops	Country	Average
1	✗	✗	0.290	0.402	0.197	0.243	0.090	0.096	0.219
	✗	✓	0.210	0.358	0.129	0.248	0.114	0.119	0.196
	✓	✗	0.548	**0.609**	**0.343**	**0.353**	**0.203**	**0.159**	**0.369**
	✓	✓	**0.556**	0.608	0.335	0.344	0.169	0.152	0.360
4	✗	✗	0.606	0.650	0.392	0.501	0.298	**0.268**	0.452
	✗	✓	0.607	0.650	0.390	0.467	0.262	0.239	0.436
	✓	✗	**0.628**	0.654	0.384	0.505	**0.308**	0.267	**0.457**
	✓	✓	0.627	**0.658**	**0.393**	**0.508**	0.272	0.261	0.453
12	✗	✗	**0.642**	**0.674**	**0.417**	0.605	**0.372**	**0.389**	**0.516**
	✗	✓	0.623	0.667	0.413	**0.610**	0.356	0.365	0.505
	✓	✗	0.632	0.669	0.405	0.591	0.359	0.335	0.498
	✓	✓	0.631	0.665	0.409	0.584	0.337	0.327	0.492

Table A.4
Comparison of ground image aggregation strategies. "Single image" uses a single randomly selected image from four available images, while "Average Pooling" computes the mean embedding across all four images.

Pooling	Land Cover	Land Use	Habitat	Crops	Bioregion	Average
Average	60.85	58.65	29.02	23.53	34.46	41.30
Single image	60.07	57.25	30.23	24.47	34.17	41.24

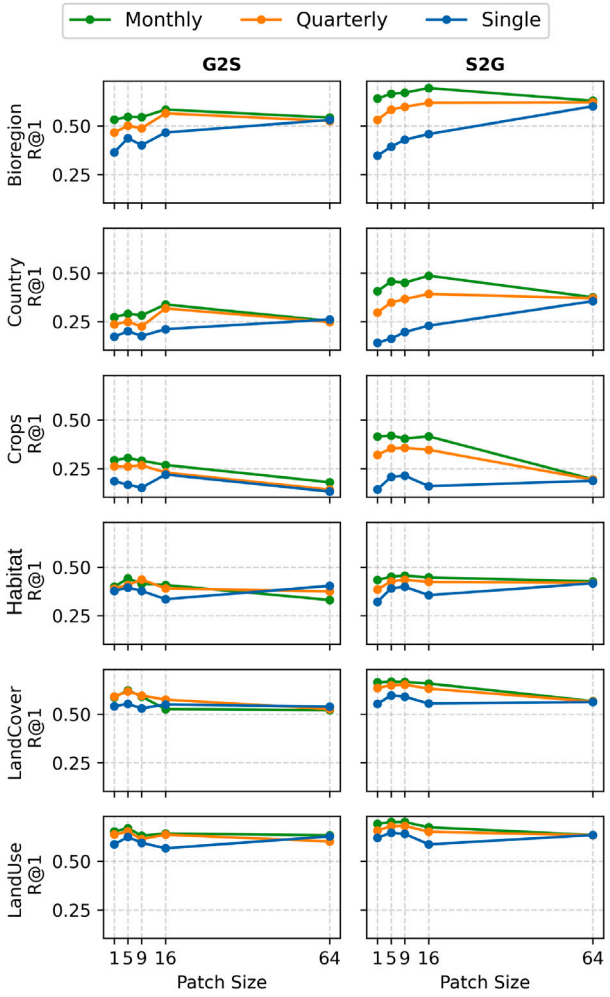


Fig. A.3. Impact of spatial information on cross-modal retrieval for S2G and G2S tasks across different semantic categories. Trends mirror zero-shot classification, with single-pixel sequences remaining competitive.

benefit and is mainly advantageous for coarse or spatially homogeneous classes, where broader context improves stability rather than fine-grained discrimination.

