LEVERAGING ECOLOGICAL CONTEXT IN BIOLOGICAL FOUNDATION MODELS AT SCALE

Anonymous authors
Paper under double-blind review

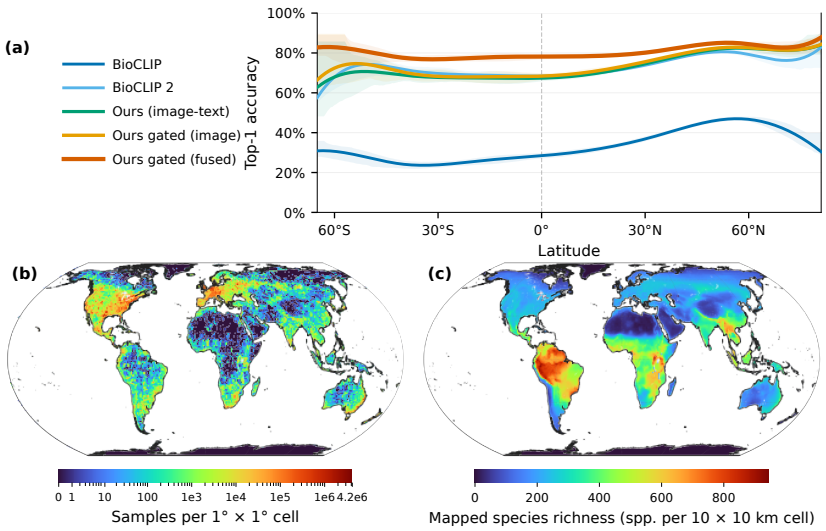


Figure 1: Geographic variation in model performance, available training data, and biodiversity richness. **(a)** Top-1 taxonomic retrieval accuracy across latitudes for BioCLIP, BioCLIP 2, and variants of our model. Smoothed curves highlight geographic variation in accuracy and gains from contextual information. **(b)** Global distribution of GBIF observation data, binned into $1^\circ \times 1^\circ$ grid cells. **(c)** Estimated terrestrial species richness derived from expert-curated biodiversity maps (production described in Appendix A.1). Differences between **(b)** and **(c)** illustrate the (especially North-South) mismatch between data availability and regions of high biodiversity richness.

ABSTRACT

Understanding the natural world often requires more than visual information: where and when an organism is observed provides critical evidence about its identity. Yet biological foundation models are trained predominantly from images and text, without exploiting readily available ecological context. We introduce a framework that incorporates geographic location and time of day and year directly into the semantic space of biological vision–language models. Using a novel dataset of more than 50 million multimodal observations, we train an ecological context encoder aligned to the BioCLIP 2 embedding space and an adaptive fusion mechanism combining ecological and visual evidence learned per-observation. Unlike other species-prior approaches, our method does not require a fixed species-specific classification head, preserving a flexible text-based retrieval interface in which candidate taxa are represented through their text embeddings. Across held-out species-recognition benchmarks, including novel biodiversity-weighted evaluation sets designed to better reflect biodiversity-rich regions underrepresented by observation effort, ecological context consistently improves recognition performance when combined with visual representations. Improvements are largest in geographically under-observed regions where visual training data are comparatively sparse. Our results demonstrate that large-scale ecological metadata provides a complementary modality for biological foundation models and offers a scalable route toward geographically robust biodiversity recognition.

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

1 INTRODUCTION

As ecosystems face increasing anthropogenic pressures, scalable technologies for extracting information from the natural world are becoming critical (IPBES, 2019; Pollock et al., 2025). Recent advances in vision–language models (VLMs) have enabled the development of biological foundation models trained on large-scale observational data (Stevens et al., 2024; Gu et al., 2025). They learn transferable visual and semantic representations that can be applied across a wide range of downstream tasks, and are now being used to support tasks across biology (Pahuja et al., 2026; Vendrow et al., 2024) and ecological monitoring (Gardiner et al., 2025; Vyskočil & Picek, 2025).

A particular challenge arises from the uneven global distribution of biodiversity observations. Biodiversity itself is highly heterogeneous, with species richness generally increasing from the poles toward the tropics (Rohde, 1992). In contrast, major biodiversity repositories such as GBIF (the Global Biodiversity Information Facility) are geographically biased towards regions with greater observation effort, particularly in the Global North (Beck et al., 2014; Meyer et al., 2016; Boakes et al., 2010). So, many biodiversity-rich regions remain comparatively underrepresented in the datasets used to train machine-learning systems. Models therefore often have the least training data available in locations where scalable identification tools may be most valuable. One source of information that may help address these limitations is ecological context. Species recognition is inherently multimodal: human experts identify organisms not only by morphology, but also by contextual information such as geographic range, season, and time of day (Alderfer & Dunn, 2007). This is useful when visual evidence is ambiguous, e.g. through poor image quality, convergent evolution, mimicry, camouflage, or when an assigned taxonomic label is not strongly attributed to appearance.

Recent multimodal approaches have demonstrated the value of complementary signals including those as disparate as geography, environmental variables, audio, satellite imagery, and DNA (Terry et al., 2020; Daroya et al., 2025; Sastry et al., 2025; Klemmer et al., 2025; Gong et al., 2025). Yet biological VLMs, those which model the greatest diversity of taxa, remain predominantly trained through image–text supervision, with recent performance gains driven in large part by increasing model and dataset scale. At the same time, many of the biodiversity observations used to construct these training sets already contain rich contextual metadata, including geographic coordinates and observation timestamps (Global Biodiversity Information Facility, 2026). This represents a substantial source of ecological information that remains largely unused by current biological VLMs.

In this work, we investigate how observation-level ecological context can be integrated directly into the semantic space of pretrained biological VLMs and used to improve species recognition. We focus on geographic location, time of year, and time of day; these signals are particularly attractive because they encode fundamental ecological information, are widely available and inexpensive to capture. We study how context can be represented within a model's existing taxonomic semantic space and combined with visual evidence. This presents two challenges. First, geographic coordinates and temporal variables differ fundamentally from images and text, yet must be mapped into a representation compatible with the pretrained embedding space. Second, the usefulness of ecological context is observation dependent: it may strongly constrain the identity of a visually ambiguous organism, but contribute little when visual evidence is already enough. We therefore learn an ecological representation aligned with model's existing semantic space and develop an adaptive mechanism for combining ecological and visual evidence on a per-sample basis.

To study this opportunity at scale, we carefully construct a multimodal biodiversity corpus of more than 50 million image–taxonomy–context triplets over 151.7K taxa. We also produce biodiversity-weighted benchmarks which are designed to specifically evaluate performance in regions where biodiversity is most prevalent. Our results show that ecological context provides substantial complementary information to images over all tests. This demonstrates that readily available metadata provides a practical additional axis for scaling biological recognition across Earth.

2 RELATED WORK

2.1 BIOLOGICAL FOUNDATION MODELS

Large-scale pretrained models have become increasingly effective for biological image understanding. There are many prior approaches (Van Horn et al., 2021; Nguyen et al., 2024) which demon-

strate the value of large-scale pretraining and general self-supervised models such as DINO (Oquab et al., 2024) also exhibit strong representations for natural imagery (Gu et al., 2025). Increasingly, the CLIP paradigm has become influential in biology because contrastive image–text training provides a flexible alternative to fixed classification heads as the amount and diversity of biological training data continue to grow. BioCLIP adapts CLIP to organismal biology by contrastively aligning images with taxonomic text, enabling zero-shot recognition and downstream transfer (Stevens et al., 2024). Subsequent work has continued to scale this paradigm: BioCLIP 2 trains on more than 200 million organism images (Gu et al., 2025), while BioTrove develops large-scale biological CLIP models from over 160 million curated images (Yang et al., 2024). These models substantially improve biological visual representations over generic pretrained CLIP baselines. Building on this foundation, we ask how the same architectural advantages can be extended with complementary ecological information.

