Audio-to-Image Bird Species Retrieval without Audio-Image Pairs via Text Distillation

Ilyass Moummad
Inria, LIRMM, Université de Montpellier
Montpellier, France
ilyass.moummad@inria.fr

Marius Miron
Earth Species Project
Spain
marius@earthspecies.org

Lukas Rauch
University of Kassel
Germany
lukas.rauch@uni-kassel.de

David Robinson
Earth Species Project
Australia
david@earthspecies.org

Alexis Joly
Inria, LIRMM, Université de Montpellier
Montpellier, France
alexis.joly@inria.fr

Olivier Pietquin
Earth Species Project
Belgium
olivier@earthspecies.org

Emmanuel Chemla
Earth Species Project
France
emmanuel@earthspecies.org

Matthieu Geist
Earth Species Project
France
matthieu@earthspecies.org

***Abstract*—Audio-to-image retrieval offers an interpretable alternative to audio-only classification for bioacoustic species recognition, but learning aligned audio–image representations is challenging due to the scarcity of paired audio–image data. We propose a simple and data-efficient approach that enables audio-to-image retrieval without any audio–image supervision. Our proposed method uses text as a semantic intermediary: we distill the text embedding space of a pretrained image–text model (BioCLIP-2), which encodes rich visual and taxonomic structure, into a pretrained audio–text model (BioLingual) by fine-tuning its audio encoder with a contrastive objective. This distillation transfers visually grounded semantics into the audio representation, inducing emergent alignment between audio and image embeddings without using images during training. We evaluate the resulting model on multiple bioacoustic benchmarks. The distilled audio encoder preserves audio discriminative power while substantially improving audio–text alignment on focal recordings and soundscape datasets. Most importantly, on the SSW60 benchmark, the proposed approach achieves strong audio-to-image retrieval performance exceeding baselines based on zero-shot model combinations or learned mappings between text embeddings, despite not training on paired audio–image data. These results demonstrate that indirect semantic transfer through text is sufficient to induce meaningful audio–image alignment, providing a practical solution for visually grounded species recognition in data-scarce bioacoustic settings.**[1]

***Index Terms*—Bird species recognition, audio-to-image retrieval, cross-modal alignment, contrastive distillation.**

I. Introduction

Automated species recognition from sensory data is a core problem in biodiversity monitoring, ecological research, and wildlife conservation. In recent years, visual species recognition has seen rapid progress driven by large-scale annotated image datasets [1]–[5] and powerful pretrained vision–language models [6], [7]. In contrast, bioacoustic species recognition remains substantially more challenging. Animal vocalizations are highly variable, often corrupted by background noise and recording conditions, and typically lack the rich spatial structure available in images [8]–[14]. As a result, audio-based models are generally trained on smaller datasets and encode weaker semantic structure than their visual counterparts [15].

In this work, we focus on *audio-to-image retrieval* for species recognition, where an audio recording—such as a bird vocalization—is used to retrieve representative images of the corresponding species from a visual corpus. Compared to conventional audio classification, which outputs a discrete label, audio-to-image retrieval produces visually grounded and interpretable results that are particularly useful in real-world biodiversity monitoring. Retrieved images allow human users to visually verify predictions, compare similar species, and reason about uncertain or ambiguous acoustic observations. Beyond its practical relevance, this task poses a fundamental representation learning question: can species-specific acoustic cues be embedded into a visually grounded semantic space, enabling meaningful cross-modal reasoning between sound and vision?

Audio-to-image retrieval is especially appealing from a knowledge transfer perspective. Image-based species recognition models are typically trained on orders of magnitude more data than their audio-based counterparts. For instance, BioCLIP-2 is trained on hundreds of millions of biological images paired with text, spanning hundreds of thousands of taxa across diverse biological groups [7], whereas bioacoustic models such as BioLingual rely on millions of audio recordings covering only thousands of species, primarily limited to vocalizing taxa. As a result, vision–language models encode rich, fine-grained biological semantics that extend well beyond the species observed in audio datasets. Aligning audio representations with this visually grounded embedding space therefore offers a principled way to transfer semantic structure from a strong, data-abundant modality to a weaker, data-scarce one. However, achieving such cross-modal alignment is challenging in practice due to the scarcity of paired audio–image observations.

