

Dual-Channel Contrastive Encoding of Sperm Whale Codas

Emma Virnelli, João Quintanilha, & Shou Zhang
Tufts University, Medford, MA, USA



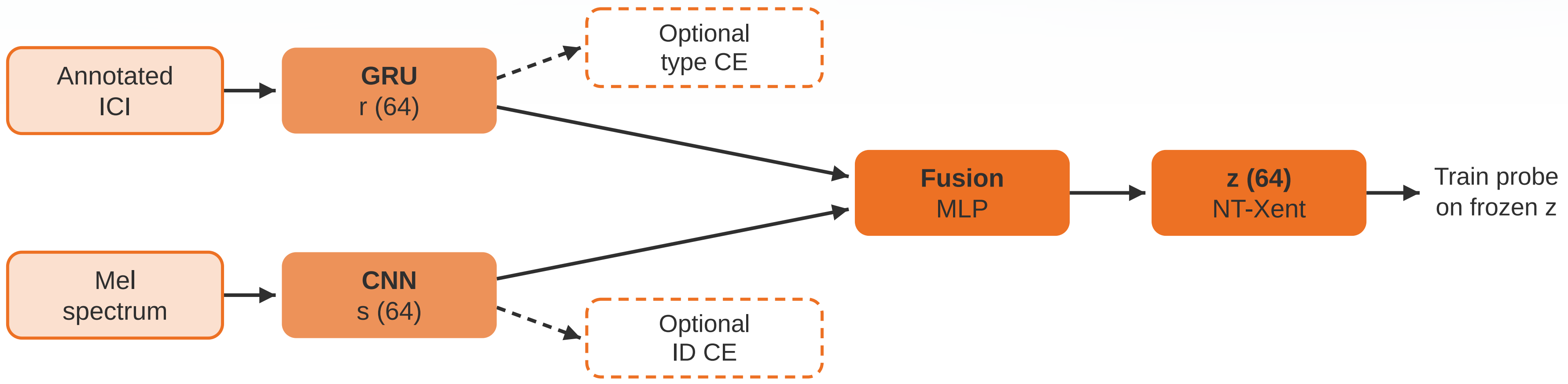
INTRODUCTION

Sperm whales (*Physeter macrocephalus*) live in stable, matrilineal family units that belong to vocal clans, each clan sharing a distinct vocal dialect [1], [2]. That dialect is built from **codas**: short, stereotyped sequences of clicks that whales exchange during social contact.

The **rhythm channel**, the pattern of inter-click intervals (ICIs), contributes to coda type: a category defined by click count and timing [2], [4]. The **spectral channel**, the frequency content within each click, includes vowel-like patterns that vary between individuals. [5], [6]. In this paper, we propose the **Dual-Channel Contrastive Encoder (DCCE)**, which explicitly encodes rhythm and spectral information with separate encoders before fusing them.

Our guiding question: Does encoding rhythm and spectral texture as separate channels improve what a model learns from sperm whale codas?

