# PHENOMAP: PHENOLOGICAL NEURAL FIELDS FOR ECOLOGICAL MAPPING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Foundation models in scientific domains have predominantly organized representations around a single dominant axis of supervision. In biology, this axis has been species identity, typically captured via image–text contrastive pretraining or multimodal systems that use species as the binding modality. Yet for plants, visual appearance changes dynamically across developmental stages, a process known as *phenology*, making species an incomplete signature. Dedicated phenology models reduce this variability to narrow classification targets, while standard taxonomic foundation models encode species identity with bias to phenophases overrepresented in training distributions. To close this gap, we introduce **PhenoMap**, a multimodal model that leverages time, location, climate, and species to produce factorized species and phenology embeddings aligned with plant imagery. To train and benchmark our model, we curate **Pheno3M**, pairing citizen-science plant imagery and phenophase annotations with per-observation climate and satellite drivers, and an external expert-survey benchmark. We extensively evaluate PhenoMap on new phenological benchmarks built from Pheno3M, where it excels on compositional species $\times$ phenophase and fine-grained phenophase retrieval. We further demonstrate its utility by qualitatively showing phenological processes emerging when the spatiotemporal field is queried as a map, and evaluating those learned embeddings on USA-NPN expert observations. All code and data will be released upon acceptance.

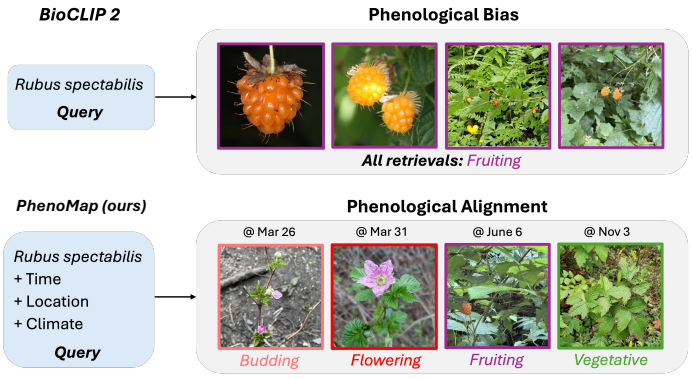



Figure 1: **Overcoming phenological bias in taxonomic foundation models.** Standard taxonomic foundation models (e.g. BioCLIP 2 **?**) exhibit dataset-driven phenological bias *(top)*. Our PhenoMap framework leverages time, location, and climate context to generate phenologically aware embeddings, achieving phenological alignment across a species' entire lifecycle *(bottom)*.

## 1 INTRODUCTION

Web-scale pretraining based on cross-modal matching has been a dominant paradigm in machine learning literature, with early works like CLIP (Radford et al., 2021), inspiring similarly trained foundation models in other domains (Zhang et al., 2023; Liu et al., 2024a; Wu et al., 2023). Ecology has been no exception, with ecological foundation models (Stevens et al., 2024; Yang et al., 2024)



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.

showing strong performance on the longstanding species recognition task, with recent work scaling to hundreds of millions of pretraining images (**?**). However, in the ecological domain, the matching task between visual appearance and species identity is a complex function largely governed by *phenology*: the biologically grounded process by which species change across seasons. Understanding the range of appearances a species takes on is a major component of the formal study of plant ecology (West-Eberhard, 1989; Des Roches et al., 2021). Phenological states and transitions are among the most sensitive biological indicators of climate change (Parmesan & Yohe, 2003; Cleland et al., 2007), modulate ecosystem-scale carbon, water, and energy fluxes (Richardson et al., 2013), and propagate timing shifts across food-web hierarchies (Edwards & Richardson, 2004).

Despite phenology's foundational role in these applications, state-of-the-art models for ecology still fail to encode these dynamics. This shortcoming can be largely attributed to misalignment between dominant pretraining objectives and the structure of phenological variation. Models like BioCLIP (Stevens et al., 2024) optimize a matching loss between species identity and ecological imagery, which encourages the model to discard variance in imagery that cannot be explained by coarse species labels. Standard contrastive losses only preserve the information shared between paired views (Liang et al., 2023). As a result, phenological signals, which the species label does not encode, are among the variance discarded. BioCLIP 2 (**?**) responds by scaling pretraining to hundreds of millions of images to residually encode intra-species structure. While scale erodes the effect, it does not close it. The residual remains entangled with species identity and lacks cross-modal alignment, meaning species text cannot query specific phenophases (Figure 1). Furthermore, because observational datasets are heavily biased toward salient developmental phases like *flowering*, we show in § 5.1 that text embeddings absorb phenological bias, consistent with other findings in representation learning that show it is difficult to factorize these underlying dynamics without inductive biases (Locatello et al., 2019).

In contrast, works that target phenology treat it as an isolated prediction task rather than an axis of a shared representation with species. Image-based phenophase classifiers operate at small scale or with binary supervision (Reeb et al., 2022), climate-time-series models predict phenological dates without imagery (Sainte Fare Garnot et al., 2025; Liu et al., 2024b), and pixel-scale satellite models forecast vegetation indices as a community-level proxy (Benson et al., 2023). None of these approaches produce a unified understanding of the visual and taxonomic traits that define ecological tasks.

In this work, we aim to bridge gaps in ecological representation learning by injecting phenologically-grounded indicators into foundation models. We make three contributions:

- We introduce *PhenoMap*, a multimodal model that produces factorized species and phenology embeddings. The species channel is aligned with species text, and the phenology channel with a species-conditioned spatiotemporal neural field.
- We curate *Pheno3M*, pairing $\sim$3M citizen-science plant observations (Van Horn et al., 2018) and their phenophase annotations with per-observation climate and satellite drivers, and an external benchmark of USA National Phenology Network expert records (Schwartz et al., 2012).
- We show analytically and empirically that species-supervised foundation models absorb the phenological bias of their training data, and that PhenoMap removes it. PhenoMap outperforms species-aligned baselines on cross-modal phenophase retrieval, on compositional species$\times$phenophase queries, and on USA-NPN expert observations. Its embeddings also recover species-specific phenology across space and time.

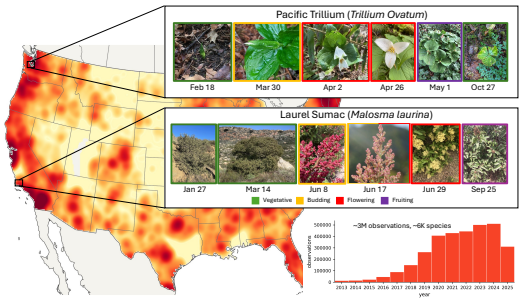
## 2 Related Work

**Foundation models for taxonomic imagery.** Foundation models for taxonomic imagery largely organize around species identity, either through image–text contrastive pretraining over taxonomic labels (Stevens et al., 2024; **?**; Yang et al., 2024) or by treating species imagery as the binding modality across geography, satellite, audio, DNA and text (Sastry et al., 2024; Huynh et al., 2024; Sastry et al., 2025b; Gong et al., 2024; Daroya et al., 2025; Sastry et al., 2026; Zhang et al., 2026; Khanal et al., 2024). Adjacent lines of work have modeled the hierarchical structure of the tree of life (Sastry et al., 2025a; Gong et al., 2025) and works like PlantTraitNet (Sharma et al., 2026) predict plant traits from imagery in a multi-task framework. Recently, BioCLIP 2 (**?**) showed that scaling species-aligned objectives to hundreds of millions of images yields emergent capabilities

within vision encoders, capturing animal life stages and sex in intra-species residual subspaces. This emergent intra-species structure is reported on animal life-stage and sex axes. Whether it transfers to plant phenophase is a distinct empirical question we investigate in § 5. As this emergence is confined to the image space, cross-modal phenological querying is unsupported.

**Spatiotemporal grounding via joint embeddings.** Another line of work grounds visual representations in spatial and temporal context. Early work incorporated spatio-temporal priors to enhance species classification (Mac Aodha et al., 2019). Early coordinate-encoding work introduced multi-scale spatial representations (Mai et al., 2020) while others learned species-conditioned spatial priors from location and satellite imagery (Cole et al., 2023; Hamilton et al., 2024; Sastry et al., 2023; Teng et al., 2023b; Dollinger et al., 2024; Lange et al., 2025). Image-location contrastive pretraining methods yielded general-purpose location encoders (Vivanco Cepeda et al., 2023; Klemmer et al., 2025; Mai et al., 2023; Dhakal et al., 2025), and other work jointly embeds images, time, and location for compositional retrieval (Zhai et al., 2019; Shatwell et al., 2025). Foundation-scale satellite embeddings extend this to dense, globally available spatial features (Brown et al., 2025). These models ground visual representations in time and space as *universal* axes, leaving species-conditioned axes like phenology unsupervised: two plants at the same phenophase but different locations and dates remain far apart in the latent space. PhenoMap follows this joint-embedding design pattern but is, to our knowledge, the first to align plant imagery with phenologically relevant field modalities (climate history, location, time, and satellite context) through an explicitly supervised, species-conditioned phenology channel.

**Phenology modeling and benchmarks.** Phenology has long been studied through process models (Hufkens et al., 2018; Basler, 2016) and standardized growth-stage scales (Meier, 2018), and, recently, through deep models predicting phenological dates from climate time series (Sainte Fare Garnot et al., 2025; Liu et al., 2024b). Pixel-scale satellite work models vegetation indices over time as a community-level phenological proxy (Benson et al., 2023). Phenology benchmarks remain site-specific, species-specific (Reeb et al., 2022), pixel-scale, or binary (Dinnage et al., 2025) in nature. Citizen-science platforms (Van Horn et al., 2018) and expert networks (Schwartz et al., 2012) provide rich raw data, but have not been packaged into a large-scale, individual-image, phenophase-labeled benchmark. Recent ecology benchmarks focus on species identification or distribution rather than phenophase (Goëau et al., 2025; Teng et al., 2023a; Picek et al., 2024a;b). Pheno3M fills this gap with a large-scale, image-grounded phenophase corpus, and PhenoMap is the first model to tie individual plant imagery to phenophase under joint species and spatiotemporal supervision.




| Benchmark | Size | Phenology |
|---|---|---|
| iNat-21 (Plants) | 1.2M | — |
| PlantCLEF 2022 | 4M | — |
| GeoPlant | 5M | — |
| BioTrove (Plants) | 67M | — |
| TreeOfLife-200M (Plants) | 101M | — |
| PlantTraitNet | 307K | — |
| Phenology-2.4K | 2.4K | 4-stage, 3 species |
| PhenoVision | 1.5M | binary flower/fruit |
| Pheno3M (ours) | 3M | 4-stage + 25 concepts |

Figure 2: **Pheno3M.** A large-scale plant phenology dataset of ∼**3M citizen-science observations** spanning ∼6K species, each annotated with a 4-class multi-hot phenophase label (vegetative, budding, flowering, fruiting). **Left:** observation density (spatially and temporally) with per-location phenophase progressions. The spring-flowering *Trillium ovatum* and the summer-flowering *Malosma laurina* illustrate within-species visual transitions. 2025 includes only a partial year of observations. **Right (table):** Comparison to existing plant-image benchmarks.

## 3 PHENO3M

**Source and phenophase labels.** Pheno3M consists of approximately 3M plant observations spanning ∼6K species, curated from research-grade iNaturalist records with phenology annotations. Each observation $(\mathbf{x}, s, \ell, t, \mathbf{y})$ pairs a ground-level photograph $\mathbf{x}$ with a species identifier $s$, a geographic location $\ell$ (lat, lon), an observation date $t$, and a phenophase label $\mathbf{y} \in \{0, 1\}^4$ over four cyclically

ordered phases: *no phenology / vegetative*, *budding*, *flowering*, and *fruiting*. The label is multi-hot with a single observation possibly having co-occurring phenophases. Adjacent phenophases are visually and ecologically more similar than distant ones. Pheno3M is restricted to the continental United States, 2013–2025. Figure 2 summarizes spatial and temporal coverage, and example progressions.

**Phenological concepts.** The four phenophases are too coarse to capture within-phase variation. We therefore define **25** fine-grained phenological concepts, each visually grounded and linked to the four coarse phenophases, drawing on how ecological analyses partition seasonal development, how observer networks stage visible phenology (Cleland et al., 2007; Richardson et al., 2013; Schwartz et al., 2012), and the principal growth stages of the BBCH scale (Meier, 2018). We describe the VLM-based pipeline for synthesizing concept annotations in App. A.2 and validate them in App. A.4.

**Generalization splits.** We construct two observation-level held-out splits: **(i) random**, i.i.d. held-out observations measuring in-distribution performance, and **(ii) spatial**, held-out geographic clusters measuring transfer to unseen regions. Refer to App. A.7 for more details.

**Spatiotemporal signals.** To support the field branch (§4), every observation is augmented with a year-long daily PRISM climate history (Daly et al., 2008) preceding the observation date and an AlphaEarth satellite embedding (Brown et al., 2025) sampled at the observation coordinates.

**USA-NPN (external evaluation).** We curate a visit-level benchmark from the USA National Phenology Network (USA-NPN) Status & Intensity product. Trained observers repeat site visits following standardized protocols (Denny et al., 2014; Schwartz et al., 2012). After canonical filters (Crimmins et al., 2017; Rosemartin et al., 2018), the corpus contains about 3M observation rows across 1,000 plant species (2013–2025, Plantae, North America). We collapse growth-form-dependent phenophases into 13 USA-NPN phenophases and represent each visit as a multi-label vector. Reported cells are binary. Unreported and non-applicable cells are masked (Rosemartin et al., 2018; Sainte Fare Garnot et al., 2025). This benchmark is disjoint from Pheno3M and evaluates transfer to expert surveys rather than opportunistic photos.

## 4 METHODOLOGY

We learn cross-modal representations of phenology by factorizing visual embeddings into *static* and *dynamic* components corresponding to species and phenological supervision, respectively. We summarize the architecture of PhenoMap in Figure 3.

