BioVITA: Biological Dataset, Model, and Benchmark for Visual-Textual-Acoustic Alignment

Supplementary Material

In this supplementary material, we provide

A. implementation details of baseline models and prompts,
B. details on the process of collecting the image dataset,
C. detailed analyses including additional comparisons, visualizations, and training data size, and
D. more dataset details, including visualizations of examples and distributions.

A. Implementation Details

A.1. Model Details

- **CLIP** [4]: CLIP is a large-scale image-text contrastive model trained on diverse web data. It projects images and texts into a shared embedding space using a ViT-based image encoder and a Transformer-based text encoder. As a unimodal-pair model, it supports only image–text alignment but provides a strong vision-language baseline for cross-modal retrieval tasks.
- **CLAP** [8]: CLAP extends the CLIP framework to audio-text modalities by introducing an audio encoder trained jointly with a text encoder. It enables zero-shot recognition and retrieval across audio and text domains. Since CLAP does not model images, it serves as the primary baseline for evaluating audio-text alignment in biological settings.
- **ImageBind** [1]: ImageBind is a unified multi-modal model that binds six different modalities, including audio, image, and text, into a single representation space. It leverages image embeddings as the central hub, learning cross-modal correspondences through large-scale contrastive training. As a tri-modal model, ImageBind provides a comprehensive reference point for evaluating unified audio–image–text alignment.
- **BioCLIP 2** [2]: BioCLIP 2 is a biology-specialized vision-language model based on ViT-L/14 for images and a 12-layer Transformer for text. Trained on large-scale curated biological datasets, it achieves strong fine-grained species-level discrimination. In our evaluation, BioCLIP 2 serves both as a strong image-text baseline and as the vision-language foundation for our BioVITA model.

A.2. Prompt Details

To incorporate taxonomy information into audio-text embeddings, we train the CLAP model with prompts augmented by 1) Common name (Com), 2) Scientific name (Sci), and 3) Taxonomic sequence (Tax) following the BioCLIP 2 setting. The augmentation function ϕ randomly selects one of the five prompts (Com, Sci, Tax, Sci+Com, and Tax+Com) defined in Table 9. For instance, given the common species' name (*e.g.*, 'Anianiau), the augmented prompts include the scientific name (*e.g.*, *Magumma parva*) and its taxonomic order (*e.g.*, *Aves Passeriforme*).

Template	Example ('Anianiau)
Common Name (Com)	'Anianiau
Scientific Name (Sci)	Magumma Parva
Taxonomic Sequence (Tax)	Aves Passeriformes, Fringillidae Magumma, Magumma Parva
Sci + Com	Magumma Parva with a common name 'Anianiau
Tax + Com	Aves Passeriformes, Fringillidae Magumma, Magumma Parva, with a common name 'Anianiau

Table 9. Examples of textual descriptions following the five templates used in training.

A.3. Model Parameter Sizes

Our tri-modal model BioVITA consists of three encoders: a BioCLIP2 image tower, a BioCLIP2 text tower, and a CLAP audio encoder, together with a linear audio-to-vision/text projection layer. We report the exact parameter counts obtained from the instantiated PyTorch model.

- BioCLIP2 image encoder (visual tower): 303.97M parameters. All weights are frozen during training.
- BioCLIP2 text encoder: 123.65M parameters. We freeze the entire transformer stack; only 0.65M parameters remain trainable.
- CLAP audio encoder: 153.49M parameters. All parameters are trainable.
- Audio projection layer: A linear adapter with 0.39M parameters.

In total, the model contains 581.5 M parameters , of which 154.5M parameters are trainable.

We trained our model on 8×V100 GPUs (32 GB each). Because the dataset is stored on a separate storage server, the data transfer overhead increases the overall training time. Stage 1 training required approximately two days, while Stage 2 took about one day.

B. Test Dataset Creation

B.1. Audio

After collecting the audio data for 14K species, we split it into training and test sets with a 9:1 ratio, holding out 325 species that remain completely unseen during training. This results in approximately 1.3M training samples and 44K test samples. For efficient evaluation, we limit the number of audio clips per species to around ten.

