Model Merging Enables In-Context Learning for Bioacoustics Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

General-purpose foundation models capable of generalizing across species and tasks represent a promising new frontier in bioacoustics, with NatureLM-audio being one of the most prominent examples. While its domain-specific finetuning yields strong performance on bioacoustic benchmarks, we observe that it also introduces trade-offs in instruction-following flexibility. For instance, NatureLM-audio achieves high accuracy when prompted for either the common or scientific name individually, but its accuracy drops significantly when both are requested in a single prompt. These effects limit zero- and few-shot generalization to novel tasks. We address this by applying a simple model merging strategy that interpolates NatureLM-audio with its base language model, recovering instruction-following capabilities with minimal loss of domain expertise. Finally, we show that this enables effective few-shot in-context learning, a key capability for real-world scenarios where labeled data for new species or environments are scarce.

1 Introduction

Bioacoustics, the study of sound production, transmission, and perception in animals, is a critical tool for understanding biodiversity, monitoring ecosystems, and informing conservation efforts [26, 24, 27]. Recent advances in machine learning (ML) have transformed the field, enabling automated detection, classification, and analysis of acoustic events at unprecedented scales [44].

Early work in ML for bioacoustics typically relied on *species-specific models*, trained and optimized for a single species and task [1]. However, as in other domains of ML, there is now a shift towards *general-purpose foundation models* that can support a broad range of downstream species and/or tasks with minimal retraining [23, 12, 15, 36, 37, 46]. One of the most prominent examples of this trend is NatureLM-audio [38], the first bioacoustics audio–language model, designed for zero-shot generalization to unseen tasks via text-based prompting.

In this paper, we examine the capabilities of NatureLM-audio as a *general* foundation model for bioacoustics. Despite its strong performance on tasks and prompts closely matching its training distribution, we find that its intense domain-specific finetuning has led to a severe reduction in the *instruction-following capabilities* of its base model (Llama-3.1-8B-Instruct), a trade-off commonly observed in other specialized models [54]. This limits its ability to generalize in zero- or few-shot settings to new tasks. We show that *model merging* with the base model can help restore these capabilities, achieving a balance between domain-specific knowledge and general instruction-following. Finally, we demonstrate that this restoration enables NatureLM-audio to perform *few-shot in-context learning*, a scenario of particular importance in bioacoustics where practitioners often have only a handful of labeled examples for new species, habitats, or acoustic conditions.

Common Name Prompt

```
What is the common name for the focal species
in the audio?
```

Scientific Name Prompt

```
What is the scientific name for the focal
species in the audio?
```

Combined Prompt

```
Identify the focal species in the audio and
provide its scientific name, followed by a
colon and its common name.
```

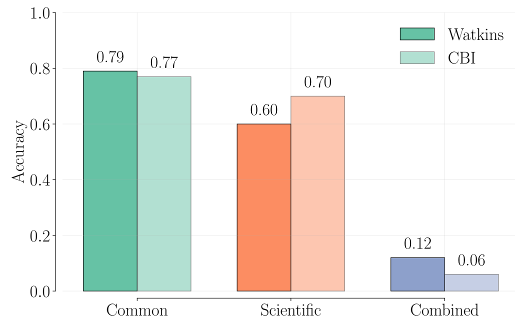


Figure 1: NATURELM-AUDIO classification accuracy for different prompts on WATKINS and CBI.

id aecd4535-ebef-438d-811b-1ffb7be5c22e

```
GT: Odobenus rosmarus: Walrus
Common: Walrus
Scientific: Odobenus rosmarus
Combined: Odobenus rosmarus - male courtship ...
```

id 4cbdd583-23c6-408d-aea5-3719dc6d9654

```
GT: Lagenodelphis hosei: Frasers Dolphin
Common: Fraser's Dolphin
Scientific: Lagenodelphis hosei
Combined: Lagenodelphis hosei
```

id 185c974a-d905-475c-8fc0-0cbab1e383b9

```
GT: Eubalaena australis: Southern Right Whale
Common: Fin- Finback Whale
Scientific: Balaenoptera physalus
Combined: Balaenoptera physalus: 52 Hz Pulses
```

id 4186fc55-0ac2-4f44-9026-009f239fcf96

```
GT: Stenella clymene: Clymene Dolphin
Common: Clymene Dolphin
Scientific: Stenella clymene
Combined: Stenella clymene: Clymene Dolphin
```

Figure 2: Example model predictions for the common name, scientific name, and combined-name prompts, compared to ground truth. Correct predictions in green, incorrect in red.