A key obstacle is the scarcity of paired audio–image observations. In natural settings, animal sounds are frequently

[1]Code is available at: https://github.com/ilyassmoummad/audiobioclip

recorded without visual confirmation, while images rarely include synchronized audio [14]. This lack of paired data makes direct alignment between audio and image encoders difficult. Consequently, existing approaches to multimodal alignment often rely on large-scale multimodal datasets and complex training pipelines that jointly optimize multiple cross-modal objectives [16]–[18]. While effective, these methods are computationally expensive and poorly suited to bioacoustic domains, where data availability and resources are often limited.

An alternative line of work aligns pretrained audio–text and image–text models through their textual representations [19]. Although promising, these approaches typically rely on explicit cross-modal losses, require access to both image and audio data during training, and involve careful balancing of multiple objectives while training, which increases training complexity.

In contrast, we propose a simple and data-efficient approach that enables audio-to-image retrieval *without* paired audio–image data and *without* using images during audio model training. Our key idea is to use text as a semantic intermediary. Modern vision–language models trained on biological data, such as BioCLIP-2 [7], encode rich visual and taxonomic structure in their text embedding spaces. We exploit this property by distilling the text embeddings of a pretrained image–text model into a pretrained audio–text model (BioLingual [15]). Concretely, we fine-tune the audio encoder to match the BioCLIP-2 text embedding space using a contrastive distillation objective, while keeping the image–text model frozen.

Through this distillation process, the audio encoder indirectly acquires visually grounded semantic structure, resulting in *emergent alignment* between audio and image representations. After training, audio recordings can be projected directly into the shared image–text embedding space, enabling audio-to-image retrieval via simple similarity search. Importantly, this alignment emerges without explicit audio–image supervision or multi-objective cross-modal training, demonstrating that strong audio–image retrieval capabilities can arise from indirect semantic transfer through text alone.

Our contributions are summarized as follows:

- We formulate audio-to-image retrieval as a practical and interpretable task for bioacoustic species recognition, highlighting its relevance for biodiversity monitoring.
- We introduce a simple contrastive distillation framework that transfers visually grounded semantic structure into audio representations via text, without requiring paired audio–image data or image inputs during training.
- We demonstrate that this approach induces emergent audio–image alignment and enables strong species-level audio-to-image retrieval, outperforming baselines based on zero-shot model combinations or text embedding mappings.
- We show that the resulting audio model preserves audio discriminative capabilities while improving text-to-audio retrieval on both focal and soundscape recordings, on average.

II. Related Work

A. Multimodal Models for Species Recognition

Multimodal representation learning has become increasingly important for species recognition and biodiversity analysis [6], [7], [15], [18], [20]. Vision–language models such as BioCLIP and BioCLIP-2 are trained on large-scale biological image–text pairs and encode rich, fine-grained visual and taxonomic semantics that support downstream tasks, including species classification and retrieval. In parallel, audio–language models such as BioLingual learn semantic representations of bioacoustic signals through contrastive alignment with textual descriptions. While these models align each modality to text, their embedding spaces are not directly aligned across modalities, limiting their ability to support cross-modal retrieval tasks such as audio-to-image retrieval.

B. Cross-Modal Alignment and Multimodal Embeddings

Several works aim to align multiple modalities into a shared embedding space. ImageBind [16], LanguageBind [17], and TaxaBind [18] unify modalities such as image, audio, and text by jointly training on large multimodal datasets where different modality pairs share a common anchor. These approaches achieve strong cross-modal alignment but rely on extensive paired data across modalities and complex joint training procedures. In bioacoustic settings, where paired audio–image observations are rare and data collection is costly, such requirements are often impractical.

C. Audio–Image Alignment via Text

More closely related to our work, recent approaches attempt to bridge audio and image modalities through text. Multimodal Contrastive Retrieval (MCR) [19] explicitly aligns pretrained image–text and audio–text models by optimizing multiple intra- and inter-modal contrastive objectives. While effective, this framework requires access to both audio and image data during training, as well as careful balancing of multiple losses over large-scale datasets, which increases computational complexity and limits applicability in resource-constrained domains.

D. Positioning of Our Approach

In contrast to existing methods, our approach enables audio-to-image retrieval without requiring paired audio–image data, explicit cross-modal losses, or image inputs during audio model training. A key enabling factor is that the pretrained audio–text and image–text models are grounded in a shared textual vocabulary, notably species names and taxonomic descriptions, which provides a common semantic anchor across modalities.