2.2 Contrastive Multimodal Representation Learning

Contrastive learning provides a natural mechanism for aligning heterogeneous sources of information (Radford et al., 2021; Girdhar et al., 2023; Cepeda et al., 2023). Within biology, CLIBD jointly represents images, DNA barcodes, and taxonomic text (Gong et al., 2025), while TaxaBind aligns imagery with geographic, environmental, satellite, audio, and textual information (Sastry et al., 2025). BioCAP extends BioCLIP with textual descriptions (Zhang et al., 2026), and SatCLIP learns geographic representations by contrastively aligning coordinates with multispectral satellite imagery (Klemmer et al., 2025). These works demonstrate the value of information beyond image–text supervision, but many additional modalities require dedicated acquisition or processing pipelines. Satellite and environmental data depend on external data products, while DNA barcodes require biological sampling and sequencing (Hajibabaei et al., 2007). In contrast, geographic coordinates and timestamps are routinely recorded alongside biodiversity observations, making them available at very large scale with little additional collection cost. They also provide excellent value as they directly capture three fundamental ecological dimensions: where an organism occurs, when within the seasonal cycle it occurs, and when within the diel cycle it is active.

2.3 Geospatial Priors for Fine-Grained Recognition

Geographic and temporal information have long been used as complementary signals for fine-grained species recognition. BirdSnap and subsequent work combine visual predictions with spatiotemporal species priors, showing that presence-only observations can substantially improve recognition when appearance is ambiguous (Berg et al., 2014; Mac Aodha et al., 2019; Chu et al., 2019). More recent approaches such as SINR learn continuous representations of species distributions from large-scale occurrence data, while related work connects geographic representations with semantic or visual embedding spaces (Cole et al., 2023; Cepeda et al., 2023). These approaches typically associate geographic representations with species-specific outputs over a training taxonomy, which is more challenging to scale as taxonomic classes increase. Here, we learn ecological context directly within the semantic space of a biological VLM, allowing ecological context and visual evidence to be compared and combined within the same semantic representation space without introducing a separate species-specific classification head.

3 Method

3.1 Problem Formulation

Each sample i includes an image x_i, taxonomic label y_i, and spatiotemporal metadata m_i, comprising latitude lat_i, longitude lon_i, day of year d_i, and hour of day h_i. We start from a pretrained biological VLM with image encoder f_V and text encoder f_T, producing visual and taxonomic text representations v_i and q_i in a shared D-dimensional semantic space. We introduce an ecological encoder f_E that maps m_i into the same space, yielding ecological representation e_i.

Our central design choice is to formulate ecological recognition as retrieval in the shared semantic space rather than classification over a fixed set of taxa. Candidate taxa are represented by the text encoder and ranked by similarity to the ecological embedding e_i. This offers three advantages over a

conventional prior: taxonomic and ecological structure can be shared across taxa rather than learned through independent class outputs; the ecological encoder retains a fixed output dimensionality (D) instead of requiring an output layer that grows with the number of taxa; and candidate taxa remain represented through text embeddings without modifying the ecological encoder or adding a species-specific classification head. These properties are particularly useful for large biological datasets containing many taxa. Figure 2 summarises the architecture.

3.2 Ecological Context Encoder

The ecological encoder captures geographic variation across multiple spatial scales together with seasonal and diel structure. Following GeoCLIP (Cepeda et al., 2023), we first project latitude and longitude using the Equal Earth projection (Šavrič et al., 2019),

$$p_i = \pi(\mathrm{lat}_i, \mathrm{lon}_i) \in \mathbb{R}^2. \tag{1}$$

We represent p_i at three spatial scales, $\sigma \in \{1, 16, 256\}$, using independent Gaussian random Fourier feature mappings (Rahimi & Recht, 2007),

$$\phi_i^{(\sigma)}(p_i) \in \mathbb{R}^{512}. \tag{2}$$

Each mapping samples 256 Fourier frequencies, producing 512 sine–cosine features. The three frequency scales allow the encoder to represent both broad and fine-grained geographic variation.

Day of year d_i and hour of day h_i are cyclic variables with known periodicities, and are therefore encoded using paired sine and cosine components, with a 366-day period to accommodate leap years:

$$c_i = \left[\sin\left(\frac{2\pi h_i}{24}\right), \cos\left(\frac{2\pi h_i}{24}\right), \sin\left(\frac{2\pi d_i}{366}\right), \cos\left(\frac{2\pi d_i}{366}\right)\right]. \tag{3}$$

For each spatial scale, the temporal encoding is concatenated with the corresponding Fourier representation,

$$z_i^{(\sigma)} = \left[\phi_i^{(\sigma)}(p_i); c_i\right] \in \mathbb{R}^{516}. \tag{4}$$

Each $z_i^{(\sigma)}$ is processed by an independent branch comprising three fully connected 1024-unit layers with ReLU activations, followed by a linear projection to the 768-dimensional BioCLIP 2 embedding space. The three scale-specific outputs are summed to produce the final ecological representation e_i. This preserves separate processing of geographic structure at different spatial frequencies while incorporating seasonal and diel context into each branch.

3.3 Vision–Context Fusion

The relative reliability of visual and ecological evidence varies across observations. We L_2-normalise v_i and e_i to obtain $\hat{v}_i$ and $\hat{e}_i$, then predict a scalar gate $g_i \in (0, 1)$ from their concatenation. The gate is learned during multimodal training and controls their weighted combination, f_i:

$$g_i = \mathrm{sigmoid}\left(\mathrm{MLP}\left([\hat{v}_i; \hat{e}_i]\right)\right). \tag{5}$$

$$f_i = g_i \hat{v}_i + (1 - g_i)\hat{e}_i. \tag{6}$$

As $g_i \to 1$, f_i relies more on visual evidence; as $g_i \to 0$, it relies more on ecological context. Our scalar gate keeps the fusion simple and interpretable. Because both modalities share the same semantic space, f_i remains directly comparable with taxonomic text embeddings for retrieval. If ecological context is unavailable, inference can use the visual representation alone.

3.4 Training Procedure

Stage 1: Ecological–text alignment. We first train the ecological encoder while keeping the BioCLIP 2 image and text encoders frozen; no image information is used. Following (Sastry et al., 2025), we optimise a multi-positive InfoNCE (van den Oord et al., 2018) objective $\mathcal{L}(e, q)$, treating observations of the same species within a minibatch as positives (Appendix A.2). This aligns the randomly initialised ecological encoder to BioCLIP 2's existing taxonomic semantic space.

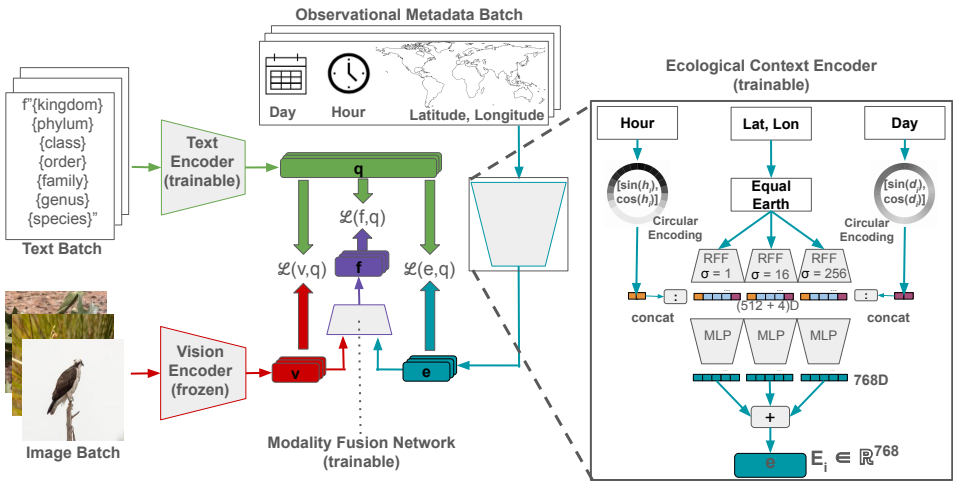


Figure 2: Architecture overview of the proposed multimodal framework. Taxonomic text is passed through the trainable text encoder to produce text embeddings (q), images are encoded by the frozen vision encoder to produce visual embeddings (v). Geographic and temporal metadata are processed by the ecological context encoder to produce 768-dimensional context embeddings (e). Our fusion network adaptively combines (v) and (e) to produce fused embeddings (f) per sample. Contrastive training objectives between the visual, ecological, fused, and text representations are denoted by $\mathcal{L}$.