Unlike zero-shot evaluation, which remains largely symmetric, retrieval reveals mild differences between G2S and S2G. G2S performance is relatively insensitive to patch size, with temporal depth accounting for most gains and larger patches helping only in coarse categories (e.g., Bioregion or Country). In contrast, S2G retrieval shows greater sensitivity to spatial context: small patches can improve alignment when ground queries contain meaningful structure, whereas overly large patches may introduce irrelevant variation and slightly reduce accuracy. Nevertheless, temporally rich P1 sequences remain competitive across all tasks and directions.

The variation across patch sizes is modest compared with the influence of temporal coverage, confirming temporal dynamics as the primary factor governing cross-view retrieval. In several settings, sufficiently long P1 time series approach the performance of substantially larger spatial contexts, demonstrating that strong temporal information can compensate for limited spatial extent and enabling scalable retrieval without large image patches.

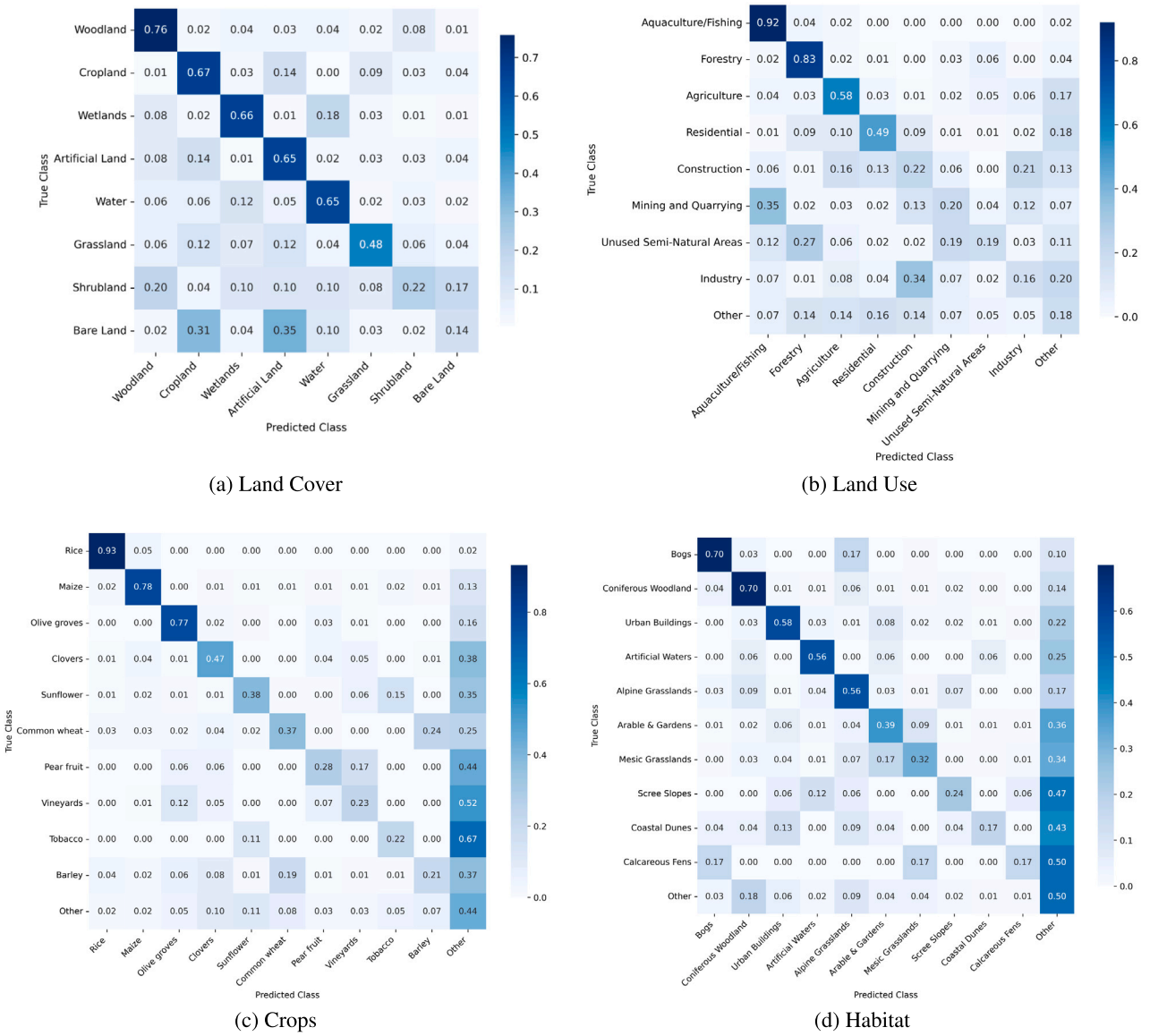
A.4.2. Impact of dropout strategies

For cross-modal retrieval, we extend the analysis to include both single, quarterly, and monthly temporal aggregations. Retrieval performance, averaged across Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G) directions, follows the same trends observed in zero-shot classification (Table A.3). Temporal dropout consistently improves alignment when temporal coverage is limited, with Recall@1 increasing from 0.219 to 0.369 for single timestamps. Incorporating MS Drop does not yield complementary benefits and can even harm performance in tasks reliant on spectral information, such as Crops and Habitat.

While TS Drop provides the largest gains for single timestamps, its effect diminishes as temporal depth increases to quarterly or monthly sequences. Single-pixel multispectral time series still achieve competitive retrieval performance, demonstrating that temporal information is the dominant factor for cross-view alignment, while spatial and spectral cues provide additional task-dependent gains.

A.4.3. Impact of ground image aggregation

We evaluate two strategies for incorporating multiple ground-level images per location: Average Pooling, which aggregates embeddings across all available images, and Single Image, which uses one randomly



(a) Land Cover (b) Land Use