For benchmarking, we construct 100-option multiple-choice questions, yielding roughly 30K species-level questions, 10K genus-level questions, and 1.7K family-level questions for each task type. The dataset covers a total of 9,725 species.

B.2. Image

To ensure fairness in building our model, we carefully curate the training and test data for the image modality while preventing any leakage.

For training, we collect images from ToL-200M for all species that appear in the audio training set. This results in 12,916 species, covering 91.4% of the species in the audio training split.

For testing, we collect images from iNaturalist using the iNaturalist API by querying species names that appear in the audio test set. Since ToL-200M provides the original image source URLs, we can extract the corresponding iNaturalist observation IDs directly from these URLs. During test image collection, we exclude all images whose observation IDs match those extracted from ToL-200M, ensuring that no images overlap between the training and test splits. We also apply GroundingDINO [3] to filter images: if no animal is detected when using "animal" as the text prompt, we exclude the corresponding image. We additionally verify within each species that the retrieved test images are distinct from the training set. Finally, we obtain a clean set of 128,645 images from 9,487 species that do not overlap with the ToL-200M dataset.

C. Further Analysis

C.1. Model Comparison

In the main paper, we did not include a comparison with BioLingual [5] because its training split may contain samples from our test set. To avoid this potential data leakage, we construct a new test subset that contains only the 2024 split, which is not included in the BioLingual training data. This test set consists of 2,710 species and 4,483 recordings.

In this setting, we evaluate classification performance by averaging over classes. Table 10 shows the retrieval results using species-, family-, and genus-level prompts, meaning that the input prompt specifies the species name, family name, or genus name of the target class.

The results indicate that, for species-level classification, BioVITA achieves better accuracies compared with BioLingual. For family- and genus-level prompts, BioVITA clearly outperforms BioLingual. This suggests that BioVITA benefits from the taxonomy-aware prompting strategy inherited from BioCLIP, enabling the model to generalize more effectively beyond the species level.

C.2. Evaluation on Other Datasets

Table 11 presents a direct comparison with TaxaBind on species retrieval evaluated by top-5 accuracy. The gray text indicates an in-dataset evaluation. Our model consistently outperforms TaxaBind, demonstrating that the larger dataset and VITA training strategy improve retrieval accuracy.

Table 12 shows results for top-1 accuracy for zero-shot species-level retrieval averaged over both retrieval directions on CUB-200 [7], BioCLIP-Rare [6], iSoundNat. Our model demonstrates generality.

C.3. t-SNE

We visualize the t-SNE embeddings from the audio encoder in Fig. 10, clustered by the top six categories in each taxonomy level (species, family, and order). The results show that in Stage 1 (text–audio training), the model learns well-aligned audio–text representations. Moreover, with careful tri-modal learning, Stage 2 successfully preserves the inherent structure of the audio clusters.

C.4. Training Data Size

We investigate how the size of the training set affects our model's performance. In Table 13, we present results for two reduced dataset settings, one-fourth and one-half of the original training size. If downsampling would remove a species entirely, we retain at least one sample to avoid collapsing the taxonomy structure. We maintain consistency in the training protocol.

These results demonstrate that the size of the training audio dataset has a substantial impact on model performance. Our BioVITA model benefits greatly from the large-scale dataset and learns robust audio representations from extensive training data.

D. Dataset Details

D.1. Annotation Example

We present annotation examples for Tokay Gecko and Schlegel's Green Tree Frog in Fig. 11. For each animal species, we collect images, audio recordings, taxonomic information, and trait annotations.

D.2. Dataset Distribution

We illustrate the genus-level distribution of our dataset in Fig. 12 through Fig. 15. The blue bars represent the training set, while the orange bars represent the test set.

Table 10. Classification results of BioLingual and BioVITA. Since BioLingual may include part of our test data in its training split, we construct a new 2024 audio test set and evaluate the classification performance on it.

Model	Species				Genus				Family				Average			
	Audio→Text		Text→Audio		Audio→Text		Text→Audio		Audio→Text		Text→Audio		Audio→Text		Text→Audio	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
BioLingual [5]	24.0	45.9	26.5	50.3	23.2	42.8	51.7	68.2	5.4	15.9	20.2	38.9	17.5	34.9	32.8	52.5
BioVITA	24.4	49.2	27.7	56.6	32.5	51.4	58.0	74.7	17.4	42.9	52.3	77.3	24.8	47.9	46.0	69.5

Table 11. BioVITA vs. TaxaBind.

