

ProM3E: Probabilistic Masked MultiModal Embedding Model for Ecology

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Jiayu Lin, Dan Cher,
 Phoenix Jarosz, Nathan Jacobs
 Washington University in St. Louis

{s.sastry, k.subash, a.dhakal, jiayu.lin, cher, jarosz, jacobsn}@wustl.edu

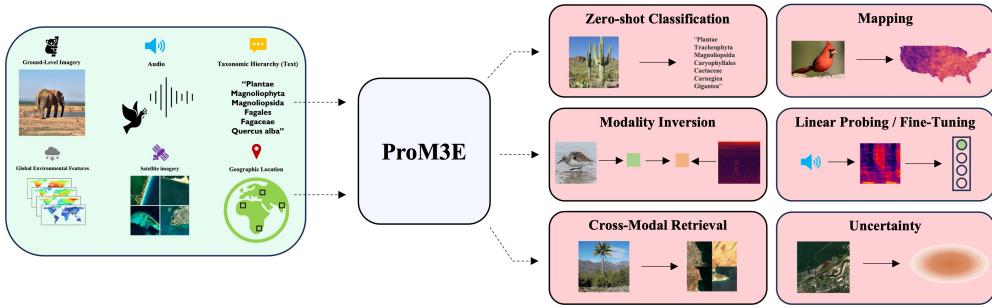


Figure 1. **ProM3E Overview.** The versatile capabilities of our model with the ability to accept arbitrary input modalities.

Abstract

We introduce *ProM3E*, a probabilistic masked multimodal embedding model for any-to-any generation of multimodal representations for ecology. *ProM3E* is based on masked modality reconstruction in the embedding space, learning to infer missing modalities given a few context modalities. By design, our model supports modality inversion in the embedding space. The probabilistic nature of our model allows us to analyse the feasibility of fusing various modalities for given downstream tasks, essentially learning what to fuse. Using these features of our model, we propose a novel cross-modal retrieval approach that mixes inter-modal and intra-modal similarities to achieve superior performance across all retrieval tasks. We further leverage the hidden representation from our model to perform linear probing tasks and demonstrate the superior representation learning capability of our model. All our code, datasets and model will be released at <https://vishu26.github.io/prom3e>.

1. Introduction

In the past decade, the field of multimodal* learning has undergone significant advancements, enabling the development

of generalist frameworks [20, 59] capable of solving a wide range of tasks. This progress led to the development of domain-specific multimodal models such as in remote sensing [30, 46, 76], ecology [11, 54] or medicine [35, 61]. Although such models are versatile, they suffer from either of these two limitations: 1) assume some/all modalities are available during inference; and 2) cannot infer missing modalities. These limitations lead to the development of "Any-to-Any" models [5, 41, 49, 57, 74] that can generate any modality given a few modalities in the input. Recently, in remote sensing, such any-to-any models [3, 63] have been created. These models are trained using a student-teacher framework similar to the joint-embedding predictive architecture (JEPA) [2].

However, such any-to-any models are trained using massive amounts of "paired" data which are difficult to acquire with growing number of modalities. To address this, Mizrahi et al. [49] used powerful off-the-shelf models to synthesize paired data given RGB images. Yet, in domains like remote sensing and medicine, certain modalities, such as hyperspectral or MRI imagery, are challenging to acquire or even synthesize. This challenge becomes more difficult when dealing with multimodal data without one-to-one correspondence. For instance, in remote sensing, a single satellite image can correspond to multiple ground-level images.

In this paper, we propose *ProM3E*, a probabilistic masked multimodal embedding model to address the previously discussed challenges with any-to-any models. We design

*With a slight abuse of notation, we use the term "multimodal learning" to refer to frameworks learning from more than two modalities (unless specified). This excludes bimodal models such as CLIP.

ProM3E as a two-stage framework with a focus on optimizing and scaling paired multimodal data required for training any-to-any models. This is done by aligning all representations before fusing them. First, we obtain modality-specific encoders using imagebind-style training on massive-scale image-paired datasets. Then, we train a lightweight multimodal embedding-based masked variational autoencoder (MVAE) by incorporating embeddings obtained from freezing the encoders. Since, the modalities are already aligned, we require only small-scale paired data of all modalities for training the MVAE. After our model is trained, we analyze various aspects of our model, including the uncertainty captured by our model and the modality gap present in various modalities. As our model supports modality inversion, we propose a novel retrieval strategy combining inter-modal and intra-modal similarities.

The contributions of our work are fourfold:

1. **ProM3E:** We introduce a framework for any-to-any generation of representations. Our model learns a joint probability distribution over input modalities, which is then used to reconstruct the embeddings of unavailable modalities.
2. **Modality Inversion:** We introduce a novel cross-modal retrieval strategy that combines the benefits of intra-modal and inter-modal interactions of the modalities using the modality inversion feature of our model.
3. **Uncertainty:** We present an extensive qualitative and quantitative analysis of the uncertainty captured by our model to identify the most informative modalities and determine if combining multiple modalities reduces uncertainty in the model.
4. **Modality Gap:** We present an analysis of the modality gap present in the modalities before and after training. Furthermore, we analyse whether modality gap is related to the uncertainty captured by our model.

2. Related Works

Multimodal Learning. Recent multimodal learning approaches aim to align diverse modalities—such as image, audio, text, and tactile signals—into shared embedding spaces using self-supervised contrastive learning and modality-specific encoders [1, 18, 20, 21, 36, 68, 71, 79, 82]. Some frameworks leverage pre-trained encoders with learnable routers [72] or share transformers to enable flexible fusion under task-specific supervision [37, 58, 59]. More recent approaches [22, 42, 77] utilize LLMs to unify modalities via generated textual anchors or multimodal reasoning. In parallel, multimodal frameworks for remote sensing [13, 14, 29, 30, 46, 63, 70] have emerged to integrate modalities such as satellite imagery, text or audio to learn geospatially and semantically rich representations. More recent works focus on understanding and mitigating modality gap [39]—a persistent separation between modalities in multi-modal representation spaces—attributing it to factors such as initialization [39], contrastive loss dynamics [39, 56] and modality imbalance [24, 48, 55].

Multimodal Learning for Ecology. Recent advances in multimodal learning for ecology have been driven by the availability of data through crowdsourced platforms [65] and structured benchmarks [54, 66] which provide data across multiple modalities vital for ecological tasks such as species distribution modeling (SDM) and species fine-grained visual classification (FGVC). New research has moved beyond bimodal frameworks [16, 25, 60, 78, 81] to richer models that integrate additional available modalities [11, 53, 54, 80]. All such models are provably general-purpose ecological predictors that can be utilized to address a diverse array of tasks in remote sensing and ecology.

Masking-based Learning. Masking the input signal and training a model to reconstruct the masked signal has proven to be an effective strategy for pretraining deep-learning models. Several works in the natural language domain mask input word tokens and learn a model to predict those missing tokens [12, 34, 40]. Inspired by the success of this approach, this strategy was extended to the visual domain [23, 62, 73, 75]. He et al. [23] was the earliest work that used a high masking ratio and train a vision transformer to predict the masked patches. Recent works [4, 49, 69] extended masked modeling to accept multimodal inputs. These works reconstruct missing signals from other modalities using signal from one modality, creating powerful multimodal foundational models.

Probabilistic Representation Learning. Recent works on probabilistic multimodal representations aim to capture uncertainty and enhance alignment by projecting inputs as distributions rather than point vectors. Variational Autoencoders (VAEs) introduced foundational techniques for learning latent probabilistic spaces through variational inference and the reparameterization trick [32]. In the vision-language domain, methods such as PCME [7] and PCME++ [6, 8] represent image-text pairs as Gaussian distributions and learn cross-modal similarities using Monte Carlo sampling [7] or closed-form approximations [6, 8]. Other approaches [27, 50] extend probabilistic modeling with distribution-aware objectives for contrastive, matching, or compositional learning [51, 64].

3. Method

Our overall framework is depicted in Figure 2. We design ProM3E as a two stage framework to make it data efficient, scalable and flexible in terms of modalities it can consume. Our method is based on reconstructing global embeddings of modalities since there does not exist

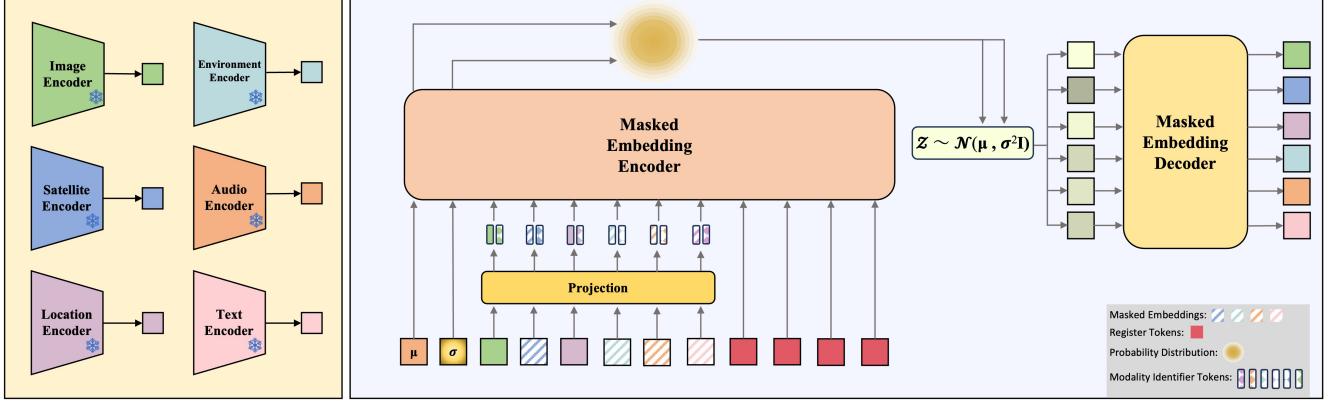


Figure 2. ProM3E Framework. Using embeddings obtained from aligned modality-specific encoders, we model the probability distribution of input modalities using a masked variational autoencoder framework. Subsequently, we utilize the predicted variational distribution of input modalities to reconstruct the embeddings of masked modalities.

one-to-one correspondence between the observations of the modalities. Below we describe our framework with additional details in the subsequent sections.