2 Problem Analysis

NATURELM-AUDIO is a LoRA [18] finetuning of LLAMA-3.1-8B-INSTRUCT [13] on ∼2M steps of audio–text pairs, predominantly bioacoustics but also music and human sounds. While the original evaluation shows that the model follows training-like instructions well, such as predicting either the common or the scientific name of the focal species in an audio in isolation, we find that requesting both in a single prompt often leads to a substantial drop in accuracy. Figure 1 shows the exact prompts and corresponding accuracies on WATKINS and CBI, two species-classification datasets from the BEANS benchmark [16] covering marine mammals and birds, respectively. On both datasets, the model performs slightly better on common names than on scientific names, yet achieves high accuracy (60–80%) in both cases. However, when prompted for both names jointly, accuracy falls to 6–12%.

The examples in Figure 2 illustrate typical failure modes. In the top left, the model outputs correct common and scientific names individually, but under the combined prompt it drifts into behavioural description ("male courtship behavior"), possibly reflecting its exposure to captioning-style data during training [38]. In the bottom left, it misidentifies the species in all cases, yet common and scientific predictions remain mutually consistent; in the combined case it again appends unrelated context ("52 Hz pulses"). In the top right, correct individual predictions degrade to only the scientific name under the combined prompt. In the bottom right, all three prompts succeed.

We additionally experiment with the ZF-INDIV dataset originally used in [38] to evaluate zero-shot task generalization (see Appendix B.2) and observe a similar pattern: NATURELM-AUDIO shows reduced robustness to even mild prompt variations. This behaviour is consistent with the effects of extensive domain-specific finetuning observed in other specialized LLMs, where overfitting to training prompt formats can narrow instruction-following flexibility and limit generalization [54].

3 Method

Section 2 shows that NATURELM-AUDIO has lost its instruction following capabilities in favor of task-specific ones acquired during finetuning. We recover these ones through model merging.

Model Merging Model merging aims to ensemble different models without incurring in additional inference or storage costs [2, 11, 49, 5]. While the non-linear nature of neural networks prevents from taking the weighted average of the models in general [48], this aggregation is well behaved when the two models exhibit linear mode connectivity [9], *i.e.* can be connected via a linear path over which the loss does not significantly increase. In this case, the merged model $\mathbf{\Theta}^{(\text{merge})}$ can be obtained from the endpoint models $\mathbf{\Theta}^{(1)}, \mathbf{\Theta}^{(2)}$ simply as $\mathbf{\Theta}^{(\text{merge})} = (1-\alpha)\mathbf{\Theta}^{(1)} + \alpha\mathbf{\Theta}^{(2)}$, where $\alpha \in [0,1]$ is a scaling parameter controlling the contribution of each model. Consistent with previous findings [48, 32, 9], we observe that linear interpolation remains effective along the finetuning trajectory, suggesting that linear mode connectivity holds when (part of) the optimization path is shared.