Model	BioVITA				TaxaBench-8k			
	A↔T	A↔I	I↔T	Avg.	A↔T	A↔I	I↔T	Avg.
Taxabind	30.5	38.6	83.5	50.9	22.9	30.9	47.8	33.9
Ours	89.3	81.5	96.8	89.2	57.0	36.4	77.2	56.9

Table 12. Evaluation on other image and audio benchmarks.

Model	CUB-200	BioCLIP-Rare	iSoundNat
Taxabind	75.0	34.1	16.2
Ours	91.1	82.9	44.4

Table 13. Training dataset size variation. The amount of training data significantly affects model performance. These results correspond to Stage 1. "Sci" and "Com" denote the prompt types used during inference: scientific name and common name, respectively.

	Model	Audio→Text		Text→Audio		Audio→Image		Image→Audio		Image→Text		Text→Image		Average	
		Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
Sci	25%	48.0	68.1	67.5	85.9	39.6	65.9	41.1	71.4	45.4	59.5	80.9	93.2	53.8	74.0
	50%	53.6	73.4	75.2	90.5	43.7	70.0	44.5	74.5	45.4	59.5	81.1	93.5	57.2	76.9
	BioVITA (Full)	60.3	80.0	79.3	93.9	47.8	72.8	48.6	78.8	65.1	80.5	84.8	95.3	64.3	83.6
Com	25%	50.0	72.1	65.8	86.7	39.6	65.9	41.1	71.4	65.1	80.4	84.5	95.1	57.7	78.6
	50%	55.7	76.3	73.0	90.6	43.7	70.0	44.5	74.5	65.1	80.5	84.6	95.2	61.1	81.2
	BioVITA (Full)	60.4	80.0	79.3	93.9	48.0	72.8	48.8	78.9	65.2	80.5	84.7	95.0	64.4	83.5