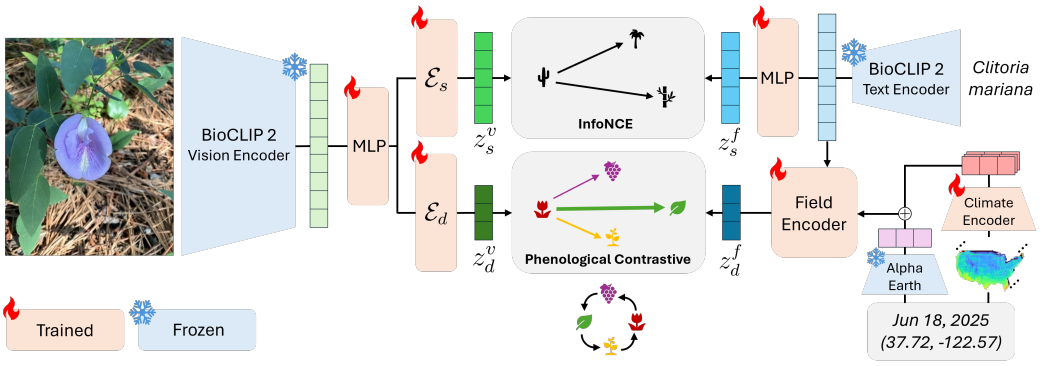



Figure 3: **PhenoMap Architecture.** Frozen BioCLIP 2 encoders process species text and plant imagery. The visual embeddings are factorized into static ($\boldsymbol{z}_s^v$) and dynamic ($\boldsymbol{z}_d^v$) components via MLPs and aligned with corresponding text and environmental field embeddings ($\boldsymbol{z}_s^f$, $\boldsymbol{z}_d^f$) using dual contrastive losses, while a decoder reconstructs the original visual embedding.

### 4.1 ARCHITECTURE

**Image Branch**. A frozen BioCLIP 2 image embedding $\boldsymbol{z}$ is passed through an MLP to a static encoder and a dynamic encoder, giving the factorized embeddings $\boldsymbol{z}_s^v = \mathcal{E}_s(\boldsymbol{z})$ and $\boldsymbol{z}_d^v = \mathcal{E}_d(\boldsymbol{z})$. To prevent representation collapse, a decoder $\mathcal{D}$ reconstructs the original embedding $\boldsymbol{z}$ from both factors.

**Field Branch**. For the static field embedding ($\boldsymbol{z}_s^f$), a frozen BioCLIP 2 species embedding is passed through an MLP. The dynamic field embedding ($\boldsymbol{z}_d^f$) captures environmental context using the observation's geolocation and day-of-year to extract an AlphaEarth embedding and a 365-day PRISM climate window. The PRISM data is processed by a learnable convolutional climate encoder into weekly tokens. These tokens, alongside the AlphaEarth embedding, are fed into a self-attention transformer. Since phenological responses are intrinsically tied to taxonomy, the frozen species embedding conditions the transformer blocks using zero-initialized adaptive layer normalization (Peebles & Xie, 2023). Because it maps any species, location, and date to a phenology embedding, the field branch is a species-conditioned spatiotemporal neural field, which we query on a CONUS grid in §5.4. Further architectural details are provided in App. E.

### 4.2 Training Objective

The composite objective is formulated as:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{recon}} \;+\; \lambda_s \mathcal{L}_{\text{static}} \;+\; \lambda_d \mathcal{L}_{\text{dynamic}}, \tag{1}$$

where $\lambda_s$ and $\lambda_d$ are the contrastive loss weights, both set to $4.0$.

To prevent the collapse of the visual embeddings, we apply a Mean Squared Error (MSE) reconstruction loss $\mathcal{L}_{\text{recon}}$ to the frozen representation $\boldsymbol{z}$. To align the static visual ($\boldsymbol{z}_s^v$) and field ($\boldsymbol{z}_s^f$) embeddings, we use a standard InfoNCE loss across the batch $B = 8192$, with temperature $\tau_s = 0.07$:

$$\mathcal{L}_{\text{static}} \;=\; -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big((\boldsymbol{z}_{s,i}^v)^\top \boldsymbol{z}_{s,i}^f / \tau_s\big)}{\sum_{k=1}^{B} \exp\big((\boldsymbol{z}_{s,i}^v)^\top \boldsymbol{z}_{s,k}^f / \tau_s\big)}. \tag{2}$$

#### 4.2.1 Phenological Contrastive Loss

To model the ordinal, repeating cycle of plant phenology, we replace standard contrastive learning with a cyclically weighted contrastive objective. Let $\boldsymbol{h} \in \mathbb{R}^{2B \times D}$ be the concatenation of the $B$ dynamic image and $B$ dynamic field embeddings in a batch, $p(i)$ the index of the cross-modal positive for anchor $i$, and $W_{ik}$ a cyclic weight that masks negatives at the anchor's phenophase and down-weights adjacent ones ($W_{ik} = 0.8$). The dynamic alignment loss is:

$$\mathcal{L}_{\text{dynamic}} = -\frac{1}{2B}\sum_{i=1}^{2B} \log \frac{\exp\Big(\boldsymbol{h}_i^\top \boldsymbol{h}_{p(i)} / \tau_d\Big)}{\sum_{k=1,k\neq i}^{2B} W_{ik} \exp\Big(\boldsymbol{h}_i^\top \boldsymbol{h}_k / \tau_d\Big)}, \tag{3}$$

where $\tau_d = 0.1$ is the temperature scale and $W_{i,p(i)} = 1$ for the matched positive pair.

### 4.3 Batch Balancing

Training must balance $\mathcal{L}_{\text{static}}$, which relies on high intra-species density, and $\mathcal{L}_{\text{dynamic}}$, which requires balanced phenological coverage. We achieve this with a species-stratified, phenology-balanced sampling strategy. To prevent dominant phenological states from biasing the dynamic loss, we cap samples per phenophase at the per-species median, up to a maximum budget $M = 32$.

## 5 Experiments

### 5.1 Phenology as a First-Class Channel

Citizen-science training data are skewed toward conspicuous phenophases (Table 1), and models trained with only species-level supervision inherit this skew, making their performance unstable across phenological states. Because general contrastive objectives are hypersensitive to batch-distribution skew, they pull the species text embedding toward the most frequent visual phenological state (Oord et al., 2018; Radford et al., 2021) (see App. D). Empirically, web-scraped datasets suffer from severe observation bias toward conspicuous phenology (Iwanycki Ahlstrand et al., 2022; Chapman et al., 2015), and as shown in Table 1, datasets like Pheno3M and its overlap with TOL-200M are heavily skewed toward flowering imagery.

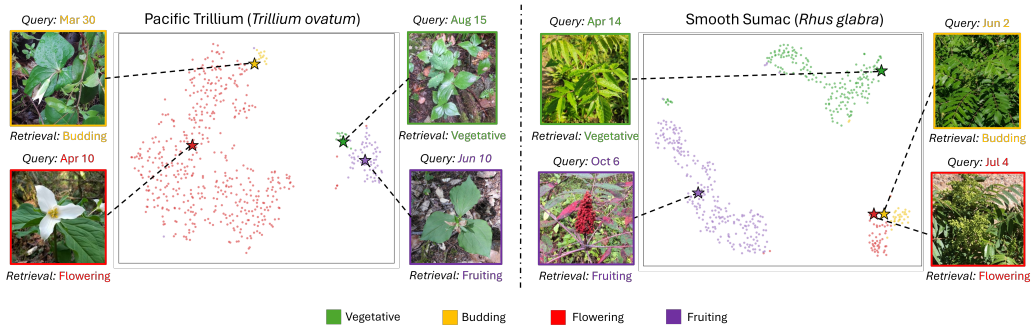



Figure 4: **Field→Image retrieval in dynamic latent space.** Dynamic image embeddings $\boldsymbol{z}_d^v$ (○) organize by phenophase, while dynamic field queries $\boldsymbol{z}_d^f$ (★) retrieve phenologically aligned imagery.

To quantify how this bias impacts taxonomic representations, we isolate and test the species axis directly. We pool our test sets to construct a perfectly balanced gallery for 344 species, each containing exactly $K = 8$ images per phenological state (32 images per species). We then query this gallery using species-name text embeddings and record the distribution of the top-8 retrievals. If a model's representation is truly agnostic to phenological change, the retrieval rate should be uniform.

Table 1: **Training Skew and Species → Image Phenological Retrieval Bias.** (Top) Phenophase proportions in Pheno3M and its overlap with TOL-200M. (Bottom) Mean retrieval rate per phenophase ± 95% CI in a balanced gallery (null = 0.25). **Bolding** indicates significant deviation from the null ($p < 0.003$), while JSD is Jensen–Shannon Divergence of each row's distribution to the null.

| Train Dataset / Model | Vegetative | Budding | Flowering | Fruiting | JSD (↓) |
|---|---|---|---|---|---|
| Pheno3M | 0.15 | 0.12 | 0.57 | 0.17 | – |
| Pheno3M ∧ TOL-200M | 0.17 | 0.10 | 0.56 | 0.17 | – |
| TaxaBind (Sastry et al., 2025b) | **0.13** ± 0.015 | **0.29** ± 0.017 | **0.33** ± 0.018 | 0.25 ± 0.017 | 0.0126 |
| BioCLIP (Stevens et al., 2024) | **0.13** ± 0.015 | **0.29** ± 0.017 | **0.33** ± 0.017 | 0.24 ± 0.017 | 0.0130 |
| BioCLIP 2 (**?**) | **0.18** ± 0.017 | 0.26 ± 0.015 | **0.29** ± 0.017 | **0.28** ± 0.017 | 0.0040 |
| Ours $\boldsymbol{z}_s$ (random batching) | **0.19** ± 0.016 | 0.27 ± 0.015 | **0.28** ± 0.016 | 0.26 ± 0.015 | 0.0026 |
| Ours $\boldsymbol{z}_s$ (balanced batching) | 0.25 ± 0.016 | 0.25 ± 0.015 | 0.24 ± 0.015 | 0.26 ± 0.016 | 0.0001 |
| *Unbiased null* | 0.25 | 0.25 | 0.25 | 0.25 | 0.0000 |

All baseline models significantly overrepresent flowering imagery, as shown in Table 1. Incorporating other modalities (e.g., TaxaBind: flowering at 0.33) and scaling up training data (e.g., BioCLIP to BioCLIP 2) fail to resolve this bias. Budding and fruiting are also over-retrieved (0.29 budding for TaxaBind and BioCLIP) despite their low training share, which we attribute to shortcut learning on their highly discriminative visual features. To isolate the impact of our balanced batching, we compare with random batching, which underrepresents the vegetative phenophase (0.19) and overrepresents flowering (0.28) like baseline models. However, after balancing, the retrieval rates stabilize across all phenophases, demonstrating the phenological invariance of our static representations.

### 5.2 Phenological and Compositional Retrieval

**Task setup.** We define two complementary retrieval benchmarks. *Compositional matched-pair retrieval* (Table 2) asks whether a query retrieves images that jointly match *species* ∧ *phenophase*. We report Recall@5, Recall@10 and mAP for field→image and image→image across splits. PhenoMap ranks gallery items by late fusion of the two channels, $s = \alpha\, s_d + (1 - \alpha)\, s_s$, the weighted sum of the dynamic- and static-channel similarities with $\alpha = 0.9$. *Fine-grained phenological retrieval* (Table 3) instead ranks with a 25-concept caption relevance signal at NDCG@5, using the full pool, same-species and cross-species galleries. The cross-species columns evaluate phenophase transfer without shared taxonomy.

**Baselines.** In addition to pretrained baselines (**?**Stevens et al., 2024; Yang et al., 2024; Sastry et al., 2025b; 2026; 2025a), we fine-tune BioCLIP 2's vision encoder on the species labels in our training dataset (BioCLIP 2-FT). We also fine-tune the BioCLIP 2 vision encoder on a probing objective

Table 2: **Compositional matched-pair retrieval (species ∧ phenophase).** Recall@5, Recall@10, and mean average precision for field→image and image→image retrieval on the random and spatial test splits. A gallery item is relevant only when species and coarse phenophase both match the query. Values are percentages; best per column in bold.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| TaxaBind (Sastry et al., 2025b) | 14.3 | 25.2 | 25.1 | 12.8 | 22.8 | 22.7 | 16.9 | 27.1 | 25.8 | 18.2 | 29.3 | 28.5 |
| RCME (Sastry et al., 2025a) | 13.9 | 25.0 | 25.9 | 12.7 | 23.2 | 24.1 | 18.5 | 30.7 | 29.8 | 20.0 | 32.9 | 32.5 |
| ProM3E (Sastry et al., 2026) | 13.8 | 24.4 | 23.9 | 12.4 | 21.8 | 21.6 | 16.7 | 27.0 | 25.6 | 18.4 | 29.4 | 28.6 |
| BioTrove-CLIP (Yang et al., 2024) | 7.2 | 14.2 | 16.1 | 7.2 | 14.0 | 15.6 | 21.2 | 36.5 | 36.9 | 22.1 | 37.4 | 37.7 |
| BioCLIP (Stevens et al., 2024) | 14.3 | 25.3 | 25.5 | 12.8 | 23.3 | 23.3 | 17.3 | 28.0 | 26.9 | 18.7 | 30.8 | 29.9 |
| BioCLIP 2 (**?**) | 20.2 | 39.4 | 44.7 | 19.9 | 39.2 | 43.9 | 27.2 | 48.8 | 51.5 | 27.9 | 49.6 | 52.3 |
| BioCLIP 2–FT (**?**) | 23.2 | 44.5 | 49.3 | 20.8 | 39.2 | 43.7 | 25.6 | 46.8 | 51.5 | 25.0 | 45.0 | 48.9 |
| BioCLIP 2–Sup (**?**) | 22.5 | 43.0 | 47.8 | 20.1 | 38.0 | 42.5 | 27.4 | 50.3 | 54.2 | 27.0 | 48.3 | 52.6 |
| PhenoMap | **29.2** | **55.1** | **60.3** | **27.7** | **52.0** | **56.7** | **28.6** | **52.0** | **56.1** | **28.4** | **51.0** | **54.9** |

Table 3: **Fine-grained phenological retrieval (25 concepts, NDCG@5).** Field→image and image→image ranking on the random and spatial splits. *Same* restricts the gallery to same-species images, *All* uses the full pool, and *Cross* restricts to cross-species images. Relevance uses caption-derived concept agreement. Best per column in bold.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* |
| TaxaBind (Sastry et al., 2025b) | 0.630 | 0.533 | 0.390 | 0.598 | 0.515 | 0.396 | 0.756 | 0.656 | 0.545 | 0.745 | 0.652 | 0.562 |
| RCME (Sastry et al., 2025a) | 0.635 | 0.532 | 0.394 | 0.596 | 0.512 | 0.402 | 0.766 | 0.665 | **0.554** | 0.752 | 0.657 | **0.570** |
| ProM3E (Sastry et al., 2026) | 0.629 | 0.469 | 0.321 | 0.592 | 0.469 | 0.339 | 0.721 | 0.617 | 0.474 | 0.706 | 0.609 | 0.494 |
| BioTrove-CLIP (Yang et al., 2024) | 0.615 | 0.506 | 0.366 | 0.585 | 0.489 | 0.387 | 0.746 | 0.635 | 0.495 | 0.735 | 0.632 | 0.516 |
| BioCLIP (Stevens et al., 2024) | 0.642 | 0.527 | 0.383 | 0.592 | 0.514 | 0.398 | 0.766 | 0.661 | 0.549 | 0.750 | 0.658 | 0.568 |
| BioCLIP 2 (**?**) | 0.623 | 0.538 | 0.385 | 0.593 | 0.519 | 0.407 | 0.765 | 0.641 | 0.514 | 0.754 | 0.638 | 0.540 |
| BioCLIP 2–FT (**?**) | 0.641 | 0.637 | 0.405 | 0.614 | 0.629 | 0.421 | 0.738 | 0.667 | 0.433 | 0.721 | 0.669 | 0.458 |
| BioCLIP 2–Sup (**?**) | 0.640 | 0.635 | 0.402 | 0.612 | 0.625 | 0.418 | 0.762 | 0.688 | 0.460 | 0.745 | 0.682 | 0.475 |
| PhenoMap $z_d$ | **0.740** | **0.671** | **0.502** | **0.697** | **0.654** | **0.498** | **0.776** | **0.702** | 0.545 | **0.761** | **0.696** | 0.558 |

for phenophase classification using a simple binary cross-entropy loss (BioCLIP 2-Sup). We use LoRA (Hu et al., 2021) for both models.