(c) Crops (d) Habitat

Fig. A.4. Normalized confusion matrices for TimeSenCLIP, showing the top 8 predicted classes for LULC and the top 10 predicted classes for Crops and Habitat. Classes beyond these top predictions are grouped as "Other". All evaluations are performed in a zeroshot setting on monthly, single-pixel TimeSenCLIP inputs using descriptive text prompts.

selected image without pooling. As summarized in Table A.4, Average Pooling consistently performs slightly better than the Single Image setting across most tasks. However, the overall performance gap is negligible, with the mean Top-1 accuracy varying only marginally from 41.24 to 41.30.

These findings suggest that, although aggregating multiple ground images can provide minor improvements, a single representative image is sufficient to achieve nearly identical retrieval performance. This observation reduces the reliance on collecting and processing multiple images per location, thereby simplifying data requirements and computational overhead. Importantly, it supports a more scalable training and deployment protocol for large-scale cross-view remote sensing applications, where dense ground-image coverage is often impractical.

A.4.4. Zero-shot class-wise classification

To provide further insight, we include a class-wise zero-shot classification behavior. Fig. A.4 shows confusion matrices for four zero-shot classification tasks: Land Cover, Land Use, Crops, and Habitat, using TimeSenCLIP-P1-MS (monthly) with descriptive prompts. For readability, we display the top classes per task (8-Land Use and 10-for Crops and Habitat) and group remaining categories as "Other". Overall, dominant and well-defined classes are predicted reliably, while most errors occur between semantically or structurally similar categories, reflecting intrinsic ambiguity rather than systematic failure. For example, in Crops, confusion is concentrated among phenologically similar cereals such as wheat and barley, whereas in Land Cover, shrubland is occasionally misclassified as woodland due to overlapping seasonal signatures. Habitat-level errors primarily arise between ecologically related classes, such as grasslands and agricultural areas, which share similar vegetation structure and temporal patterns.

References

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G., et al., 2020. Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European union. Sci. Data 7 (1), 352.