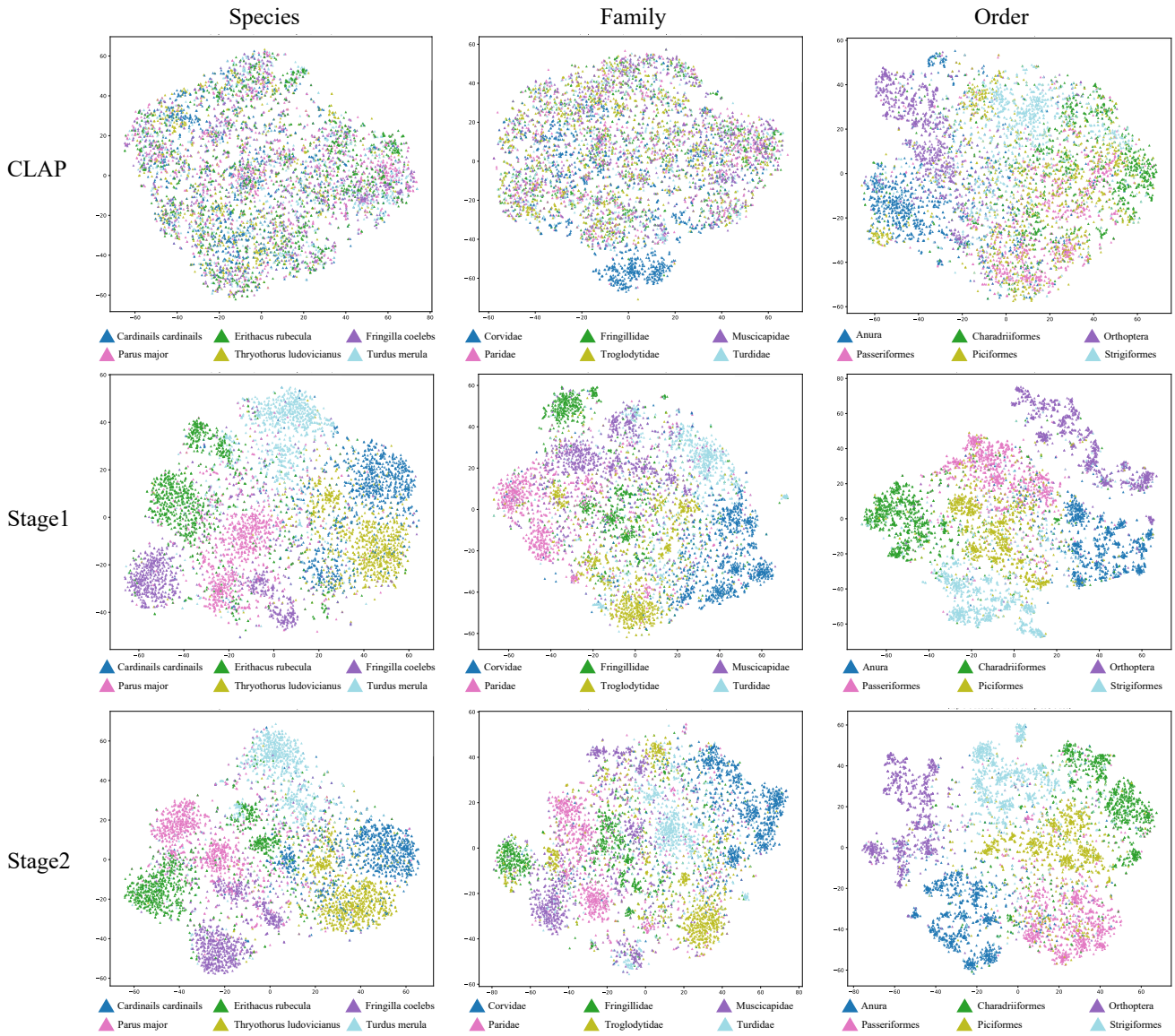

Figure 10. t-SNE visualization. Our model successfully learns all three modalities in Stage 2 without collapsing the audio feature embedding space.

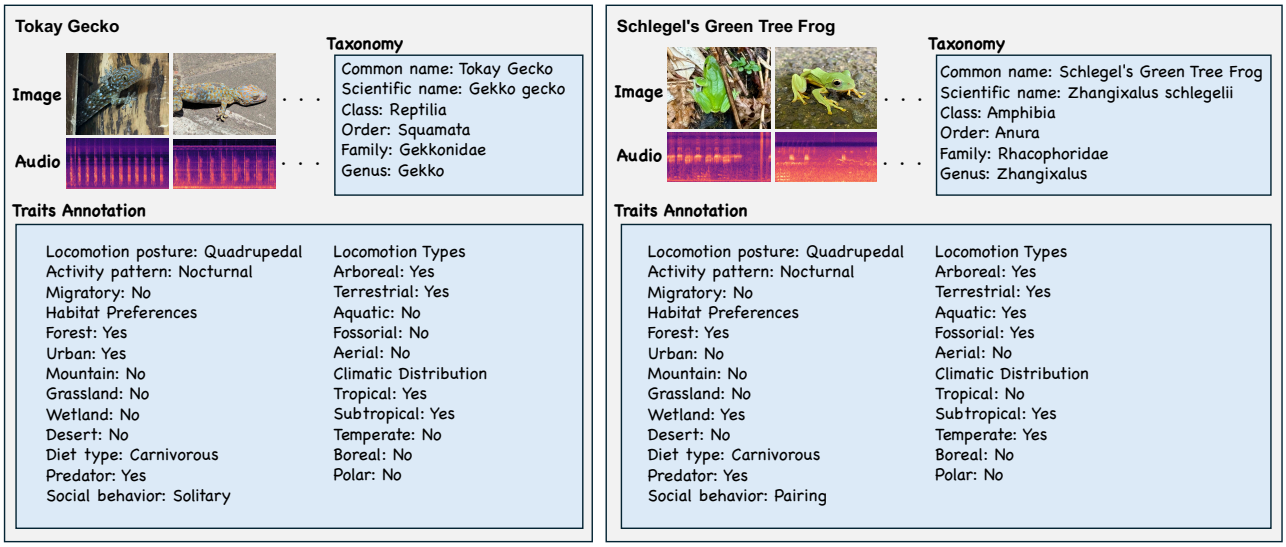

Figure 11. Dataset Example.