Merging NATURELM-AUDIO with its base model We merge LLAMA-3.1-8B-INSTRUCT with its finetuning NATURELM-AUDIO to combine the instruction following abilities of the former and the task-specific performance of the latter. In particular, being NATURELM-AUDIO a LoRa [18] finetuning, linearly interpolating between the base and the finetuned is equivalent to changing the multiplicative factor α in LoRa: Given the weight matrix $\mathbf{W}_{\text{base}}$ of the base model for some layer, LoRA updates it as $\mathbf{W}_{\text{ft}} = \mathbf{W}_{\text{base}} + \mathbf{AB}$, where $\mathbf{A}$ and $\mathbf{B}$ are two low-rank learnable matrices; thus

$$(1-\alpha)\,\mathbf{W}_{\text{base}} + \alpha\,\mathbf{W}_{\text{ft}} = (1-\alpha)\,\mathbf{W}_{\text{base}} + \alpha\,(\mathbf{W}_{\text{base}} + \mathbf{AB}) \tag{1}$$

$$= \mathbf{W}_{\text{base}} - \cancel{\alpha\,\mathbf{W}_{\text{base}}} + \cancel{\alpha\,\mathbf{W}_{\text{base}}} + \alpha\,\mathbf{AB} = \mathbf{W}_{\text{base}} + \alpha\,\mathbf{AB}. \tag{2}$$

This shows that we can interpolate between the base and the finetuned model simply by varying α.

4 Results

Combined Instruction-Following Task We evaluate the merged model over a range of interpolation coefficients α, using the three prompt variants in Figure 1. The y-axis in Figure 3 reports the accuracy on the *combined* prompt, while the x-axis shows the mean accuracy on the *training-like* prompts (common name and scientific name individually). For the combined prompt, intermediate interpolation values substantially outperform both extremes. Specifically, rescaling from $\alpha = 1$ (NATURELM-AUDIO) to $\alpha \approx 0.7$ increases combined-task accuracy from $6\% \rightarrow 45\%$ on Watkins and $12\% \rightarrow 63\%$ on CBI, reflecting a restoration of instruction-following capabilities degraded in the finetuned model. However, setting α too low ($\alpha < 0.5$) sharply reduces accuracy on combined prompts due to a loss of domain-specific audio knowledge from the finetuning stage.

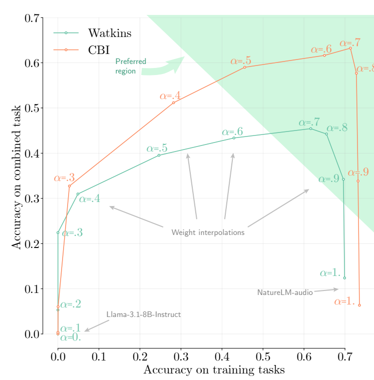


Figure 3: Accuracy on the combined prompt (y-axis) from Figure 1 versus the mean accuracy on the individual common- and scientific-name prompts (x-axis) when varying α.

The observed behaviour highlights α as a controllable *capability trade-off parameter*. At $\alpha = 1$, the model fully retains its domain adaptation but suffers in compositional instruction following. At $\alpha = 0$, it maximizes general instruction-following behaviour inherited from the base model, but discards most bioacoustic specialization. The monotonic decline in x-axis accuracy with decreasing α further confirms that domain-task performance and instruction-following ability are in tension.

In-Context Learning Task We next assess the merged model on a more challenging *one-shot in-context learning* task, where a single example for each class is provided directly in the prompt. We use the UNSEEN-CMN-FAMILY split from the BEANS-ZERO benchmark, the most difficult "unseen species" scenario evaluated in [38]. In this setting, no species from the same taxonomic family as those in the test set are present in training, and the goal is to predict the *common name* of the focal species. In the original paper, NATURELM-AUDIO performs extremely poorly when evaluated zero-shot on this challenging split (0.035 accuracy).

In-Context Learning Prompt

```
Identify the common name for the focal species
in the audio.  Output exactly one of:  Dall's
Porpoise, Spotted Elachura

Audio:  [high-pitched clicks and whistles]
Label:  Dall's Porpoise

Audio:  [bird-like chirping and trilling]
Label:  Spotted Elachura

Audio:  <Audio><AudioHere></Audio>
Label:
```

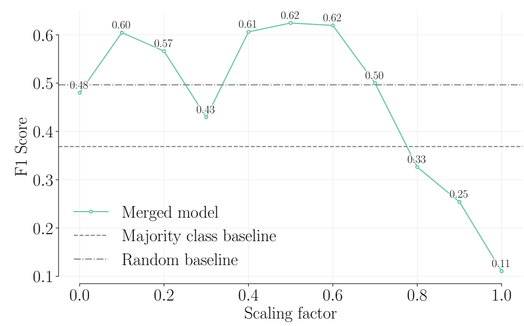


Figure 4: F1-score for 1-shot in-context classification on UNSEEN-CMN-FAMILY when varying α.