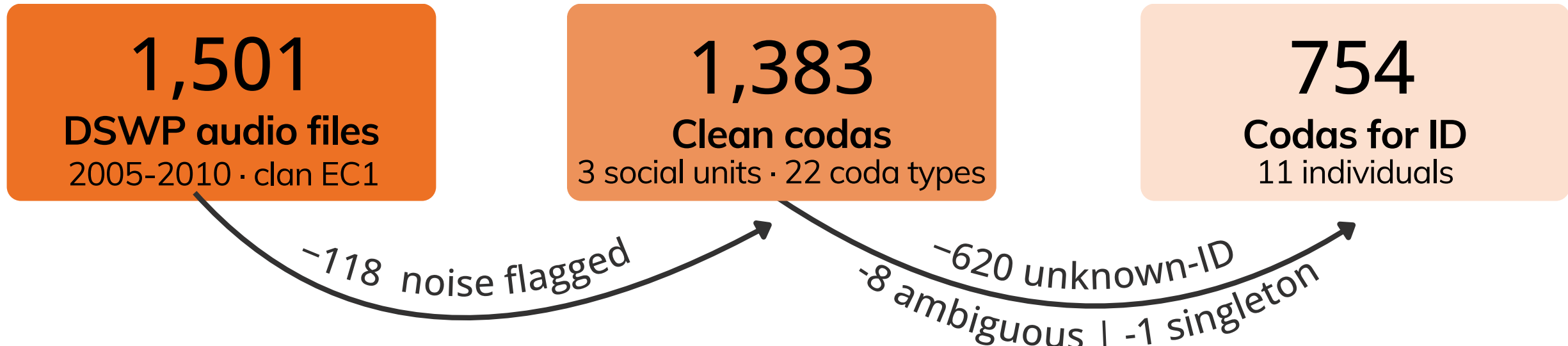
METHODOLOGY



$$L = L_{\text{NT-Xent}} + 0.5 a_T L_{\text{type}} + 0.5 a_I L_{\text{ID}}$$

We join the rhythm embedding (r) and the spectral embedding (s) and pass them through a small neural network that fuses them into one 64-dimensional representation (z). The model learns by pulling together codas that should be similar and pushing apart those that shouldn't (contrastive learning). We can also add optional extra training signals, a coda-type loss on the rhythm branch and an identity loss on the spectral branch.