Multimodal Alignment. The first stage of our framework entails aligning all modalities and projecting them into a unified embedding space. This is accomplished using ImageBind or TaxaBind training recipe. After alignment, modality-specific encoders project their respective modalities into the unified embedding space. This simple stage allows training on massive-scale image-paired datasets. It works on global alignment, representing all observations of each modality using a single global embedding. For domains with a one-to-one pixel correspondence, such as in Mizrahi et al. [49], patch-wise contrastive methods can be trained to obtain patch-wise embeddings.

Masked Modality Training. The second stage involves an SSL-based training for the objective of reconstruction of masked global embeddings. We use a Masked VAE (MVAE) based approach to simultaneously model the joint distribution of modalities and capture uncertainty in the modalities. This stage employs aligned embeddings of various modalities, thus requiring much less paired data.

Modalities. In this work, we consider six modalities: ground-level images of species, satellite images, geographic location, species audio, taxonomic text, and environmental covariates. Modalities pertaining to species observations are naturally occurring together and are easily accessible through open citizen science platforms such as iNaturalist [65]. Modalities such as satellite imagery and environmental covariates are available through various remote sensing platforms.

3.1. Multimodal Alignment

Let $\mathcal{M} = \{m_1, m_2, \dots, m_6\}$ be the set of modalities in consideration. Consider modality-specific encoders: $\mathcal{H} = \{h_1, h_2, \dots, h_6\}$. We employ these modality-specific encoders to transform each modality into a single global normalized embedding: $f_i^j = h_i(m_i^j)$, where m_i^j is the j^{th} observation for the modality m_i . For ground-level images, satellite images, audio, and taxonomic texts, we utilize transformer-based models. In contrast, for geographic locations, we employ a Random Fourier Feature-based network, while for environmental covariates, we use a feedforward network. We utilize TaxaBind’s [54] training recipe, which includes multimodal patching to align each modality. We use frozen image and text encoders and project all other modalities into the image-text embedding space. We use image-paired datasets to align all other modalities to the ground-level image modality. We use symmetric SupCon-Loss [31] to align each modality with the image modality. Each alignment training is done independently. We then patch each encoder using the multimodal patching technique in TaxaBind. In the end, we have modality-specific encoders that are in a unified embedding space.

3.2. Masked Modality Training

We employ global normalized embeddings from modality-specific encoders. We freeze these encoders during training. We use a transformer-based encoder-decoder architecture to train for masked embedding reconstruction. Our encoder learns a joint probability distribution over arbitrary input modalities, parameterized as a Gaussian distribution. We draw sample embeddings from this distribution for each modality and feed them to the decoder to reconstruct the masked embeddings. This probabilistic model handles the many-to-many correspondence problem between modalities. For this stage we require an all paired dataset of the

modalities. Below we describe the training process in detail.

Encoding. We treat each embedding as a distinct token for our transformer encoder. Modality-specific projectors transform the embeddings into a compatible dimension with the encoder. Each projector has a two-layer feedforward network with GELU activation. We add modality identifier tokens as positional encoding for the encoder. We introduce two tokens, $[\mu]$ and $[\sigma]$, to learn the joint probability distribution's mean and standard deviation. In practice, $[\sigma]$ captures the log of variance. We also incorporate four register tokens [10], useful for eliminating noise and memorizing distribution of modalities. All tokens are concatenated and passed to our transformer encoder, which has stacked self-attention blocks similar to MAE [23]. We extract $[\mu]$ and $[\sigma]$ from the encoder output to determine the encoded Gaussian distribution's mean and standard deviation. Let \mathcal{E} be our encoder and $\mathcal{F} = \{f_1, f_2 \dots, f_6\}$ be a mini-batch of embeddings. Then, the encoding function is given as follows:

$$\mu_{\mathcal{G}}, \log \sigma_{\mathcal{G}}^2 = \mathcal{E}(\mathcal{G}) \quad (1)$$

where, $\mathcal{G} \subseteq \mathcal{F}$, which represents the set of input modalities. Our encoder learns a joint gaussian distribution represented as $\mathcal{Z}_i \sim p_{\mathcal{E}}(\mathcal{Z}|\mathcal{G})$.

Masking. For pre-training our VAE, we use a masking strategy similar to MultiMAE [4]. During training, we randomly select one or two visible modalities as input to the encoder. As we will show, the model learns to incorporate additional modalities effectively during inference. The remaining masked modalities are dropped and not encoded. The encoder predicts the joint probability distribution of only the unmasked modalities. This approach aligns with real-world scenarios where most modalities are missing.

Decoding. Once we obtain the encoded $[\mu]$ and $[\sigma]$ tokens, we employ the reparameterization trick [32] to generate embedding tokens for each modality. Each sampled token is fed to our decoder for reconstruction. Our decoder learns to generate marginals of each modality from the joint distribution learned by the encoder. Our decoder comprises modality-specific decoders, one for each modality. Each modality-specific decoder is a two-layer FeedForward network with GELU activation. Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2 \dots, \mathcal{D}_6\}$ be the set of modality-specific decoders. Then the decoding function is given by the following expressions:

$$\mathcal{Z}_i(\mathcal{G}) = \mu_{\mathcal{G}} + \sigma_{\mathcal{G}} \cdot \epsilon_i \quad (2)$$

$$\hat{f}_i(\mathcal{G}) = \mathcal{D}_i(\mathcal{Z}_i(\mathcal{G})) \quad (3)$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ denotes noise used for sampling the latent embedding $\mathcal{Z}_i(\mathcal{G})$.

Objective Function. Our training objective is similar to the traditional VAE objective with a modification to its reconstruction loss term. We first calculate the reconstruction loss between the predicted and ground-truth embeddings. This is simply the Euclidean distance between the embeddings. Let $\hat{f}_i^j(\mathcal{G})$ and f_i^j represent the j^{th} predicted and the ground-truth embedding. Then their distance is calculated as $d_i^{\mathcal{G}}(j, j) = \|\hat{f}_i^j(\mathcal{G}) - f_i^j\|_2$. We then use this distance to compute an InfoNCE-based contrastive loss. By employing a contrastive-based loss, we ensure that the model effectively learns the intra-modal distribution of embeddings without merely predicting its centroid. After obtaining the distances, we scale and shift them. This operation acts similarly to the temperature parameter in the InfoNCE loss. The contrastive objective is given as follows:

$$\mathcal{L}_{\text{recon}}(m_i) = \frac{1}{N} \sum_{j=1}^N \frac{e^{[\alpha \cdot d_i^{\mathcal{G}}(j, j) + \beta]}}{\sum_{p=1}^N e^{[\alpha \cdot d_i^{\mathcal{G}}(j, p) + \beta]}} \quad (4)$$

where, α and β are scaling and shifting parameters respectively. N is the size of the mini-batch used in training. We use the variational information bottleneck (VIB) loss [6, 8] to regularize our training and prevent the σ term from collapsing to zero. The loss is formulated as the KL-divergence between the variational distribution predicted by the model and the Gaussian normal distribution. This loss is given by the closed form as follows:

$$\mathcal{L}_{\text{VIB}} = -\frac{1}{2}(1 + \log \sigma_{\mathcal{G}}^2 - \mu_{\mathcal{G}}^2 - \sigma_{\mathcal{G}}^2) \quad (5)$$

The final loss is a combination of Equations 4 and 5:

$$\mathcal{L}(m_i) = \mathcal{L}_{\text{recon}}(m_i) + \lambda \mathcal{L}_{\text{VIB}} \quad (6)$$

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \mathcal{L}(m_i) + \lambda \mathcal{L}_{\text{VIB}} \quad (7)$$

3.3. Training Datasets

We primarily rely on species observation data from iNaturalist [65] for training. For multimodal alignment, we use the same training datasets as TaxaBind, including iSatNat and iSoundNat. In the second stage, we compile an all-paired dataset called **MultiNat**. We download all species observations from iNaturalist with ground-level images and sound, then filter out observations in the TaxaBench-8k dataset, since this dataset will be used for evaluation. For each observation, we retrieve 256x256 Sentinel-2 imagery from the sentinel-2-cloudless platform and extract bioclimatic variables from the WorldClim-2 dataset. Our dataset contains 79,317 samples. For more details, refer to the appendix.

3.4. Multimodal Embeddings for Downstream Transfer

Our trained model can handle various downstream tasks. Two crucial tasks are linear probing and cross-modal

Model	Modality	Birds525	CUB-200-2011	BioCLIP-Rare	Inquire (iNat-2024)	iNat-2021	TaxaBench-8k
BioCLIP [60]		82.92	77.51	34.52	62.36	68.24	32.88
ArborCLIP [78]		65.84	82.41	27.58	53.02	68.00	31.34
TaxaBind [54]		83.74	78.22	35.84	64.66	70.09	34.45
ProM3E		86.89	82.85	37.49	66.96	75.83	39.45
<hr/>							
ImageBind [20]	+	-	-	-	-	71.02	36.40
	+	-	-	-	-	72.62	36.30
	+	-	-	-	-	71.96	36.59
	+	-	-	-	-	-	35.91
<hr/>							
TaxaBind [54]	+	-	-	-	-	72.73	36.59
	+	-	-	-	-	73.20	37.54
	+	-	-	-	-	72.02	36.51
	+	-	-	-	-	-	36.27
<hr/>							
ProM3E	+	-	-	-	-	78.27	47.00
	+	-	-	-	-	77.63	47.05
	+	-	-	-	-	78.36	46.42
	+	-	-	-	-	-	44.84

Table 1. Zero-shot classification performance on various fine-grained species classification datasets using the taxonomic description of species.

retrieval, which require careful embedding selection, especially in multimodal settings. We provide detailed information on embedding generation for these tasks.