We leverage the observation that text embedding spaces of vision–language models encode visually grounded semantic and taxonomic structure. By distilling the text embeddings of a pretrained image–text model (BioCLIP-2) into a pretrained audio–text model (BioLingual), we transfer visual semantic structure into the audio representation through a simple contrastive distillation objective. This transfer induces emergent

alignment between audio and image embeddings, enabling data-efficient audio-to-image retrieval using only pretrained models and audio–text data.

III. Method

A. Problem Formulation

Let $\mathcal{A}$, $\mathcal{I}$, and $\mathcal{T}$ denote the input domains of audio recordings, images, and textual descriptions, respectively. Our goal is to learn audio representations that are aligned with image representations without audio–image pairs, enabling cross-modal tasks such as audio-to-image retrieval.

We assume access to pretrained audio–text and image–text models whose text encoders are grounded in overlapping semantic vocabularies, such as shared species names and taxonomic descriptions, though not necessarily identical label sets. This shared textual grounding provides a semantic bridge that enables indirect alignment between audio and image representations via text.

Rather than attempting to align audio and image modalities directly, we exploit the fact that modern multimodal models independently align audio and images to text. We treat text as a semantic intermediary and transfer visually grounded semantic structure from an image–text model into an audio encoder, thereby inducing indirect alignment between audio and image embeddings.

B. Pretrained Multimodal Models

BioCLIP-2 (Image–Text). BioCLIP-2 is a vision–language model trained via contrastive learning on large-scale biological image–text pairs. It consists of an image encoder $f_I : \mathcal{I} \to \mathbb{R}^{d_I}$ and a text encoder $f_T^I : \mathcal{T} \to \mathbb{R}^{d_I}$. Textual descriptions used during training typically include taxonomic information spanning multiple biological ranks (e.g., species, genus, family), resulting in a text embedding space that captures rich visual and hierarchical biological semantics shared with the image embeddings.

BioLingual (Audio–Text). BioLingual is an audio–language model trained via contrastive alignment between bioacoustic recordings and textual descriptions. It consists of an audio encoder $f_A : \mathcal{A} \to \mathbb{R}^{d_A}$ and a text encoder $f_T^A : \mathcal{T} \to \mathbb{R}^{d_A}$. While BioLingual encodes semantic structure relevant to bioacoustics, its embedding space is not aligned with that of BioCLIP-2, preventing direct comparison between audio and image representations.

C. Visually Grounded Semantics in Text

A key observation underlying our approach is that the text embedding space of BioCLIP-2 is not merely a label space, but a carrier of visually grounded and taxonomically structured semantics. For example, textual descriptions of the form:

> *"Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia (black-billed magpie)."*

encourage the model to organize text embeddings such that biologically related species occupy nearby regions. Because BioCLIP-2 aligns text and images in a shared space, this structure is also reflected in its image embeddings. We leverage this property to transfer visual semantic structure into the audio modality.

D. Contrastive Distillation into the Audio Encoder

To align audio representations with the BioCLIP-2 embedding space, we distill BioCLIP-2 text embeddings into the BioLingual audio encoder. Since the two models operate in different embedding dimensions ($d_A \neq d_I$), we introduce a learnable linear projection head:

$$g : \mathbb{R}^{d_A} \to \mathbb{R}^{d_I}. \tag{1}$$

Given a batch of N audio recordings $\{a_i\}_{i=1}^N$ and their associated textual descriptions $\{t_i\}_{i=1}^N$, we compute:

$$\mathbf{z}_i^A = g(f_A(a_i)), \tag{2}$$

$$\mathbf{z}_i^T = f_T^I(t_i), \tag{3}$$

where $\mathbf{z}_i^A \in \mathbb{R}^{d_I}$ is the projected audio embedding and $\mathbf{z}_i^T \in \mathbb{R}^{d_I}$ is the corresponding BioCLIP-2 text embedding.

We optimize a contrastive distillation objective that encourages each audio embedding to match its corresponding BioCLIP-2 text embedding:

$$\mathcal{L}_{\text{distill}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_i^A, \mathbf{z}_i^T)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^A, \mathbf{z}_j^T)/\tau)}, \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

During training, gradients are applied only to the BioLingual audio encoder f_A and the projection head g. The text encoder of the BioCLIP-2 model is entirely frozen. This one-sided optimization transfers visually grounded semantic structure into the audio representation without modifying the image–text model or using any image data. Fig. 1 illustrates our distillation pipeline.

