Bioacoustic Multi-Step Attention: Underwater Ecosystem Monitoring in Climate Change Context

Amine Razig[1,5], Youssef Soulaymani[2], Loubna Benabbou[3], Pierre Cauchy[4]
[1] MILA – Quebec AI Institute, [2] Université de Montréal, [3] Université du Québec à Rimouski, [4] ISMER, [5] Institut Polytechnique de Paris

Context

- The St. Lawrence Estuary is an acoustic habitat where protected marine mammal species must maintain essential biological functions, communication, navigation, and foraging, in the presence of increasing anthropogenic noise. Ship noise can mask calls and echolocation, disrupt essential behavioral sequences, and induce physiological stress Erbe et al. (2019) with ecosystem-level consequences when behaviors change over space and time.
- These impacts have motivated concrete mitigation and policy efforts (e.g., quieter ship design, operational routing, and speed management) and targeted recovery planning for St. Lawrence species such as beluga.
- Our focus in this work is to turn raw hydrophone data into reliable presence signals that support biodiversity protection, monitoring, and adaptation actions in this sensitive region.

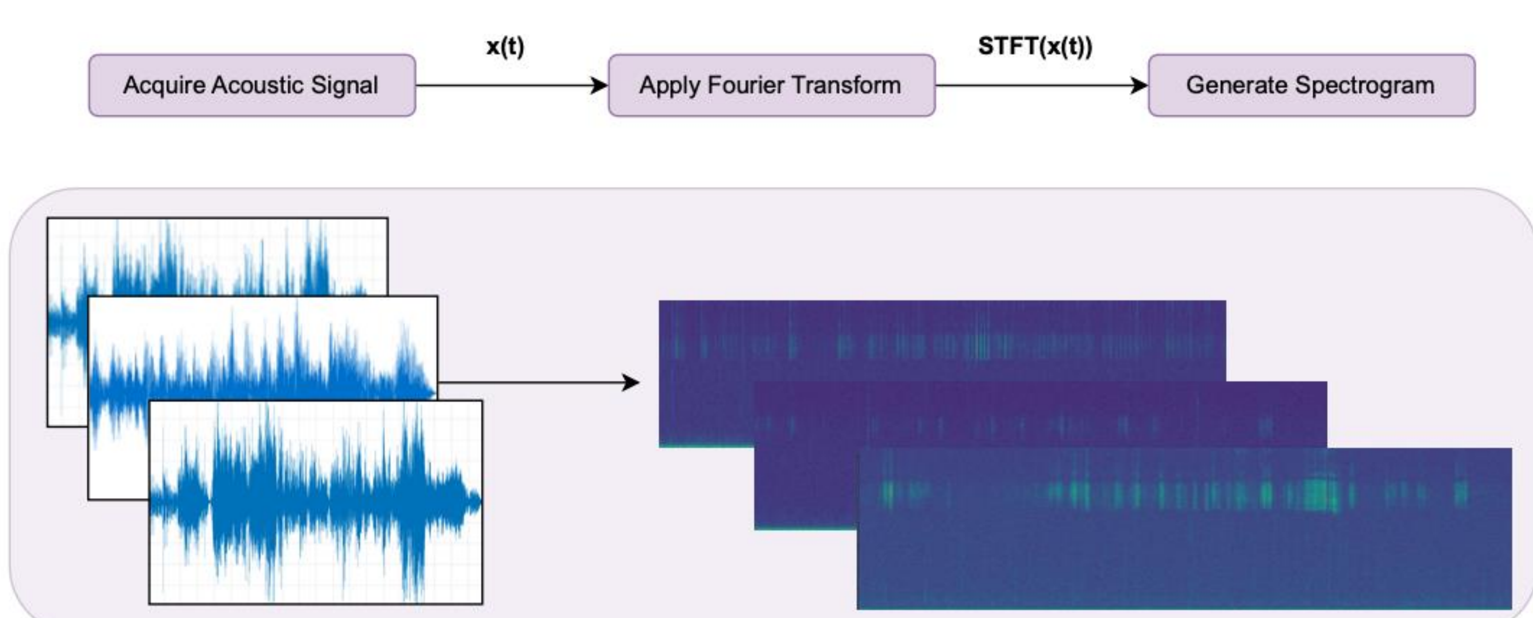
Data description & problem setup

The Saguenay–St. Lawrence Marine Park Dataset

- Long-term monitoring of endangered belugas and harbour porpoises in the Saguenay–St. Lawrence Marine Park.
- Hydrophones (PAM): ~1,500 h of continuous underwater recordings
- Annotations: species presence + sound types (whistles, clicks)
- Frequency range: 10 Hz – 150 kHz (biological + vessel + ambient noise)

The Challenge

Highly variable soundscape leads to overlapping calls, strong ship noise, seasonal & sensor shifts. The aim is to **Automatically detect and classify marine mammal vocalizations in noisy, multi-species recordings** (Porpoise and Beluga Clicks, Whistles)



STEP 1 : From audio to spectrogram representation

Framework overview

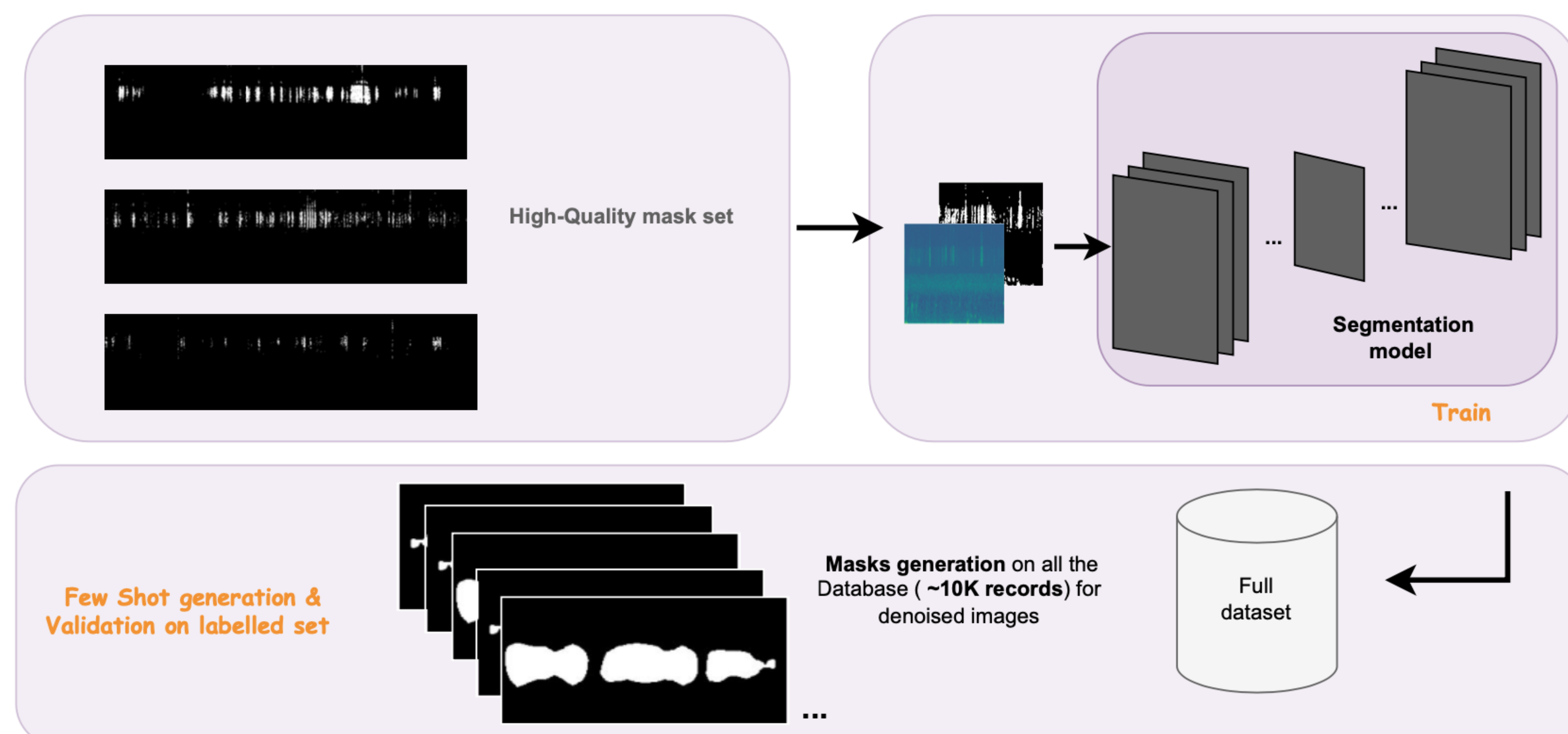
Our framework combines spectrogram and segmentation-mask embeddings through mid-level fusion to enhance biologically relevant signals and predict species presence probabilities as follows:

$$\hat{\mathbf{y}} = \mathcal{C}\Big(\mathrm{Fuse}\big(\mathcal{E}_{\mathrm{spec}}(\mathcal{T}(\mathbf{x}(\mathbf{t}))), \mathcal{E}_{\mathrm{mask}}(\mathcal{M}_{\mathrm{seg}}(\mathcal{T}(\mathbf{x}(\mathbf{t}))))\big)\Big), \quad \hat{\mathbf{y}} \in \mathbb{R}^{\mathbf{K}}$$

where $\mathcal{C}$ is the classification head mapping fused embeddings $\mathcal{E}$ to species-level outputs

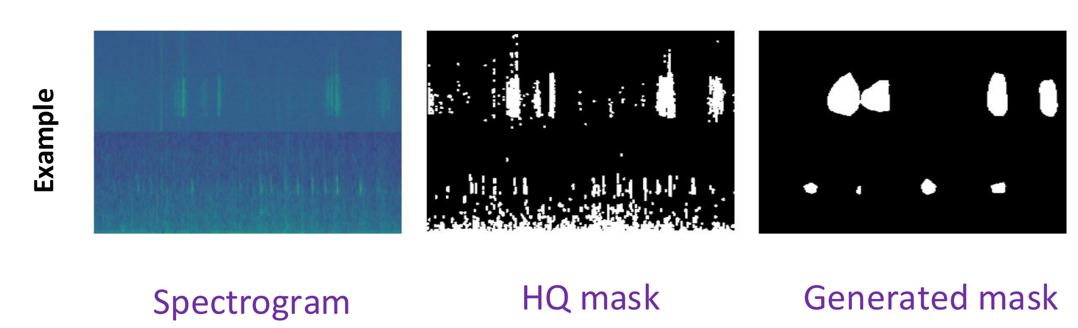
Attention mask generation

Semi-automatic labeling & segmentation-based denoising