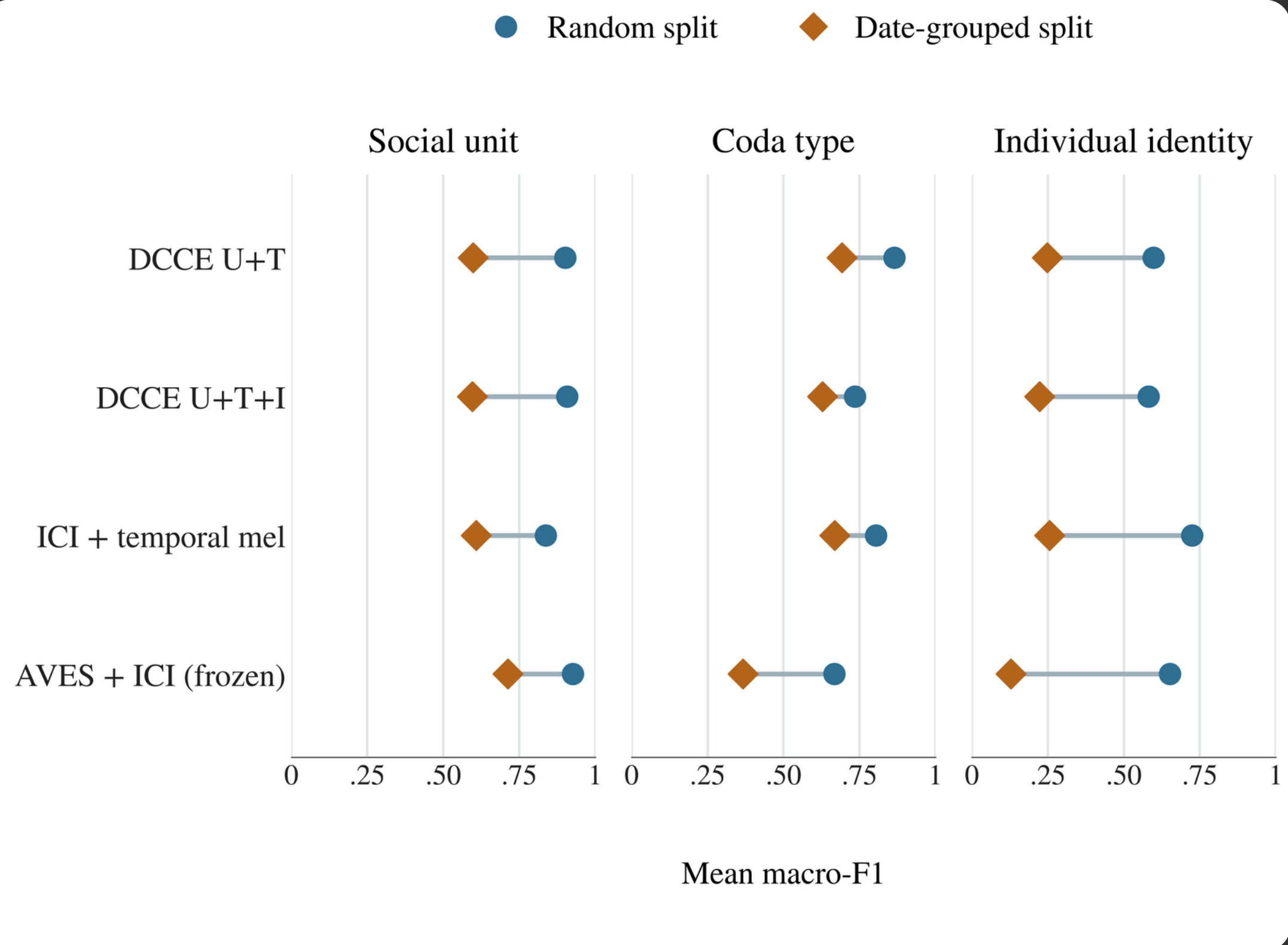
DATA & EVALUATION



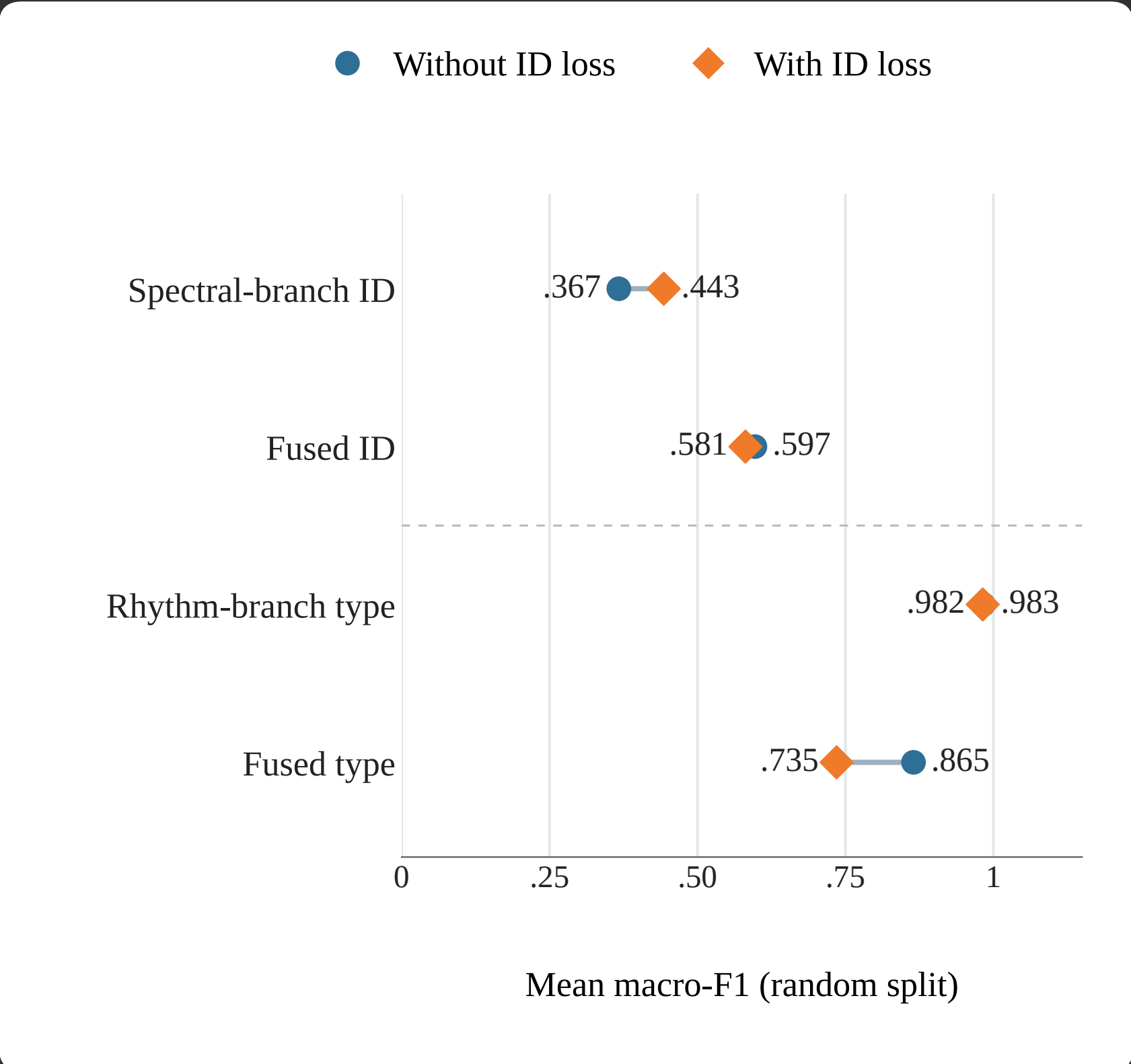
We split the data at random three times (seeds 0–2) into about 64% training, 16% validation, and 20% test, keeping individual whales balanced across the splits where possible. We compared our model against three existing models, WhAM, AVES, and BirdNET. Every model was evaluated the same way, with a simple classifier (balanced logistic regression) predicting social unit, coda type, or individual whale.