**Results.** On matched-pair retrieval (Table 2), PhenoMap exceeds all baselines with the clearest margins over BioCLIP 2 when queries are field metadata (e.g. +9.0 R@5 and +15.6 mAP on Random field→image). On concept retrieval (Table 3), PhenoMap exceeds most baselines, especially in field→image cross-species cells where species priors are absent (+0.091 NDCG@5 vs. BioCLIP 2 on Spatial). BioCLIP 2 improves upon BioCLIP most where species×phase binding dominates, but gains are smaller on fine-grained phenology, indicating that scaling species-first contrastive pretraining does not by itself align the representation to within-phase retrieval. Supervising on phenophase naively (BioCLIP 2-Sup) produces similar performance, motivating the need for explicitly disentangled learning. RCME leads image→image cross-species retrieval, which we attribute to its explicit taxonomic hierarchy bridging visual gaps through shared morphology.

### 5.3 USA-NPN MULTI-LABEL PROBING

We evaluate frozen foundation features on the USA-NPN visit benchmark from §3. We use a species-masked binary cross-entropy over a 13-way multi-label presence of `phenophase`, matching species-conditioned phenophase heads in prior models (Sainte Fare Garnot et al., 2025; Körschens et al., 2025). In Figure 6 we report macro average precision (macro AP), across three fixed splits: **random** (stratified visit holdout), **spatial** (leave-region-out) (Brun et al., 2024), and **temporal** (train on early years, test on held-out recent years), and under partial budget scenarios. All methods use a two-layer MLP probe head (width 128, depth 2, dropout 0.1) on top of frozen features. Baselines lack a temporal encoder, so we pair them with growing-degree-days at the observation date, the canonical thermal-time phenology covariate. Under spatial and temporal shift, baselines decay sharply as supervision shrinks, while the PhenoMap field probe degrades gradually.

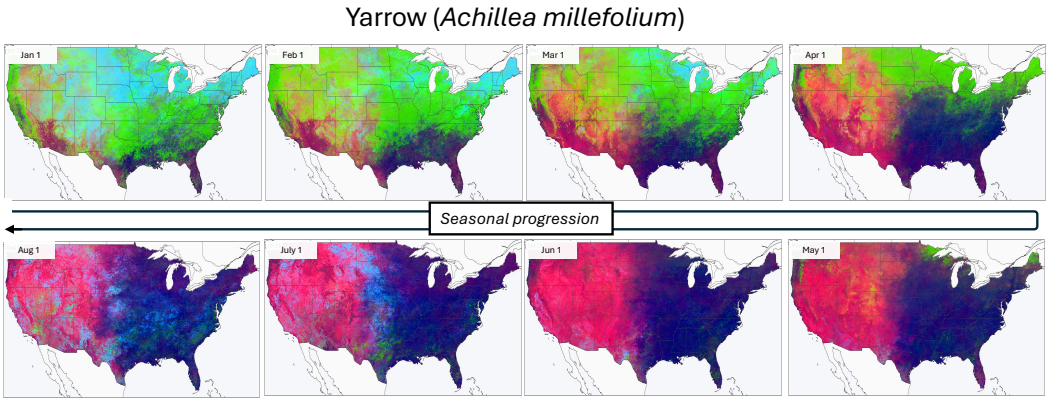



Figure 5: **Bioclimatic structure in our field encoder.** CONUS field embeddings for *Achillea millefolium* across twelve months, projected via ICA to three components and rendered as RGB. Recovered axes show trends of south→north shift tracking seasonal progression, and a longitudinal component aligned with the ∼100th meridian (the canonical East/West bioclimatic divide).

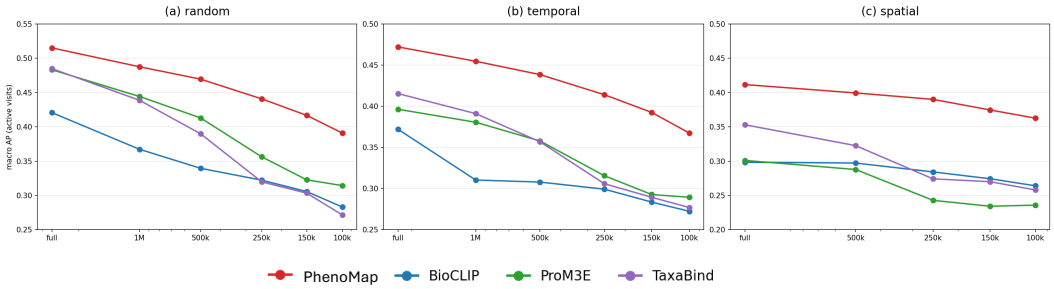



Figure 6: **USA-NPN evaluation.** Macro AP over 13 phenophase classes vs. training-set size across random, spatial, and temporal splits. Identical MLP probes on frozen features; baselines paired with growing-degree-days. PhenoMap degrades more gradually under spatial and temporal shift.

Figure 7 shows our neural field has internalized phenological structure by comparing predicted and observed phenophase activity curves for Vanilla leaf (*Achlys californica*) across three latitude bands at two phenophases. Spring onset shifts later as latitude increases, while fall senescence shifts earlier, moving in opposite directions across the cycle. This is consistent with latitude-photoperiod-temperature coupling. Predicted curves track the USA-NPN observed rates within each band in timing and amplitude. Construction details are in App. A.8.

### 5.4 Spatiotemporal Neural Field Representations

Querying the neural field for a single species on a 12-month CONUS grid and decomposing the embeddings with ICA yields phenology maps (Figure 5). We observe visible latitudinal trends tied to seasonal development and a longitudinal separation aligned with the 100th meridian, the canonical North American bioclimatic boundary between the humid East and semi-arid West (Seager et al., 2018). This longitudinal boundary is consistent with ecological findings that eastern US phenology is temperature-driven and western US phenology is water-driven (Forrest & Miller-Rushing, 2010).

### 5.5 Ablations

Table 4 shows that each component of PhenoMap contributes to strong phenological retrieval. Replacing the BioCLIP 2 backbone with other visual backbones significantly reduces retrieval accuracy, emphasizing the importance of a strong pretrained backbone. Using an unfactorized approach also leads to a performance drop, highlighting the need to separate static taxonomic information from dynamic phenological information. Field encoder ablations further show that

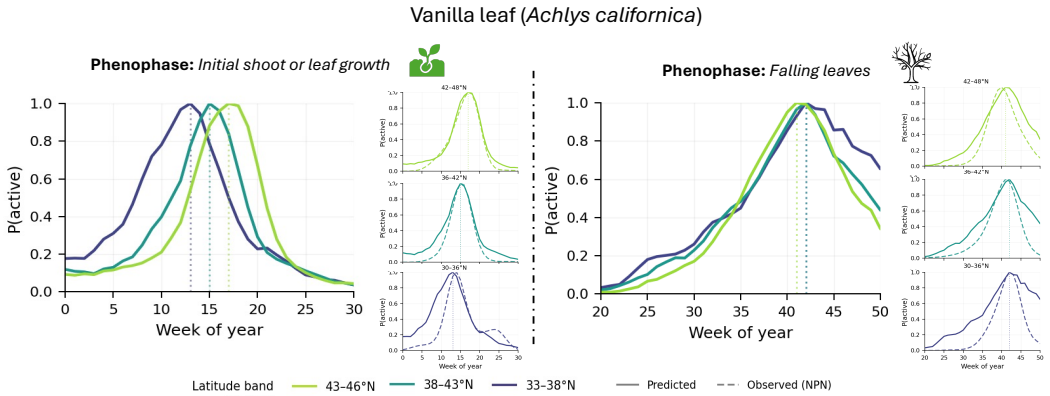



Figure 7: **Predicted vs. observed phenophase activity curves. Left:** Initial shoot or leaf growth. **Right:** Falling leaves. Predicted curves (solid) come from a linear probe on frozen PhenoMap field embeddings. Observed curves (dashed) are USA-NPN ground-truth rates. Spring onset shifts later moving north. Fall senescence shifts earlier moving north.

Table 4: **PhenoMap Ablations.** We evaluate on matched-pair retrieval. We use a scaled down version of PhenoMap (PhenoMap-S), trained for 15k steps.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| PhenoMap-S | 26.2 | 49.6 | 54.3 | 25.4 | 47.4 | 51.9 | 27.3 | 49.5 | 53.0 | 27.9 | 50.0 | 53.5 |
| *Architecture* | | | | | | | | | | | | |
| w/ TaxaBind backbone | 19.7 | 35.2 | 36.7 | 18.0 | 32.2 | 33.6 | 19.1 | 32.2 | 32.2 | 20.1 | 33.9 | 34.3 |
| w/ BioCLIP backbone | 19.2 | 34.4 | 35.3 | 17.6 | 31.7 | 32.8 | 18.1 | 30.4 | 30.1 | 19.4 | 32.6 | 32.7 |
| w/ DINOv3 (ViT) | 20.9 | 37.3 | 38.7 | 20.6 | 37.1 | 38.4 | 18.7 | 31.0 | 30.5 | 21.0 | 34.7 | 35.0 |
| w/ no factorization | 21.8 | 37.7 | 37.7 | 21.9 | 38.0 | 38.6 | 24.4 | 42.1 | 42.7 | 25.2 | 43.2 | 44.5 |
| *Field Encoder Modalities* | | | | | | | | | | | | |
| w/o AlphaEarth | 26.1 | 49.5 | 54.3 | 25.2 | 47.5 | 51.8 | 27.3 | 49.6 | 53.0 | 27.9 | 49.9 | 53.5 |
| w/o PRISM | 21.5 | 42.7 | 47.4 | 21.3 | 41.3 | 46.1 | 26.4 | 47.8 | 51.4 | 27.1 | 48.4 | 52.1 |
| w/o Species | 24.8 | 47.2 | 51.8 | 24.0 | 45.0 | 49.4 | 26.8 | 48.6 | 52.0 | 27.3 | 48.8 | 52.2 |

temporal climate information is the most important contextual signal. In addition, removing the species-conditioning signal also reduces performance, confirming that phenology must be modeled jointly with taxonomy. For additional ablations, please refer to App. F.

## 6 Conclusions

In this work, we proposed **PhenoMap**, a multimodal model that produces factorized species and phenology embeddings by aligning plant images with species text and a spatiotemporal field encoder. To train PhenoMap, we curated **Pheno3M**, a large-scale dataset of plant observations with phenophase and concept annotations, paired with an external benchmark of expertly surveyed USA-NPN records. We showed that PhenoMap exceeds species-aligned baselines on compositional and fine-grained phenology retrieval, retains accuracy under spatial and temporal shift on USA-NPN, recovers a debiased species channel, and qualitatively shows intuitive trends in its field-encoder latent space. The results showed that phenology can be a first-class representational channel, queryable by species and phenological context. PhenoMap provides a strong foundation for phenology-aware ecological research and points the way toward multimodal foundation models that go beyond taxonomy to the within-species axes it suppresses.

### AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data: a vision-language model (InternVL3-38B) produced the 25 fine-grained phenological concept ratings used only to build the fine-grained retrieval evaluation (App. A.2), whose reliability we quantify against community annotations and a human agreement study conducted by the authors (App. A.4). We have not used generative AI tools for any other task in the required-disclosure categories of the ICLR 2027 AI Policy. Additionally, we used generative AI tools for drafting and editing the paper for readability. All AI-assisted text was checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

Pheno3M is curated from research-grade iNaturalist observations contributed by citizen scientists. We do not redistribute photographs. We release annotations, splits, and image URLs subject to each contributor's license, and retain coordinates at the precision published by iNaturalist, including records obfuscated for endangered species (App. A.10). The human agreement study (App. A.4) was performed by the authors on public images and involved no external participants or personal data. Citizen-science data carry distributional and acquisition biases. We quantify the phenophase skew and its effect on species-supervised models (§5.1) and discuss remaining limitations in App. H. The intended use is ecological monitoring and phenology research, and we foresee no direct harmful applications.

### REPRODUCIBILITY STATEMENT

Dataset curation, label assembly, and split construction are described in App. A.1 and App. A.7, the concept annotation pipeline with the verbatim prompt in App. A.2, and its validation in App. A.4. Architecture, losses, and hyperparameters are given in §4 and App. E, model scales in Table 12, and baseline protocols in App. G.2. Pheno3M annotations, splits, and the code to reproduce feature extraction, training, and evaluation will be released under CC-BY 4.0 upon acceptance.

## REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

David Basler. Evaluating phenological models for the prediction of leaf-out dates in six temperate tree species across central europe. *Agricultural and Forest Meteorology*, 217:10–21, 2016. doi: 10.1016/j.agrformet.2015.11.007.

Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. *arXiv preprint arXiv:2303.16198*, 2023.

Peggy A. Bevan, Omiros Pantazis, Holly A. I. Pringle, Guilherme Braga Ferreira, Daniel J. Ingram, Emily K. Madsen, Liam Thomas, Dol Raj Thanet, Thakur Silwal, Santosh Rayamajhi, Gabriel J. Brostow, Oisin Mac Aodha, and Kate E. Jones. Deep learning-based ecological analysis of camera trap images is impacted by training data quality and quantity. *Remote Sensing in Ecology and Conservation*, 12(3):494–508, 2026. doi: 10.1002/rse2.70052.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Philipp Brun, Dirk N Karger, Damaris Zurell, Patrice Descombes, Lucienne C de Witte, Riccardo de Lutio, Jan Dirk Wegner, and Niklaus E Zimmermann. Multispecies deep learning using citizen

science data produces more informative plant community models. *Nature Communications*, 15(1): 4421, 2024.

Daniel S Chapman, Sandra Bell, Stephan Helfer, and David B Roy. Unbiased inference of plant flowering phenology from biological recording data. *Biological Journal of the Linnean Society*, 115(3):543–554, 2015.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Elsa E Cleland, Isabelle Chuine, Annette Menzel, Harold A Mooney, and Mark D Schwartz. Shifting plant phenology in response to global change. *Trends in Ecology & Evolution*, 22(7):357–365, 2007.

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

Theresa M. Crimmins, Michael A. Crimmins, Katherine L. Gerst, Alyssa H. Rosemartin, and Jake Weltzin. Usa national phenology network's volunteer-contributed observations yield predictive models of phenological transitions. *PLOS ONE*, 12(8):e0182919, 2017. doi: 10.1371/journal.pone.0182919.

Christopher Daly, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, 28(15):2031–2064, 2008.

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 (ICCV)*, pp. 6143–6154, October 2025.

Riccardo de Lutio, John Y Park, Kimberly A Watson, Stefano D'Aronco, Jan D Wegner, Jan J Wieringa, Melissa Tulig, Richard L Pyle, Timothy J Gallaher, Gillian Brown, et al. The herbarium 2021 half–earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, 12:787127, 2022.

Ellen G. Denny, Katherine L. Gerst, Abraham J. Miller-Rushing, Gail L. Tierney, Theresa M. Crimmins, Carolyn A. F. Enquist, Patricia Guertin, Alyssa H. Rosemartin, Mark D. Schwartz, Kimberly A. Thomas, and Jake F. Weltzin. Standardized phenology monitoring methods to track plant and animal activity for science and resource management applications. *International Journal of Biometeorology*, 58(4):591–601, 2014. doi: 10.1007/s00484-014-0789-5.

Simone Des Roches, Linwood H Pendleton, Beth Shapiro, and Eric P Palkovacs. Conserving intraspecific variation for nature's contributions to people. *Nature ecology & evolution*, 5(5): 574–582, 2021.

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

Russell Dinnage, Erin Grady, Nevyn Neal, Jonn Deck, Ellen Denny, Ramona Walls, Carrie Seltzer, Robert Guralnick, and Daijiang Li. Phenovision: A framework for automating and delivering research-ready plant phenology data from field images. *Methods in Ecology and Evolution*, 16(8): 1763–1780, 2025.

Johannes Dollinger, Philipp Brun, Vivien Sainte Fare Garnot, and Jan Dirk Wegner. Sat-sinr: High-resolution species distribution models through satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:41–48, 2024.

Martin Edwards and Anthony J Richardson. Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature*, 430(7002):881–884, 2004.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

G. A. Finn, A. E. Straszewski, and V. Peterson. A general growth stage key for describing trees and woody plants. *Annals of Applied Biology*, 151(1):127–131, 2007. doi: 10.1111/j.1744-7348.2007.00159.x.

Jessica Forrest and Abraham J. Miller-Rushing. Toward a synthetic understanding of the role of phenology in ecology and evolution. *Philosophical Transactions of the Royal Society B*, 365(1555):3101–3112, 2010.

Camille Garcin, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, Joseph Salmon, et al. Pl@ ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Hervé Goëau, Pierre Bonnet, and Alexis Joly. Overview of plantclef 2022: Image-based plant identification at global scale. *arXiv preprint arXiv:2509.17632*, 2025.

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. *arXiv preprint arXiv:2405.17537*, 2024.

ZeMing Gong, Chuanqi Tang, Xiaoliang Huo, Nicholas Pellegrino, Austin T Wang, Graham W Taylor, Angel X Chang, Scott C Lowe, and Joakim Bruslund Haurum. Hyperbolic multimodal representation learning for biological taxonomies. *arXiv preprint arXiv:2508.16744*, 2025.

Max Hamilton, Christian Lange, Elijah Cole, Alexander Shepard, Samuel Heinrich, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Combining observational data and language for species range estimation. *Advances in Neural Information Processing Systems*, 37:17719–17742, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Koen Hufkens, David Basler, Tom Milliman, Eli K Melaas, and Andrew D Richardson. An integrated phenology modelling framework in r. *Methods in Ecology and Evolution*, 9(5):1276–1285, 2018.

Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *European Conference on Computer Vision*, pp. 173–190. Springer, 2024.

Natalie Iwanycki Ahlstrand, Richard B Primack, and Anders P Tøttrup. A comparison of herbarium and citizen science phenology datasets for detecting response of flowering time to climate change in denmark. *International Journal of Biometeorology*, 66(5):849–862, 2022.

Subash Khanal, Eric Xing, Srikumar Sastry, Aayush Dhakal, Zhexiao Xiong, Adeel Ahmad, and Nathan Jacobs. PSM: Learning probabilistic embeddings for multi-scale zero-shot soundscape mapping. In *ACM Multimedia*, October 2024. doi: 10.1145/3664647.3681620.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

Matthias Körschens, Solveig Franziska Bucher, Paul Bodesheim, Joachim Denzler, and Christine Römermann. Plantcapnet: A deep learning system for image-based plant cover and phenology analysis. *Ecological Informatics*, pp. 103413, 2025.

Christian Lange, Max Hamilton, Elijah Cole, Alexander Shepard, Samuel Heinrich, Angela Zhu, Subhransu Maji, Grant Van Horn, and Oisin Mac Aodha. Feedforward few-shot species range estimation. *arXiv preprint arXiv:2502.14977*, 2025.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024a. doi: 10.1109/TGRS.2024.3390838. URL https://doi.org/10.1109/TGRS.2024.3390838.

Guohua Liu, Mirco Migliavacca, Christian Reimers, Basil Kraft, Markus Reichstein, Andrew D. Richardson, Lisa Wingate, Nicolas Delpierre, Hui Yang, and Alexander J. Winkler. Deepphenomem v1.0: Deep learning modelling of canopy greenness dynamics accounting for multi-variate meteorological memory effects on vegetation phenology. *Geoscientific Model Development*, 17: 6683–6701, 2024b. doi: 10.5194/gmd-17-6683-2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

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.

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

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning (ICML)*, 2023.

Uwe Meier. *Growth Stages of Mono- and Dicotyledonous Plants: BBCH Monograph*. Julius Kühn-Institut (JKI), Quedlinburg, Germany, 2 edition, 2018. doi: 10.5073/20180906-074619.

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

Camille Parmesan and Gary Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421(6918):37–42, 2003.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4199–4209, 2023.

Lukas Picek, Christophe Botella, Maximilien Servajean, César Leblanc, Rémi Palard, Théo Larcher, Benjamin Deneu, Diego Marcos, Pierre Bonnet, and Alexis Joly. GeoPlant: Spatial plant species prediction dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024a.

Lukas Picek, Christophe Botella, Maximilien Servajean, César Leblanc, Rémi Palard, Théo Larcher, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, Pierre Bonnet, and Alexis Joly. Overview of GeoLifeCLEF 2024: Species composition prediction with high spatial resolution at continental scale using remote sensing. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pp. 1966–1977, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Arjun Rao, Marc Rußwurm, Konstantin Klemmer, and Esther Rolf. Measuring the intrinsic dimension of earth representations. *arXiv preprint arXiv:2511.02101*, 2025.

Rachel A. Reeb, Naeem Aziz, Samuel M. Lapp, Justin Kitzes, J. Mason Heberling, and Sara E. Kuebbing. Using convolutional neural networks to efficiently extract immense phenological data from community science images. *Frontiers in Plant Science*, 12:787407, 2022. doi: 10.3389/fpls.2021.787407.

Andrew D Richardson, Trevor F Keenan, Mirco Migliavacca, Youngryel Ryu, Oliver Sonnentag, and Michael Toomey. Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agricultural and Forest Meteorology*, 169:156–173, 2013.

Alyssa H. Rosemartin, Ellen G. Denny, Katharine L. Gerst, R. Lee Marsh, Erin E. Posthumus, Theresa M. Crimmins, and Jake F. Weltzin. USA national phenology network observational data documentation. Open-File Report 2018-1060, U.S. Geological Survey, 2018.

Vivien Sainte Fare Garnot, Lynsay Spafford, Jelle Lever, Christian Sigg, Barbara Pietragalla, Yann Vitasse, Arthur Gessler, and Jan Dirk Wegner. Deep learning meets tree phenology modelling: PhenoFormer versus process-based models. *Methods in Ecology and Evolution*, 2025.

Srikumar Sastry, Xin Xing, Aayush Dhakal, Subash Khanal, Adeel Ahmad, and Nathan Jacobs. LD-SDM: Language-driven hierarchical species distribution modeling. *arXiv preprint arXiv:2312.08334*, 2023.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 7136–7145, 2024.

Srikumar Sastry, Aayush Dhakal, Eric Xing, Subash Khanal, and Nathan Jacobs. Global and local entailment learning for natural world imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15770–15780, 2025a.

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

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Jiayu Lin, Dan Cher, Phoenix Jarosz, and Nathan Jacobs. Prom3e: Probabilistic masked multimodal embedding model for ecology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20564–20574, 2026.

Mark D Schwartz, Julio L Betancourt, and Jake F Weltzin. From Caprio's lilacs to the USA national phenology network. *Frontiers in Ecology and the Environment*, 10(6):324–327, 2012.

Richard Seager, Nathan Lis, Jamie Feldman, Mingfang Ting, A. Park Williams, Jennifer Nakamura, Haibo Liu, and Naomi Henderson. Whither the 100th meridian? the once and future physical and human geography of america's arid–humid divide. Part I: The story so far. *Earth Interactions*, 22 (5):1–22, 2018.

Ayushi Sharma, Johanna Trost, Daniel Lusk, Johannes Dollinger, Julian Schrader, Christian Rossi, Javier Lopatin, Etienne Laliberté, Simon Haberstroh, Jana Eichel, et al. Planttraitnet: An uncertainty-aware multimodal framework for global-scale plant trait inference from citizen science data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39239–39248, 2026.

David G Shatwell, Ishan Rajendrakumar Dave, Sirnam Swetha, and Mubarak Shah. GT-Loc: Unifying when and where in images through a joint embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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, et al. Bioclip: A vision foundation model for the tree of life. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424. IEEE, 2024.

Brian J. Stucky, Rob Guralnick, John Deck, Ellen G. Denny, Kjell Bolmgren, and Ramona Walls. The plant phenology ontology: A new informatics resource for large-scale integration of plant phenology data. *Frontiers in Plant Science*, 9:517, 2018. doi: 10.3389/fpls.2018.00517.

Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. Satbird: A dataset for bird species distribution modeling using remote sensing and citizen science data. In *NeurIPS Datasets and Benchmarks Track*, 2023a.

Mélisande Teng, Amna Elmustafa, Benjamin Akera, Yoshua Bengio, Hager Radi, Hugo Larochelle, and David Rolnick. SatBird: a dataset for bird species distribution modeling using remote sensing and citizen science data. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023b.

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

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 (NeurIPS)*, 2023.

Mary Jane West-Eberhard. Phenotypic plasticity and the origins of diversity. *Annual review of Ecology and Systematics*, 20:249–278, 1989.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Chih-Hsuan Yang, Ben Feuer, Zaki Jubery, Zi K Deng, Andre Nakkab, Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. *arXiv preprint arXiv:1909.07499*, 2019.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Ziheng Zhang, Xinyue Ma, Arpita Chowdhury, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Samuel Stevens, Hilmar Lapp, Tanya Berger-Wolf, Yu Su, et al. Biocap: exploiting synthetic captions beyond labels in biological foundation models. In *International Conference on Learning Representations*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

# APPENDIX

## A PHENO3M: ADDITIONAL DETAILS

### A.1 CURATION PIPELINE

We curate Pheno3M directly from the iNaturalist API. We retain only research-grade observations carrying phenology annotations, restricted to plant taxa, with non-null geographic coordinates and observation dates, at least one accessible image URL, and at least one phenophase annotation. Figure 8 shows the temporal, spatial, and label information.

**Phenophase label assembly.** We prompt InternVL3-38B (Zhu et al., 2025) with each image and output a ternary score per concept (absent / present / prominent) plus a short caption. We generate VLM outputs for *25,000* observations *per test split*. iNaturalist exposes phenology as a set of community-contributed annotation flags per observation, drawn from *flowering*, *flower budding*, *fruiting*, and *no evidence of flowering*. We treat each flag independently and aggregate across annotators by inclusive disjunction, yielding a multi-hot label over four phases: *vegetative* (the implicit complement, set when only *no evidence of flowering* is reported), *budding*, *flowering*, and *fruiting*. Because the four phases follow a natural seasonal cycle (vegetative $\rightarrow$ budding $\rightarrow$ flowering $\rightarrow$ fruiting $\rightarrow$ vegetative), adjacent phases co-occur far more often than opposite phases, a property we exploit during training (§4).

**Deduplication.** A single botanical observation can yield multiple iNaturalist records — e.g., several photos of the same individual taken minutes apart, or a contributor uploading the same plant from different angles. We deduplicate at the (species_id, lat, lon, day-of-year, year) level, with coordinates discretized to four decimal places ($\sim$10 m). When duplicates collapse to a single entry, we take the union of their multi-hot phenophase labels.