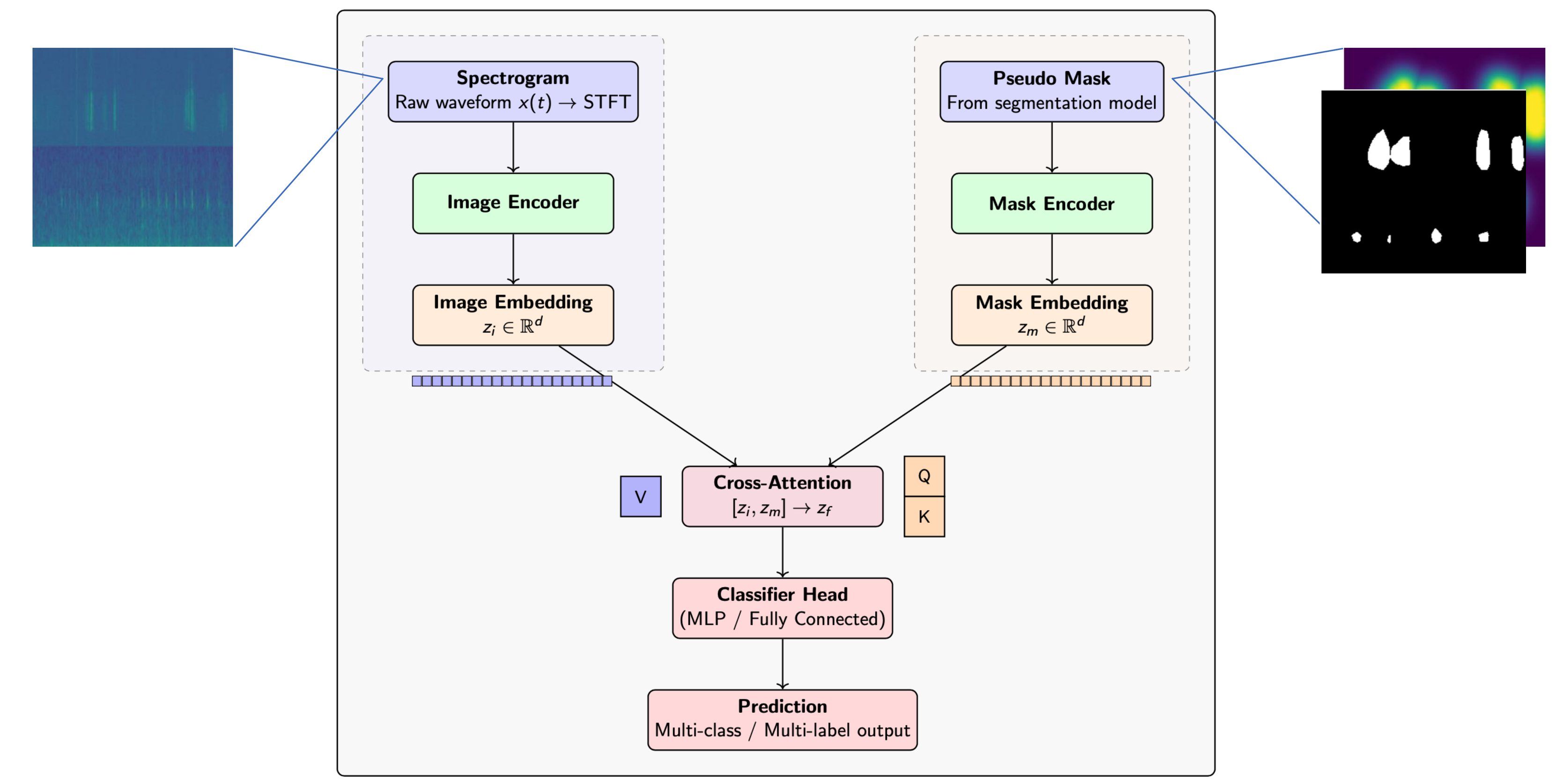


STEP 2 : Generation of the masks

- We generate candidate acoustic regions using signal processing (edge detection, adaptive thresholding), then refine them via minimal human correction through an interactive interface producing **a high-quality set** of ~200 annotated spectrograms.
- Using these attention-like masks, a **few-shot segmentation** model (DeepLabV3) learns to identify biologically relevant time–frequency patterns and generalize mask prediction across the full dataset.