Linear Probing. Probing the learned representation of our model is crucial to assess the effectiveness of our pretraining strategy. Several design choices exist for generating embeddings for linear probing. We find that using the hidden representations of our model outperforms using the reconstructed representations. Furthermore, incorporating all encoded tokens, including the register tokens, yields superior performance. For a detailed comparison of linear probing performance across various design choices, please refer to the appendix.

Cross-Modal Retrieval. Cross-modal retrieval requires robust representations. Many design choices exist for query and target modality embeddings. Since our model supports modality inversion, we combine the input query embedding with the reconstructed target embedding. This merges inter-modal and intra-modal interactions for retrieval. Let m_q and m_t denote the query and target modalities, respectively. Let f_q and f_t represent the query and target embeddings, respectively. Suppose $\mathcal{G} = \{f_q\}$ denotes the input embeddings. When \mathcal{G} is input, the model reconstructs the target embeddings as $\hat{f}_t(\mathcal{G})$. Finally, we combine the input query embedding with the reconstructed target embedding to generate the final query embedding as follows:

$$f_q = (1 - \delta).f_q + \delta.\hat{f}_t(\mathcal{G}) \quad (8)$$

where, δ is a mixing coefficient found using optimal performance on our validation split of MultiNat dataset. We then use these embeddings to compute the cosine similarity with f_t to perform the retrieval.

4. Experiments

We assess the effectiveness of our trained model on retrieval and linear probing tasks spanning all modalities. For ease of learning, we initialize our modality-specific encoders using pretrained TaxaBind [54] encoders. We then train our MVAE with 27M parameters on the MultiNat dataset on a single NVIDIA H-100 GPU with a batch size of 1024. Our MVAE model takes merely 2.5 GPU hours to train, proving to be time and cost effective. For exact implementation details and hyperparameters used, please refer to the appendix. Below we present results on three distinct tasks. We also analyze the uncertainty captured by our model and the modality gap observed. Please see appendix for additional experimental results.

Image Classification. In Table 1, we present the zero-shot species image classification performance of our model across various datasets. We use full taxonomic text for classification encompassing labels from the kingdom level upto the species level. We compare our model against BioCLIP [60], ArborCLIP [78], ImageBind [20], and TaxaBind [54]. Notably, we observe significant improvements of up to 5% in the unimodal setting and 10% in the multimodal setting compared to TaxaBind. ProM3E emerges as the superior model, outperforming all others in all six datasets.

Method	Modality	R@1	R@5	R@10	Method	Modality	R@1	R@5	R@10
<i>Random Baseline</i>	-	0.01	0.05	0.11	<i>Random Baseline</i>	-	0.01	0.05	0.11
ImageBind [20]	📍 → 🌳	8.79	22.72	30.84	ImageBind [20]	📍 → 🌳	1.09	4.34	7.32
	📍 → 🌸	9.32	24.16	32.24		📍 → 🌸	1.27	5.24	8.87
	📍 → 🌸	1.94	6.68	10.56		📍 → 🌸	11.92	26.81	35.30
TaxaBind [54]	📍 → 🌸	1.86	5.33	9.05	TaxaBind [54]	📍 → 🌸	0.05	1.12	1.57
	📍 → 🌳	8.43	21.72	30.42		📍 → 🌳	1.03	4.30	7.59
	📍 → 🌳	9.62	24.60	33.42		📍 → 🌳	1.34	5.42	9.26
ProM3E	📍 → 🌸	2.05	7.03	11.05	ProM3E	📍 → 🌸	12.27	27.63	36.13
	📍 → 🌸	2.36	5.98	9.50		📍 → 🌸	0.06	1.12	1.57
	📍 → 🌸	17.87	43.16	54.53		📍 → 🌸	2.26	8.21	13.72
ImageBind [20]	📍 → 🌳	13.18	32.29	42.74	ImageBind [20]	📍 → 🌳	2.51	9.00	14.89
	📍 → 🌸	3.71	10.56	15.22		📍 → 🌸	14.25	29.42	37.65
	📍 → 🌸	2.24	7.48	12.02		📍 → 🌸	0.07	2.01	2.21

Table 2. **Cross-Modal Retrieval.** We present cross-modal retrieval performance of our model on the TaxaBench-8k dataset with comparison to ImageBind and TaxaBind.

Model	Modality	BirdCLEF'22 (%)	BirdCLEF'23 (%)	BirdCLEF'24 (%)
CLAP [17]	🔊	42.33	32.85	39.72
MGACLAP [38]	🔊	56.05	44.03	47.36
ImageBind [20]	🔊	47.11	37.46	45.04
TaxaBind [54]	🔊	52.60	42.19	49.31
ProM3E (ours)	🔊	58.94	52.30	56.66
ImageBind	🔊 + 🌳	60.22	44.04	51.64
TaxaBind	🔊 + 🌳	65.07	46.97	56.24
ProM3E (ours)	🔊 + 🌳	71.56	59.06	62.43

Table 3. Top-1 linear probing results on the task of bird species audio classification.

Cross-modal Retrieval. We evaluate the performance of our models’ cross-modal alignment on the task of cross-modal retrieval. We use the all paired TaxaBench-8k dataset and perform retrieval given various input and target modalities as shown in Table 2. Our model outperforms TaxaBind and ImageBind in all settings. This demonstrates our model can better understand the interactions between various modalities than the other models in consideration.

Audio Classification. One challenging ecological task is the fine-grained classification of audio of species. We evaluate our model’s linear probing capability to predict bird species given audio recording from three geographically distinct datasets. Table 3 showcases the superior performance of our model in this task with gains upto $\sim 12\%$. Our model outperforms other state-of-the-art audio encoders in both unimodal and multimodal setting.

4.1. What is captured by σ ?

We answer this question by analyzing $\|\sigma\|_1$ values upon adding modalities together and comparing them to reconstruction loss of our model. Essentially, $\|\sigma\|_1$ depends on the pretraining strategy used. Our pretraining strategy involves predicting masked modalities from a few input modalities. Therefore, $\|\sigma\|_1$ captures the informativeness of input modalities in predicting the masked modalities. Lower $\|\sigma\|_1$ values depict high informativeness. We compute mean $\|\sigma\|_1$ values on the TaxaBench-8k dataset for all the modalities.

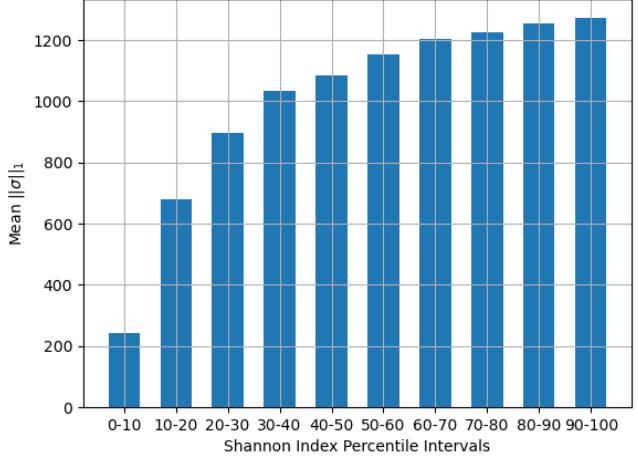


Figure 3. Mean $\|\sigma\|_1$ values of geographic locations at various percentile intervals of the Shannon Diversity Index derived from the iNaturalist dataset.

Figure 4(a) shows the mean $\|\sigma\|_1$ values of each modality. It depicts that geographic location and ground-level imagery are the most informative. Figure 4(b) shows correlation between MSE objective and $\|\sigma\|_1$ values in Figure 4(a). In Figure 4(c), we add a single modality at a time to the ground-level image modality. In Figure 4(d) we progressively add modalities from left to right. As we combine multiple modalities, the mean $\|\sigma\|_1$ values decrease, as shown in Figures 4(d) and (f) (with the exception of text). The correlation between $\|\sigma\|_1$ and MSE increases as modalities are added together. We find that combining text with other modalities does not add new information. Overall, the figure suggests that the task of inferring masked modalities becomes easier with more input modalities.

Species Diversity and $\|\sigma\|_1$. Intuitively, we anticipate that the $\|\sigma\|_1$ values of geographic locations reflect the species diversity of those locations. This is because if a specific location or habitat harbors multiple species, the task of predicting the other modalities such as text or ground-level imagery becomes challenging. To this end, we generate a 250x500 species diversity map over the USA using iNaturalist observations and compute Shannon Diversity index at each grid cell. We then compute the $\|\sigma\|_1$ value at each of those cell and finally calculate the correlation between the biodiversity and $\|\sigma\|_1$ map. We find a spearman correlation of 0.401 with p-value=0.0, indicating significant positive correlation. Figure 3 depicts a positive correlation between $\|\sigma\|_1$ and Shannon Diversity index at various percentile intervals. Please see appendix for additional details and visualizations.

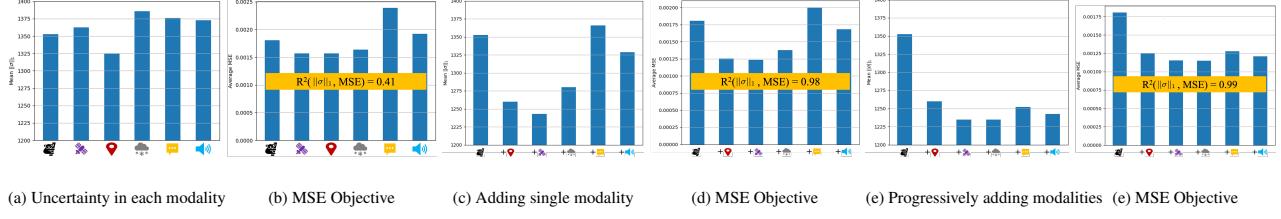


Figure 4. **Effect on $\|\sigma\|_1$ values when adding modalities.** We show mean $\|\sigma\|_1$ predicted and compare it to reconstruction loss of our model. This figure demonstrates a correlation between uncertainty and predictive power of input modalities. In general, adding additional modalities improves reconstruction and the corresponding loss is positively correlated with $\|\sigma\|_1$.