Daroya, R., Cole, E., Mac Aodha, O., Van Horn, G., Maji, S., 2024. WildSAT: Learning satellite image representations from wildlife observations. arXiv preprint arXiv:2412.14428.

Dhakal, A., Ahmad, A., Khanal, S., Sastry, S., Jacobs, N., 2023. Sat2cap: Mapping fine-grained textual descriptions from satellite images. arXiv preprint arXiv:2307.15904.

Elgendy, H., Sharshar, A., Aboeitta, A., Ashraf, Y., Guizani, M., 2024. Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing. arXiv preprint arXiv:2410.19552.

European Environment Agency, 2016. Biogeographical regions. European Environment Agency Datahub, Dataset published on 26 Jan 2016, last modified 30 Apr 2025.

European Environment Agency, 2019. Ecosystem types of europe 2012 – terrestrial habitats – version 3 revision 1, feb. 2019. European Environment Agency Spatial Data Catalogue, Dataset published on 31 Dec 2018, last modified on 27 Apr 2021.

Foumani, N.M., Tan, C.W., Webb, G.I., Salehi, M., 2024. Improving position encoding of transformers for multivariate time series classification. Data Min. Knowl. Discov. 38 (1), 22–48.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Jain, P., Ienco, D., Interdonato, R., Berchoux, T., Marcos, D., 2025. Senclip: Enhancing zero-shot land-use mapping for sentinel-2 with ground-level prompting. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 5656–5665.

Keith, D.A., Ferrer-Paris, J.R., Nicholson, E., Bishop, M.J., Polidoro, B.A., Ramirez-Llodra, E., Tozer, M.G., Nel, J.L., Mac Nally, R., Gregr, E.J., et al., 2022. A function-based typology for Earth's ecosystems. Nature 610 (7932), 513–518.

Kiranyaz, S., Ince, T., Gabbouj, M., 2015. Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Trans. Biomed. Eng. 63 (3), 664–675.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M., 2025. Satclip: Global, general-purpose location embeddings with satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 4347–4355.

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., Burnaev, E., 2024. Challenges in data-driven geospatial modeling for environmental research and practice. Nat. Commun. 15 (1), 10700.

Lausch, A., Bannehr, L., Beckmann, M., Boehm, C., Feilhauer, H., Hacker, J., Heurich, M., Jung, A., Klenke, R., Neumann, C., et al., 2016. Linking earth observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. Ecol. Indic. 70, 317–339.

Levering, A., Marcos, D., Jacobs, N., Tuia, D., 2024. Prompt-guided and multimodal landscape scenicness assessments with vision-language models. PLoS One 19 (9), e0307083.

Levering, A., Marcos, D., Tuia, D., 2021. On the relation between landscape beauty and land cover: A case study in the UK at sentinel-2 resolution with interpretable AI. ISPRS J. Photogramm. Remote Sens. 177, 194–203.

Li, Z., Xu, D., Guo, X., 2014. Remote sensing of ecosystem health: opportunities, challenges, and future perspectives. Sensors 14 (11), 21117–21139.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Zhou, J., 2023. RemoteCLIP: A vision language foundation model for remote sensing. arXiv preprint arXiv:2306.11029.

Liu, Z., Zhang, F., Jiao, J., Lao, N., Mai, G., 2025. GAIR: Improving multimodal geo-foundation model with geo-aligned implicit representations. arXiv preprint arXiv:2503.16683.

Mall, U., Phoo, C.P., Liu, M.K., Vondrick, C., Hariharan, B., Bala, K., 2023. Remote sensing vision-language foundation models without annotations via ground remote alignment. arXiv preprint arXiv:2312.06960.

Marimo, C.T., Blumenstiel, B., Nitsche, M., Jakubik, J., Brunschwiler, T., 2025. Beyond the visible: Multispectral vision-language learning for earth observation. arXiv preprint arXiv:2503.15969.