E. emergent Audio–Image Alignment

Because BioCLIP-2 aligns images and text in a shared embedding space, aligning audio embeddings to BioCLIP-2 text embeddings implicitly aligns audio and image representations. After training, semantically related audio, text, and images satisfy:

$$g(f_A(a)) \approx f_T^I(t) \approx f_I(i). \tag{5}$$

As a result, audio recordings are embedded according to the visual and taxonomic semantics encoded by BioCLIP-2. Audio samples whose textual descriptions share higher-level biological attributes (e.g., order or family) are placed near images of species with the same attributes, even when fine-grained acoustic cues are weak or ambiguous. This emergent structure enables meaningful audio-to-image retrieval without explicit audio–image supervision.

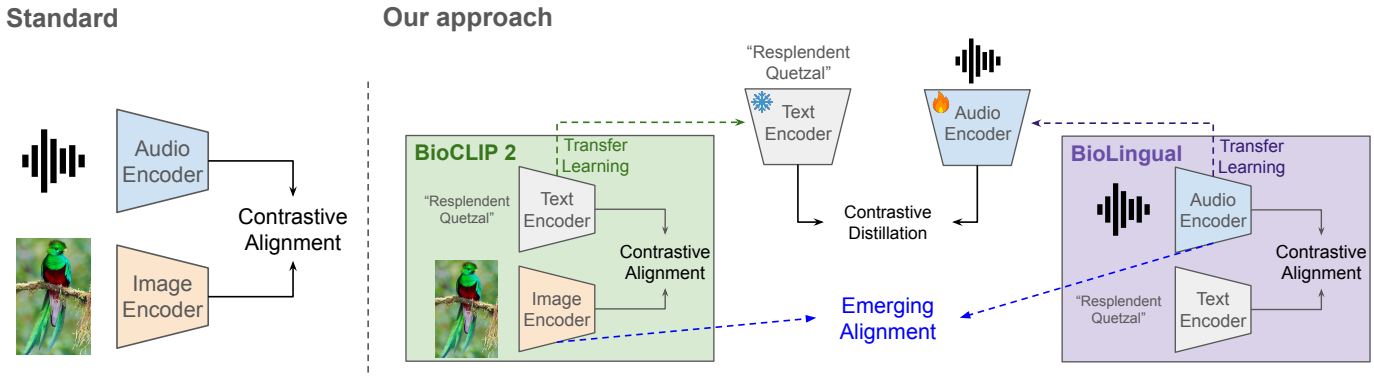

Fig. 1. Comparison of audio–image alignment approaches. Left: direct contrastive alignment using paired audio–image data. Right (ours): text-based distillation from BioCLIP-2 into the BioLingual audio encoder, inducing emergent audio–image alignment without audio–image training pairs.

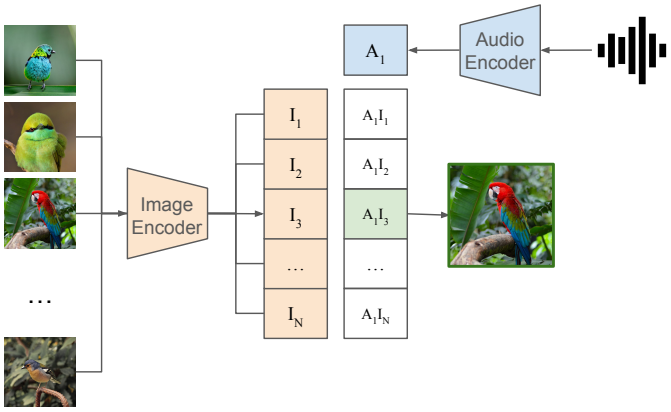

Fig. 2. Audio-to-image retrieval: Images are embedded using BioCLIP-2 image encoder, while the audio query is embedded using BioLingual-FT audio encoder. Retrieval is performed by ranking images via cosine similarity in the shared embedding space.

TABLE I
DATASETS USED FOR TRAINING AND EVALUATION. ZS DENOTES ZERO-SHOT CLASSIFICATION, T→A TEXT-TO-AUDIO RETRIEVAL, AND A→I AUDIO-TO-IMAGE RETRIEVAL.