**Final dataset statistics.** After filtering and deduplication, Pheno3M contains approximately 3M observations across approximately 6K plant species. Training, validation and testing splits are shown in Figure 9.

**Existing datasets.** In Table 5, we compare our curated dataset with other plant-based benchmarks.

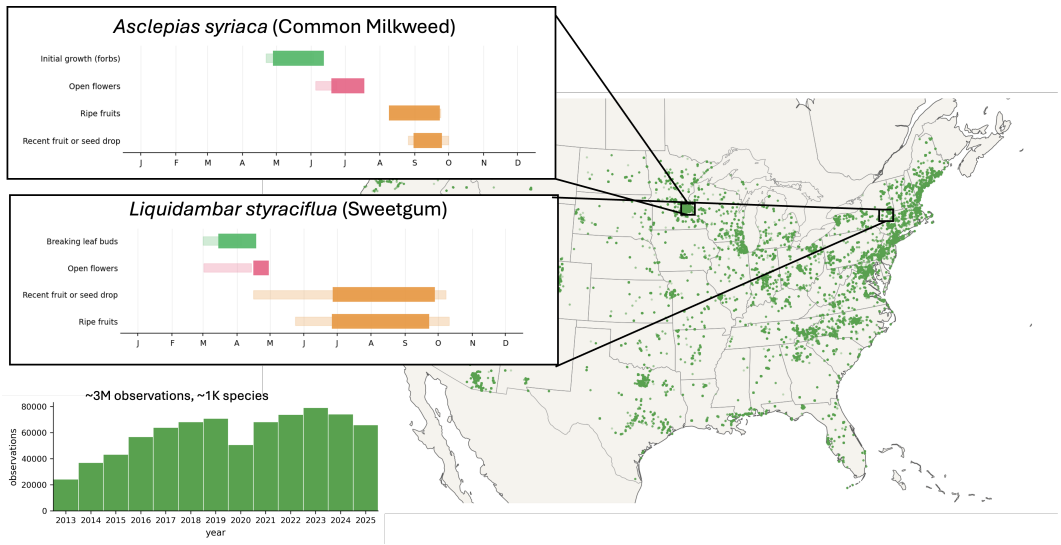



Figure 8: **USA-NPN data.** Spatial, and Temporal coverage of the USA-NPN dataset. We show observation density across the contiguous United States in addition to recorded species-specific phenophases gathered through systematic surveys.




Figure 9: **Pheno3M Splits.** Dataset splits of Pheno3M. We show observation density across the Contiguous United States of the sampled citizen science phenology observations.

Table 5: **Existing plant-image benchmarks.** Most large benchmarks are designed for species identification, while existing phenology datasets either provide binary supervision at scale or finer-grained supervision at much smaller size. Pheno3M is designed for fine-grained phenology learning at larger scale.

| Benchmark | Size | # Species | Primary task | Phenology |
|---|---|---|---|---|
| *General plant benchmarks* | | | | |
| Pl@ntNet-300K (Garcin et al., 2021) | 306K | 1K | species ID | — |
| iNat-21 (Plants) (Van Horn et al., 2018) | 1.2M | 4K | species ID | — |
| Herbarium-21 (de Lutio et al., 2022) | 2.5M | 64K | species ID | — |
| PlantCLEF 2022 (Goëau et al., 2025) | 4M | 80K | species ID | — |
| GeoPlant (Picek et al., 2024a) | 5M | 11K | SDM | — |
| TreeOfLife-10M (Plants) (Stevens et al., 2024) | 3.6M | 38K | species ID | — |
| BioTrove (Plants) (Yang et al., 2024) | 67M | 42K | species ID | — |
| TreeOfLife-200M (Plants) (**?**) | 101M | 104K | species ID | — |
| PlantTraitNet (Sharma et al., 2026) | 307K | 5K | Trait Regression | — |
| *Phenology benchmarks* | | | | |
| Phenology-2.4K (Reeb et al., 2022) | 2.4K | 3 | phenology classification | 4-stage |
| PhenoVision (Dinnage et al., 2025) | 1.5M | 4K | phenology classification | binary flower/fruit |
| Pheno3M (ours) | 3M | 6.8K | phenology retrieval | 4-stage + 25 concepts |

### A.2 Phenological concept generation

We derive $C$=25 per-image concept scores used during evaluation in two stages: a hand-designed ontology, and VLM annotations on selected observations.

**Ontology.** The vocabulary lists 25 fine-grained phenology concepts grouped under the same four cyclic phenophase classes as the iNaturalist-derived multi-hot label (vegetative, budding, flowering, fruiting), plus a small set of cross-class leaf traits (e.g. mature, coloring, senescent leaves) that often co-occur with reproductive stages. The concepts were defined with reference to the principal growth stages of the BBCH scale (Meier, 2018) and to USA-NPN phenophase definitions (Denny et al., 2014). We align to principal stages rather than two-digit BBCH codes because the latter are defined by proportions (e.g. code 65, half of all flowers open) that a single opportunistic photograph cannot support. The concept-to-stage mapping is given in App. A.3.

**VLM annotations.** We use InternVL3-38B in bfloat16 with dynamic-resolution tiling (at most six multi-crop tiles per image), generating up to 512 new tokens per call with temperature 0.6 and nucleus top-$p$=0.9. The prompt asks for two outputs in order: (1) a JSON object assigning each of the 25 concepts a rating in $\{0, 1, 2\}$ (absent, present, prominent), with instructions to rate only visible evidence and to use 0 when unsure; (2) a one- to two-sentence caption of visible phenophase (growth, reproduction, leaves, transitions) without naming the species. The iNaturalist phenophase multi-hot for that image is included in a short preamble to disambiguate structurally similar stages (e.g. flower vs. fruit), while still requiring descriptions grounded in pixels.

**Concept list (25).** BUD_DORMANT, BARK_ONLY, LEAF_FALLEN, WITHERED_STEMS; BUD_SWELLING, BUD_BURST, LEAF_EMERGING, LEAF_EXPANDING, NEW_GROWTH; FLOWER_BUD, FLOWER_OPEN, FLOWER_WILTING, PETAL_FALL; DEAD_FLOWERS, SEED_HEAD; FRUIT_GREEN, FRUIT_COLORED, FRUIT_SET, FRUIT_DEVELOPING, FRUIT_MATURE, FRUIT_DISPERSING, SEED_POD; LEAF_MATURE, LEAF_COLORING, LEAF_SENESCENT.

**Prompt.** The prompt sent to InternVL3-38B is reproduced verbatim below. The second paragraph is present only when the observation carries an iNaturalist phenophase flag. It is built from a fixed template in which `{stages}` joins the active flags with `AND`, using the strings "No phenology present (no buds, flowers, or fruit observed)", "Budding (new buds or leaves emerging)", "Flowering (flowers present)", and "Fruiting/Ripe (fruit or seed development)". The example below shows an observation flagged as flowering and fruiting. The no-context variant used in the prompt ablation omits this paragraph and is otherwise identical.

```
You are a plant phenology expert. Examine this image of a plant and
    describe its phenological state.

Expert phenological observers have classified this plant as: Flowering
    (flowers present) AND Fruiting/Ripe (fruit or seed development). Use
    this context to guide your interpretation of ambiguous structures
    (e.g. distinguishing fruit from flowers), but still describe only
    what you can actually see.

First, output a JSON object rating each phenological attribute (0 =
    absent, 1 = present, 2 = prominent):

{"bud_dormant", "bark_only", "leaf_fallen", "withered_stems",
    "bud_swelling", "bud_burst", "leaf_emerging", "leaf_expanding",
    "new_growth", "flower_bud", "flower_open", "flower_wilting",
    "petal_fall", "dead_flowers", "seed_head", "fruit_green",
    "fruit_colored", "fruit_set", "fruit_developing", "fruit_mature",
    "fruit_dispersing", "seed_pod", "leaf_mature", "leaf_coloring",
    "leaf_senescent"}

Rate ONLY what you can see in the image. Use 0 if unsure.

Then write a 1-2 sentence caption describing the visible phenological
    state, focusing on: growth stage, reproductive status, leaf
    condition, and any transitional features. Be specific about what you
    observe, not what you expect. Do NOT mention or guess the species
    name.

Output format (no other text):
'''json
{...ratings...}
Caption: <your caption>
```

### A.3 CONCEPT-TO-BBCH MAPPING

Table 6 maps the 25 concepts to BBCH principal growth stages (Meier, 2018). We map to principal stages only following Finn et al. (2007) which adopt the same principal-stage-only key for woody plants. Boundary cases reflect ambiguity intrinsic to single images: a swelling bud can be vegetative (stage 0) or floral (stage 5), which is why concept scores are multi-hot rather than forced to one code. USA-NPN phenophases are status-based rather than BBCH-coded, but the Plant Phenology Ontology (Stucky et al., 2018) reconciles the two vocabularies.

Table 6: **Concept to BBCH principal growth stage.** 14 concepts map to a single stage, 9 are boundary or type-dependent cases listed with both candidate stages, and 2 do not correspond to a growth stage.

| Concept | Stage | Note |
|---|---|---|
| *Clean single principal stage* | | |
| BUD_DORMANT | 0 | 00 dormancy |
| LEAF_EXPANDING | 1 | 11–19 |
| FLOWER_BUD | 5 | 55 flower buds visible, still closed |
| FLOWER_OPEN | 6 | 60–65 |
| FLOWER_WILTING | 6 | 67 flowers fading |
| PETAL_FALL | 6 | 67–68 |
| FRUIT_GREEN | 7 | |
| FRUIT_DEVELOPING | 7 | |
| FRUIT_COLORED | 8 | 81 beginning of colouration |
| FRUIT_MATURE | 8 | 89 fully ripe |
| LEAF_COLORING | 9 | 92–93 |
| LEAF_SENESCENT | 9 | |
| LEAF_FALLEN | 9 | 97 end of leaf fall |
| WITHERED_STEMS | 9 | 97, herbaceous above-ground parts dead |
| *Boundary or type-dependent* | | |
| BUD_SWELLING | 0 \| 5 | vegetative vs floral bud |
| BUD_BURST | 0 \| 5 | vegetative vs floral bud |
| LEAF_EMERGING | 0 \| 1 | 09 green tips $\rightarrow$ 10 leaves separating |
| NEW_GROWTH | 1 \| 3 | leaf vs shoot development |
| FRUIT_SET | 6 \| 7 | 69 fruit set visible $\rightarrow$ 71 |
| DEAD_FLOWERS | 6 \| 7 | 69 end of flowering, all flowers dry |
| SEED_HEAD | 7 \| 8 | maturity dependent |
| SEED_POD | 7 \| 8 | maturity dependent |
| FRUIT_DISPERSING | 8 \| 9 | |
| *Unmapped* | | |
| BARK_ONLY | — | visibility artifact, encodes framing not stage |
| LEAF_MATURE | — | persistent state spanning stages 1–8 |

### A.4 Validation of the concept annotations

The 25 concepts are used only to build the fine-grained retrieval evaluation (Table 3). We validate the annotations in the following ways:

**Agreement with iNaturalist flags.** We cross-check all 62,924 annotated test observations against the community phenophase flags. Due to the multi-hot nature of the concepts, we score each flag against the rating mass of its associated concepts and report AUC (Table 7). Agreement is high on every recorded phenophase (mean AUC 0.86). Budding is the hardest stage visually, since buds are small and transient, and is the softest here.

Table 7: **Agreement between concept ratings and iNaturalist phenophase flags** over the 62,924 annotated test observations.

| iNaturalist phenophase | Prevalence | Concept–flag AUC |
|---|---|---|
| No phenology | 0.14 | 0.97 |
| Flowering | 0.63 | 0.89 |
| Fruiting | 0.19 | 0.85 |
| Budding | 0.20 | 0.74 |

**Human agreement study.** Six of the authors independently labeled the 25 concepts on a stratified 200-image sample from the test splits, balanced across the four phenophase classes and boosted so that every concept appears in at least four images. Annotators saw the image and species name but not the iNaturalist flag or the VLM output. The VLM agrees with the annotators about as closely as they agree with one another, and its $\kappa$ against the six-annotator consensus is higher than against any single annotator (Table 8). Absolute $\kappa$ is modest because most concepts are rare, the low-base-rate regime in which high raw agreement coexists with low $\kappa$. Where the VLM disagrees with the annotators, the annotators also disagree with each other, so residual error reflects genuine visual ambiguity rather than annotator failure.

Table 8: **Human agreement study.** 200 images, 25 concepts, six author annotators. Agreement is mean per-concept percent agreement; $\kappa$ is computed over concepts present in at least 10% of the sample, with bootstrap 95% confidence intervals.

| **Comparison** | **Agreement (%)** | $\kappa$ **(95% CI)** |
|---|---|---|
| Human–human (mean over 15 annotator pairs) | 93.0 | 0.236 (0.209–0.264) |
| Human–VLM (mean over 6 annotators) | 91.8 | 0.305 (0.269–0.338) |
| Consensus–VLM (six-annotator majority) | 93.6 | 0.414 |

### A.5 Per-species observation distribution

Like all citizen-science datasets, Pheno3M exhibits a long-tailed species distribution: a small number of charismatic or widely-distributed species account for a disproportionate share of observations, while the majority of species have far fewer. Figure 10 shows the per-species observation count on log-log axes.

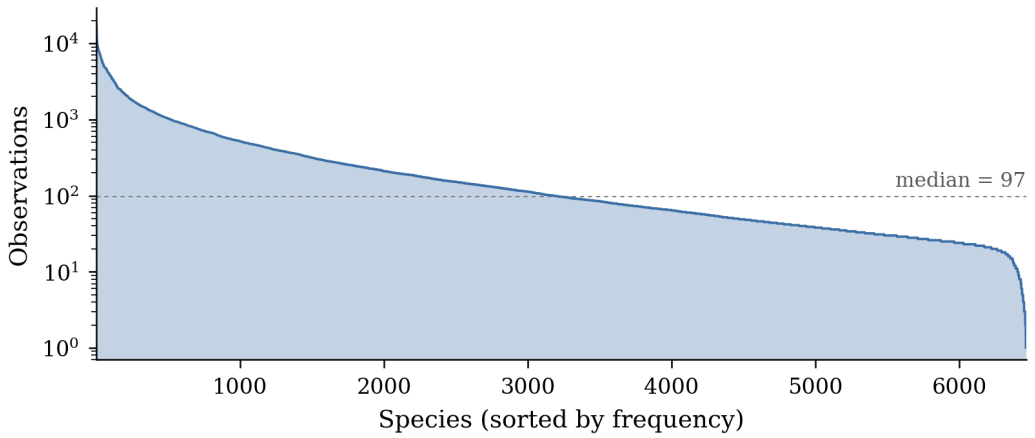



Figure 10: **Species abundance distribution in Pheno3M**. We show species abundance in a log-scale distribution. Similar to other large taxonomic datasets, Pheno3M exhibits a long-tail distribution with regards to species counts.