Menon, S., Vondrick, C., 2022. Visual classification via description from large language models. arXiv preprint arXiv:2210.07183.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Pelletier, C., Webb, G.I., Petitjean, F., 2019. Temporal convolutional neural network for the classification of satellite image time series. Remote. Sens. 11 (5), 523.

Pesaresi, S., Mancini, A., Quattrini, G., Casavecchia, S., 2022. Functional analysis for habitat mapping in a special area of conservation using sentinel-2 time-series data. Remote. Sens. 14 (5), 1179.

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T.P., Fernandez, M., Lucas, R., Schaepman, M.E., Wang, T., O'Connor, B., et al., 2016. Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. Remote. Sens. Ecol. Conserv. 2 (3), 122–131.

Qiu, C., Zhang, X., Tong, X., Guan, N., Yi, X., Yang, K., Zhu, J., Yu, A., 2024. Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. ISPRS J. Photogramm. Remote Sens. 209, 368–382.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, F., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566 (7743), 195–204.

Roth, K., Kim, J.M., Koepke, A., Vinyals, O., Schmid, C., Akata, Z., 2023. Waffling around for performance: Visual classification with random words and broad concepts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15746–15757.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS Int. J. Geo-Inform. 7 (4), 129.

Sainte Fare Garnot, V., Landrieu, L., 2020. Lightweight temporal self-attention for classifying satellite images time series. arXiv preprint arXiv:2007.00586.

Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., Jacobs, N., 2025. Taxabind: A unified embedding space for ecological applications. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 1765–1774.

Seresinhe, C.I., Preis, T., Moat, H.S., 2015. Quantifying the impact of scenic environments on health. Sci. Rep. 5 (1), 16899.

Shabbir, A., Zumri, M., Bennamoun, M., Khan, F.S., Khan, S., 2025. GeoPixel: Pixel grounding large multimodal model in remote sensing. arXiv preprint arXiv:2501.13925.

Sharma, S., Sedona, R., Riedel, M., Cavallaro, G., Paris, C., 2024. Sen4map: Advancing mapping with sentinel-2 by providing detailed semantic descriptions and customizable land-use and land-cover data. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..

Soni, S., Dudhane, A., Debary, H., Fiaz, M., Munir, M.A., Danish, M.S., Fraccaro, P., Watson, C.D., Klein, L.J., Khan, F.S., et al., 2024. Earthdial: Turning multi-sensory earth observations to interactive dialogues. arXiv preprint arXiv:2412.15190.

Soubry, I., Doan, T., Chu, T., Guo, X., 2021. A systematic review on the integration of remote sensing and GIS to forest and grassland ecosystem health attributes, indicators, and measures. Remote. Sens. 13 (16), 3262.

Tan, X., Xi, B., Li, J., Zheng, T., Li, Y., Xue, C., Chanussot, J., 2024. Review of zero-shot remote sensing image scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens..

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al., 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786.

Vivanco Cepeda, V., Nayak, G.K., Shah, M., 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Adv. Neural Inf. Process. Syst. 36, 8690–8701.

Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R., 2024. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5805–5813.

Wang, Y., Xiong, Z., Liu, C., Stewart, A.J., Dujardin, T., Bountos, N.I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., et al., 2025. Towards a unified copernicus foundation model for earth vision. arXiv preprint arXiv:2503.11849.

Workman, S., Souvenir, R., Jacobs, N., 2017. Understanding and mapping natural beauty. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5589–5598.

Zermatten, V., Castillo-Navarro, J., Jain, P., Tuia, D., Marcos, D., 2025. EcoWikiRS: Learning ecological representation of satellite images from weak supervision with species observations and wikipedia. arXiv preprint arXiv:2504.19742.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. IEEE Geosci. Remote. Sens. Mag. 4 (2), 22–40.

Zhang, Z., Zhao, T., Guo, Y., Yin, J., 2023. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. arXiv preprint arXiv:2306.11300.

Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sens. Environ. 221, 430–443.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote. Sens. Mag. 5 (4), 8–36.