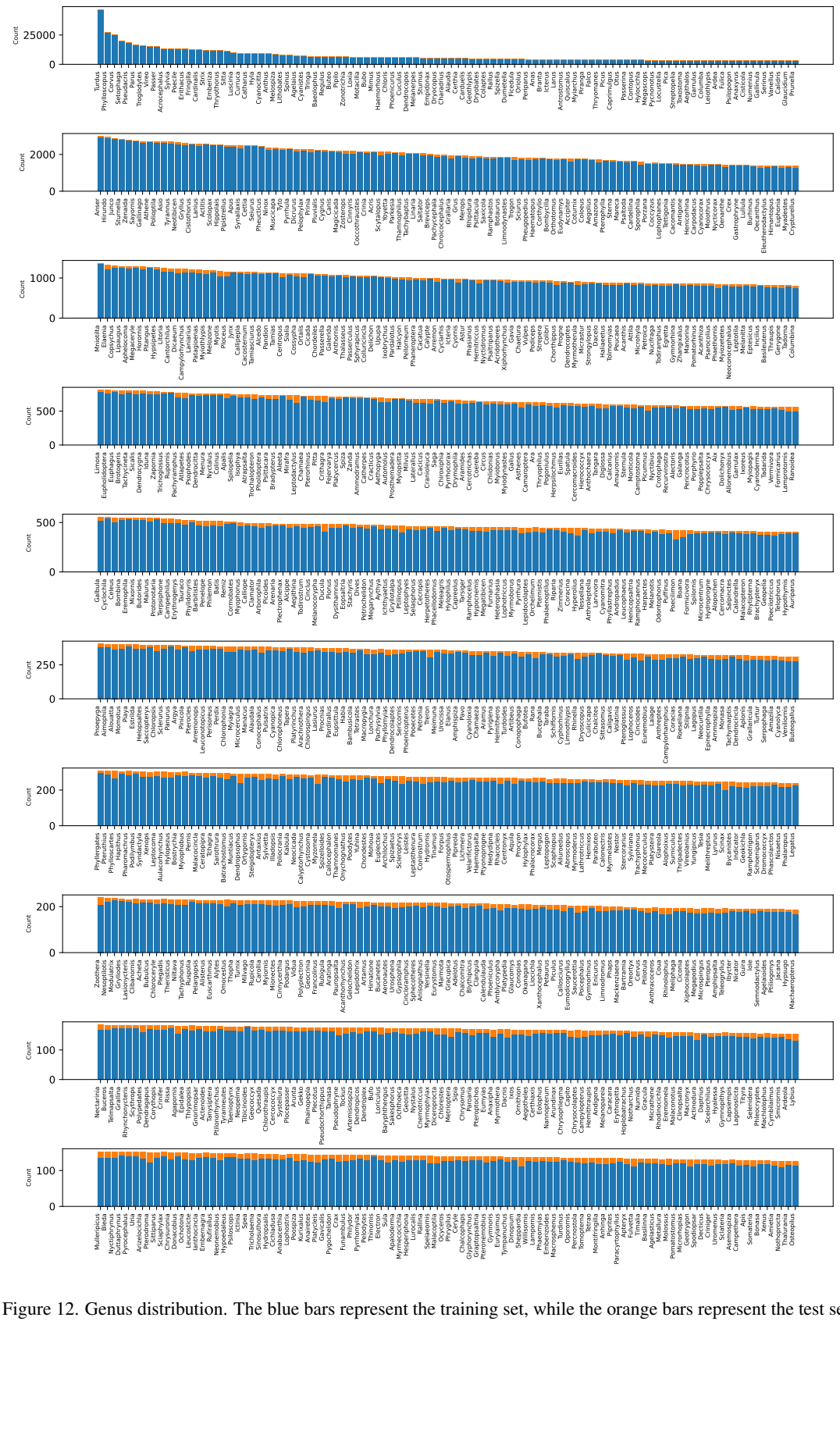

Figure 12. Genus distribution. The blue bars represent the training set, while the orange bars represent the test set.