Stage 2: Multimodal adaptation. We initialise the ecological encoder from Stage 1 and jointly optimise it with the text encoder and fusion module, while keeping the visual encoder frozen. We use a summative loss, $\mathcal{L} = (\mathcal{L}(e,q) + \mathcal{L}(v,q) + \mathcal{L}(f,q))$. The ecological–text loss preserves the alignment learned in Stage 1, the visual–text loss keeps the adapted text space anchored to the frozen visual representation, and the fused–text loss trains the representation used for multimodal recognition. The fusion gate is initialised to ($g_i = 0.5$) for every observation by zero-initialising its final layer, while the hidden layer is Xavier initialised. Training then learns observation-specific departures from this equal weighting.

4 Dataset

4.1 Foundation-scale Biodiversity Corpus

Existing biodiversity datasets provide complementary combinations of scale, taxonomic coverage, and metadata, but none jointly provide the number of observations, species diversity, and observation-level spatial and temporal information required for our study (Table 1). We therefore construct a new corpus from image-bearing occurrence records provided by GBIF. We initially retrieve approximately 160 million observations with taxonomic annotations, geographic coordinates, observation date, and observation time. Because GBIF sampling effort is highly non-uniform, we spatially and temporally downsample the corpus to reduce prominent sampling biases while retaining broad geographic and taxonomic coverage; full details of this and the entire filtering and balancing procedure are provided in Appendix A.3. Our resulting corpus contains 50.4 million image–taxonomy–metadata triplets spanning 151.7K taxa. Each observation includes geographic coordinates, day of year, and time of day in addition to its image and taxonomic annotation.

GBIF also contributes substantially to TreeOfLife-200M, the pretraining dataset used by BioCLIP 2 (Gu et al., 2025). We therefore preserve GBIF occurrence identifiers throughout dataset construction and cross-reference our corpus against TreeOfLife-200M. Overlapping occurrences are permitted during training but excluded from all validation and test sets, ensuring that evaluation is performed on occurrences not matched to the BioCLIP 2 pretraining corpus. Also, because TreeOfLife-200M is constructed from a large fraction of image-bearing GBIF records, and GBIF itself aggregates substantial numbers of observations originating from platforms such as iNaturalist, prior established biodiversity benchmarks derived from the same sources (e.g. Sastry et al. (2025);

Table 1: Comparison with existing large-scale biodiversity datasets. A checkmark indicates that the corresponding modality is available at the observation level.

Dataset	# Samples	# Taxa	Taxonomic	Image	Location	Day-of-year	Time-of-day
Insect-1M	1M	34.2K	✓	✓	–	–	–
TreeOfLife-10M	10.1M	454.1K	✓	✓	–	–	–
BioTrove	161.9M	366.6K	✓	✓	–	–	–
TreeOfLife-200M	213.9M	867.5K	✓	✓	–	–	–
BIOSCAN-5M	5.15M	22.6K	✓	✓	✓	–	–
TaxaBind / TaxaBench-8k	8.8K	2.2K	✓	✓	✓	–	–
CrypticBio	166.5M	67.1K	✓	✓	✓	✓	–
iNaturalist-2021	2.7M	10K	✓	✓	✓	✓	–
Ours	50M	151.7K	✓	✓	✓	✓	✓

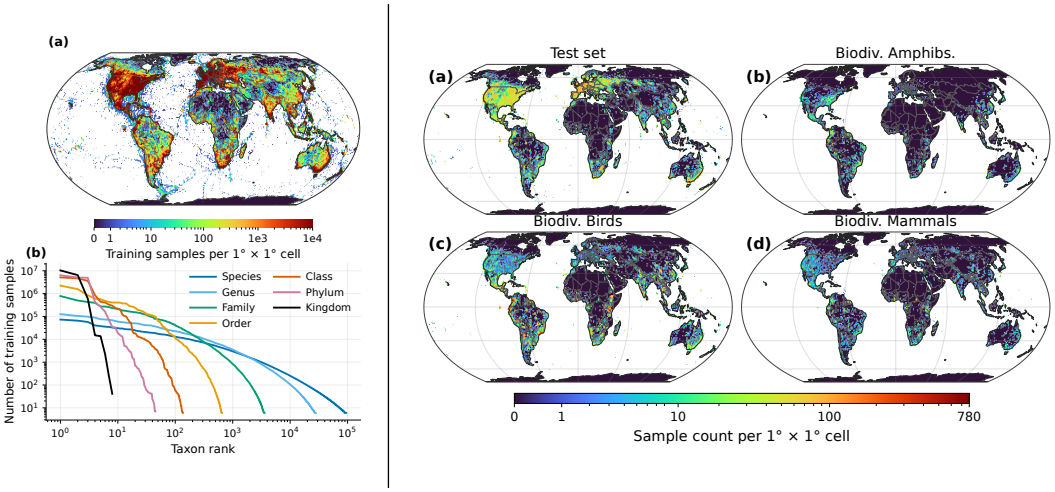


Figure 3: Distributions of training and evaluation data. **Left:** **(a)** Geographic and **(b)** taxonomic (showing number of samples at each level of the taxonomic tree) distribution of the training data. **Right:** Geographic distribution of samples in the evaluation partitions: **(a)** Test set, **(b)** Biodiv. Amphibs, **(c)** Biodiv. Birds, and **(d)** Biodiv. Mammals. Geographic distributions show the number of samples per world grid cell.

Van Horn et al. (2021); Vendrow et al. (2024)) may also contain overlap with pretraining data. We do not attempt to quantify this for external benchmarks here. Instead, we construct our own dedicated evaluation sets from the non-overlapping portion of our corpus.

4.2 Evaluation Sets

We construct validation and test partitions reserving approximately 1% of eligible observations each for random held-out validation and test sets. However, a random sample continues to reflect the distribution of GBIF observation effort providing limited coverage of many biodiversity-rich regions. We construct additional biodiversity-weighted evaluation sets for birds, mammals, and amphibians using global species-richness maps from `BiodiversityMapping.org` (Jenkins et al., 2013; Pimm et al., 2014), derived from BirdLife International range maps for birds (BirdLife International and Handbook of the Birds of the World, 2018) and IUCN range maps for mammals and amphibians (IUCN, 2018). For each taxonomic group, held-out occurrences are sampled preferentially from regions of high mapped species richness, producing complementary evaluation distributions that are less dominated by the geographic concentration of observation effort. The complete sampling procedure and benchmark statistics are provided in Appendix A.3. Figure 3 illustrates some distributions of the training and evaluation partitions.

5 EXPERIMENTS

We evaluate four questions: (1) does ecological context alone capture taxonomic information? (2) does combining ecological and visual information improve recognition? (3) where does ecological context provide the greatest benefit? and (4) does ecological structure emerge in the learned representation? Retrieval ranks L_2-normalised candidate taxonomic text embeddings by cosine similarity to the query. R@1 and R@5 denote the proportion of observations for which the correct taxon appears first or within the top five candidates, respectively.

5.1 DOES ECOLOGICAL CONTEXT ALONE CAPTURE TAXONOMIC INFORMATION?

We first test whether ecological context alone contains sufficient taxonomic signal to support retrieval in the BioCLIP 2 semantic space. After Stage 1 training, we retrieve among all 31,463 taxonomic names in the test set using cosine similarity between ecological and taxonomic text embeddings (Table 2). We ablate the information available to the ecological encoder, progressing from location alone to location with day-of-year and time-of-day. We additionally compare our multi-scale RFF spatial encoding with a capacity-matched MLP receiving the same location and cyclic temporal inputs; full architectural details are provided in Appendix A.4.