Ideally, the in-context examples would include *audio tokens* from the encoded waveforms. For simplicity, we instead provide a short textual description of each audio clip. To further simplify evaluation, we restrict to the two most frequent classes, reducing the task to binary classification.

We show the prompt and results in Figure 4. The original NATURELM-AUDIO ($\alpha = 1$) reaches an F1-score of just 0.11, performing *worse than random guessing* and often producing labels outside the provided ones, indicating a failure in instruction following. In contrast, the merged model achieves F1-scores above 0.6 for $\alpha \in [0.4, 0.6]$. These results show that model merging can *restore in-context learning*, enabling bioacoustic models to adapt quickly to new tasks with few labeled examples.

5 Related Work

Catastrophic forgetting in multi-modal finetuning Catastrophic forgetting is a well-known challenge when fine-tuning large language models [39]. A common training-time mitigation is to freeze the LLM and update only the projection layer that maps visual or audio features into the text embedding space, often with fewer fine-tuning steps [54] or using PEFT methods [55, 34]. In contrast, post-training approaches aim to restore forgotten skills in already fine-tuned models [57, 35].

Model merging Model merging provides an efficient alternative to ensembling, producing a single model combining multiple models' capabilities without increasing inference cost. Early work, inspired by linear mode connectivity [9, 11, 30, 8], focused on aligning independently trained models, often by solving a neuron permutation problem [2, 22, 5, 40, 41, 14, 31, 17]. Closer to our work, Wortsman et al. [48] produce robust finetuned models by linearly interpolating them with their base model, while Ilharco et al. [20] use interpolations to improve specific tasks without waiving others.

6 Conclusions

We investigated the instruction-following limitations of NATURELM-AUDIO and found that even small changes in prompt structure can significantly degrade performance, reducing its utility as a general-purpose model. To address this, we applied a lightweight model-merging strategy that interpolates the finetuned NATURELM-AUDIO with its base model. Intermediate interpolation weights restore much of the lost instruction-following capability while preserving most domain-specific accuracy. This recovery further enables few-shot in-context learning, a critical feature in bioacoustics where only a few labeled examples are often available. In our experiments, $\alpha \approx 0.6$ provided a strong balance between instruction following and domain expertise, though the optimal value remains task- and dataset-dependent.

Limitations and Future Work Our current evaluation of in-context learning is limited in scope. In future work, we plan to incorporate raw audio tokens directly into prompts and extend the evaluation to multiple datasets and many-class settings. Convex weight interpolation may not be optimal, we intend to explore alternative strategies for restoring instruction-following abilities, including more advanced model-merging methods (e.g. evolutionary merging [3, 29], subspace-based merging [10, 25, 42]) and activation-steering techniques [4, 43]. We believe these directions will further enhance the adaptability of bioacoustic foundation models, especially in real-world, low-resource scenarios.

References

[1] T.M̃itchell Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013. doi: 10.7717/peerj.103. URL https://doi.org/10.7717/peerj.103.

[2] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[3] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 2025. ISSN 2522-5839.

[4] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

[5] Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodolà. C^2M^3: Cycle-consistent multi-model merging. In *Advances in Neural Information Processing Systems*, volume 37, 2025.

[6] Nico Daheim, Thomas Möllenhoff, Edoardo Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *The Twelfth International Conference on Learning Representations*.

[7] MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*. Springer, 2025.

[8] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[9] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, 2020.

[10] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proc. CVPR*, 2025.