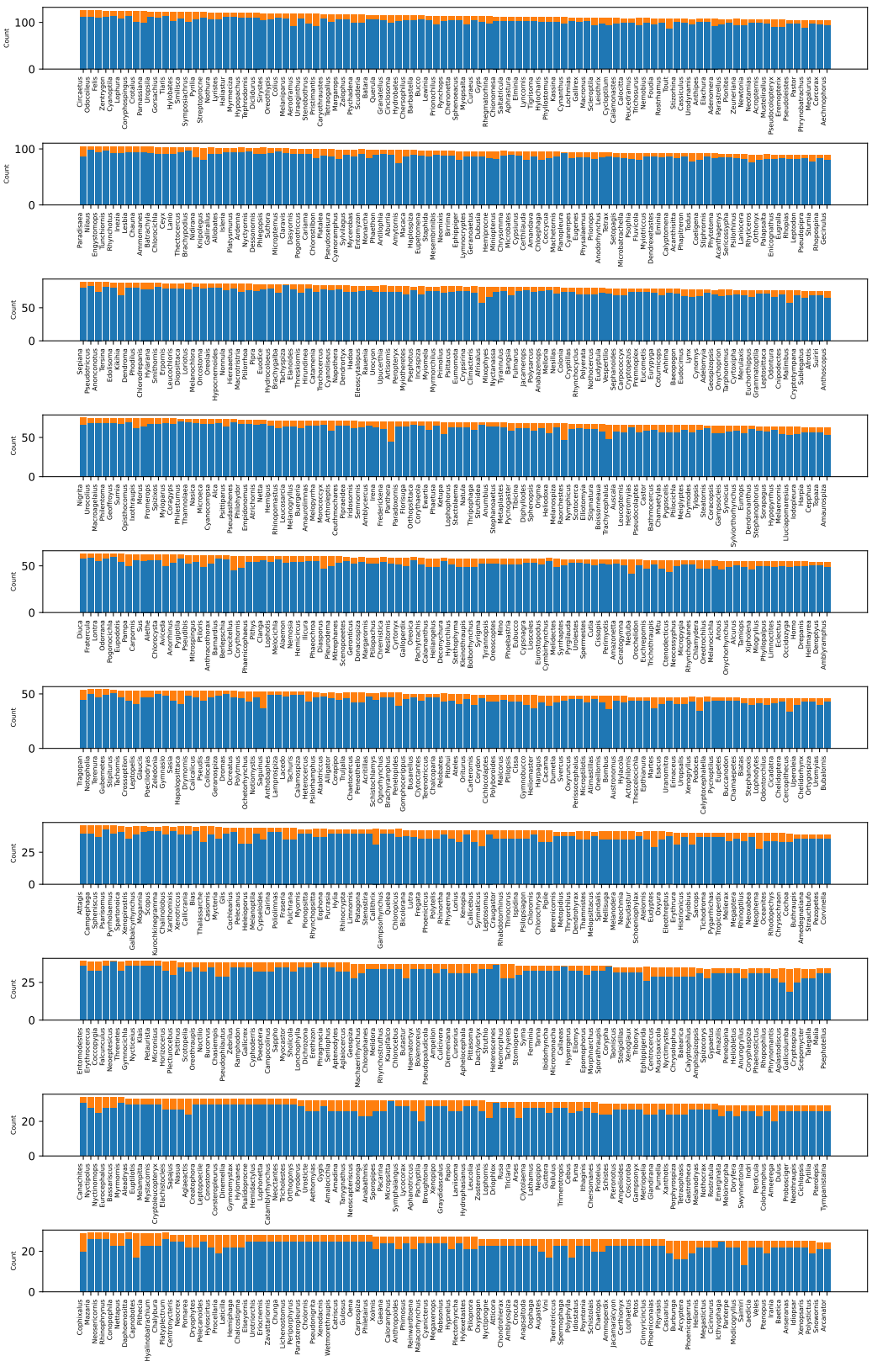

Figure 13. Genus distribution. The blue bars represent the training set, while the orange bars represent the test set.