### A.6 Spatiotemporal context: curation details

**PRISM climate window.** For each observation on date $t$, we extract a 365-day daily climate history from the PRISM gridded climate dataset (Daly et al., 2008) over the window $[t-365, t-1]$, sampled at the nearest 800m grid cell to the observation's coordinates. The day-of-observation itself is excluded from the window to prevent label leakage through same-day weather features.

**AlphaEarth landcover embedding.** We sample a 64-dimensional landcover embedding from AlphaEarth (Brown et al., 2025) at the observation's coordinates, at the model's native 10 m spatial resolution. Embeddings are sourced from Source Cooperative.

### A.7 GENERALIZATION SPLITS: CONSTRUCTION

Both held-out splits are constructed at the deduplicated observation level on the full curated dataset, then frozen and reused across all experiments and baselines.

**Random split.** A uniform random partition with fixed seed.

**Spatial split.** We cluster all observation coordinates with $k$-means ($k = 30$) on $(\text{lat}, \text{lon})$, then hold out 8 of the 30 clusters as the test set. Held-out clusters were selected to balance test-set observation count while spanning multiple biogeographic regions.

**Per-split observation counts.** Table 9 reports the size of each split.

We additionally test phenological retrieval on cross-species galleries (e.g. no imagery of queried species in gallery). Under this regime, the model must recover phenophase from imagery characteristics seperate from the dominant species signal. We observe that approximately 40% of all observations in Pheno3M overlap with TreeOfLife-200M (**?**). To ensure fair comparison against foundation baselines pretrained on TOL-200M, which mirrors a substantial portion of public iNaturalist records, we restrict all evaluation to observations *not* present in TOL-200M. All numbers reported in §5 are computed on this contamination-filtered evaluation set.

Table 9: **Dataset split statistics.** A single training set is used across all evaluations; we report counts for each test protocol.

| Split | Observations | Species |
|---|---|---|
| Train | 2.3M | 6.4K |
| Val | 351K | 6.4K |
| Random test | 178K | 6.2K |
| Spatial test | 178K | 3.3K |

### A.8 USA-NPN ACTIVITY CURVE CONSTRUCTION

In ecological monitoring, dense labeled datasets are sparse: expert and volunteer effort limits how many visits receive complete phenophase reporting, and models are often deployed under partial budgets or narrow regions. We therefore stress-test representations with nested training subsets. Because USA-NPN supervision is attached to repeat visits (taxonomy, place, time, climate, and remote-sensing context) rather than opportunistic photographs, we probe a series of *field-aligned* encoders and baselines.

Observed rates in Figure 7 (main text) are empirical fractions of active visits per calendar week within each latitude band, computed from USA-NPN Status & Intensity rows after the filters in §3. Predicted curves apply the same binning to probe outputs on held-out visits: we aggregate sigmoid probabilities per week and band, smoothing with a trailing three-week moving average for visualization only.

### A.9 TOL-200M CONTAMINATION

We perform photo-level matching between Pheno3M and TreeOfLife-200M (**?**). Of the 214M rows in TOL-200M, 134.9M (63%) carry iNaturalist source URLs spanning 79.7M unique photo IDs. We extract these photo IDs and match them against the photo IDs of every observation in Pheno3M's evaluation splits. **Caveat: photo-level vs. observation-level matching.** A single iNaturalist observation can contain multiple photos, of which TOL-200M may have stored only a subset. Matching at the photo level therefore gives a *lower bound* on observation-level contamination: a TOL-200M-trained model may have seen one photo of an observation in training while we evaluate on a different photo of the same observation. We report photo-level numbers and apply photo-level filtering, but note that the true contamination rate is likely modestly higher than reported.

### A.10 LICENSING AND ACCESS

**Annotations and code.** Our curated annotations (phenophase labels, splits, contamination flags) and data-processing code are released under CC-BY 4.0.

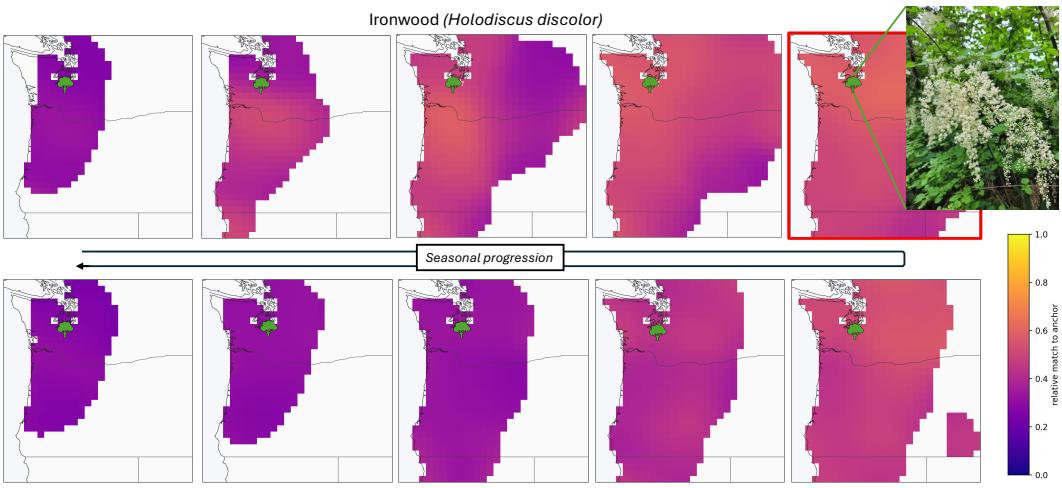



Figure 11: **Image→Field similarity visualization.** Ironwood image→field embedding similarity across shared calendar dates. Similarity peaks where and when the anchor's phenophase is expected for that species, indicating state-conditioned spatiotemporal structure in the embedding. Red borders mark the anchor-date panel.

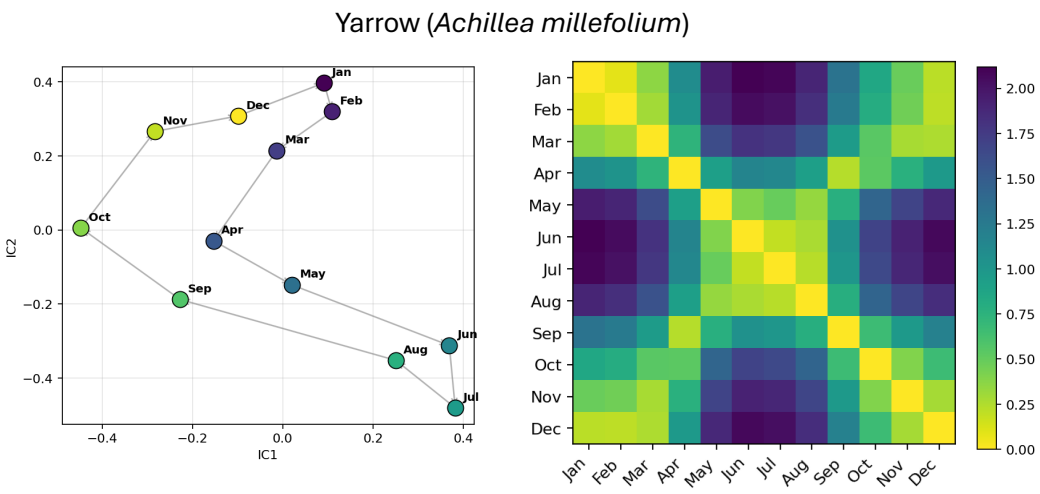



Figure 12: **ICA dynamic time visualization.** ICA decomposition of species Yarrow, CONUS grid monthly average embeddings shows a cyclical nature in-line with changing seasons.

**Photographs.** Photographs are governed by per-contributor licenses set on iNaturalist, ranging from CC0 to CC-BY-NC. We do not redistribute pixel data; users access photographs via the URLs we provide, subject to each photograph's individual license.

**Coordinates.** Geographic coordinates are user-reported. iNaturalist obfuscates the locations of some records for endangered species, and we retain those records at their published precision.

**Underlying data.** PRISM data are distributed by the PRISM Climate Group at Oregon State University and used under their public terms. AlphaEarth embeddings are obtained from Source Cooperative. We release code that reproduces the full feature-construction pipeline from these sources.

Table 10: **Informativeness of Embeddings.** Evaluating the total, mutual and unique information of image and text embeddings using CLUB estimate.

| Model | $H(\mathcal{I})$ | $H(\mathcal{T})$ | $I(\mathcal{I};\mathcal{T})$ | $H(\mathcal{I}) - I(\mathcal{I};\mathcal{T})$ | $H(\mathcal{T}) - I(\mathcal{I};\mathcal{T})$ | $H(\mathcal{I},\mathcal{T})$ |
|---|---|---|---|---|---|---|
| BioCLIP | 27.66 | 32.68 | 24.49 | 3.17 | 8.19 | 35.85 |
| BioCLIP 2 | 26.88 | 27.44 | 21.74 | 5.14 | 5.70 | 32.58 |
| **Ours** | **62.65** | **54.92** | **26.00** | **36.65** | **28.92** | **91.57** |

Table 11: **Intrinsic Dimensionality (ID) of Embeddings**. We compute intrinsic dimension of image and text embeddings using the TwoNN method. Lower ID indicates compact and task-specific representations that are highly generalizable.

| Model | $\mathcal{I}$ | $\mathcal{T}$ |
|---|---|---|
| BioCLIP | 20.37 | 20.46 |
| BioCLIP 2 | 13.45 | 13.48 |
| **Ours** | 10.19 | 10.31 |

## B Qualitative Results

Figure 11 provides qualitative support for the retrieval results by visualizing image-embedding similarity from each anchor across the test set in $\boldsymbol{z}_d^v$. Similarity concentrates in species-appropriate regions and seasons: Oceanspray peaks on Jul 01 over the Puget lowlands and Willamette Valley and weakens into Pacific Northwest winter. Across Mar–Jul, oceanspray also shows a south-to-north progression in high-similarity regions, consistent with latitudinal phenology gradients in bloom timing. The non-overlapping spatial modes and approximately 80-day peak offset indicate state-conditioned spatiotemporal structure rather than a shared seasonal prior.

## C Semantic Embedding Richness

Table 10 evaluates the informativeness of the learned image and text embeddings using entropy and mutual information estimates computed with CLUB (Cheng et al., 2020; Liang et al., 2023). The entropy terms $H(\mathcal{I})$, $H(\mathcal{T})$ and $H(\mathcal{I},\mathcal{T})$ measure the total information content captured by the image and text embeddings, respectively, while the mutual information term $I(\mathcal{I};\mathcal{T})$ quantifies the amount of shared cross-modal information. The residual terms $H(\mathcal{I}) - I(\mathcal{I};\mathcal{T})$ and $H(\mathcal{T}) - I(\mathcal{I};\mathcal{T})$ correspond to unique information retained in each embedding space. Compared to BioCLIP and BioCLIP-2, our method produces substantially higher embedding entropy while also achieving the highest mutual information, indicating that the learned representations preserve richer semantic structure without sacrificing image–text alignment. Importantly, the larger residual information values suggest that the model retains more modality-specific information beyond the shared alignment signal, yielding embeddings that are both highly informative and complementary across modalities.

Table 11 reports the intrinsic dimensionality (ID) of the learned embeddings using the TwoNN (Facco et al., 2017) estimator. Intrinsic dimensionality provides an effective tool to measure representation compactness and redundancy. While BioCLIP and BioCLIP-2 exhibit higher intrinsic dimensionalities, our method achieves significantly lower IDs for both image and text embeddings. This indicates that the learned representations from our model lie on a more compact manifold containing task-specific rich information. This is consistent with other works which found that lower ID corresponds to linearly separable structures, higher task performance, and better generalization (Rao et al., 2025; Ansuini et al., 2019).

## D Formalizing Phenological Bias in InfoNCE

Consider a species-image foundation model trained using the InfoNCE loss. Let $s$ be a species and $\phi$ denote its phenophase (e.g., vegetative, flowering). The training distribution of images for a given

species can be decomposed by phenophase: $p(v|s) = \sum_\phi p(\phi|s)p(v|s,\phi)$, where $v$ is the image embedding and $p(\phi|s)$ is the marginal probability of observing a specific phenophase in the dataset.

**Standard InfoNCE.** In practice, InfoNCE optimizes contrastive alignment over a finite batch rather than the global population. However, under uniform random sampling, the empirical distribution of phenophases within any randomly sampled batch $\mathcal{B}_s$ for species $s$ reflects the global training distribution in expectation. Thus, at the population limit, minimizing the InfoNCE loss is equivalent to maximizing the mutual information $I(V;S)$. The optimal text embedding $t_s$ for species $s$ satisfies:

$$t_s \propto \mathbb{E}_{p(v|s)}[v] = \sum_\phi p(\phi|s)\mathbb{E}[v|s,\phi]$$

This demonstrates that the ideal text embedding $t_s$ is a convex combination of the phase-conditional image centroids $\mu_{s,\phi} = \mathbb{E}[v|s,\phi]$. If $p(\phi|s)$ is highly skewed in the training data (e.g., $p(\text{flowering}|s) \gg 0.25$), the batch will consistently be dominated by that phenophase. Consequently, $t_s$ will be pulled significantly closer to $\mu_{s,\text{flowering}}$ than to the centroids of underrepresented phases, inducing a systematic phenological bias.

**Hierarchical InfoNCE.** BioCLIP and BioCLIP 2 use hierarchical InfoNCE, which encodes hierarchy by passing a flattened taxonomic string (from Kingdom down to Species) through an autoregressive text encoder, or by randomly sampling from a mixture of text types (taxonomic, scientific, and common names) during training (**?**Stevens et al., 2024). Let $T_s$ be the set of valid hierarchical text descriptions for species $s$. For a given batch, a text embedding $t \sim T_s$ is paired with an image $v \sim p(v|s)$.