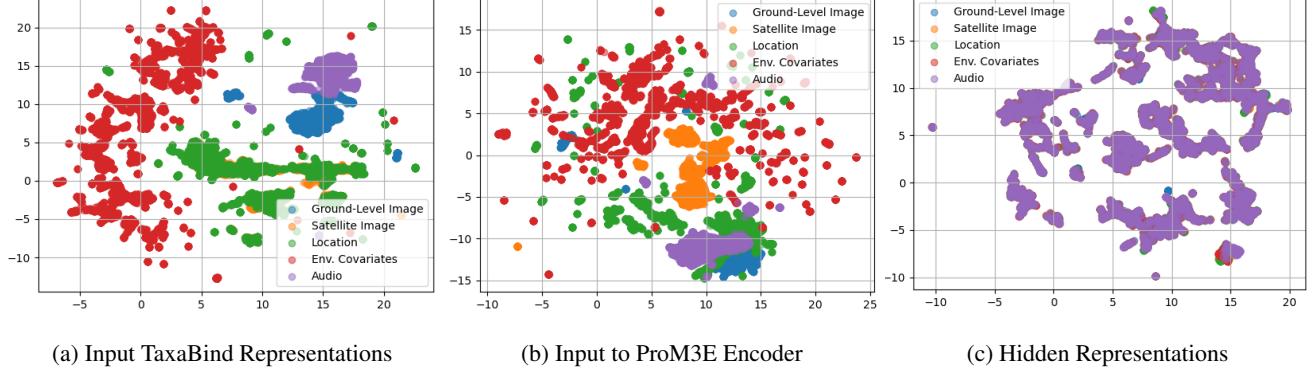


Figure 5. **Crossing the Modality Gap.** Our training strategy minimizes the existing modality gap in the hidden representation space of the masked embedding model. This allows our model to predict masked modalities in the input.

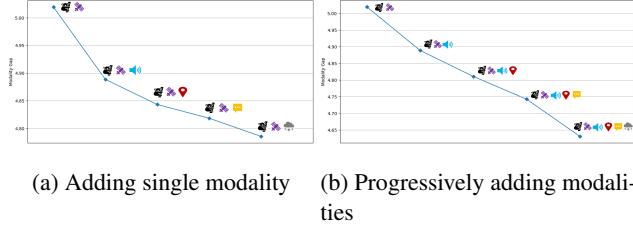


Figure 6. **Effect on modality gap between two modalities in presence of other modalities.** Quantitative evaluation of the modality gap between ground-level image and satellite image when other modalities are provided as input. It is noticed that the modality gap reduces as more modalities are added.

4.2. What happens to the modality gap?

To the best of our knowledge, we are the first work to analyze modality gap in a model with more than two modalities. In Figure 6, we quantitatively evaluate the effect on modality gap when additional modalities are present. Specifically, we evaluate the modality gap between ground-level and satellite image in the hidden representation space. Following the procedure outlined in Liang et al. [39], we calculate the modality gap by taking the distance between centroid of each modality’s embeddings. We find that the modality gap reduces when additional modalities are introduced. This

phenomenon is consistent across all modalities. We investigate the modality gap in the input representation and the hidden representation space of our model in Figure 5 using UMAP [47]. We notice that the modality gap reduces as the representations pass through the encoder. For further analysis, please refer to the appendix.

5. Ablations

Here we conduct ablation on dataset size for pretraining ProM3E and various architectural choices. Note that all ablations are done for the MVAE component of our model. The modality-specific encoders from TaxaBind are kept frozen and utilized as is. Additional ablations are in the appendix.

Scaling ProM3E. We investigate whether our model can scale well in low data regimes since acquiring multiple modalities is time consuming and expensive. ProM3E is based on the idea of aligning representation before fusing them in order to reduce the amount of paired data required for training. In Table 4, we report the performance of our model trained with varying amounts of training data. We evaluate the performance of the models on various tasks and report their average performance. We notice that ProM3E scales very well and performs well in low data regimes. For instance, ProM3E trained with only 7,931 (10%) samples shows a performance drop of about 3% on average. Training

Task	Dataset	Modality	TaxaBind [54]	10%	20%	50%	75%	100%
Image Classification	TaxaBench-8k		34.45	36.51	37.31	37.42	38.34	39.45
Image Classification	TaxaBench-8k		37.54	41.50	42.31	44.01	44.53	47.05
Retrieval	TaxaBench-8k		8.43	10.68	10.78	10.82	15.35	17.87
Retrieval	TaxaBench-8k		9.62	9.62	9.44	9.77	11.70	13.18
Loc. Classification	EcoRegions		73.75	80.26	80.67	80.59	81.13	81.35
Loc. Classification	Biome		71.73	79.54	80.71	81.10	81.85	82.30
Audio Classification	BirdCLEF-2023		42.19	52.06	51.88	53.32	52.48	52.30
Audio Classification	BirdCLEF-2023		46.97	59.96	59.30	60.44	59.96	59.06
Average			40.58	46.27	46.55	47.18	48.17	49.06

Table 4. **Data Scaling.** We show that our model can be trained with much less all paired dataset and the performance across various dataset sizes and tasks remain consistent. For instance training with 10% of the dataset (7,913 samples) only results in a performance drop of $\sim 3\%$ on average.

dim	cls	lin	depth	cls	lin	# Registers	cls	lin	Loss	cls	lin
256	36.28	74.24	1	39.45	81.35	0	39.21	78.51	MSE	33.92	80.27
512	38.26	78.00	2	38.59	80.68	1	39.20	79.35	Contrastive	39.45	81.35
1024	39.45	81.35	3	38.47	80.73	2	38.61	80.21			
2048	39.23	82.52	4	38.36	80.82	4	39.45	81.35			

(a) Encoder dimension

(b) Encoder depth

(c) # Register tokens

(d) Loss function

Table 5. **Ablations.** We perform various ablations related to our architecture and loss function. Here, *cls* denotes species image classification on TaxaBench-8k and *lin* denotes linear probing on EcoRegions classification.

with the full MultiNat dataset demonstrates the best performance. In all settings, our model outperforms TaxaBind.

Architecture and Loss. We perform several ablations to determine the most optimal architecture and loss function as shown in Table 5. We use species image classification on TaxaBench-8k and linear probing on EcoRegions classification to benchmark each model. From our experiments, we find a single encoder layer and higher embedding dimensions to work the best. We also find that including large number of register tokens improves downstream performance. As suspected, our contrastive objective outperforms vanilla VAE’s MSE objective. We find that using MSE loss leads to representation collapse and prevents the model from learning intra-modal distributions.

6. Conclusions

In this paper, we introduced ProM3E, a probabilistic masked multimodal embedding model for ecology. Our model learns a joint probability distribution of arbitrary input modalities and reconstructs the embeddings of unavailable modalities. The probabilistic nature of our model allows us to capture the uncertainty of modalities. This is especially useful to quantify the uncertainty of geographic locations which we

discovered to be correlated with the species diversity at those locations. The representations generated from our model show excellent performance in downstream tasks beating the state-of-the-art. Our future lines of work will focus on integrating additional modalities such as camera trap imagery to further enhance species understanding and mapping.

Acknowledgements

This research used the TGI RAILS advanced compute and data resource which is supported by the National Science Foundation (award OAC-2232860) and the Taylor Geospatial Institute.

References

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021. 2
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 1
- [3] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024. 1
- [4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2, 4
- [5] Roman Bachmann, Oguzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *Advances in Neural Information Processing Systems*, 37:61872–61911, 2024. 1
- [6] Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 4
- [7] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 2
- [8] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. *International Conference on Learning Representations*, 2025. 2, 4
- [9] 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*, pages 6320–6342. PMLR, 2023. 5, 6
- [10] Timothée Darzet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*. 4
- [11] Rangel Daroya, Elijah Cole, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Wildsat: Learning satellite image representations from wildlife observations. *arXiv preprint arXiv:2412.14428*, 2024. 1, 2
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2
- [13] Ayush Dhakal, Adeel Ahmad, Subash Khanal, Sri Kumar Sastry, Hannah Kerner, and Nathan Jacobs. Sat2cap: Mapping fine-grained textual descriptions from satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 533–542, 2024. 2
- [14] Ayush Dhakal, Subash Khanal, Sri Kumar Sastry, Adeel Ahmad, and Nathan Jacobs. Geobind: Binding text, image, and audio through satellite images. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2729–2733, 2024. 2
- [15] Ayush Dhakal, Sri Kumar 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*, 2025. 2
- [16] Johannes Dollinger, Damien Robert, Elena Plekhanova, Lukas Drees, and Jan Dirk Wegner. Climplicit: Climatic implicit embeddings for global ecological tasks. *International Conference on Learning Representations (ICLR) Workshops*, 2025. 2, 6
- [17] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6, 5
- [18] Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*, 2025. 2
- [19] Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In *International Conference on Learning Representations*, 2019. 6
- [20] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 2, 5, 6
- [21] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 2
- [22] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 2
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 4
- [24] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2024. 2
- [25] 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*, pages 173–190. Springer, 2024. 2
- [26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal

- Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5
- [27] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23262–23271, 2023. 2
- [28] Alexis Joly, Lukáš Picek, Stefan Kahl, Hervé Goëau, Vincent Espitalier, Christophe Botella, Benjamin Deneu, Diego Marcos, Joaquim Estopinan, Cesar Leblanc, et al. Lifeclef 2024 teaser: Challenges on species distribution prediction and identification. In *European Conference on Information Retrieval*, pages 19–27. Springer, 2024. 2
- [29] Subash Khanal, Srikumar Sastry, Aayush Dhakal, and Nathan Jacobs. Learning tri-modal embeddings for zero-shot soundscape mapping. In *British Machine Vision Conference (BMVC)*, 2023. 2
- [30] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1361–1369, 2024. 1, 2
- [31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3
- [32] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 2, 4
- [33] 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*, pages 4347–4355, 2025. 6
- [34] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. 2
- [35] Ana Lawry Aguila, James Chapman, and Andre Altmann. Multi-modal variational autoencoders for normative modelling across multiple imaging modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer, 2023. 1
- [36] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26647–26657, 2024. 2
- [37] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024. 2
- [38] Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. Advancing multi-grained alignment for contrastive language- audio pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7356–7365, 2024. 6
- [39] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 2, 7
- [40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2
- [41] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2023. 1
- [42] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26752–26762, 2024. 2
- [43] 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*, pages 9596–9606, 2019. 6
- [44] 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*, pages 23498–23515. PMLR, 2023. 6
- [45] Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023. 6
- [46] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv preprint arXiv:2312.06960*, 2023. 1, 2
- [47] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7
- [48] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. *International Conference on Learning Representations*, 2025. 2
- [49] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023. 1, 2, 3
- [50] Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4547–4557, 2022. 2
- [51] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14711–14721, 2022. 2
- [52] 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*, pages 8748–8763. PMLR, 2021. 5
- [53] 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*, pages 7136–7145, 2024. 2
- [54] 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)*, pages 1765–1774. IEEE, 2025. 1, 2, 3, 5, 6, 8, 7
- [55] Simon Schrödi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: on the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. *International Conference on Learning Representations*, 2025. 2
- [56] Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in clip. In *ICLR 2023 workshop on multimodal representation learning: perks and pitfalls*, 2023. 2
- [57] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. Unival: Unified model for image, video, audio and language tasks. *Transactions on Machine Learning Research*, 2023. 1
- [58] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1236–1248, 2024. 2
- [59] Siddharth Srivastava and Gaurav Sharma. Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27412–27424, 2024. 1, 2
- [60] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 2, 5
- [61] Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multimodal biomedical observations. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [62] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2
- [63] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025. 1, 2
- [64] Uddesha Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023. 2
- [65] 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 conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 3, 4
- [66] Edward Vendrow, Omilos Pantazis, Alexander Shepard, Gabriel Brostow, Kate E Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *NeurIPS*, 2024. 2
- [67] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2023. 5, 6
- [68] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. Onepeace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 2
- [69] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 2
- [70] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. *arXiv preprint arXiv:2503.11849*, 2025. 2
- [71] Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023. 2
- [72] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnidbind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024. 2
- [73] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. [2](#)
- [74] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024. [1](#)
- [75] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. [2](#)
- [76] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024. [1](#)
- [77] Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pages 1–7, 2024. [2](#)
- [78] Chih-Hsuan Yang, Benjamin Feuer, Zaki Jubery, Zi K Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh K Singh, et al. Arboretum: A large multimodal dataset enabling ai for biodiversity. *arXiv preprint arXiv:2406.17720*, 2024. [2](#), [5](#)
- [79] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, et al. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26340–26353, 2024. [2](#)
- [80] Robin Zbinden, Nina van Tiel, Gencer Sumbul, Chiara Vanalli, Benjamin Kellenberger, and Devis Tuia. Masksdm with shapley values to improve flexibility, robustness, and explainability in species distribution modeling. *arXiv preprint arXiv:2503.13057*, 2025. [2](#)
- [81] Valerie Zermatten, Javiera Castillo-Navarro, Pallavi Jain, Devis Tuia, and Diego Marcos. Ecowikirs: Learning ecological representation of satellite images from weak supervision with species observations and wikipedia, 2025. [2](#)
- [82] Ziang Zhang, Zehan Wang, Luping Liu, Rongjie Huang, Xize Cheng, Zhenhui Ye, Huadai Liu, Haifeng Huang, Yang Zhao, Tao Jin, et al. Extending multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 37:91880–91903, 2024. [2](#)

ProM3E: Probabilistic Masked MultiModal Embedding Model for Ecology

Supplementary Material

7. Mapping and Visualization

7.1. ICA Visualization of Geo-Embeddings

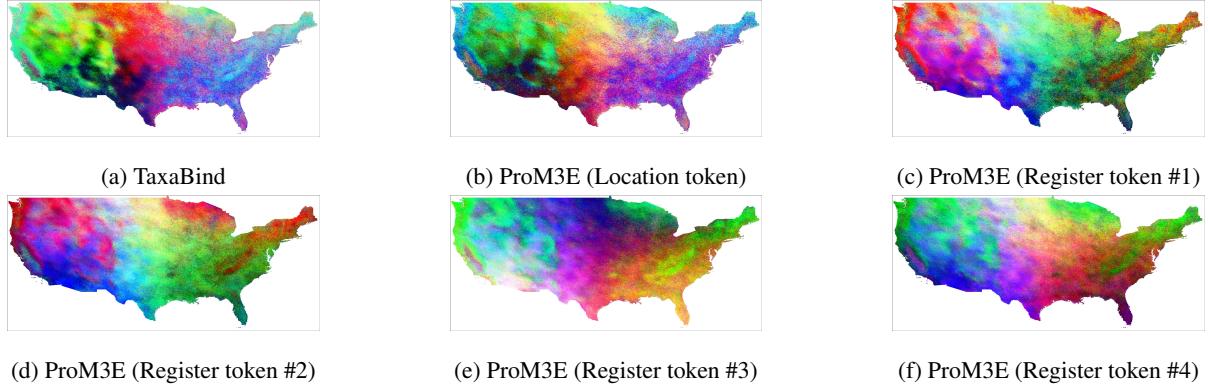


Figure 7. ICA Plot of Location Embeddings. We visually compare embeddings obtained from various tokens in the hidden representation of our model with the representation from TaxaBind. We notice that each register token captures different information.

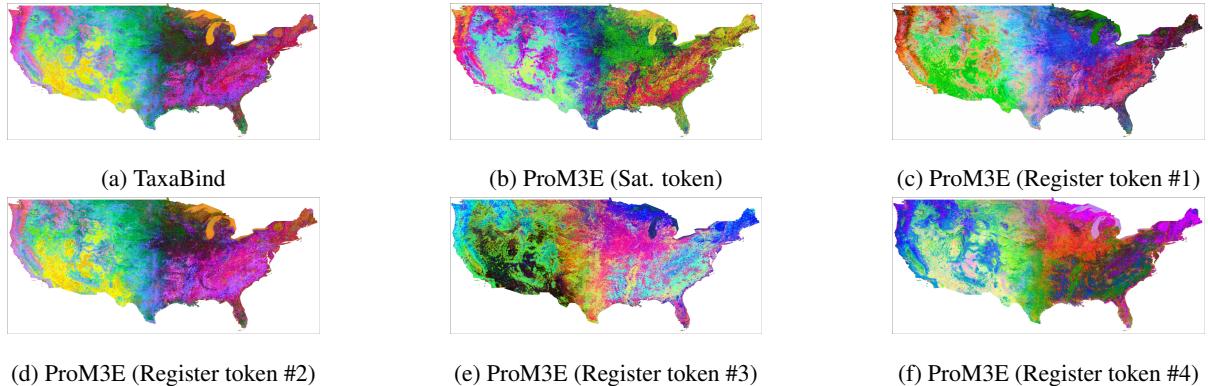


Figure 8. ICA Plot of Satellite Image Embeddings. Similary, we compare satellite image embeddings with TaxaBind and notice register tokens capture diverse information.

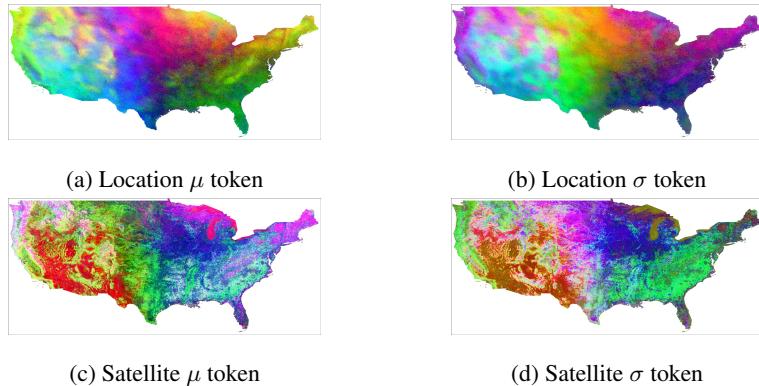


Figure 9. ICA Plot of $[\mu]$ and $[\sigma]$ Tokens. We plot the representations obtained from $[\mu]$ and $[\sigma]$ tokens for geographic location and satellite images across the USA.

7.2. Species Distribution Mapping

In Figure 10, we show species distribution maps generated using our model given a query ground-level image depicting a species. We create a dense grid of satellite imagery over the USA and compute ProM3E embeddings for each location on the grid. We then compute cosine similarity of the query image and the embeddings of each location.

7.3. Generating an iNaturalist Species Diversity Map

We create species diversity and richness maps of the USA using iNaturalist observations. To create the maps, we first filtered observations to include only those within the contiguous United States (excluding Alaska and Hawaii). We then employed a spatial analysis technique that divided the US territory into a 250×500 grid based on the geographic bounding coordinates of the USA. We then filtered out all cells falling outside the USA. For each grid cell, we identified and counted the number of unique species observed by mapping latitude and longitude coordinates to their corresponding grid indices. For calculating species diversity, we used the Shannon index, which computes the entropy in the species distribution. The species richness is calculated as the number of unique species present within each grid cell. To each of the maps, we applied a kernel density estimation (KDE) based Gaussian smoothing with a sigma parameter of 2.0, which smoothed the discrete data across neighboring cells.