[11] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

[12] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[14] Fidel A. Guerrero-Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

[15] Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[16] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] Stefan Horoi, Albert Manuel Orozco Camacho, Eugene Belilovsky, and Guy Wolf. Harmony in Diversity: Merging Neural Networks with Canonical Correlation Analysis. 2024.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[19] Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging, 2024.

[20] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[21] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[22] Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: renormalizing permuted activations for interpolation repair. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[23] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[24] Paola Laiolo. The emerging significance of bioacoustics in animal species conservation. *Biological conservation*, 143(7):1635–1645, 2010.

[25] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*.

[26] Peter R Marler and Hans Slabbekoorn. *Nature's music: the science of birdsong*. Elsevier, 2004.

[27] Tiago A Marques, Len Thomas, Stephen W Martin, David K Mellinger, Jessica A Ward, David J Moretti, Danielle Harris, and Peter L Tyack. Estimating animal population density using passive acoustics. *Biological reviews*, 88(2):287–309, 2013.

[28] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[29] Tommaso Mencattini, Robert Adrian Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. Merge^3: Efficient evolutionary merging on consumer-grade gpus. In *Forty-second International Conference on Machine Learning*, 2025.

[30] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[31] Aviv Navon, Aviv Shamsian, Ethan Fetaya, Gal Chechik, Nadav Dym, and Haggai Maron. Equivariant deep weight space alignment, 2023.

[32] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[33] Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[34] Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms. *ArXiv preprint*, 2024.

[35] Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27011–27033. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

[36] Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds?, 2025. URL https://arxiv.org/abs/2504.12880.

[37] David Robinson, Adelaide Robinson, and Lily Akrapongpisak. Transferable models for bioacoustics with human language supervision, 2023. URL https://arxiv.org/abs/2308.04978.

[38] David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=hJVdwBpWjt.

[39] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024. URL https://api.semanticscholar.org/CorpusId:269362836.

[40] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[41] George Stoica, Daniel Bolya, Jakob Brandt Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Zipit! merging models from different tasks without training. In *The Twelfth International Conference on Learning Representations*, .

[42] George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In *The Thirteenth International Conference on Learning Representations*, .

[43] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

[44] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

[45] Anke Tang, Li Shen, Yong Luo, Yibing Zhan, Han Hu, Bo Du, Yixin Chen, and Dacheng Tao. Parameter-efficient multi-task model fusion with partial linearization. In *The Twelfth International Conference on Learning Representations*.

[46] Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2025. URL https://arxiv.org/abs/2508.04665.

[47] Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[48] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2021.

[49] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, 2022.

[50] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[51] Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[52] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[54] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning (Proceedings Track)*, 2023. URL https://openreview.net/forum?id=g7rMSiNtmA.

[55] Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. Sapt: A shared attention framework for parameter-efficient continual learning of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *ACL (1)*, pages 11641–11661. Association for Computational Linguistics, 2024. ISBN 979-8-89176-094-3. URL `http://dblp.uni-trier.de/db/conf/acl/acl2024-1.html#ZhaoWHZQZYXC24`.

[56] Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Fabrizio Silvestri, and Emanuele Rodolà. Atm: Improving model merging by alternating tuning and merging, 2024. URL `https://arxiv.org/abs/2411.03055`.

[57] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A Extended related work

Foundation Models in Bioacoustics Recent advances have introduced large-scale bioacoustic foundation models designed for cross-species and cross-task generalization. NATURELM-AUDIO [38] integrates a self-supervised audio encoder with a LLaMA-based language decoder. BIOLINGUAL [37] adapts CLAP-style audio–text contrastive learning to bioacoustics, while audio-only models such as BIRDMAE [36], AVES [15] and PERCH 2.0 [46] pretrain large models on extensive birdsong or multi-taxa datasets to produce broadly transferable acoustic features. Although these models surpass species-specific baselines, they remain susceptible to domain shifts and catastrophic forgetting, limiting their robustness in real-world deployment.