While this hierarchical text strategy robustly regularizes the text embedding space and acts as a strong prior for zero-shot generalization to unseen taxa, it does not alter the underlying image sampling distribution $p(v|s)$. The expected gradient updating any sampled hierarchical text representation $t \in T_s$ is still strictly governed by the weighted sum $\sum_\phi p(\phi|s)\mu_{s,\phi}$. Therefore, while higher-order taxonomic tokens help position the species globally within the tree of life, the local fine-grained alignment remains inherently biased toward the phenophases most frequently sampled into the batch.

## E Implementation Details

**Compute and resources.** All PhenoMap models are trained on a single GPU (NVIDIA A100, 80 GB) in bfloat16 mixed precision, batch size 8,192, with the AdamW optimizer (lr $5 \times 10^{-4}$, weight decay $10^{-4}$, gradient clipping at 1.0).

**Architecture Details.** We provide architecture details for our small, medium, and large model in Table 12.

**Field Branch Details.** As detailed in Figure 13, the PhenoMap field branch produces the static field embedding $\boldsymbol{z}_s^f$ by passing a frozen BioCLIP 2 text embedding of species text through a learnable MLP. For the dynamic field embedding, 365-day PRISM climate windows are tokenized into 52 weekly tokens through 1D convolution with kernel size 7 and stride=7. These tokens, along with the spatiotemporally localized AlphaEarth embeddings are linearly projected into the hidden dimension of the field encoder. These tokens are concatenated with a learned CLS token and processed by Transformer blocks. Simultaneously, a frozen BioCLIP-2 species embedding modulates these layers via AdaLN-Zero to output the final dynamic field embedding, $\boldsymbol{z}_d^f$.

**Image Decoder Details.** To prevent representation collapse, we reconstruct the frozen image embedding $\boldsymbol{z}$ from the factorized embeddings ($\boldsymbol{z}_s$, $\boldsymbol{z}_d$). This is achieved using a learned 2-layer transformer decoder with self-attention, where $\boldsymbol{z}_s$ and $\boldsymbol{z}_d$ are treated as input tokens.

**Cyclic phenophase weighting.** Phases are indexed on the four-cycle vegetative $\rightarrow$ budding $\rightarrow$ flowering $\rightarrow$ fruiting $\rightarrow$ vegetative, and $\delta(a,b) = \min(|a-b|,\, 4-|a-b|)$ is their cyclic distance.

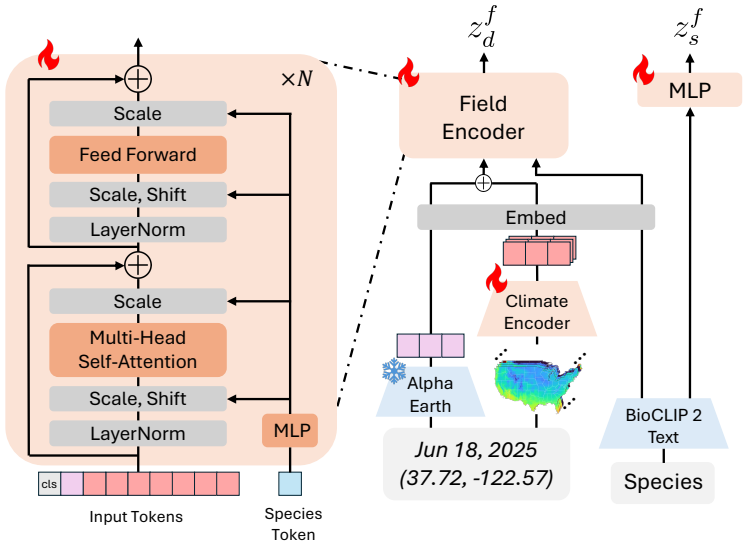



Figure 13: **Field Branch Architecture.**

For anchor $i$ and instance $k$ with phases $\phi_i$ and $\phi_k$, the weight on the negative pool in Eq. 3 is

$$W_{ik} = \begin{cases} 0 & \delta(\phi_i, \phi_k) = 0 \quad \text{(identical phase, masked)} \\ 0.8 & \delta(\phi_i, \phi_k) = 1 \quad \text{(adjacent)} \\ 1 & \delta(\phi_i, \phi_k) = 2 \quad \text{(opposite)} \end{cases} \qquad W_{i,p(i)} = 1. \tag{4}$$

For multi-hot labels, $\delta$ is the minimum cyclic distance over all pairs of active phases, so instances sharing any phase are masked. It is defined on phenophase rather than day-of-year because the same date maps to different phenophases across species, latitudes, and climates. Related supervised and ordinal contrastive losses (SupCon (Khosla et al., 2020), Rank-N-Contrast (Zha et al., 2023)) are compared in Table 13.

Table 12: PhenoMap scale variants. Static and dynamic latent dimensions sum to the total field embedding dimension. Trainable parameter counts exclude the frozen BioCLIP 2 image and species text encoders. Steps and wall-clock are measured under our species-stratified sampler with a per-species cap.

| **Component** | **PhenoMap-S** | **PhenoMap-M** | **PhenoMap-L** |
|---|---|---|---|
| Backbone (frozen) | BioCLIP 2 | BioCLIP 2 | BioCLIP 2 |
| Image / species emb dim | 768 | 768 | 768 |
| Static latent dim ($z_s$) | 192 | 384 | 576 |
| Dynamic latent dim ($z_d$) | 64 | 128 | 192 |
| Total field embedding dim | 256 | 512 | 768 |
| Field encoder hidden dim | 128 | 256 | 512 |
| Field encoder layers / heads | 2 / 4 | 3 / 4 | 4 / 4 |
| Decoder hidden / layers / heads | 256 / 2 / 4 | 512 / 2 / 4 | 768 / 2 / 4 |
| Trainable parameters | 8.7M | 22M | 42.3M |
| Optimizer steps | 15k | 15k | 30k |

## F Additional Ablation Studies

**Static-Dynamic Dimension Ratio.** We investigate the impact of the hidden dimension allocation between the static and dynamic components. Our baseline Small PhenoMap ($d = 256$) model utilizes a 3:1 ratio (192 for the static factor, 64 for the dynamic factor). We ablate this against 1:1 (128, 128) and 6:1 (224, 32) dimension ratios. As shown in Table 13, our 3:1 baseline achieves slight

Table 13: **More PhenoMap Ablations.** We evaluate based on matched-pair retrieval at $\alpha = 0.9$. We use Small PhenoMap ($d = 256$) for ablations, trained for 15k steps. 1:1 is (128, 128) and 6:1 is (224, 32) hidden dimension ratio for (static, dynamic) factors.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| PhenoMap-S | 26.2 | 49.6 | 54.3 | 25.4 | 47.4 | 51.9 | 27.3 | 49.5 | 53.0 | 27.9 | 50.0 | 53.5 |
| *Static-Dynamic Dimension Ratio* | | | | | | | | | | | | |
| 1:1 | 25.9 | 49.3 | 53.9 | 25.0 | 47.1 | 51.4 | 27.2 | 49.3 | 52.8 | 27.7 | 49.8 | 53.3 |
| 6:1 | 26.0 | 49.5 | 54.2 | 24.9 | 47.3 | 51.6 | 27.2 | 49.4 | 52.9 | 27.8 | 49.7 | 53.4 |
| *Losses* | | | | | | | | | | | | |
| w/ Dynamic InfoNCE | 25.1 | 47.8 | 52.5 | 24.5 | 46.2 | 50.7 | 27.0 | 48.9 | 52.4 | 27.6 | 49.3 | 53.0 |
| w/ Dynamic SupCon | 23.1 | 44.5 | 49.2 | 22.5 | 43.1 | 47.6 | 28.0 | 50.7 | 54.1 | 28.5 | 51.3 | 54.8 |
| w/ Dynamic Cyclic SupCon | 24.0 | 46.0 | 50.7 | 23.4 | 44.2 | 48.8 | 27.3 | 49.4 | 52.9 | 28.0 | 50.0 | 53.6 |
| w/ Dynamic RnC | 25.3 | 48.6 | 52.8 | 24.7 | 47.0 | 51.1 | 27.7 | 50.5 | 53.7 | 28.4 | 51.2 | 54.5 |
| w/o reconstruction | 23.5 | 44.1 | 48.0 | 22.2 | 41.7 | 44.7 | 24.6 | 44.1 | 46.9 | 24.5 | 43.4 | 46.1 |

performance gains over the alternative configurations. Based on these results, we adopt the 3:1 ratio for all subsequent scaling to larger models.

**Losses.** We also evaluate various contrastive loss formulations for learning the dynamic representations. We compare our approach against Dynamic InfoNCE, a standard instance-discrimination contrastive loss, as well as Dynamic SupCon (Khosla et al., 2020), a supervised contrastive approach where representations of the same phenophase are weighted as positives and differing states are treated as negatives. Additionally, we test Dynamic Cyclic SupCon, a variation of SupCon where negative samples are weighted according to their cyclical phenological distance from the anchor, and Dynamic RnC, an adaptation of Rank-N-Contrast (Zha et al., 2023) that utilizes the ordinal phenophase as continuous ranking supervision. While Dynamic SupCon and Dynamic RnC yield stronger Image → Image retrieval performance, our baseline formulation outperforms the alternative losses on Field → Image retrieval. The cyclic weighting matters because the four phases form a cycle rather than a line, whereas a linear or day-of-year ordering would place nearing phenophases maximally far apart (§4.2.1). We opt for our approach because optimizing cross-modal representations is our primary objective, and our method secures larger gains in the Field → Image setting. Furthermore, we demonstrate that removing the reconstruction loss causes a noticeable drop in performance across all metrics, highlighting its necessity in preventing representation collapse.

**LoRA Fine-tuning.** To determine if tuning the vision encoder's weights improves representation learning, we apply LoRA finetuning to the BioCLIP 2 ViT backbone. Due to computational constraints, this necessitates lowering the training batch size to 256. For a direct comparison, we train a PhenoMap-S baseline at this identical, lower batch size. As observed in Table 14, while LoRA finetuning provides marginal gains in Field → Image retrieval over the low-batch baseline, it ultimately degrades the visual representations, resulting in worsened Image → Image retrieval. More importantly, both low-batch configurations are still outperformed by our standard high-batch ($B = 8192$) PhenoMap-S.

**Model Scaling.** Finally, we scale the capacity of the PhenoMap model by increasing the total embedding dimension $d$. We evaluate PhenoMap-S ($d = 256$), PhenoMap-M ($d = 512$), and PhenoMap-L ($d = 768$). As demonstrated in Table 15, scaling the embedding dimension yields steady, consistent improvements across both retrieval tasks and corresponding metrics.

# G Additional Experiments

## G.1 Day-of-year linear probe

To probe whether PhenoMap's dynamic factor $\boldsymbol{z}_d$ encodes phenophase directly, we fit a post-hoc probe from frozen image features to day-of-year (DOY). DOY targets are encoded cyclically as $(\sin(2\pi\,\mathrm{DOY}/365), \cos(2\pi\,\mathrm{DOY}/365))$ to handle the year-boundary discontinuity, and the regression is solved in closed form via ridge-stabilized normal equations on standardized features. Predicted

Table 14: **LoRA Finetuning BioCLIP 2 ViT.** Due to computational cost, we lower the batch size with LoRA Finetuning. For direct comparison, we also evaluate PhenoMap-S with the same batch size.

| Method | Batch size | Field → Image Retrieval | | | | | | Image → Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Random | | | Spatial | | | Random | | | Spatial | | |
| | | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| PhenoMap-S | 8192 | 26.2 | 49.6 | 54.3 | 25.4 | 47.4 | 51.9 | 27.3 | 49.5 | 53.0 | 27.9 | 50.0 | 53.5 |
| PhenoMap-S | 256 | 24.2 | 46.3 | 50.5 | 23.5 | 44.4 | 48.7 | 26.6 | 47.8 | 51.1 | 27.0 | 48.4 | 51.7 |
| PhenoMap-S w/ LoRA | 256 | 25.0 | 46.8 | 50.9 | 24.3 | 45.1 | 49.0 | 26.0 | 46.8 | 50.1 | 26.6 | 47.2 | 50.5 |

Table 15: **PhenoMap size matters.** Scaling PhenoMap to larger embedding dimensions improves compositional retrieval.

| Method | Dim | Field → Image Retrieval | | | | | | Image → Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Random | | | Spatial | | | Random | | | Spatial | | |
| | | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| PhenoMap-S | 256 | 26.2 | 49.6 | 54.3 | 25.4 | 47.4 | 51.9 | 27.3 | 49.5 | 53.0 | 27.9 | 50.0 | 53.5 |
| PhenoMap-M | 512 | 27.7 | 52.8 | 57.7 | 26.8 | 50.3 | 55.0 | 27.8 | 50.4 | 54.1 | 28.2 | 50.4 | 54.2 |
| PhenoMap-L | 768 | **29.2** | **55.1** | **60.3** | **27.7** | **52.0** | **56.7** | **28.6** | **52.0** | **56.1** | **28.4** | **51.0** | **54.9** |

DOY is recovered with $\arctan_2$ and error is reported as the cyclic absolute difference, capped at $182.5$ days.