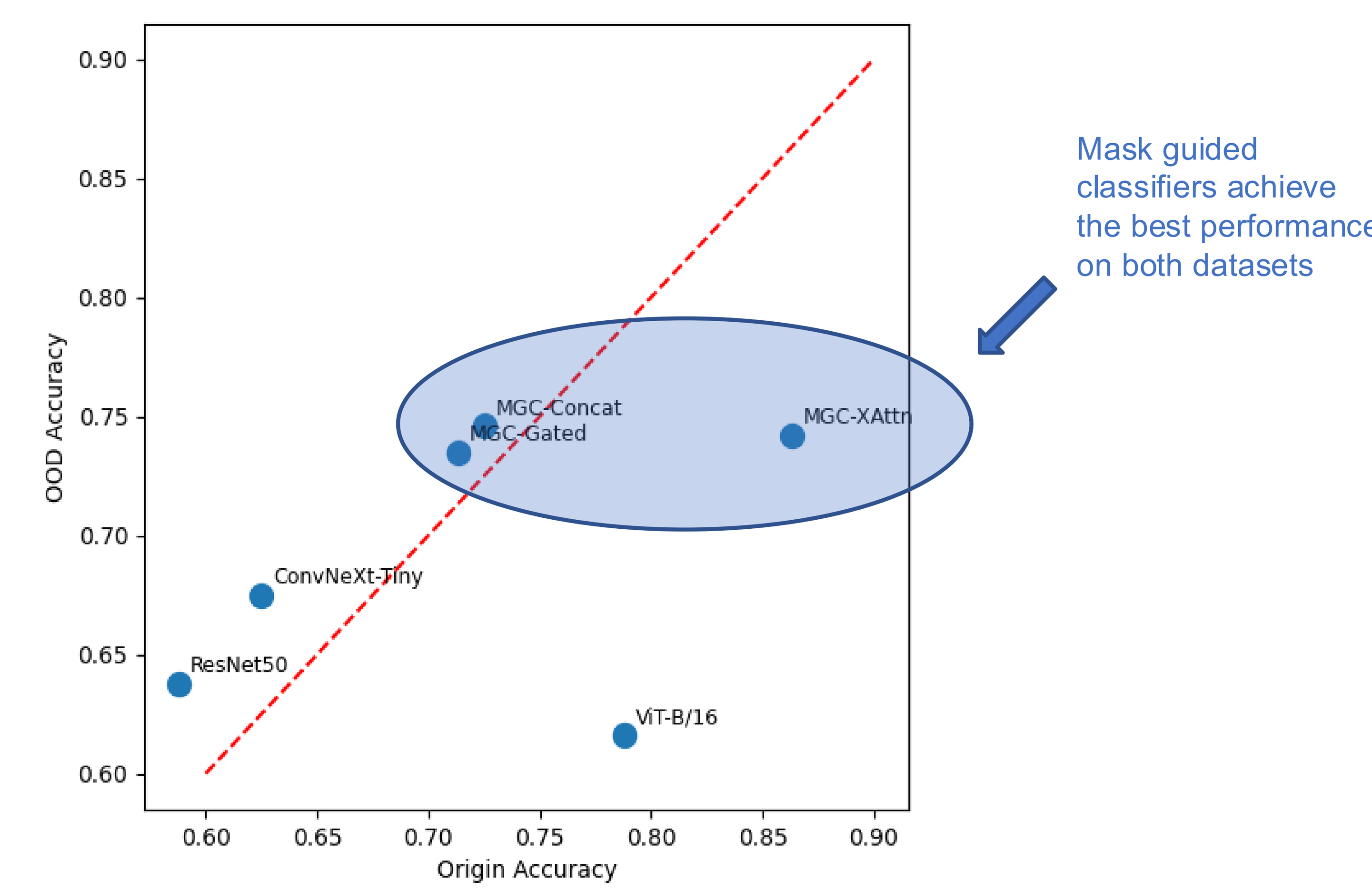
Fusion method and classification



STEP 3 : Cross Attention Fusion and classification

Results

The classification framework outperforms all baselines, reaching 89.7% accuracy, confirming that attention-guided fusion and structural priors significantly enhance denoising and classification robustness, even if an out-of-distribution setting.



References & Acknowledgement

We gratefully thank the AI for Supply Chains Chair, XST, Mila and GERAD for their support and contribution to this research.

- Y. Zhang and J. Li (2022). "Birdsoundsdenoising: Deep visual audio denoising for bird sounds."
- C. Bernier-Breton (2025). "Écouter et observer les mammifères marins pour les étudier sans déranger." *Master's thesis in Oceanography*
- Z. Jiang, A. Soldati, I. Schamberg, A. R. Lameira, and S. Moran (2024). "Automatic sound event detection and classification of great ape calls using neural networks."
- B. Ishachandar and S. Vairamuthu (2021). "Diverse ocean noise classification using deep learning." *Applied Acoustics181, 108141.*