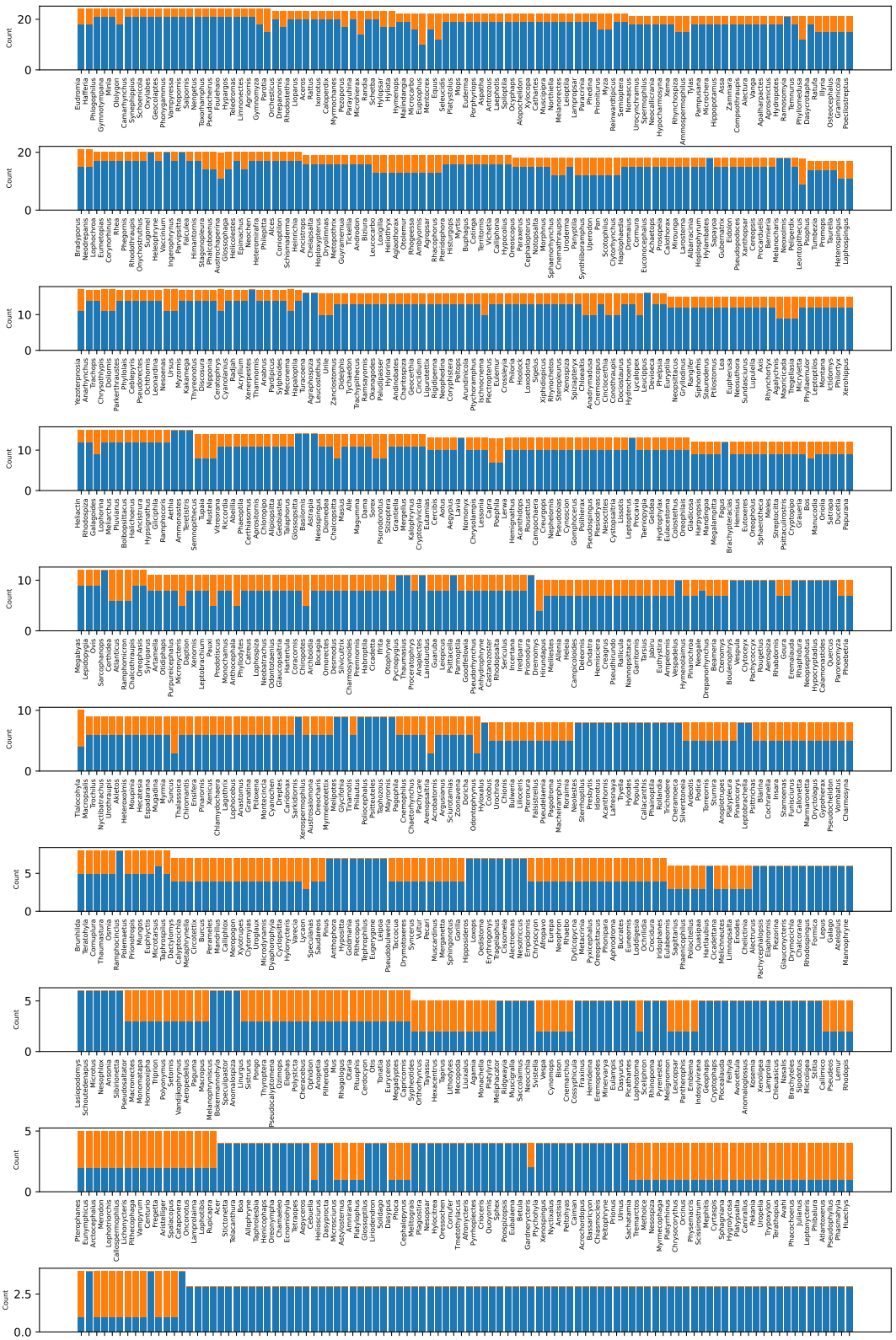

Figure 14. Genus distribution. The blue bars represent the training set, while the orange bars represent the test set.