Additionally, we generate an uncertainty map of the USA by computing the $\|\sigma\|_1$ value at each grid cell. We then compare all the generated maps visually. We visualize the maps in Figure 11. Remember, in section, we conducted a quantitative comparison between $\|\sigma\|_1$ and Shannon diversity index and found a significant positive correlation between them. The maps in the figure show similarities visually. This is in agreement with the quantitative analyses conducted in the previous sections.

8. Dataset Details

8.1. Training Datasets

ProM3E has a flexible two-stage framework that can be trained independently. The first stage allows for training on large-scale image-paired datasets while the second stage requires an all paired dataset of all modalities for training. The first stage involves training modality-specific encoders using TaxaBind recipe. Below we present the details for the pretraining datasets used in stage one.

TreeofLife-10M. This dataset is composed of 10M pairs of species images and their corresponding taxonomic labels derived from open databases such as . It was introduced by Stevens et al. [60], which was used to train BioCLIP. Here,

we utilize BioCLIP’s image-text frozen embedding space and project all other modalities to this space.

iNaturalist-2021. We use the iNaturalist-2021 dataset primarily to train for aligning geographic location and species images. This dataset consists of 2.7M images across 10k species categories captured around the globe. Each image is associated with metadata including geographic location, timestamp, etc.

iSatNat. Sastry et al. [54] curated a paired dataset of satellite and species images using the iNaturalist-2021 dataset. For each ground-level image, they download a 256x256 Sentinel-2 imagery. This dataset is used to align satellite image with species images. There are 2.55M samples for training, 134k samples for validation and 100k samples for testing.

iSoundNat. This dataset [54] consists of paired species images and audio downloaded from the iNaturalist platform. There are 74k samples for training, 4k samples for validation and 8k samples for testing.

WorldClim-2. Climatic variables derived from WorldClim-2 are used to align environmental covariates and species images. These are environmental covariates are curated for each species in the iNaturalist-2021 dataset.

8.2. Evaluation Datasets

Below we provide details on the evaluation datasets used in the paper.

Taxabench-8k. This dataset consists of 8813 observations from the iNaturalist platform including all modalities paired for each observation. This dataset is used primarily for evaluating the models for the task of cross-modal retrieval.

BirdClef series. These datasets, released annually as part of the LifeClef [28] competition, contain geographically confined audio recordings of rare bird species. These datasets are used to identify bird species based on their soundscapes. We use the training, validation and testing split from TaxaBind for this task of bird species audio classification.

EcoRegions & Biome. We follow Range [15] and use their curated dataset for ecoregion and biome classification of given geographic locations. The dataset was curated by randomly sampling 100k geographic locations across the globe. Each geographic location was assigned a ecoregion label and a biome label. In total, there exist 846 ecoregions and 14 biomes.

9. Implementation Details

Below we provide all the implementation details that were used to train our model.

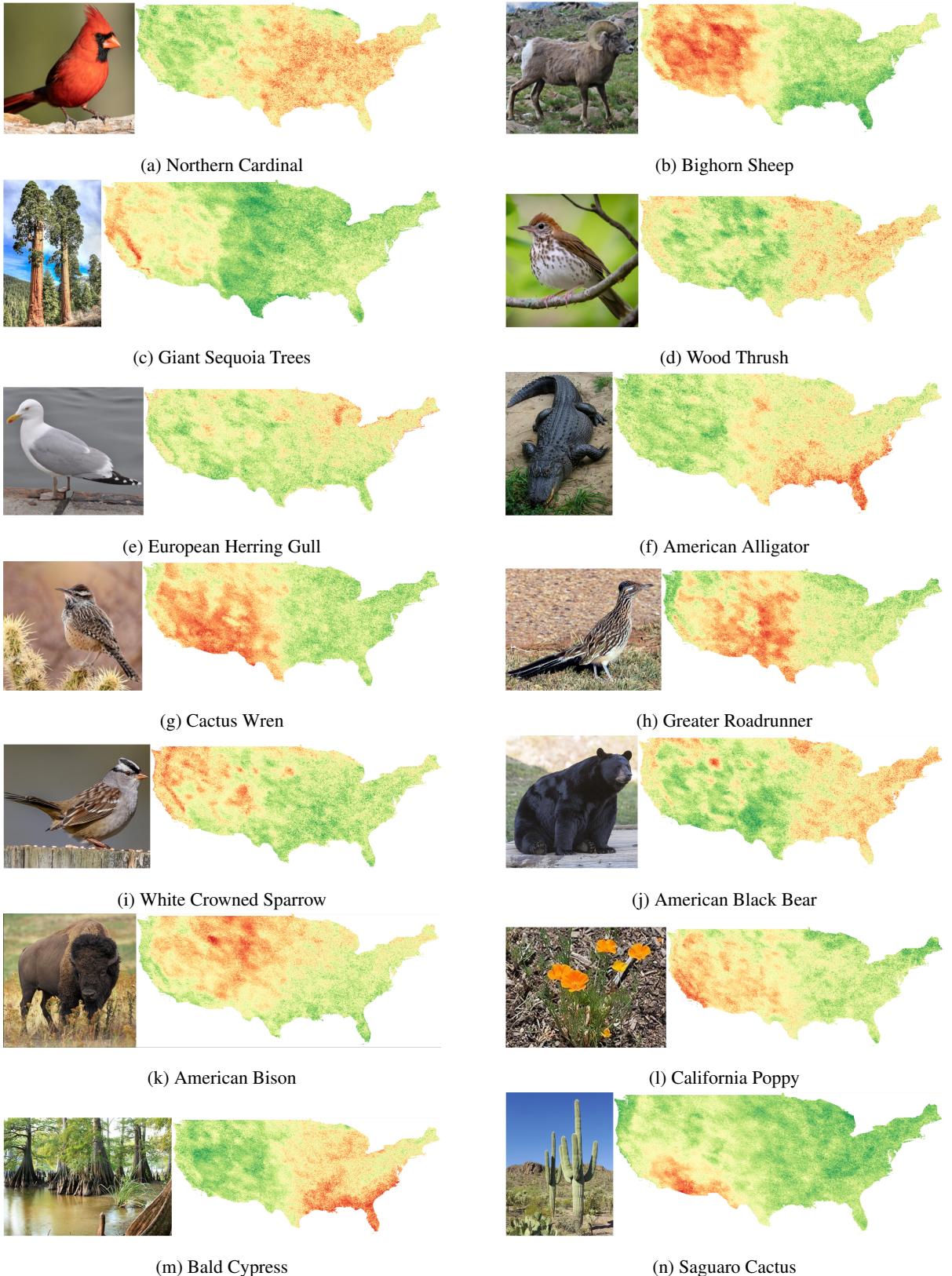


Figure 10. High Resolution Species Distribution Mapping using Ground-Level Imagery. We create species distribution maps by computing the similarity between query ground-level image and geo-locations sampled uniformly across USA.

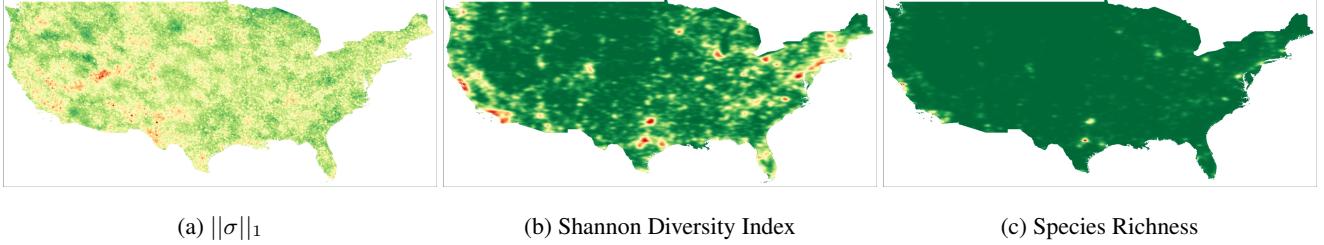


Figure 11. Species Biodiversity Maps. We plot $\|\sigma\|_1$ values predicted by our model and compare it with shannon diversity index and species richness maps derived from iNaturalist observations. The maps are plotted using a rectangular grid of 250x500 points over USA.

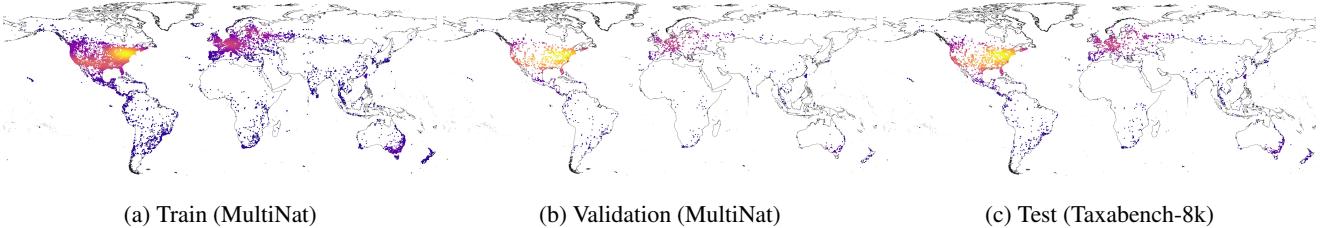


Figure 12. Spatial Distribution of Data. The spatial distribution of our MultiNat dataset covering the globe.

Hyperparameter	Value
batch size	1024
max training epochs	500
optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$
learning rate	1e-4
scheduler	cosine decay
weight decay	0.01
gpu type	NVIDIA H-100
num gpus	1

Table 6. Configuration used for training.

Configuration	Value
VIB term weight (λ)	0.001
base scale parameter (α)	-5.0
base shift parameter (β)	5.0

Table 7. Base configuration for our loss function for MVAE.