Dataset	# Classes	Type	Usage
iNatSounds [21]	5,569	Focal	Training, ZS, T→A
SSW60 [14]	60	Focal	kNN, A→I
HSN [8]	24	Soundscape	ZS, T→A
NES [9]	51	Soundscape	ZS, T→A
PER [10]	21	Soundscape	ZS, T→A
SNE [11]	20	Soundscape	ZS, T→A
SSW [12]	35	Soundscape	ZS, T→A
UHH [13]	24	Soundscape	ZS, T→A

F. Audio-to-Image Retrieval

At inference time, text encoders are discarded. Audio recordings are embedded using the fine-tuned audio encoder followed by the projection head, while images are embedded using the frozen BioCLIP-2 image encoder. Audio-to-image retrieval is performed by ranking images according to cosine similarity with the audio query embedding, as illustrated in Fig. 2.

IV. EXPERIMENTS

A. Experimental Setup

Training protocol. We initialize our model from the pretrained audio encoder of BioLingual and fine-tune it using the contrastive distillation objective described in Sec. III. During training, audio embeddings are aligned to the frozen BioCLIP-2 text embedding space. Training uses only audio–text pairs from the iNatSounds dataset [21]; no image data or paired audio–image observations are used at any stage.

For each species, textual descriptions are sampled at random from a mixture of text types supported by BioCLIP-2, including common names, scientific (Latin) names, and taxonomic descriptions (optionally combined with common names). At evaluation time, we use common-name prompts for consistency across datasets and tasks.

Evaluation. We evaluate the resulting model, denoted *BioLingual-FT*, along three complementary dimensions:

1) emergent audio–image alignment,
2) preservation of audio-specific representations, and
3) quality of audio–text semantic alignment.

Unless otherwise specified, all comparisons are against the original pretrained BioLingual model.

Datasets. We evaluate our approach across multiple bioacoustic benchmarks covering both focal recordings and soundscapes. A summary of datasets and their usage is provided in Table I.

B. Audio-to-Image Retrieval

We first evaluate the central task of this work: audio-to-image retrieval on SSW60. Image embeddings are computed using the frozen BioCLIP-2 image encoder, while audio queries are embedded using the fine-tuned BioLingual-FT audio encoder.

We compare our approach against several baselines that reflect different strategies for bridging audio and image modalities without paired audio–image supervision:

Random Projection applies a random linear projection to BioLingual audio embeddings to match the dimensionality of BioCLIP-2 image embeddings and serves as a lower bound.

TABLE II
MEAN AVERAGE PRECISION (MAP) FOR AUDIO-TO-IMAGE RETRIEVAL ON THE SSW60 DATASET.

Method	mAP
Random Projection	3.79
Text Embeddings Mapping	51.39
Cascaded Zero-Shot (Image+Audio)	39.85
BioLingual-FT (Ours)	**70.47**

Text Embeddings Mapping learns a non-linear mapping between the text embedding spaces of BioLingual and BioCLIP-2 via the contrastive loss using iNatSounds annotations spanning 5,569 classes. Text captions of iNatSounds classes are embedded with BioLingual and BioCLIP-2 text encoders and then are aligned via the contrastive loss. This baseline tests whether translating BioLingual's text embedding to BioCLIP-2's text embedding without audio is sufficient for cross-modal retrieval.

Cascaded Zero-Shot decomposes audio-to-image retrieval into two independent zero-shot classification steps: audio is first classified into one of 3,846 bird classes from iNatSounds using BioLingual, and images are independently classified using BioCLIP-2. Retrieval is then performed by matching predicted class embeddings. This baseline reflects a common cascade approach but is susceptible to error propagation across stages.

BioLingual-FT, trained via our proposed text-based distillation strategy, substantially outperforms baselines, despite never observing images or paired audio–image data during training. These results demonstrate that strong audio–image alignment can emerge purely through indirect semantic transfer via text, validating the effectiveness and data efficiency of the proposed method.

C. Preservation of Audio Representations

A key concern when distilling visual semantics into an audio encoder is potential degradation of audio-discriminative features. To assess this, we evaluate k-nearest neighbor (kNN) classification accuracy on the SSW60 dataset, a standard benchmark for fine-grained categorization of bird species.

TABLE III
KNN CLASSIFICATION ACCURACY ON SSW60.