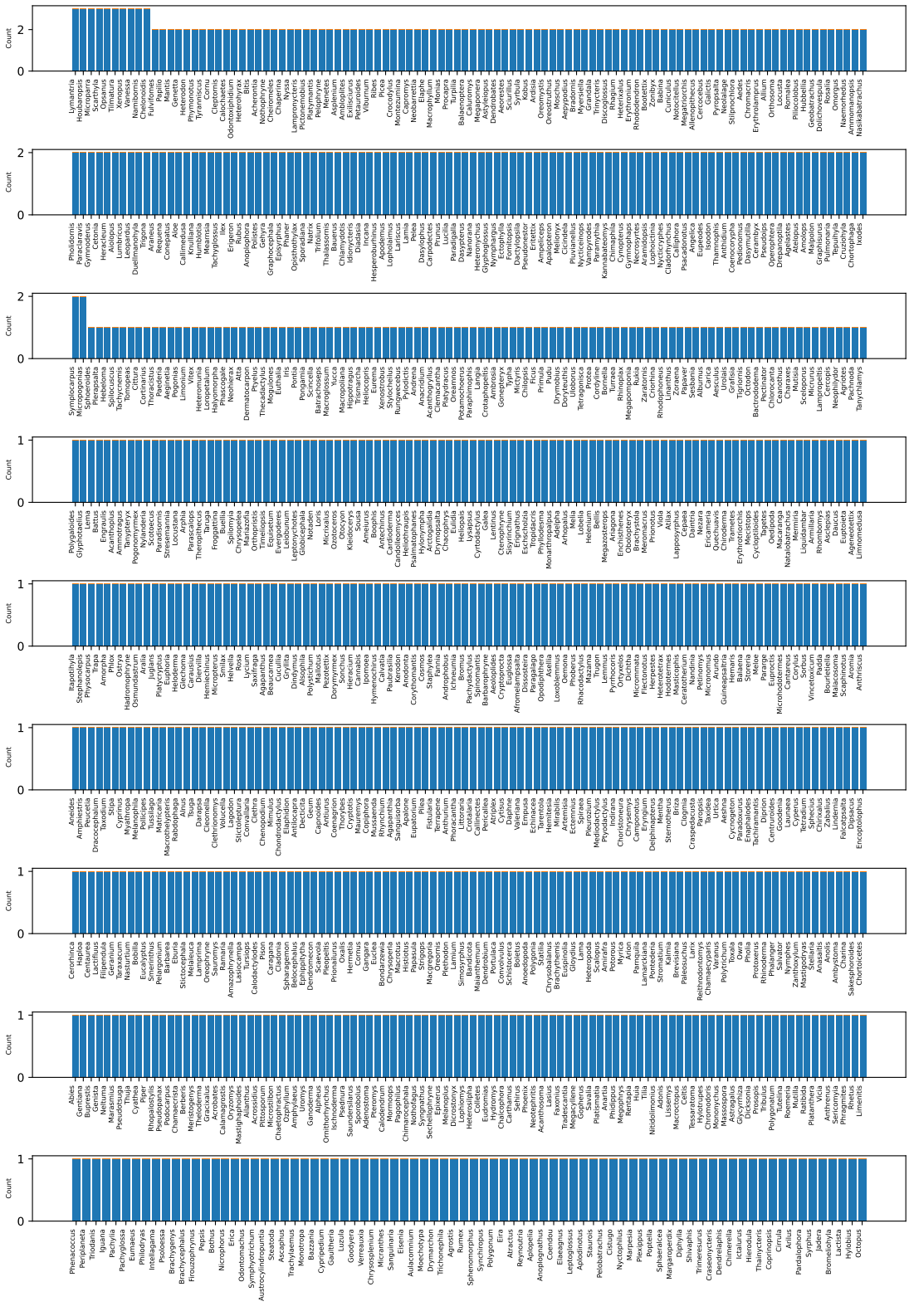

Figure 15. Genus distribution. The blue bars represent the training set, while the orange bars represent the test set.

References

[1] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[2] Jianyang Gu, Sam Stevens, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun (Harry) Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 1

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proc. European Conference on Computer Vision (ECCV)*, pages 38–55, 2025. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 1

[5] David Robinson, Adelaide Robinson, and Lily Akrapongpisak. Transferable models for bioacoustics with human language supervision. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1316–1320, 2024. 2, 3

[6] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Rare species, 2023. 2

[7] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2

[8] 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 *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023. 1