Table 16: **Day-of-year linear probe MAE (days, lower is better).** *Overall* is the micro-average across all test samples. *Macro* averages the per-phase MAEs across the four phenophases, equally weighted, and is robust to class imbalance. *Balanced* subsamples the test set to a cap of $10$ per species$\times$phase cell. PhenoMap's $\boldsymbol{z}_d$ uses $192$ dimensions, $4\times$ fewer than BioCLIP-2's image features.

| Model | Dim | Unbalanced | | | | Balanced | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Random | | Spatial | | Random | | Spatial | |
| | | Overall | Macro | Overall | Macro | Overall | Macro | Overall | Macro |
| BioCLIP | 512 | 39.10 | 41.27 | 46.56 | 47.41 | 43.59 | 45.07 | 47.06 | 48.32 |
| BioCLIP-2 | 768 | 32.66 | 35.37 | 39.62 | 41.04 | 36.82 | 38.70 | 40.15 | 41.86 |
| PhenoMap $\boldsymbol{z}_d$ | 192 | **28.28** | **30.30** | **35.40** | **36.14** | **31.79** | **33.18** | **35.74** | **36.99** |

Table 17: **Per-phase DOY MAE (days) on the unbalanced test set.** Phases are vegetative, budding, flowering, fruiting. Vegetative imagery carries no visible phenology cue and is the hardest class for every model; PhenoMap retains its margin in every phase.

| Model | Dim | Random | | | | Spatial | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Veg. | Bud. | Flow. | Fruit | Veg. | Bud. | Flow. | Fruit |
| BioCLIP | 512 | 49.62 | 34.61 | 35.99 | 44.86 | 54.32 | 41.19 | 44.60 | 49.53 |
| BioCLIP-2 | 768 | 45.53 | 28.44 | 28.87 | 38.64 | 50.41 | 34.58 | 36.52 | 42.63 |
| PhenoMap $\boldsymbol{z}_d$ | 192 | **39.22** | **24.22** | **25.48** | **32.30** | **44.15** | **30.55** | **33.55** | **36.29** |

Table 18: **Pheno3M 4-class phenophase retrieval (NDCG@5).** Field→image and image→image ranking on the random and spatial splits. *Same* restricts the gallery to same-species images, *All* uses the full pool, and *Cross* restricts to cross-species images. Relevance uses iNat's 4-class label (no_phenology, budding, flowering, fruiting). Best per column in bold.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* |
| TaxaBind (Sastry et al., 2025b) | 0.433 | 0.373 | 0.334 | 0.508 | 0.458 | 0.443 | 0.609 | 0.501 | 0.505 | 0.761 | 0.641 | 0.672 |
| RCME (Sastry et al., 2025a) | 0.436 | 0.371 | 0.344 | 0.512 | 0.459 | 0.429 | 0.624 | 0.508 | 0.522 | 0.770 | 0.645 | 0.677 |
| ProM3E (Sastry et al., 2026) | 0.452 | 0.347 | 0.266 | 0.545 | 0.429 | 0.342 | 0.565 | 0.464 | 0.447 | 0.724 | 0.593 | 0.602 |
| BioTrove-CLIP (Yang et al., 2024) | 0.418 | 0.354 | 0.315 | 0.513 | 0.430 | 0.400 | 0.610 | 0.487 | 0.473 | 0.759 | 0.623 | 0.623 |
| BioCLIP (Stevens et al., 2024) | 0.447 | 0.373 | 0.336 | 0.511 | 0.452 | 0.428 | 0.621 | 0.504 | 0.513 | 0.766 | 0.647 | 0.663 |
| BioCLIP 2 (**?**) | 0.438 | 0.374 | 0.335 | 0.527 | 0.470 | 0.445 | 0.631 | 0.498 | 0.494 | 0.781 | 0.634 | 0.656 |
| BioCLIP 2–FT (**?**) | 0.457 | 0.545 | 0.365 | 0.550 | 0.659 | 0.472 | 0.596 | 0.593 | 0.410 | 0.734 | 0.728 | 0.529 |
| BioCLIP 2–Sup (**?**) | 0.440 | 0.380 | 0.335 | 0.525 | 0.470 | 0.445 | **0.690** | 0.535 | 0.505 | 0.801 | 0.665 | 0.660 |
| PhenoMap $z_d$ | **0.597** | **0.622** | **0.530** | **0.684** | **0.712** | **0.623** | 0.688 | **0.675** | **0.618** | **0.808** | **0.796** | **0.750** |

Table 19: **Factorized channel ablation.** The static channel $z_s$ is supervised toward species identity, the dynamic channel $z_d$ toward phenology, and late fusion combines the similarity scores from both ($\alpha = 0.9$ weight on $z_d$). $z_d$ is strongest on phenophase NDCG@5; late fusion is strongest on compositional matched-pair retrieval, which requires matching both species and phenophase. Best per column in bold within each block.

| | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| *Phenophase NDCG@5 (Pheno3M 4-class)* | | | | | | | | | | | | |
| | Same | All | Cross | Same | All | Cross | Same | All | Cross | Same | All | Cross |
| $z_s$ (static) | 0.430 | 0.528 | 0.331 | 0.528 | 0.642 | 0.452 | 0.587 | 0.586 | 0.406 | 0.734 | 0.720 | 0.543 |
| $z_d$ (dynamic) | **0.597** | **0.622** | **0.530** | **0.684** | **0.712** | **0.623** | **0.688** | **0.675** | **0.618** | **0.808** | **0.796** | **0.750** |
| Late $\alpha = 0.9$ | 0.574 | 0.597 | 0.426 | 0.666 | 0.712 | 0.548 | 0.654 | 0.631 | 0.506 | 0.788 | 0.765 | 0.650 |
| *Compositional matched-pair retrieval* | | | | | | | | | | | | |
| | R@5 | R@10 | mAP | R@5 | R@10 | mAP | R@5 | R@10 | mAP | R@5 | R@10 | mAP |
| $z_s$ (static) | 21.2 | 43.3 | 48.3 | 21.5 | 42.4 | 46.7 | 26.0 | 47.7 | 51.9 | 25.8 | 46.7 | 50.4 |
| $z_d$ (dynamic) | 17.2 | 26.4 | 23.9 | 16.1 | 25.2 | 23.1 | 15.7 | 23.7 | 21.2 | 16.4 | 25.2 | 22.8 |
| Late $\alpha = 0.9$ | **29.2** | **55.1** | **60.3** | **27.7** | **52.0** | **56.7** | **28.6** | **52.0** | **56.1** | **28.4** | **51.0** | **54.9** |

PhenoMap's 192-d dynamic factor achieves the lowest MAE on every split and sampling condition (Table 16), beating BioCLIP-2's 768-d image features by 4.4 days on the random split and 4.2 days on the spatial split. The advantage holds in the balanced setting, where each species×phase cell contributes equally to test-time error, ruling out class imbalance as an explanation. Per-phase errors (Table 17) follow the expected pattern. The vegetative class is the hardest for every model because the imagery contains no visible phenology cue, and budding and flowering are the easiest because the visual signal is sharpest. PhenoMap preserves its lead in every phase.

The result complements the entanglement diagnosis in §5.1. The same factorization that produces an unbiased species channel concentrates phenological structure in $z_d$ tightly enough that a linear map recovers the date of capture more accurately than 4× the feature dimension of a strong general-purpose plant foundation model.

### G.2 Metadata and self-supervised baselines

Tables 20 and 21 extend the main-text comparison. ⊕ **fields** tests whether PhenoMap's gain comes from access to time, location, climate, and satellite metadata rather than from factorization. BioCLIP 2 receives the same inputs through the same field encoder as PhenoMap, the static and dynamic factors are concatenated on both the image and field sides, and a single InfoNCE loss aligns them. All ⊕ fields variants are trained for 15k steps in the PhenoMap-S setting of Table 4. We also compare with **DINOv3** (Siméoni et al., 2025) which tests whether phenological signal is already present in generic self-supervised features, frozen and with a phenophase-supervised head. It has no text or field encoder, so field→image cells are marked –.

**Metadata alone does not close the gap.** On matched-pair retrieval, frozen ⊕ fields falls far short (21.8 vs. 29.2 R@5 on Random field→image) and even degrades mAP relative to plain frozen BioCLIP 2 (37.7 vs. 44.7). Appending metadata entangles species and phenology and factorization recovers the joint structure.

**Tuning the backbone does not substitute for factorization.** Finetuned ⊕ fields improves on frozen ⊕ fields in field→image matched-pair retrieval and supervised ⊕ fields improves further, but both remain well short of PhenoMap and their image→image gains are modest. We hypothesize that tuning the BioCLIP 2 ViT collapses the visual embedding around the coarser metadata target, making it less discriminative within a phase.

**Generic features encode phenophase but not its composition with species.** Supervised DINOv3 is the strongest method on fine-grained image→image retrieval, but our main capability is to capture the composition of phenophase with species, on which it is weakest. On matched-pair retrieval, frozen DINOv3 reaches only 18.6 R@5 (image→image, Random) against PhenoMap's 28.6, and supervision makes it worse (12.0). DINOv3 also cannot be queried by species and context.

Table 20: **Matched-pair retrieval (species ∧ phenophase) for metadata and self-supervised baselines.** Same protocol and columns as Table 2. ⊕ fields: BioCLIP 2 with the same field inputs as PhenoMap, concatenated rather than factorized (see text). DINOv3 has no field encoder, so field→image cells are marked –. Best per column in bold.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ | $R@5$ | $R@10$ | $mAP$ |
| BioCLIP 2 (**?**) | | | | | | | | | | | | |
| frozen | 20.2 | 39.4 | 44.7 | 19.9 | 39.2 | 43.9 | 27.2 | 48.8 | 51.5 | 27.9 | 49.6 | 52.3 |
| finetuned | 23.2 | 44.5 | 49.3 | 20.8 | 39.2 | 43.7 | 25.6 | 46.8 | 51.5 | 25.0 | 45.0 | 48.9 |
| supervised | 22.5 | 43.0 | 47.8 | 20.1 | 38.0 | 42.5 | 27.4 | 50.3 | 54.2 | 27.0 | 48.3 | 52.6 |
| frozen ⊕ fields | 21.8 | 37.7 | 37.7 | 21.9 | 38.0 | 38.6 | 24.4 | 42.1 | 42.7 | 25.2 | 43.2 | 44.5 |
| finetuned ⊕ fields | 23.7 | 43.3 | 46.5 | 23.2 | 42.6 | 45.8 | 25.6 | 45.5 | 48.2 | 26.4 | 46.7 | 49.7 |
| supervised ⊕ fields | 24.3 | 44.5 | 47.8 | 24.2 | 43.9 | 47.2 | 25.7 | 45.9 | 48.3 | 26.5 | 46.7 | 49.5 |
| DINOv3 (Siméoni et al., 2025) | | | | | | | | | | | | |
| frozen | – | – | – | – | – | – | 18.6 | 29.9 | 28.4 | 21.5 | 34.4 | 33.7 |
| supervised | – | – | – | – | – | – | 12.0 | 17.8 | 15.2 | 14.6 | 21.7 | 19.4 |
| PhenoMap | **29.2** | **55.1** | **60.3** | **27.7** | **52.0** | **56.7** | **28.6** | **52.0** | **56.1** | **28.4** | **51.0** | **54.9** |

Table 21: **Fine-grained phenological retrieval (25 concepts, NDCG@5) for metadata and self-supervised baselines.** Same protocol and columns as Table 3; ⊕ fields and – as in Table 20. Best per column in bold.

| Method | Field→Image Retrieval | | | | | | Image→Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Spatial | | | Random | | | Spatial | | |
| | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* | *Same* | *All* | *Cross* |
| BioCLIP 2 (**?**) | | | | | | | | | | | | |
| frozen | 0.623 | 0.538 | 0.385 | 0.593 | 0.519 | 0.407 | 0.765 | 0.641 | 0.514 | 0.754 | 0.638 | 0.540 |
| finetuned | 0.641 | 0.637 | 0.405 | 0.614 | 0.629 | 0.421 | 0.738 | 0.667 | 0.433 | 0.721 | 0.669 | 0.458 |
| supervised | 0.640 | 0.635 | 0.402 | 0.612 | 0.625 | 0.418 | 0.762 | 0.688 | 0.460 | 0.745 | 0.682 | 0.475 |
| frozen ⊕ fields | 0.733 | 0.660 | 0.483 | **0.705** | 0.648 | 0.491 | 0.776 | 0.706 | 0.545 | 0.760 | 0.704 | 0.563 |
| finetuned ⊕ fields | 0.721 | 0.587 | 0.421 | 0.694 | 0.569 | 0.432 | 0.748 | 0.637 | 0.455 | 0.723 | 0.629 | 0.474 |
| supervised ⊕ fields | 0.730 | 0.579 | 0.429 | 0.699 | 0.549 | 0.432 | 0.765 | 0.668 | 0.502 | 0.745 | 0.668 | 0.523 |
| DINOv3 (Siméoni et al., 2025) | | | | | | | | | | | | |
| frozen | – | – | – | – | – | – | 0.799 | 0.692 | 0.603 | 0.787 | 0.688 | 0.618 |
| supervised | – | – | – | – | – | – | **0.810** | **0.760** | **0.644** | **0.796** | **0.756** | **0.655** |
| PhenoMap $z_d$ | **0.740** | **0.671** | **0.502** | 0.697 | **0.654** | **0.498** | 0.776 | 0.702 | 0.545 | 0.761 | 0.696 | 0.558 |

### G.3 Retrieval error analysis

Matched-pair retrieval is demanding: a gallery item is relevant only when species and phenophase both match, and the gallery is balanced across cells to remove the flowering skew of Table 1, so absolute recall is modest for every model. Table 22 groups the phenophase of PhenoMap's top-10 field→image retrievals by query phenophase on the random split. The main confusion is between adjacent stages, and distant-stage confusions are rare.

Table 22: **Phenophase of PhenoMap top-10 field→image retrievals by query phenophase** (random split, row-normalized).

| Query ↓ / Retrieved → | Vegetative | Budding | Flowering | Fruiting |
|---|---|---|---|---|
| Vegetative | 0.62 | 0.10 | 0.10 | 0.18 |
| Budding | 0.07 | 0.42 | 0.39 | 0.12 |
| Flowering | 0.03 | 0.23 | 0.62 | 0.11 |
| Fruiting | 0.14 | 0.10 | 0.17 | 0.59 |

## H LIMITATIONS

Pheno3M currently focuses on the contiguous United States, providing a strong foundation for large-scale phenological modeling. Expanding the framework globally with additional imagery and climate data offers an exciting direction for future work and would enable broader ecological coverage across diverse ecosystems. Beyond phenology, the dual-axis factorization framework could naturally be extended to capture other within-species dimensions such as habitat, microhabitat, and ecological context. The longitudinal structure of Pheno3M also creates a promising pathway toward phenological forecasting.

While citizen-science datasets inevitably contain geographic, taxonomic, and observer-driven biases, Pheno3M explicitly addresses important entanglement effects through the species-channel debiasing strategy described in § 5.1. Acquisition-level variability cannot be controlled at collection time, so fine-grained cues are not uniformly visible; recent work has shown that ecological models remain robust to such noise at the aggregate level even as rare and visually similar cases become harder (Bevan et al., 2026).