Configuration	Value
input embedding dim	512
num projection layers	2
encoder dim	1024
feedforward activation function	GeLU
num encoder layers	1
decoder dim	1024
num decoder layers	1

Table 8. Base architecture configuration for MVAE.

Modality	Model Architecture
ground-level image	OpenCLIP [26] (ViT-B/16)
satellite image	CLIP [52] (ViT-B/16)
geographic location	GeoCLIP [67]
environment covariates	SINR [9]
text	OpenCLIP [26]
audio	CLAP [17]

Table 9. Architecture configuration for modality-specific encoders.

10. Additional Experiments & Ablations

10.1. Linear Probing

10.1.1. Embedding Generation Ablation

In this experiment, we evaluate several design choices for curating embedding effective for linear probing tasks. The first choice is to utilized all the reconstructed embeddings. This is done by concatenating all embeddings at the output of our MVAE decoder. The rest of the choices involve using the hidden representations of our MVAE encoder. We could use the $[\mu]$, $[m_i]$ or register tokens. Additionally, we can concatenate all the tokens from the hidden representations of our MVAE encoder. Table 10 presents results of all these choices. We perform linear probing on the BirdClef-2023, EcoRegions and Biome datasets. Except the experiment with $[\mu]$ token, all the choices outperform TaxaBind. Using all the hidden tokens results in the best performance. We find that register tokens are essential and result in a significant gain in performance.

10.1.2. Probing Geo-Location Embeddings

Our models can serve as general purpose ecological predictors over space. Generating insights about habitat and climatic conditions of various geographic locations around the world is crucial in understanding global ecological trends. In this experiment, we compare the performance of several pretrained geographic location encoders in predicting various ecological indicators over space. In Table 11, we show the performance of our location encoder in predicting Biome, EcoRegion, Temperature and Elevation at a given geographic location. Climplicit is considered the absolute SOTA since it is specifically training on rich spatio-temporal climate data. We find that our model has the best performance beating TaxaBind, SINR and SatCLIP on all the tasks. We also conduct linear probing for the task of predicting several climatic variables in the ERA5 dataset. The results are presented in Table 12. We find that our model beats TaxaBind by a large margin and achieves the second best performance on average after SINR. We believe that high-frequency geographic location features may not be necessary for these tasks. Climatic variables are typically low-frequency and often do not vary significantly across large regions. SINR is a

simple feedforward-based model that outputs low-frequency geo-location embeddings. As a result, it achieves superior performance over other location encoding frameworks.

10.1.3. Habitat Classification

In this experiment, our aim is to classify the habitat of species represented using a given ground-level image. To achieve this, we use the iNat-2021 dataset that includes over 2.7M images of species with corresponding geographic location information. For each sample, we extract the Biome and EcoRegion label. We then obtain the image embeddings for each sample using our model and train a single layer linear classification model to predict the Biome/EcoRegion label given the image embedding. We note that this is a single positive multi label (SPML) problem. For training, we use the assume negative loss which is a common loss used in SPML problems. We evaluate the trained model on the testing split of iNat-2021 dataset.

10.2. Cross-Modal Retrieval

10.2.1. Embedding Generation Ablation

In this section, we investigate an optimal procedure to generate embeddings for effective cross-modal retrieval. There are several design choices one could use. We compare these design choices in Table 14. We find that using the representations from the hidden $[m_i]$ leads to poor performance. We suspect that the representations useful for reconstruction are not necessarily useful for retrieval. We compare the reconstructed embeddings alone for retrieval and find that its performance is better than simply using the TaxaBind representations. We get the best performance using our proposed hybrid approach.

11. Uncertainty & Modality Gap

12. Broader Impact

12.1. Limitations

We acknowledge that the datasets used for training and evaluation in this paper suffer from various biases including geographic, socio-economic and human biases. The aim of this paper is to demonstrate the benefits of fusing multiple

Task	Dataset	Modality	TaxaBind [54]	Recons.	$[\mu]$	$[m_i]$	Reg. Tokens	All
Loc. Classification	EcoRegions	📍	73.75	75.96	74.06	74.38	79.44	81.35
Loc. Classification	Biome	📍	71.73	76.45	71.19	73.80	78.81	82.30
Audio Classification	BirdCLEF-2023	🔊	42.19	41.17	43.20	45.30	51.65	52.30
Audio Classification	BirdCLEF-2023	🔊 + 📍	46.97	49.25	42.55	54.09	58.10	59.06
Average			58.66	60.71	57.75	61.89	67.00	68.73

Table 10. **Embedding Generation Ablation.** We investigate different choices for generating embeddings for linear probing. We find using all hidden tokens to achieve the best performance on average.

	Modality	Biome	EcoRegions	Temperature	Elevation
Direct	📍	29.1	0.6	0.381	0.025
Cartesian_3D	📍	30.2	1.8	0.362	0.030
Wrap [43]	📍	34.4	1.1	0.861	0.085
Theory [19]	📍	33.5	1.0	0.849	0.093
SphereM [45]	📍	36.4	27.3	0.629	0.139
SphereM ⁺ [45]	📍	58.7	50.1	0.886	0.294
SphereC [45]	📍	36.3	52.9	0.461	0.185
SphereC ⁺ [45]	📍	53.2	61.6	0.842	0.260
CSP-INat [44]	📍	61.1	57.1	0.717	0.388
CSP-FMoW [44]	📍	61.4	58.0	0.865	0.399
SINR [9]	📍	67.9	54.9	0.942	0.644
GeoCLIP [67]	📍	70.2	71.6	0.916	0.604
SatCLIP [33]	📍	68.9	69.3	0.825	0.666
TaxaBind [54]	📍	71.7	73.7	0.915	0.601
ProM3E (ours)	📍	82.3 +10.6	81.3 +7.6	<u>0.918</u> +0.003	0.772 +0.171
Climplictit[†] [16] (Absolute SOTA)	📍	83.3	78.4	0.985	0.898

Table 11. Comparison of various pretrained location encoders on predicting four ecological indicators. [†]Note that climplicit is pretrained on rich spatio-temporal climate data.

Models	temp_mean	temp_min	temp_max	dew_temp	precipitation	pressure	u_wind	v_wind	Avg
CSP	0.944	0.933	0.940	0.918	0.610	0.427	0.499	0.550	0.727
CSP-INat	0.987	0.897	0.886	0.857	0.534	0.307	0.413	0.386	0.658
SINR	0.982	0.975	0.976	0.977	<u>0.758</u>	<u>0.706</u>	0.726	0.694	0.849
GeoCLIP	0.960	0.953	0.948	0.954	0.591	0.651	0.502	0.529	0.761
SatCLIP	0.904	0.900	0.887	0.894	0.497	0.743	0.488	0.455	0.721
TaxaBind	0.965	0.954	0.955	0.957	0.637	0.662	0.525	0.560	0.777
ProM3E (Ours)	0.978 +0.013	<u>0.971</u> +0.017	<u>0.970</u> +0.015	<u>0.972</u> +0.015	0.730 +0.093	0.758 +0.096	<u>0.630</u> +0.105	<u>0.638</u> +0.078	<u>0.830</u> +0.058

Table 12. We show the linear probe results on real-world climate data from ERA5. Our model consistently beats TaxaBind and achieves the second best performance on average after SINR.

Method	Modality	Biome			EcoRegions		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
TaxaBind [54]		45.07	83.00	93.03	8.05	29.16	43.10
ProM3E (ours)		57.75	89.51	93.70	25.37	54.06	61.72
TaxaBind [54]		65.37	90.75	93.56	35.33	65.43	74.93
ProM3E (ours)		76.57	93.64	93.96	54.79	69.63	70.18
TaxaBind [54]		60.10	90.49	93.36	26.19	55.13	60.90
ProM3E (ours)		83.05	93.90	94.00	64.95	73.76	73.79

Table 13. **Habitat Classification.** We perform Biome and EcoRegion classification given species images as input. This is a challenging task and requires robust alignment of species images with geographic location and satellite images. We also test other inputs such as satellite images and environmental covariates.

Task	Dataset	Modality	TaxaBind [54]	Recons.	Hybrid
Image Classification	TaxaBench-8k	 → 	34.45	33.23	39.45
Image Classification	TaxaBench-8k	 +  → 	37.54	42.34	47.05
Retrieval	TaxaBench-8k	 → 	8.43	17.19	17.87
Retrieval	TaxaBench-8k	 → 	9.62	12.64	13.18
Average			58.66	60.71	68.73

Table 14. **Embedding Generation Ablation.** Here we investigate choices for embeddings useful for cross-modal retrieval.