Model	Accuracy
BioLingual	77.37
BioLingual-FT	77.29

As shown in Table III, BioLingual-FT matches the performance of the original BioLingual model, indicating that the proposed distillation preserves core audio representations while injecting additional semantic structure.

D. Audio–Text Alignment on Focal Recordings

We next evaluate audio–text alignment on focal recordings from iNatSounds. We report both text-to-audio retrieval performance and zero-shot classification accuracy, which directly probe the semantic consistency between audio and textual embeddings.

TABLE IV
INATSOUNDS TEXT-TO-AUDIO RETRIEVAL (MAP@1000) AND ZERO-SHOT CLASSIFICATION ACCURACY ON VALIDATION AND TEST SPLITS.

Model	Text→Audio		Zero-Shot	
	Val	Test	Val	Test
BioLingual	60.89	58.77	38.45	36.29
BioLingual-FT	**76.45**	**77.97**	**64.88**	**63.32**

BioLingual-FT substantially outperforms the base model, improving retrieval performance by over 19% in mAP and nearly doubling zero-shot classification accuracy. These gains demonstrate that aligning audio embeddings to the BioCLIP-2 text space significantly strengthens semantic grounding for focal bioacoustic recordings.

A potential concern is that the strong gains observed on iNatSounds could partially stem from incidental overlap between the species vocabulary or underlying observations used to train BioCLIP-2 (on images) and the iNatSounds audio recordings. While exact audio–image overlap is highly unlikely in practice, both datasets are drawn from large biodiversity repositories and may share species-level semantic context.

To assess whether our improvements reflect genuine audio–text alignment rather than dataset-specific effects, we evaluate both the original BioLingual model and BioLingual-FT on the CBI dataset [22], a focal bioacoustic dataset that is disjoint from iNatSounds and not used during training.

TABLE V
ZERO-SHOT CLASSIFICATION ACCURACY ON THE CBI DATASET.

Model	Zero-Shot Acc. (%)	A→I mAP (%)
BioLingual	58.43	51.44
BioLingual-FT	**58.77**	**72.30**

As shown in Table V, BioLingual-FT substantially improves text-to-audio retrieval on CBI (+20.9 mAP) while preserving zero-shot classification accuracy. Because CBI is independent of iNatSounds and was not used during training, these results indicate that the audio–text improvement on focal recordings does not stem from dataset overlap.

E. Audio–Text Alignment on Soundscape Recordings

We further test generalization to real-world soundscapes, which are more challenging due to background noise and overlapping species. Table VI reports zero-shot classification accuracy and text-to-audio retrieval across six soundscape datasets.

While zero-shot classification accuracy shows mixed behavior across datasets, BioLingual-FT yields a large average mAP increase in text-to-audio-retrieval. This suggests that the distilled model prioritizes semantic alignment over strict class discrimination, a trade-off that is beneficial for retrieval-based and exploratory tasks.

TABLE VI
SOUNDSCAPE ZERO-SHOT CLASSIFICATION ACCURACY AND TEXT-TO-AUDIO RETRIEVAL (MAP@1000).

Dataset	Zero-Shot (Acc.)		Text→Audio (mAP)	
	BioLingual	BioLingual-FT	BioLingual	BioLingual-FT
HSN	**19.69**	17.93	27.85	**45.40**
NES	**32.18**	18.46	24.32	**33.50**
PER	**7.68**	2.95	**9.73**	6.43
SNE	42.70	**42.89**	21.49	**42.75**
SSW	35.98	**47.13**	33.64	**59.29**
UHH	**16.65**	8.34	**32.00**	27.54
AVG	25.81	22.95	24.83	35.81

V. CONCLUSION

We introduced a simple and data-efficient approach for audio-to-image species retrieval in bioacoustic settings where paired audio–image data are unavailable. By using text as a semantic intermediary, we distill visual–text embeddings from BioCLIP-2 into a pretrained BioLingual audio encoder, inducing direct audio–image alignment without requiring image supervision.

Experiments show that the proposed model achieves strong audio-to-image retrieval performance, substantially outperforming zero-shot and text-embedding mapping baselines. These results demonstrate that text-based distillation alone is sufficient to induce effective audio–image alignment, providing a practical and lightweight solution for visually grounded bioacoustic analysis in data-scarce biodiversity applications. Importantly, this emergent cross-modal alignment does not come at the expense of audio modeling: the distilled audio encoder preserves audio discriminative capabilities and improves text-to-audio retrieval.