RESULTS

Model	Unit	Random split		Recording-date grouped		
		Type	ID	Unit	Type	ID
Raw ICI	0.570 ± 0.023	0.987 ± 0.009	0.439 ± 0.007	0.540 ± 0.036	0.956 ± 0.036	0.379 ± 0.103
Temporal mel	0.773 ± 0.007	0.202 ± 0.025	0.429 ± 0.045	0.569 ± 0.094	0.069 ± 0.012	0.130 ± 0.074
ICI + temporal mel	0.837 ± 0.013	0.804 ± 0.016	0.724 ± 0.039	0.608 ± 0.150	0.668 ± 0.051	0.255 ± 0.062
WhAM (frozen)	0.906 ± 0.011	0.258 ± 0.046	0.450 ± 0.052	0.667 ± 0.155	0.066 ± 0.016	0.101 ± 0.075
AVES (frozen)	0.910 ± 0.010	0.253 ± 0.028	0.491 ± 0.029	0.700 ± 0.100	0.061 ± 0.022	0.074 ± 0.051
BirdNET (frozen)	0.887 ± 0.020	0.215 ± 0.053	0.480 ± 0.022	0.612 ± 0.133	0.063 ± 0.011	0.085 ± 0.022
WhAM + ICI (frozen)	0.921 ± 0.005	0.722 ± 0.027	0.598 ± 0.028	0.768 ± 0.057	0.376 ± 0.151	0.153 ± 0.135
AVES + ICI (frozen)	0.926 ± 0.011	0.667 ± 0.044	0.652 ± 0.044	0.712 ± 0.111	0.366 ± 0.084	0.128 ± 0.103
BirdNET + ICI (frozen)	0.884 ± 0.018	0.497 ± 0.028	0.580 ± 0.031	0.628 ± 0.095	0.287 ± 0.061	0.100 ± 0.055
DCCE U+T	0.901 ± 0.020	0.865 ± 0.115	0.597 ± 0.051	0.598 ± 0.078	0.692 ± 0.088	0.248 ± 0.130
DCCE U+T+I	0.907 ± 0.010	0.735 ± 0.091	0.581 ± 0.012	0.596 ± 0.052	0.628 ± 0.074	0.223 ± 0.175
AVES FT	0.924 ± 0.007	0.243 ± 0.025	0.414 ± 0.078	0.626 ± 0.157	0.072 ± 0.050	0.060 ± 0.026
AVES + ICI FT	0.913 ± 0.011	0.466 ± 0.059	0.544 ± 0.017	0.641 ± 0.158	0.211 ± 0.069	0.073 ± 0.038



Mean macro-F1, three runs. Scores drop on unseen recording dates. Grouped ID covers different classes, so gaps aren't a pure recording-context penalty.