Location alone provides substantial taxonomic information, achieving 3.875% R@5 compared with 0.0156% under random retrieval. Adding day-of-year improves both R@1 and R@5, while time-of-day further improves R@5 to 4.361% with a small decrease in R@1. The full RFF encoder also substantially outperforms the capacity-matched MLP (4.361% versus 3.080% R@5), supporting the use of multi-scale spatial features for representing ecological context.

Table 2: Ablation of encoder architecture and inputs showing R@1 and R@5 on our test set.

Encoder	Location	Day of year	Time of day	R@1 (%)	R@5 (%)
Random chance	–	–	–	0.00318	0.0159
RFF	✓	–	–	1.388	3.875
RFF	✓	✓	–	1.480	4.170
RFF	✓	✓	✓	1.459	4.361
MLP	✓	✓	✓	0.8858	3.080

5.2 DOES COMBINING ECOLOGICAL AND VISUAL INFORMATION IMPROVE RECOGNITION?

We next evaluate whether ecological context improves recognition when combined with visual evidence. We report R@1 and R@5 under two candidate vocabularies: a restricted vocabulary (RV), containing only taxa represented in each evaluation set, and a full vocabulary (FV), containing all taxa represented during training. We compare against BioCLIP, BioCLIP 2, BioTrove-CLIP, BioCAP, and TaxaBind. For BioCLIP, BioCAP and BioCLIP 2, we use the standard OpenAI CLIP prompt-template ensemble, which outperformed encoding taxonomic names directly; prompt ensembling did not improve our models, for which we use taxonomic names directly. We additionally train an image–text-only version of our framework (*Ours: Image–Text*) without ecological metadata. We omit BioTrove-CLIP results for amphibians and mammals because these groups were not represented in its training subset.

To separate ecological alignment from adaptive fusion, we evaluate two multimodal variants. *Ours: Ungated* jointly aligns image, context, and text representations but includes no learned fusion module or fused contrastive objective. Because image and context embeddings share the same semantic space, we combine them using a single global fusion weight selected on the validation set by evaluating weights at 0.01 intervals and choosing the value with highest validation accuracy. We apply the same selection procedure to TaxaBind. *Ours: Gated* instead learns an observation-specific gate together with the fused contrastive objective. For each model, we report image, context/location, and fused retrieval where applicable. Table 3 reports performance across all benchmarks under both RV and FV vocabularies.

Across both candidate vocabularies, incorporating ecological context substantially improves taxonomic retrieval (Table 3). Relative to the strongest image-only result on each benchmark, fused

Table 3: R@1, R@5 accuracy (%) across evaluation benchmarks. **RV** restricts candidate taxa to those represented in the corresponding evaluation set, while **FV** considers all taxa represented during training. Retrieval denotes the representation used as the query. Best results are shown in **bold**.

Model	Retrieval	Gate	Test (RV)		Amphib. (RV)		Birds (RV)		Mammals (RV)		Test (FV)		Amphib. (FV)		Birds (FV)		Mammals (FV)	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
BioTrove-CLIP	Image	–	27.87	42.27	–	–	19.24	33.13	–	–	23.19	35.79	–	–	18.76	32.41	–	–
BioCLIP	Image	–	38.48	56.45	5.34	12.10	14.95	27.82	11.26	24.35	36.29	53.68	5.01	11.44	14.88	27.64	10.62	22.64
BioCAP	Image	–	39.73	67.17	9.36	24.42	24.13	47.71	18.56	41.23	29.06	54.63	8.73	22.85	23.97	47.33	17.64	39.06
TaxaBind	Image	–	43.29	64.62	8.43	20.70	21.72	41.01	15.98	35.47	38.75	59.27	7.98	19.73	21.62	40.73	15.20	33.57
BioCLIP 2	Image	–	75.79	91.51	**28.15**	**57.35**	**50.73**	**76.90**	**37.24**	**67.02**	66.31	85.13	**27.42**	**55.99**	**50.49**	**76.58**	**36.38**	**65.44**
Ours: Image–Text	Image	–	**77.14**	**91.65**	25.86	53.00	48.13	75.11	35.00	64.95	**71.75**	**86.70**	25.13	51.63	47.90	74.58	34.24	63.55
Ours: Ungated	Image	–	76.21	91.33	23.50	50.55	45.93	73.96	34.52	64.41	71.03	86.24	22.70	48.93	45.68	73.40	33.80	63.04
Ours: Gated	Image	–	76.34	91.36	23.48	50.82	46.13	74.15	34.90	64.92	70.95	86.16	22.81	49.35	45.87	73.63	34.12	63.51
TaxaBind	Location	–	0.19	0.67	1.21	4.10	0.45	1.60	1.89	5.78	0.083	0.27	0.14	0.57	0.19	0.67	0.26	0.68
Ours: Ungated	Context	–	**3.06**	**9.34**	**10.60**	**30.24**	**5.11**	**16.24**	**14.77**	**36.90**	**1.95**	**5.89**	2.75	**9.11**	**2.65**	**8.32**	**4.38**	**12.10**
Ours: Gated	Context	–	2.96	8.96	10.18	29.29	4.94	15.36	13.77	35.32	1.93	5.69	**2.76**	8.85	2.50	7.93	4.31	11.12
TaxaBind	Fused	Selected (0.85)	45.71	67.70	10.41	24.86	25.22	46.13	19.00	40.51	40.51	61.87	9.88	23.63	25.09	45.81	18.08	38.07
Ours: Ungated	Fused	Selected (0.75)	**83.68**	**94.81**	45.82	77.59	65.31	**87.61**	56.28	**83.32**	**78.49**	**91.42**	44.67	75.69	64.85	**86.86**	**55.15**	**81.59**
Ours: Gated	Fused	Learned	81.64	93.21	**48.23**	**78.79**	**65.64**	87.09	**56.53**	83.22	75.72	89.29	**46.78**	**76.63**	**64.90**	85.86	54.75	80.44

retrieval improves R@1 by 6.54 percentage points on the random test set under RV, compared with 20.08, 14.91, and 19.29 points on the amphibian, bird, and mammal biodiversity-weighted benchmarks, respectively. Similar gains are observed under FV. The substantially larger improvements on the biodiversity-weighted sets indicate that ecological context is particularly informative under these distributions.

These gains cannot alone be explained by improved visual representations as the image representations of our multimodal models perform similarly to, and on the biodiversity-weighted benchmarks below, BioCLIP 2, while context alone remains substantially weaker than image retrieval. Their combination nevertheless yields large improvements, demonstrating that the two modalities provide complementary information. Context is also substantially more informative on the biodiversity-weighted benchmarks than on the random test set, reaching R@1 values of 10.60% for amphibians and 14.77% for mammals, compared with 3.06% on the random test set - this pattern is also true of the TaxaBind location encoder.

The validation-selected global fusion weight and the learned sample-specific gate perform strongly; neither uniformly dominates across benchmarks. Global fusion performs best on the random test set, while learned gating performs best on amphibians and remains competitive on birds and mammals. These results indicate that no single fusion strategy is optimal across evaluation distributions; this supports the use of adaptive mechanisms rather than relying exclusively on a fixed weighting.

5.3 Where does Ecological Context Provide the Greatest Benefit?

Species diversity and observation effort vary substantially with latitude, motivating our evaluation of whether model performance is uniform geographically. We group observations into 2° latitude bins and compute R@1 for BioCLIP, BioCLIP 2, *Ours: Image–Text*, the image representation from *Ours: Gated*, and the fused representation from *Ours: Gated*. We fit smooth spline curves to the resulting binned accuracies. Figure 1(a) shows model performance as a function of latitude; a plot without these splines is also available (Appendix A.6).