Mode connectivity and model merging Mode connectivity investigates the weight configurations that define local minima. Frankle et al. [9] examines the linear mode connectivity of models trained for only a few epochs from the same initialization, linking this phenomenon to the lottery ticket hypothesis. Relaxing the shared-initialization requirement, Entezari et al. [8] argues that, after resolving neuron permutations, all trained models may reside in a single connected basin. Model merging pursues a different goal: combining multiple models into one that inherits their capabilities without the cost and complexity of ensembling. In this direction, Singh and Jaggi [40] introduced an optimal-transport–based weight-matching method, while `Git Re-Basin` [2] proposes optimizing a linear assignment problem (LAP) for each layer. More recently, `REPAIR` [22] shows that substantial gains in the performance of interpolated models can come from renormalizing activations, while C^2M^3 [5] proposes matching and merging many models jointly through cycle-consistent permutations. When the models to merge are fine-tuned from a shared backbone, task-vector-based methods are most effective [21, 50, 53, 28, 48, 7, 47, 56, 10, 33, 29, 19, 6, 42, 51, 45]. These involve taking the parameter-level difference between the finetuned model and its pretrained base, termed a task vector. Improvements can be obtained by optimizing task-vector combinations [52], mitigating sign disagreement [50], randomly dropping updates [53], or applying evolutionary methods [3, 29]. Finally, techniques employing layer-wise task vectors [42, 10, 25] obtain state-of-the-art results by leveraging layer-level structures through SVD of the parameter differences.

B Additional Experiments

B.1 Combined Instruction Following Task

B.2 Zero-Shot Generalization Task

In Robinson et al. [38], zero-shot generalization was evaluated on the ZF-INDIV dataset, part of the BEANS-ZERO benchmark [38], which tests the ability to infer the number of zebra finch individuals in an audio recording. This task was not included in the model's training set.

With the original prompt from Robinson et al. [38] (Figure 6), NATURELM-AUDIO achieves 0.66 accuracy (random baseline: 0.5), indicating partial generalization to this unseen task. However, reversing the order of the class names in the prompt or removing explicit class labels, asking instead for the number of birds, reduces accuracy to 0.52, essentially random performance.

These observations suggest that the higher-than-random performance reported in Robinson et al. [38] may be sensitive to prompt formulation. While the original result remains valid for the tested prompt,

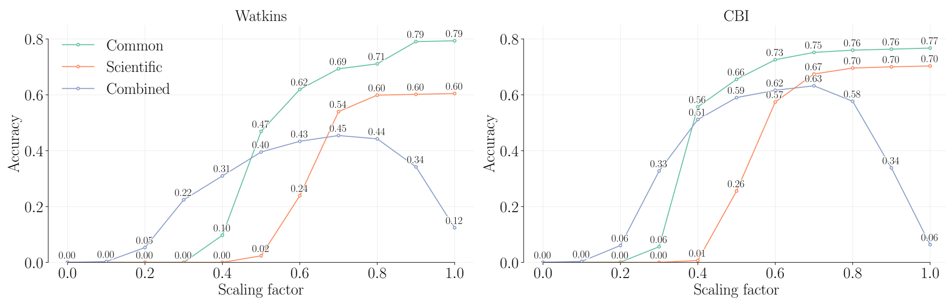


Figure 5: Accuracy on the common name, scientific name, and combined prompts from **??**, as a function of the rescaling parameter α.

Original Prompt

`Is there only one bird in the audio, or more? Reply with ‘One’ or ‘More’.`

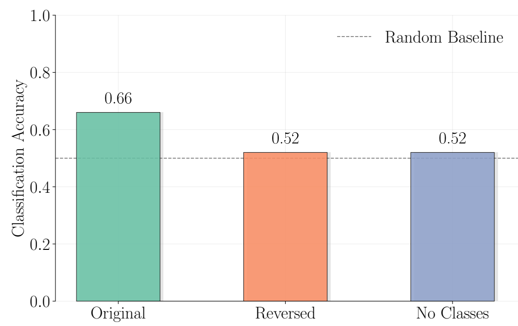
Reversed Classes Prompt

`Is there more than one bird in the audio, or just one? Reply with ‘More’ or ‘One’.`