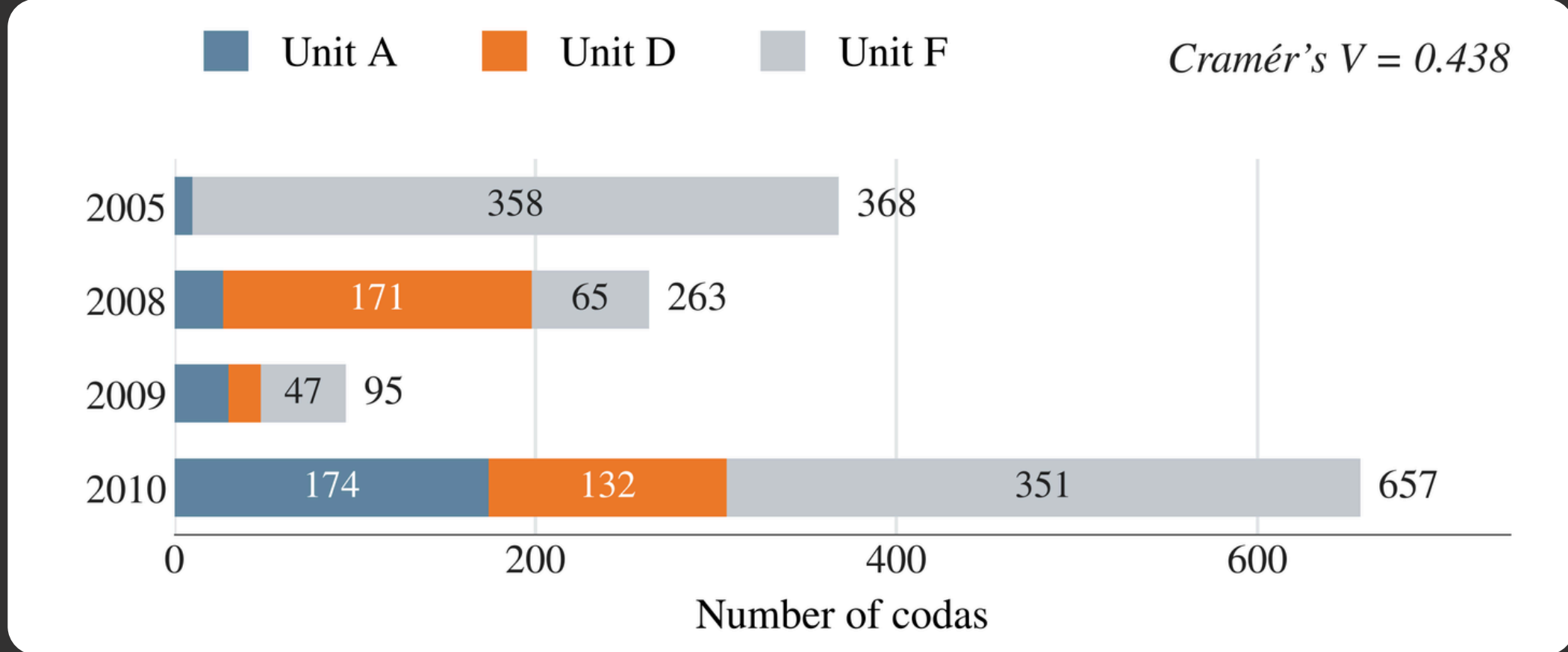


ID loss helps the spectral branch (0.37 → 0.44) but lowers fused ID (0.60 → 0.58) and fused type (0.87 → 0.74).

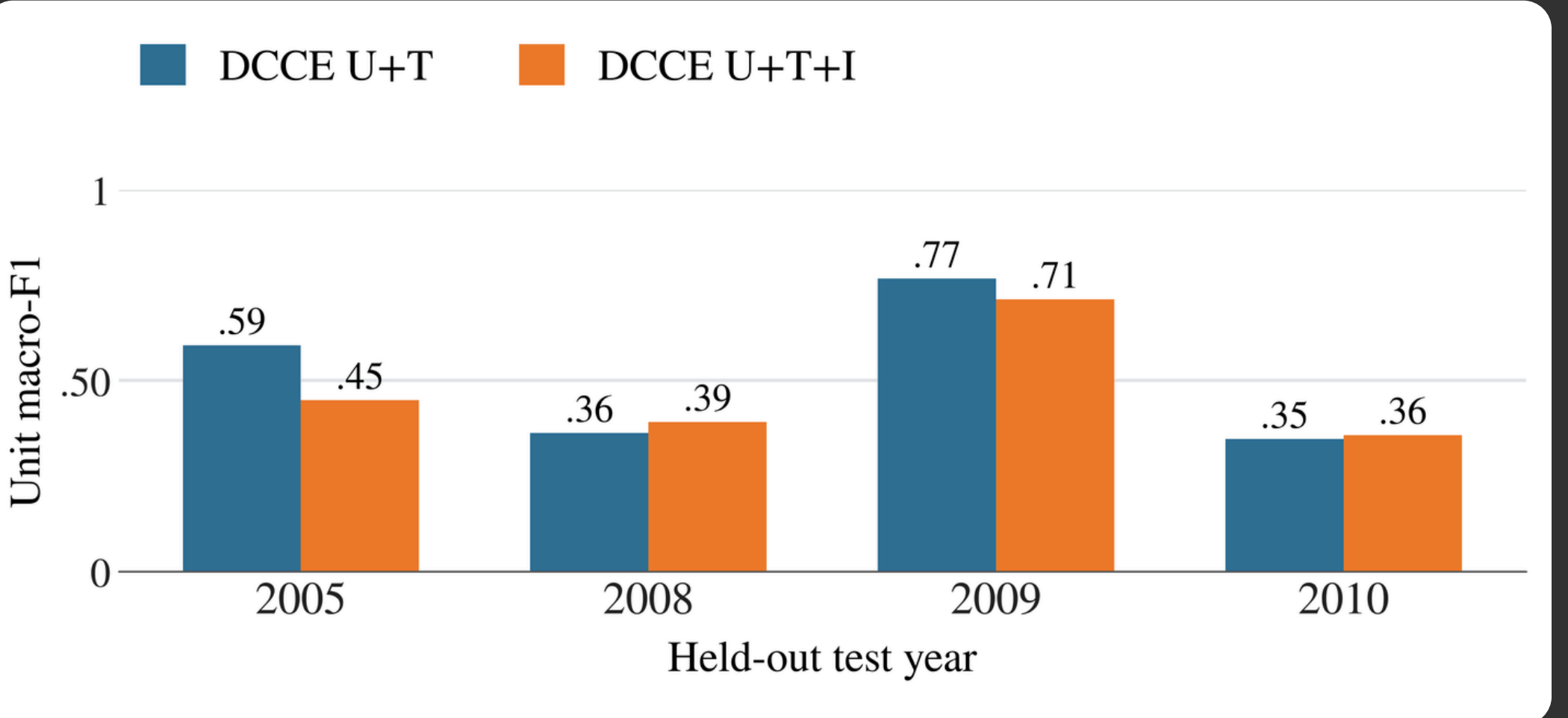
Macro-F1, mean ± SD over three runs. "+ ICI" adds annotated click timing. DCCE variants: unit pairing (U), type loss (T), ID loss (I). FT = fine-tuned AVES. All representations are frozen before the classifier is trained. Grouped ID covers different classes than random splits.