The benefit of ecological context is largest at lower and southern latitudes, where the fused model exhibits the greatest improvement over image-only recognition. These regions are also more sparsely represented in our occurrence data, suggesting that ecological context may be particularly valuable where visual training coverage is limited. One possible explanation is that densely sampled regions, particularly at northern latitudes, provide stronger visual supervision during training, reducing the requirement for additional information which can be provided by ecological context.

5.4 Does Ecological Structure Emerge in the Learned Representation?

We qualitatively examine the ecological representation by computing cosine similarity between context embeddings sampled across space and time and selected text embeddings (Figure 4). For the owl genus *Strix*, high-similarity regions broadly resemble its geographic distribution; additional examples and reference range maps are provided in Appendix A.7. Structure also emerges for concepts

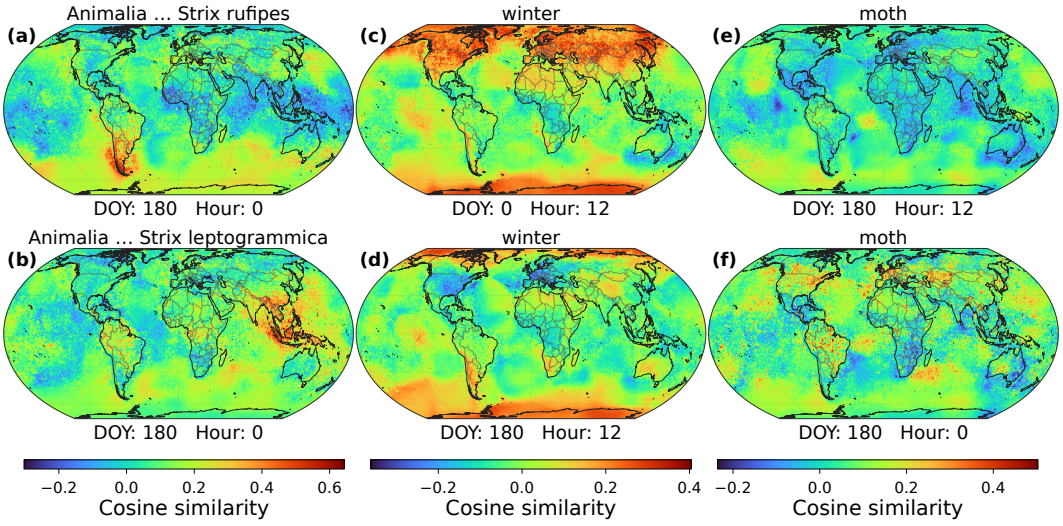


Figure 4: **Cosine similarity between context embeddings and selected text embeddings across space and time. (a–b)** Full taxonomic names of **(a)** *Strix rufipes* and **(b)** *Strix leptogrammica*; **(c–d)** "winter" in **(c)** January and **(d)** June; and **(e–f)** "moth" at **(e)** midday and **(f)** midnight.

not used as explicit ecological supervision. Similarity to "winter" shifts from the Northern Hemisphere near the beginning of the year to the Southern Hemisphere approximately six months later. Hourly temporal structure is also visible: similarity to "moth" is higher at night than during the day. These examples show how purely taxonomic supervision capture broader spatial, seasonal, and diel structure within the shared semantic space.

6 Discussion and Conclusion

We show that readily available ecological metadata can be integrated directly into the semantic space of a biological foundation model and used to improve species recognition at scale. Using more than 50 million image–taxonomy–context observations, we introduce a lightweight ecological encoder for geographic, seasonal, and diel information together with an observation-specific fusion mechanism for combining ecological and visual evidence. Across held-out benchmarks, ecological context provides substantial complementary signal, with the largest gains on biodiversity-weighted evaluation sets. These results show that improving biological recognition need not rely only on larger visual datasets or models: information already attached to biodiversity observations can provide an additional and highly informative modality.

Biodiversity observations are distributed extremely unevenly across the globe, and increasing model or dataset scale alone does not guarantee that biodiversity-rich but poorly sampled regions will be equally well represented. Our results suggest that incorporating ecological structure can improve recognition under evaluation distributions where visual data are more limited. More generally, as biological foundation models grow, it will be important to develop methods that exploit complementary sources of information rather than relying exclusively on vision, particularly for ecological monitoring and conservation applications where geographic data gaps are substantial.

Several limitations remain. We consider only coordinates, day of year, and time of day; climate, elevation, habitat, remote sensing, and other environmental variables could provide additional signal. Our experiments also evaluate only taxa represented during training and therefore do not establish generalisation to entirely unseen taxa. Moreover, contextual supervision did not improve the visual representation itself in our experiments, with gains instead arising from complementary ecological information at inference time. Future work could address these limitations by modelling uncertainty and missing metadata, incorporating richer environmental context, and evaluating transfer across taxa and geographic regions. Our model, large-scale multimodal corpus and biodiversity-weighted evaluation framework provide a foundation for studying these questions further.

7 Ethical Considerations

Our approach uses geographic coordinates and observation timestamps associated with biodiversity records. These metadata could be sensitive, particularly for threatened, commercially valuable, or otherwise vulnerable species, where precise location information may increase exploitation risks. Although this study uses occurrence records made available through existing public data, deployment of models using ecological context must respect applicable access restrictions and consider safeguards such as withholding sensitive records or obfuscating coordinates where appropriate.

AI use statement

Generative AI tools were used to assist with writing and editing the manuscript, retrieval and discovery of related literature, and research execution, including writing, debugging, and refining software code. Literature and citations identified with AI assistance were checked against the original sources before inclusion. All AI-assisted text was reviewed and edited by the authors, and AI-assisted code was inspected and tested for correctness before use. The authors made the final decisions regarding the methodology, experiments, analysis, and interpretation of results, verified the reported results and scientific claims, and take responsibility for the final content of the paper and associated artifacts.

Reproducibility statement

We provide detailed information needed to reproduce our experiments throughout the paper and appendix. The method section specifies the ecological encoder, multimodal alignment and fusion objectives, while the experimental section reports the training procedure, optimisation settings, model configurations, evaluation protocols, and random seeds. The appendix provides additional implementation details and a detailed description of dataset construction, including source datasets, filtering, geographic sampling, train/validation/test splitting, and construction of the biodiversity-weighted evaluation sets. We will release the training and evaluation code, dataset-processing pipelines, model checkpoints, and associated data or download metadata required to reproduce the experiments, in the near future.

References

Jonathan K Alderfer and Jon Lloyd Dunn. *National Geographic Birding Essentials: All the Tools, Techniques, and Tips You Need to Begin and Become a Better Birder*. National Geographic Books, 2007.

Jan Beck, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19: 10–15, 2014. doi: 10.1016/j.ecoinf.2013.11.002.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014. doi: 10.1109/CVPR.2014.259.

BirdLife International and Handbook of the Birds of the World. Bird species distribution maps of the world. `https://datazone.birdlife.org/species/requestdis`, 2018. Version 7.0.

Elizabeth H. Boakes, Philip J. K. McGowan, Richard A. Fuller, Ding Chang-qing, Natalie E. Clark, Kim O'Connor, and Georgina M. Mace. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, 8(6):e1000385, 2010. doi: 10.1371/journal.pbio.1000385.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6320–6342. PMLR, 2023.

Rangel Daroya, Elijah Cole, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Wildsat: Learning satellite image representations from wildlife observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6143–6154, 2025.

Ross J. Gardiner, Guillaume Mougeot, Sareh Rowlands, Benno I. Simmons, Flemming Helsing, and Toke Thomas Høye. Bridging domain gaps for fine-grained moth classification through expert-informed adaptation and foundation model priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 5169–5174, October 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Global Biodiversity Information Facility. Global data trends. `https://www.gbif.org/analytics/global`, 2026. Accessed 25 September 2026.

ZeMing Gong, Austin T. Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X. Chang. Clibd: Bridging vision and genomics for biodiversity monitoring at scale. In *International Conference on Learning Representations*, 2025.

