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

Risa Shinoda[1,2] Kaede Shiohara[2] Nakamasa Inoue[3] Kuniaki Saito Hiroaki Santo[1] Fumio Okura[1]
[1]The University of Osaka [2]The University of Tokyo [3]Science Tokyo [4]OMRON SINIC X

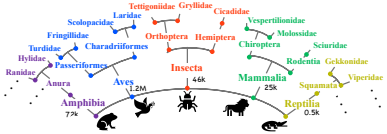
CVPR JUNE 3-7, 2026 DENVER COLORADO

1. Introduction

- Multimodal integration is essential for comprehensive species understanding.
- Existing datasets vary in taxonomy and scale, **requiring a unified multimodal dataset** in a consistent ecological context.

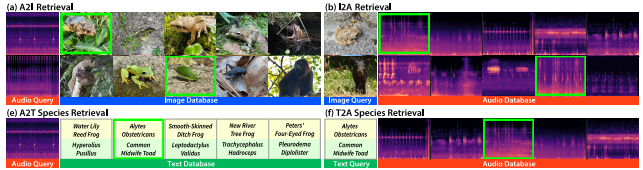
2. Data Collection

- **1.3M** audio clips and **2.3M** images
- Covering **14k species** and **34 fine grained traits**
- All data are collected from publicly available sources (iNaturalist, XenoCanto)



3. BioVITA Benchmark

We propose **BioVITA Bench**, a novel benchmark for cross-modal species-level retrieval across image, text, and audio.



4. BioVITA Model

Stage1 (Audio-Text)

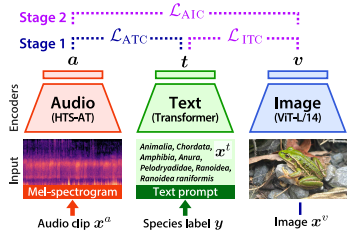
Contrastive learning between audio and species text representations

$$[S_{AT}]_{ij} = a_i^T t_j / \tau$$
$$L_{ATC} = 1/2(l(S_{AT}) + l(S_{AT}^T))$$

Stage2 (VITA)

Joint contrastive learning across audio, image, and text

$$L = L_{ATC} + \lambda(L_{AIC} + L_{ITC})$$



5. Result

- BioVITA shows strong result all retrieval scenarios.
- Stage2 indicates VITA alignment enriches Image-Text representation.

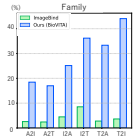
Species level retrieval Acc (100 candidate choices)

	A2T	T2A	A2I	I2A	I2T	T2I
CLIP					50.9	63.8
CLAP	0.7	1.0				
BioCLIP2					65.1	84.8
TaxaBind	9.6	11.8	13.3	16.2	56.9	66.9
Ours (Stage1)	60.3	79.3	47.8	48.6	65.1	84.8
Ours (Stage2)	63.7	81.1	50.3	57.5	86.3	91.2

6. Analysis

Error Analysis

BioVITA model more frequently predicts the correct family despite species-level errors.



Trais Prediction

BioVITA checkpoints enable models to learn ecological traits efficiently.

Category	CLAP	CLIP	BioVITA (Audio)	BioVITA (Image)
Diet Type	43.7	41.5	53.9	41.1
Act. Pattern	80.9	82.1	83.7	82.2
Locomotion	79.7	84.3	81.1	83.1
Lifestyle	82.7	92.5	85.1	89.9
Trophic Role	79.4	89.2	89.7	88.0
Habitat	62.4	58.7	64.9	64.9
Clim. Dist.	76.3	74.4	78.7	75.4
Social Behavior	76.7	78.6	77.6	79.3
Migration	22.4	0.0	45.7	10.6