Raw ICI attains 0.987 ± 0.009 random type macro-F1; ICI plus temporal mel attains 0.724 ± 0.039 ID macro-F1.

DCCE U+T+I gives 0.907/0.735/0.581: adding ID loss does not improve mean fused-ID performance in this configuration.



Units were recorded unevenly across years (Cramér's V = 0.438).



Held-out years: unit macro-F1 0.35–0.77, below frozen AVES in every year.

CONCLUSION

Separating rhythm and spectral channels helps in task-dependent ways, not across the board. Raw ICI is best for coda type (0.987), and ICI + temporal mel is best for identity (0.724). DCCE is competitive on unit (0.901) and type (0.865), but a simple concatenation baseline matched or beat learned fusion on identity. ID loss helps the spectral branch but not the fused model (0.60 → 0.58). Scores also drop on unseen recording dates and years, so models should be evaluated across recording groups.

LIMITATIONS

Our data come from one Dominica population and have no vowel annotations. Identification is closed-set, with 11 whales. Results average only three runs.

FUTURE WORK

We will test other whale populations and uncured recordings. We will also tune the strength of the identity loss.

REFERENCES

References
[1] J. Andreas et al., "Toward understanding the communication in sperm whales," *iScience*, vol. 25, no. 6, p. 104393, 2022.
[2] S. Gero, H. Whitehead, and L. Rendell, "Individual, unit and vocal clan level identity cues in sperm whale codas," *Royal Society Open Science*, vol. 3, no. 1, p. 150372, 2016.
[3] O. Paradise et al., "WhAM: Towards a translative model of sperm whale vocalization," *arXiv:2512.02206*, 2025.
[4] P. Sharma et al., "Contextual and combinatorial structure in sperm whale vocalisations," *Nature Communications*, vol. 15, no. 1, p. 3617, 2024.
[5] G. Beguš, R. L. Sprouse, A. Leban, M. Silva, and S. Gero, "Vowel- and Diphthong-Like Spectral Patterns in Sperm Whale Codas," *Open Mind*, vol. 9, pp. 1849-1874, 2025.
[6] G. Beguš, M. Dąbkowski, R. L. Sprouse, D. F. Gruber, and S. Gero, "The phonology of sperm whale coda vowels," *Proceedings of the Royal Society B*, vol. 293, no. 2069, p. 20252994, 2026.
[7] M. Hagiwara, "AVES: Animal vocalization encoder based on self-supervision," *arXiv:2210.14493*, 2022.
[8] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.