Jianyang Gu, Samuel Stevens, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila M. Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. In *Advances in Neural Information Processing Systems*, 2025.

Mehrdad Hajibabaei, Gregory A. C. Singer, Paul D. N. Hebert, and David A. Hickey. Dna barcoding: How it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, 23(4):167–172, 2007. doi: 10.1016/j.tig.2007.02.001.

IPBES. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services, 2019.

IUCN. The IUCN red list of threatened species. `https://www.iucnredlist.org`, 2018. Version 2018-1.

Clinton N. Jenkins, Stuart L. Pimm, and Lucas N. Joppa. Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences*, 110(28):E2602–E2610, 2013. doi: 10.1073/pnas.1302251110.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):4347–4355, 2025. doi: 10.1609/aaai.v39i4.32457.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019.

Carsten Meyer, Walter Jetz, Robert P. Guralnick, Susanne A. Fritz, and Holger Kreft. Range geometry and socio-economics dominate species-level biases in occurrence information. *Global Ecology and Biogeography*, 25(10):1181–1193, 2016. doi: 10.1111/geb.12483.

Hoang-Quan Nguyen, Thanh-Dat Truong, Xuan Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. Insect-foundation: A foundation model and large-scale 1m dataset for visual insect understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21945–21955, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Vardaan Pahuja, Samuel Stevens, Alyson East, Sydne Record, and Yu Su. Automatic image-level morphological trait annotation for organismal images. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oFRbiaib5Q.

Stuart L. Pimm, Clinton N. Jenkins, Robin Abell, Thomas M. Brooks, John L. Gittleman, Lucas N. Joppa, Peter H. Raven, Callum M. Roberts, and Joseph O. Sexton. The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344(6187):1246752, 2014. doi: 10.1126/science.1246752.

Laura J. Pollock, Justin Kitzes, Sara Beery, Kaitlyn M. Gaynor, Marta A. Jarzyna, Oisin Mac Aodha, Bernd Meyer, David Rolnick, Graham W. Taylor, Devis Tuia, and Tanya Berger-Wolf. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, 1(3):166–182, 2025. doi: 10.1038/s44358-025-00022-3.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Klaus Rohde. Latitudinal gradients in species diversity: The search for the primary cause. *Oikos*, 65(3):514–527, 1992. doi: 10.2307/3545569.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1765–1774, 2025.

Bojan Šavrič, Tom Patterson, and Bernhard Jenny. The equal earth map projection. *International Journal of Geographical Information Science*, 33(3):454–465, 2019. doi: 10.1080/13658816.2018.1504949.

Elena Sierra, Lauren E. Gillespie, Salim Soltani, Moises Exposito-Alonso, and Teja Kattenborn. Divshift: Exploring domain-specific distribution shifts in large-scale, volunteer-collected biodiversity datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28386–28396, 2025. doi: 10.1609/aaai.v39i27.35060.

Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M. Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

J. Christopher D. Terry, Helen E. Roy, and Tom A. August. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, 11(2):303–315, 2020. doi: 10.1111/2041-210X.13335.

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

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12884–12893, 2021.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E. Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. INQUIRE: A natural world text-to-image retrieval benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4018.

Jiří Vyskočil and Lukáš Picek. Towards zero-shot camera trap image categorization. In *Computer Vision – ECCV 2024 Workshops*, volume 15624 of *Lecture Notes in Computer Science*, pp. 37–53. Springer, 2025. doi: 10.1007/978-3-031-92387-6_3.

Sunghyeon Woo, Sol Namkung, Sunwoo Lee, Inho Jeong, Beomseok Kim, and Dongsuk Jeon. PaCA: Partial connection adaptation for efficient fine-tuning. In *International Conference on Learning Representations*, 2025.

Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K. Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K. Singh, Arti Singh, Soumik Sarkar, Nirav Merchant, Chinmay Hegde, and Baskar Ganapathysubramanian. Biotrove: A large curated image dataset enabling ai for biodiversity. In *Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL `https://openreview.net/forum?id=DFDCtGQs7S`.

Ziheng Zhang, Xinyue Ma, Arpita Chowdhury, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Samuel Stevens, Hilmar Lapp, Tanya Berger-Wolf, Yu Su, Wei-Lun Chao, and Jianyang Gu. BioCAP: Exploiting synthetic captions beyond labels in biological foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Appendix

A.1 Construction of the biodiversity richness map

To provide a reference for the geographic distribution of biodiversity, we constructed the species-richness map shown in Fig. 1**(c)** from the global biodiversity layers released by `BiodiversityMapping.org`(Jenkins et al., 2013; Pimm et al., 2014; BirdLife International and Handbook of the Birds of the World, 2018; IUCN, 2018). We used the 10 km $\times$ 10 km terrestrial species-richness rasters for birds, mammals, and amphibians. These maps are derived from digital species-range maps from BirdLife International for birds and the IUCN for mammals and amphibians, and represent native, extant species. Species richness in the source data is range-map based rather than observation based. A species is counted as present in a grid cell if any part of its mapped geographic range overlaps that cell.

The three taxonomic rasters have the same 10 km spatial resolution and projection. We therefore constructed a common grid spanning the union of the three rasters and aligned each to this grid without changing its native resolution. For each grid cell, combined richness was computed as a sum of the corresponding cells for the three rasters. If one taxonomic group contained no mapped value for a cell while at least one of the other groups contained valid data, the missing contribution was treated as zero. Cells without valid data in all three source rasters are assigned zero-values.

The resulting map therefore represents the summed mapped richness of birds, mammals, and amphibians and is intended as a broad reference for the global concentration of terrestrial vertebrate biodiversity. Since it is constructed from expert-derived geographic range polygons, it does not represent species abundance, sampling intensity, detectability, or guaranteed local occurrence within

each range, modelling all of which remain challenging problems in biodiversity monitoring. In addition to visualisation, we use these richness maps when constructing biodiversity-weighted test sets for amphibians, birds, and mammals. This is described in Appendix A.3.

A.2 Loss Function

Following TaxaBind **?**, we employ a symmetric multi-positive contrastive objective. This is important because a minibatch may contain multiple independent observations of the same taxa; under a standard paired InfoNCE objective, these would otherwise be incorrectly treated as negatives.

Let $z_i^{(a)}$ and $z_j^{(b)}$ denote embeddings from modalities a and b. After L_2-normalisation, we define their temperature-scaled similarity as

$$s_{ij}^{(a,b)} = \frac{{z_i^{(a)}}^\top z_j^{(b)}}{\tau}, \tag{7}$$

where τ is the temperature parameter. For anchor i, the positive set is therefore

$$P(i) = j : y_j = y_i \tag{8}$$

where y_i denotes its taxonomic label. The directional loss from modality a to modality b is

$$\mathcal{L}_{a\to b} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j\in P(i)} \exp\left(s_{ij}^{(a,b)}\right)}{\sum_{n=1}^{B} \exp\left(s_{in}^{(a,b)}\right)}, \tag{9}$$

where B is the size of the minibatch. The numerator aggregates all observations from the same taxa as valid positives rather than selecting only the paired instance.

We apply the loss symmetrically,

$$\mathcal{L}_{a,b} = \frac{1}{2}\left(\mathcal{L}_{a\to b} + \mathcal{L}_{b\to a}\right). \tag{10}$$

During Stage 1, the objective is applied between ecological and text representations, $\mathcal{L}(e, q)$. During Stage 2, we optimise $\mathcal{L}(e, q)$, $\mathcal{L}(v, q)$, and $\mathcal{L}(f, q)$, corresponding to ecological–text, visual–text, and fused–text alignment, respectively.

A.3 Dataset

To construct our dataset, we began with 156,771,264 GBIF occurrence records. A GBIF occurrence corresponds to a recorded observation of an organism at a particular place and time and may be associated with multiple images from the same observation event. We therefore perform dataset partitioning at the occurrence level, ensuring that correlated images from the same occurrence cannot be split across training and evaluation sets.