No Classes Prompt

`How many birds are there in the audio?`

(a)



(b)

Figure 6: **Classification accuracy for different prompt types on ZF-INDIV.** (a) Exact wording of the three evaluated prompts. (b) Accuracy of NATURELM-AUDIO on ZF-INDIV. Accuracy is above random for the original prompt from Robinson et al. [38], but drops to near-random when the prompt is slightly reworded.

our findings indicate that performance may not fully reflect broad zero-shot generalization, but could instead be partly influenced by prompt-specific biases.

B.2.1 Few-Shot In-Context Learning

Experimental Details To mitigate position bias, we randomly permute the order of the few-shot examples in the prompt for each evaluation sample. This randomization is applied independently for every sample, ensuring that any spurious correlations between class position and prediction might be minimized. As previously noted, the two species used in the experiment, Spotted Elachura (*Elachura formosa*) and Dall’s Porpoise (*Phocoenoides dalli*), were selected for being the most frequent classes in the UNSEEN-CMN-FAMILY dataset, with 73 and 53 samples respectively.

Following [38], we evaluate classification accuracy by first extracting the model’s free-form output, then computing the Levenshtein distance between this output and each possible target class name (in this case, the two species’ common names). The class with the smallest distance is selected, and the prediction is considered correct if this distance is less than a threshold t (set to $t = 5$ in our experiments).

B.2.2 Compute Resources

All experiments were conducted on a single NVIDIA A100 GPU (40 GB), using 8 CPU cores and 32 GBs of RAM, requiring, for storage, 300 GB of disk space.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction of this paper clearly state its claims, complete with the contributions of this work (described both in the abstract and at the end of the introduction).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Particularly in the last section (Conclusions) a paragraph specifically addresses the limitations of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This paper has only one equation which may be regarded as a theoretical result, namely in the Method section: the reformulation of convex interpolation between a model and its finetuned LoRa variant into the canonical LoRa formulae. Albeit a very simple proof, it may be regarded as such.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the Problem Analysis and the Results section, but also in Appendix B, all experiments are run on clearly defined datasets and prompts, furthermore the merging technique (described in the Method section) eases the reproducibility as no ad-hoc code for merging is shown to be needed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code to reproduce all the experimental results will be supplemented after acceptance, as the current submission form does not enable upload of supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the tests in the Results section and Appendix B verge on the effect of the scaling hyperparameter over the capacity of the resulting merged model to solve some task. Since such hyperparameter is the only one needed to understand the results, indeed all the necessary details are specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The current submission does not report error bars or other statistical significance measures. While we recognize their importance for evaluating experimental variability, the required compute resources (tens of hours per run) prevented running multiple folds in time for this version. We plan to include them for the camera-ready version.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix B a breakdown of the compute resources used for the experiments is given.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: All aspects of the research conform to the NeurIPS Code of Ethics. The experiments exclusively use publicly available or ethically collected animal sound recordings, with no human or sensitive personal data involved. No procedures in this work raise ethical concerns regarding privacy, safety, environmental impact, or compliance with relevant regulations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Conclusions section discusses potential positive societal impacts, such as advancing the bioacoustics landscape in low-resource scenarios, which could support biodiversity monitoring and conservation efforts. While potential negative impacts, such as misuse for wildlife surveillance in protected areas or habitat disturbance from poorly managed data collection, were considered by the authors, they are not discussed in the paper due to space limitations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any new data or models and only reuses publicly available datasets and models. As such, it poses no additional risk of misuse and does not require specific safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used models, code and datasets have been clearly stated and credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Code to replicate the experiments will be provided after acceptance, as the submission form does not enable supplementary material upload.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core methods of this paper involve the use of large language models (LLMs) as an integral component of the proposed approach. Their usage is essential to the methodology and impacts the originality and rigor of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.