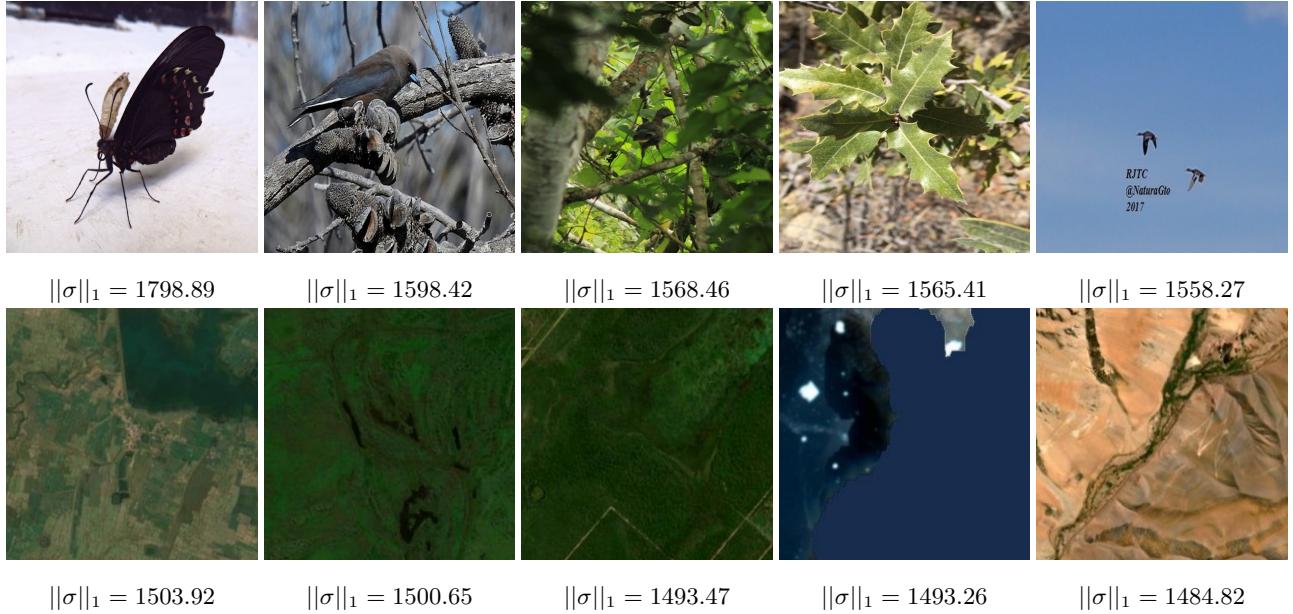


Figure 13. **Most Uncertain Images.** Most uncertain ground-level and satellite images.

modalities to improve performance of models on community accepted benchmark datasets. At present our model is limited to accept and process six modalities. However, given the simplicity of our approach, we believe it is trivial to incorporate additional modalities into the framework.

The species diversity and richness maps generated from iNaturalist observations might not accurately represent the Earth’s biodiversity. As noted above, crowdsourcing and citizen science often lead to biased observations, favoring densely populated regions and documenting a limited num-

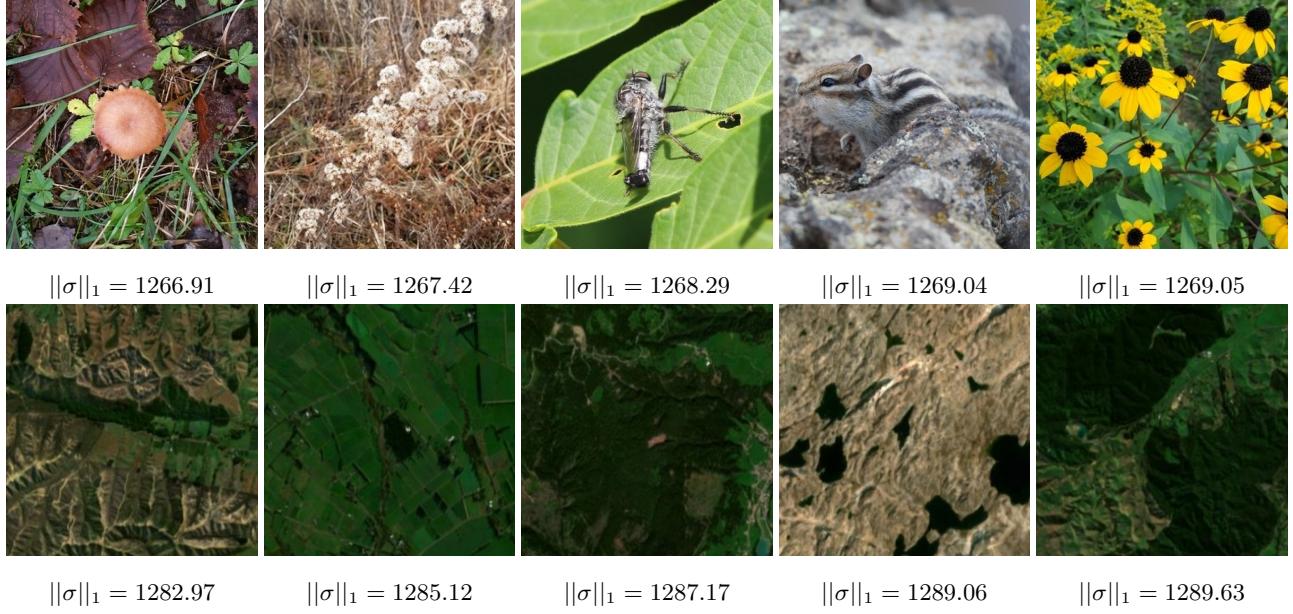


Figure 14. **Least Uncertain Images.** Least uncertain ground-level and satellite images.

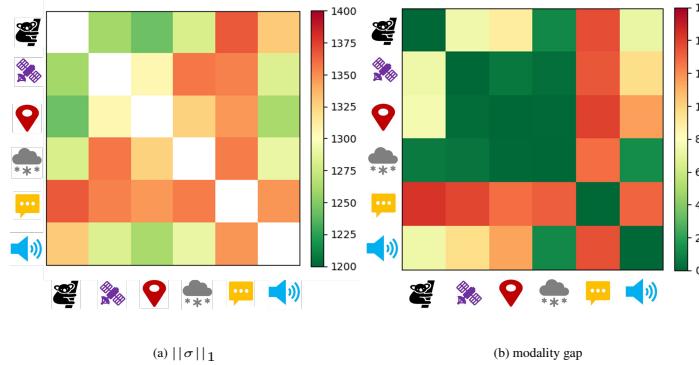


Figure 15. **Pairwise $\|\sigma\|_1$ and Modality Gap.** We compare mean $\|\sigma\|_1$ values and modality gap when pairs of modality are provided as input. We find a spearman correlation of 0.32 between uncertainty and modality gap captured by our model.

ber of species. Our study aimed to investigate whether the uncertainty captured by our model at different geographic locations correlates with the diversity of species observations in those areas. We found a significant positive correlation between these two factors. This is a promising result which we believe can form basis for future research.

12.2. Social Impact

Our models can be effectively adapted to address several remote sensing and ecological challenges. This might mean fine-tuning on additional datasets to adapt our models for specific applications. Our models can serve as a starting point from which interesting applications can emerge. However, utmost care must be taken before deploying them in the real world as is. They might need additional validation before they can be utilized for real world applications. The inher-

ent biases present in the training datasets could potentially lead to inaccurate predictions in certain cases. Consequently, the application of our models in real-world scenarios can benefit from domain expertise. Our model was trained only on openly available species observation data and does not necessarily include information about sensitive species. Yet, care must be taken when using our models for such species.

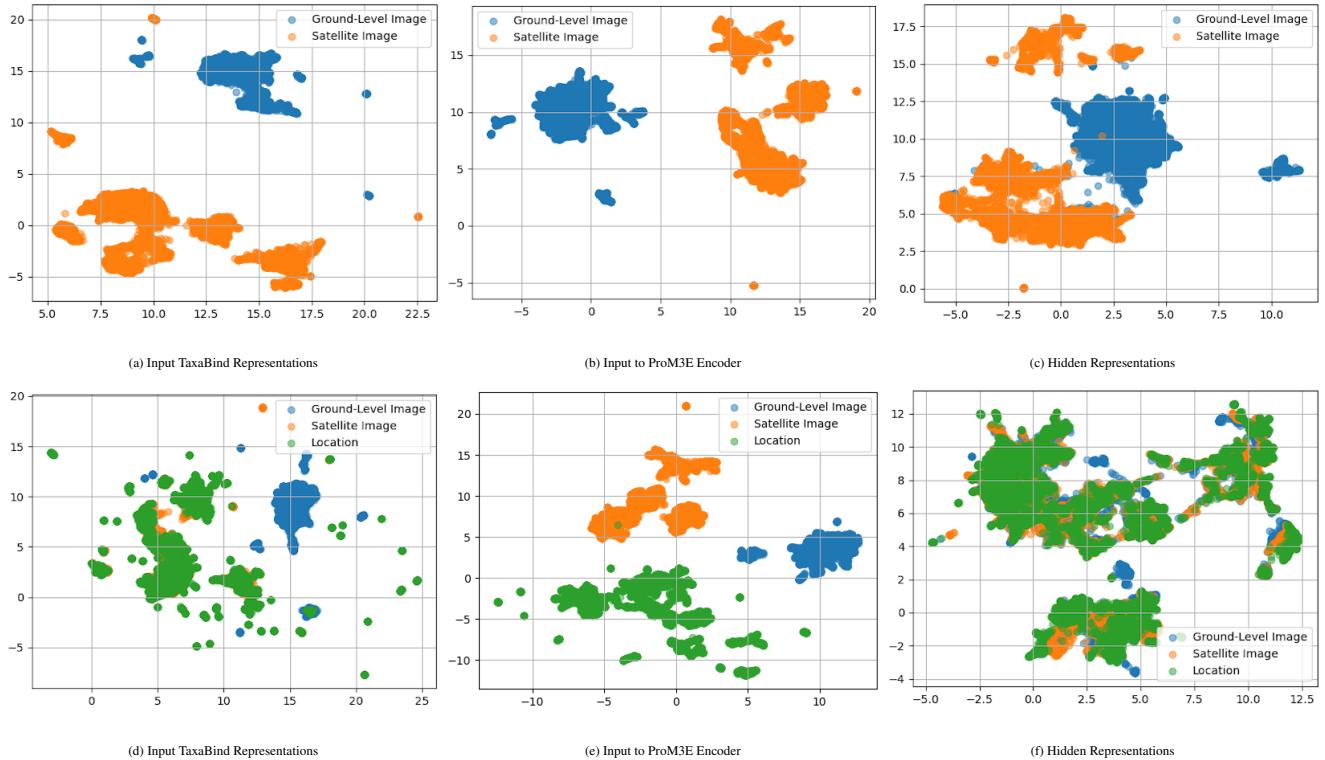


Figure 16. Adding Modalities Reduces the Modality Gap. UMAP visualization of embeddings describing the reduction in modality gap between two modalities in presence of a third modality. Top row presents ground-level and satellite image embeddings while the bottom row presents the embeddings when location is additionally provided as input.