The dataset was downloaded on 30 July 2025 using the following filters: basis of record restricted to "Machine observation" and "Human observation"; datasets restricted to iNaturalist Research-grade Observations, Observation.org, Pl@ntNet observations, NABU—naturgucker, the Norwegian Species Observation Service, Taiwan Moth Occurrence Data Collected From Social Network, Species recordings from the Danish National Portal Arter.dk, the Danish Mycological Society fungal records database, and Earth Guardians Weekly Feed; `hasCoordinate = Yes`; `hasGeospatialIssue = No`; `isInCluster = No`; media type restricted to ***StillImage***; and occurrence status restricted to ***Present***. We exclude contributions from xeno-canto because their still-image records contain spectrograms rather than photographs of organisms. A DOI for the base dataset is available: REDACTED FOR REVIEW.

We then develop and apply a series of filtering and downsampling steps to reduce sampling imbalance and obtain a tractable training corpus. First, occurrence records are randomly shuffled and

records without any taxonomic annotation are removed. Incomplete taxonomic hierarchies are permitted, but each retained occurrence must contain at least one taxonomic label.

We next partition occurrences into $1^\circ \times 1^\circ$ latitude–longitude cells and randomly subsample cells containing more than 10,000 occurrences, retaining at most 10,000 per cell. This reduces the dominance of heavily sampled geographic regions and limits the extent to which the ecological encoder can exploit highly concentrated observation patterns. This step reduces the dataset substantially to 30,387,391 occurrences.

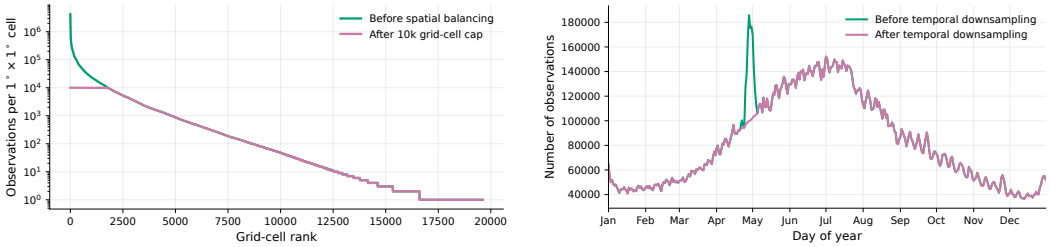


Figure 5: Distributional effects of spatial and temporal downsampling. **Left:** Rank–abundance distribution of observations across $1^\circ \times 1^\circ$ geographic grid cells before and after spatial balancing. **Right:** Distribution of observations across day of year before and after temporal downsampling.

We additionally observe a pronounced spike in the day-of-year distribution during the final week of April, coinciding with concentrated community-science recording activity such as the City Nature Challenge (Sierra et al., 2025). Because day of year is provided directly to the ecological encoder, such an event-specific sampling spike could provide an unintended shortcut. We therefore estimate a smooth baseline from the surrounding day-of-year distribution and stochastically downsample observations within the affected period toward this expected baseline. The distributions before and after correction are shown in Figure 5. This reduces the corpus to 29,850,881 occurrences.

We subsequently remove taxonomic labels represented by six or fewer retained occurrences, ensuring a minimum level of support after geographic downsampling. Applying this criterion after spatial balancing prevents taxa that were sufficiently represented in the original corpus but largely removed during downsampling from entering the final dataset with very few observations. The distributions before and after this step are shown by Figure 5. This reduces the corpus to 28,960,850 occurrences.

Finally, we cross-reference the GBIF occurrence identifiers of all remaining records against TreeOfLife-200M (ToL-200M), which forms part of the BioCLIP 2 pretraining data. We identify 17,803,090 occurrences that overlap with ToL-200M and 10,934,605 that do not. Overlapping occurrences are retained for training, where they provide valid ecological supervision, but are excluded from all validation and test partitions to prevent evaluation on occurrences previously encountered by the BioCLIP 2 visual backbone. Held-out partitions are constructed exclusively from the non-overlapping subset.

Validation set. From the non-ToL-200M portion of the corpus, we randomly sample a held-out validation partition before occurrence records are associated with their individual images. Because each row at this stage corresponds to a unique GBIF occurrence, all images associated with a given occurrence subsequently remain within the same partition. The validation set is used for model selection.

Random test set. We independently sample a random test partition from the remaining non-ToL-200M occurrences using the same occurrence-level splitting procedure. This benchmark measures recognition under the empirical distribution of the curated GBIF corpus while ensuring that no occurrence is shared with the training or validation partitions.

Occurrence-per-taxon test set. During model development, we additionally construct a taxon-balanced diagnostic set by selecting one GBIF occurrence for each represented taxon from the remaining data. All images subsequently associated with the selected occurrence remain grouped together. Although balanced at the occurrence-selection stage, variation in the number of images as-

sociated with individual occurrences means that the resulting image-level benchmark is not strictly balanced across taxa. We therefore use this set only as a development diagnostic and do not include it in the final reported evaluation.

Biodiversity-weighted sets. To construct biodiversity-weighted test sets, we use freely available spatial species-richness maps from `biodiversitymapping.org` for birds, mammals, and amphibians. These maps combine species-range information to estimate the number of species occurring within each geographic region. For each eligible held-out occurrence, we extract the richness value at its geographic location and use this value as a sampling weight. We then sample up to ten occurrences per taxon without replacement, such that occurrences from regions of greater mapped biodiversity are more likely to be represented while retaining taxonomic coverage. This produces evaluation distributions that preferentially sample high-richness regions for which suitable held-out observations are available. The resulting sets contain 132,556 bird, 32,151 mammal, and 31,761 amphibian datapoints, respectively.

All remaining occurrences not assigned to a held-out partition are assigned to training. We subsequently retrieve the media metadata associated with each occurrence to obtain image download links and construct WebDataset shards containing image–text–metadata triplets. Because a single GBIF occurrence may contain multiple images, the number of resulting training triplets exceeds the number of unique occurrences. Final partition statistics are reported in Table 4.

Table 4: Training and evaluation partitions. We report the number of unique GBIF occurrences and resulting image–text–metadata triplets for each partition.

Partition	# Occurrences	# Triplets	Purpose
Train: ToL-200M overlap	17,803,090	30,579,182	Training
Train: non-ToL-200M	10,934,605	18,677,996	Training
Total train	28,737,695	49,257,158	Training
Validation	111,577	192,471	Model selection
Test	111,577	193,380	Standard evaluation
Occurrence-per-taxon test	145,580	308,872	Development diagnostic
Bird biodiversity-weighted	75,449	132,556	Biodiversity-weighted evaluation
Mammal biodiversity-weighted	16,264	32,151	Biodiversity-weighted evaluation
Amphibian biodiversity-weighted	16,572	31,761	Biodiversity-weighted evaluation
Total	29,214,714	**50,148,349**	

A.4 MLP CONFIGURATION IN ABLATION STUDY

The MLP uses three parallel branches (each with hidden dimensions 256, 512, 1024, 1024, and 1024), followed by projection to 768-dimensional embedding space; the three branch outputs are summed. These dimensions were chosen to give the MLP an approximately similar architecture and parameter count as the RFF variant, to allow a fair comparison.

A.5 IMPLEMENTATION DETAILS

All models were trained on a single 40GB NVIDIA A100 GPU using the PyTorch library for Python. Training utilised a minibatch size of 1536 and mixed-precision computation with `bfloat16`. We optimised trainable parameters using AdamW with an initial learning rate of 5×10^{-6}, weight decay of 0.01, and gradient clipping at 1.0. The learning rate followed a cosine schedule with $T_{\max} = 50$ and a minimum learning rate of 1×10^{-7}. All models were trained for at most 80 epochs in stage 1 and 30 epochs for stage 2, checkpoint selection is described below. The contrastive temperature was fixed at $\tau = 0.07$ (Chen et al., 2020). Training and dataset randomisation were controlled with fixed independent seeds (42 and 43, respectively).