REFERENCES

[1] F. E. JAIMI, W. RABHI, W. AMARA, Z. CHAROUH, H. BENABOUD, and M. B. SAINDOU, "Lasbird: Large scale bird recognition dataset," 2023. [Online]. Available: https://dx.doi.org/10.21227/s3xd-2s66

[2] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, and Y. Su, "TreeOfLife-10M," 2023. [Online]. Available: https://huggingface.co/datasets/imageomics/TreeOfLife-10M

[3] J. Gu, S. Stevens, E. G. Campolongo, M. J. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. E. White, J. Balhoff, W. M. Dahdul, D. Rubenstein, H. Lapp, T. Berger-Wolf, W.-L. Chao, and Y. Su, "TreeOfLife-200M (revision a8f38b4)," 2025. [Online]. Available: https://huggingface.co/datasets/imageomics/TreeOfLife-200M

[4] G. V. Horn and macaodha, "iNat Challenge 2021 - FGVC8," https://kaggle.com/competitions/inaturalist-2021, 2021, kaggle.

[5] E. Vendrow, O. Pantazis, A. Shepard, G. Brostow, K. Jones, O. Mac Aodha, S. Beery, and G. Van Horn, "INQUIRE: A Natural World Text-to-Image Retrieval Benchmark," *Advances in Neural Information Processing Systems*, vol. 37, pp. 126 500–126 514, 2024.

[6] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. 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*, 2024, pp. 19 412–19 424.

[7] J. Gu, S. Stevens, E. G. Campolongo, M. J. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. E. White, J. Balhoff *et al.*, "BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning," *arXiv preprint arXiv:2505.23883*, 2025.

[8] M. Clapp, S. Kahl, E. Meyer, M. McKenna, H. Klinck, and G. Patricelli, "A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7525805

[9] A. Vega-Hidalgo, S. Kahl, L. B. Symes, V. Ruiz-Gutiérrez, I. Molina-Mora, F. Cediel, L. Sandoval, and H. Klinck, "A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7525349

[10] W. A. Hopping, S. Kahl, and H. Klinck, "A collection of fully-annotated soundscape recordings from the southwestern amazon basin," 10 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7079124

[11] S. Kahl, C. M. Wood, P. Chaon, M. Z. Peery, and H. Klinck, "A collection of fully-annotated soundscape recordings from the western united states," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7050014

[12] S. Kahl, R. Charif, and H. Klinck, "A collection of fully-annotated soundscape recordings from the northeastern united states," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7079380

[13] A. Navine, S. Kahl, A. Tanimoto-Johnson, H. Klinck, and P. Hart, "A collection of fully-annotated soundscape recordings from the island of hawaii," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7078499

[14] G. Van Horn, R. Qian, K. Wilber, H. Adam, O. Mac Aodha, and S. Belongie, "Exploring fine-grained audiovisual categorization with the SSW60 dataset," in *European Conference on Computer Vision*. Springer, 2022, pp. 271–289.

[15] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable Models for Bioacoustics with Human Language Supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1316–1320.

[16] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One Embedding Space To Bind Them All," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

[17] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, W. HongFa, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. W. Zhang, Z. Li, W. Liu, and L. Yuan, "LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=QmZKc7UZCy

[18] S. Sastry, S. Khanal, A. Dhakal, A. Ahmad, and N. Jacobs, "TaxaBind: A Unified Embedding Space for Ecological Applications," in *Winter Conference on Applications of Computer Vision*. IEEE/CVF, 2025.

[19] Z. Wang, Y. Zhao, X. Cheng, H. Huang, J. Liu, A. Yin, L. Tang, L. Li, Y. Wang, Z. Zhang, and Z. Zhao, "Connecting Multi-modal Contrastive Representations," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=IGTbT9P1ti

[20] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, P. Bonnet, and A. Joly, "GeoPlant: Spatial Plant Species Prediction Dataset," in *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[21] M. Chasmai, A. Shepard, S. Maji, and G. Van Horn, "The iNaturalist Sounds Dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 132 524–132 544, 2024.

[22] A. Howard, H. Klinck, S. Dane, S. Kahl, tom denton, and T. Denton, "Cornell Birdcall Identification," https://kaggle.com/competitions/birdsong-recognition, 2020, kaggle.