During multimodal (stage 2) training, we adapt the text encoder using Partial Connection Adaptation (PaCA) (Woo et al., 2025) with rank $r = 16$ and scaling parameter $\alpha = 16$, applied to the output projection of every attention layer. PaCA substantially reduces the number of trainable text-encoder parameters and associated VRAM requirements, allowing us to retain large contrastive minibatches

and perform multimodal adaptation at the scale of our large dataset on a single GPU. To additionally save compute time and memory requirement, we pre-compute all image encodings once before training. These are saved to disk and used throughout, as the image encoder weights are not made trainable. These aspects are critical to achieving our objective (within our compute budget): to investigate ecological context across millions of observations and many taxa. One epoch of our training set contains 32,038 minibatches, time to process 100 minibatches is reported at around 55 seconds for stage 2 and much shorter for stage 1. Time to compute one training epoch was reported at around five hours for stage 2.

For stage 1, checkpoint selection was decided based on the validation R@5 context to text score using early stopping and a patience of 5. For stage 2, checkpoint selection is based on the harmonic mean of validation image to R@1 and ecological context to text R@5. Because these retrieval metrics differ substantially in scale, each is normalised by its maximum validation value over the training run before computing the harmonic mean. This favours checkpoints that maintain visual retrieval while giving the ecological representation a chance to learn. For the image-only model, the selection of checkpoints is based directly on the validation image R@1.

A.6 Latitude-wise Accuracy Curves

To visualise latitudinal variation in recognition performance, we group predictions into 2° latitude bins and compute R@1 within each bin as the mean proportion of correctly retrieved datapoints. Bins containing fewer than 15 datapoints are excluded. We then fit an unweighted cubic spline to the resulting per-bin accuracies using seven knots. The spline is evaluated only over the latitude range containing available data, avoiding extrapolation beyond the sampled range. Uncertainty bands are estimated using 500 bootstrap resamples of the latitude bins, refitting the spline for each resample and reporting the 2.5th and 97.5th percentiles of the resulting predictions as a 95% bootstrap interval. For transparency, we additionally provide a version of Figure 1 in which the underlying per-bin accuracies are shown directly alongside the fitted curves (Figure 6).

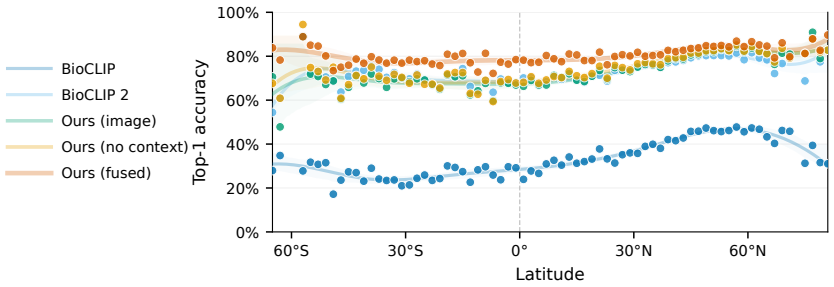


Figure 6: Top-1 taxonomic retrieval accuracy as a function of latitude for BioCLIP, BioCLIP 2, and variants of our model. Points show empirical accuracy within 2° latitude bins, while transparent curves show unweighted cubic-spline fits to the retained bins.

A.7 Further Properties

We note that the ecological encoder can learn distinct geographic representations for closely related species within the same taxonomic genus. In Figure 7, we visualise context–text similarity maps for six owl species within the genus *Strix*. Despite their close taxonomic relationship, each species exhibits a distinct spatial pattern, with high-similarity regions occurring in different parts of its estimated geographic range as shown when compared with Figure 8. This suggests that the learned ecological representation captures specific geographic structure rather than collapsing closely related taxa into a common genus-level distribution.

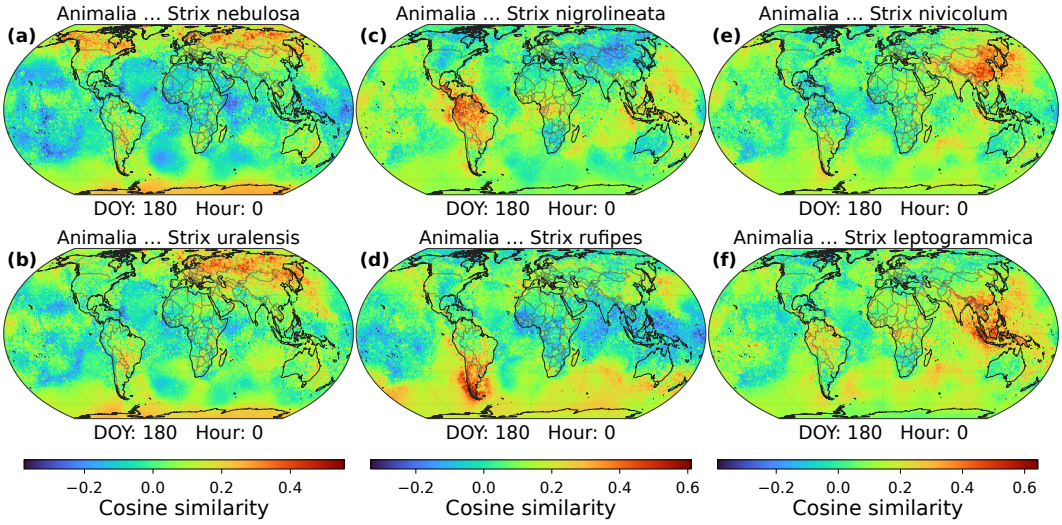


Figure 7: Cosine similarity for each $1^\circ \times 1^\circ$ latitude–longitude cell at midnight on day 180 of the year and the full taxonomic name of six species in the owl genus *Strix*: **(a)** *Strix nebulosa*, **(b)** *Strix uralensis*, **(c)** *Strix nigrolineata*, **(d)** *Strix rufipes*, **(e)** *Strix nivicolum*, and **(f)** *Strix leptogrammica*.

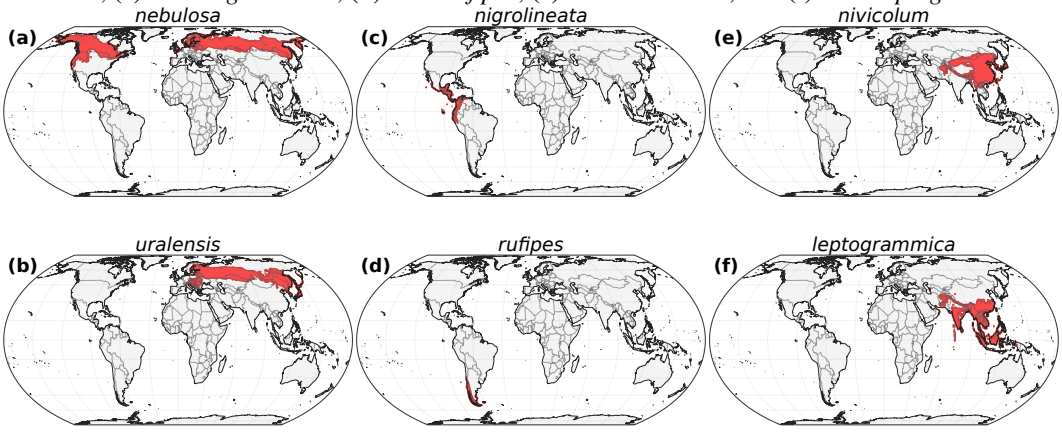


Figure 8: Estimated geographic ranges of six *Strix* species derived from iNaturalist range data: **(a)** *Strix nebulosa*, **(b)** *Strix uralensis*, **(c)** *Strix nigrolineata*, **(d)** *Strix rufipes*, **(e)** *Strix nivicolum*, and **(f)** *Strix leptogrammica*.