AUDIOFM: TOWARDS GENERAL PURPOSE AUDIO REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning audio representations that effectively bridge holistic global soundscapes with fine-grained acoustic moments is a long-standing goal in self-supervised learning. Recording-level contrastive objectives encourage agreement between pooled audio representations, but those objectives do not explicitly supervise time-frequency local representations. We introduce **AUDIOFM**, a self-supervised framework that combines recording-level contrast with patch-wise masked latent discrimination. Using visible acoustic context, a student predicts representations of masked patches and identifies their corresponding full-view teacher targets among candidates from other recordings. We further incorporate weaker supervision from eligible temporal neighbors to encourage local temporal consistency. Extensive evaluations show that AUDIOFM achieves state-of-the-art performance across diverse audio domains in linear probing, fine-tuning, and zero-shot transfer, demonstrating superior event detection, robust scene-level understanding, and strong domain generalizability.

1 INTRODUCTION

Audio carries information beyond spoken language, offering insight into our surroundings, biological activity, and human states. Developing separate supervised models for these applications requires task-specific annotations, whose collection can be costly and often depends on domain expertise (Zhang et al., 2024; Schwinger et al., 2025). Audio self-supervised learning (SSL) seeks to learn general representations from unlabeled recordings that can be transferred across downstream tasks (Huang et al., 2022a; Niizumi et al., 2024; Huang et al., 2026; Chen et al., 2023; 2024; Alex et al., 2025). These methods provide a path toward audio foundation models that support diverse acoustic applications. Progress towards this goal depends on both effective self-supervised learning and systematic evidence of transfer across acoustic domains. Yet how consistently improvements on standard audio benchmarks extend to diverse specialized domains remains insufficiently understood. Understanding how performance changes with model and data scale further informs the development of these models.

From a modeling perspective, supporting diverse acoustic applications requires representations that preserve information at different levels of detail. A recording can contain sustained backgrounds, overlapping sources, and brief events, whose relevance depends on the task. Overall acoustic context can inform scene recognition, while subtle temporal and spectral differences can distinguish animal vocalizations or patterns in physiological sounds. Information that is secondary for one task may therefore be central to another. A general-purpose representation should capture the overall content of a recording while retaining the finer structure needed to recognize and locate its constituent events. Joint global and local supervision is important for capturing recording-level context alongside fine-grained acoustic content (Chen et al., 2024; Alex et al., 2025). A key modeling question is how to make individual patch representations discriminative while maintaining recording-level consistency.

Establishing the generality of these representations, in turn, requires examining their transfer across acoustic domains and task settings. Broad audio benchmarks such as HEAR (Turian et al., 2022) and HARES (Wang et al., 2021a) already assess transfer across speech, music, and environmental sound, while more recent studies in health acoustics (Zhang et al., 2024) and bioacoustics (Schwinger et al., 2025) started to investigate specialized applications. However, differences in model coverage and evaluation protocols leave an incomplete picture of the relative strengths of audio SSL encoders

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Downstream evaluation coverage.** ✓/✗: dedicated evaluation reported/not reported in the original study; SED: frame-level sound event detection.

Study	General audio	Speech recog.	Bioacoustics	Medical	Speech emotion	SED
AudioMAE (Huang et al., 2022a)	✓	✓	✗	✗	✗	✗
BEATs (Chen et al., 2023)	✓	✓	✗	✗	✓	✗
ATST-Frame (Li et al., 2024)	✓	✓	✗	✗	✗	✓
M2D (Niizumi et al., 2024)	✓	✓	✗	✓	✓	✗
EAT (Chen et al., 2024)	✓	✓	✗	✗	✗	✗
SSLAM (Alex et al., 2025)	✓	✓	✗	✗	✗	✗
BAT (Ghaffari et al., 2026)	✓	✓	✓	✗	✗	✓
AudioMosaic (Huang et al., 2026)	✓	✓	✗	✗	✗	✗
AUDIOFM (Ours)	✓	✓	✓	✓	✓	✓

across general and specialized acoustic domains. Alongside transfer, examining performance as model capacity and pretraining data increase can help assess the potential for further improvement.

We introduce **AUDIOFM**, an audio foundation model trained through joint global-local discrimination. The model combines recording-level contrast between complementary time- and frequency-masked views with a local objective that discriminates full-view teacher targets for masked regions from targets in other recordings. We assess AUDIOFM and existing audio SSL encoders through a shared linear-probing pipeline across a diverse range of audio (Table 1). Complementary evaluations cover supervised fine-tuning, frame-level event detection, and transfer of source-trained classifiers to new datasets or speakers without target-domain training, alongside an analysis of how performance scales with model size and pretraining data volume. We summarize our contributions as follows:

- **A general-purpose audio foundation model.** AUDIOFM combines recording-level contrast with masked latent discrimination to learn an encoder for diverse acoustic tasks.
- **Systematic cross-domain evaluation.** We compare frozen audio encoders across 12 classification tasks in four acoustic domains using shared task-specific probing protocols.
- **Model and data scaling.** We observe clear and consistent performance gains as encoder capacity and pretraining data volume increase, establishing the empirical scaling laws for audio pretraining.
- **Strong downstream performance and transfer.** AUDIOFM achieves state-of-the-art fine-tuning and temporal detection results while demonstrating strong generalizability to unseen data.

2 RELATED WORK

General-purpose audio representation learning. Audio pretraining seeks to provide a common representation for diverse downstream tasks. Large-scale event classification and audio–text alignment offer supervised and language-supervised routes to transferable features (Kong et al., 2020; Elizalde et al., 2023). Within audio SSL, one line of work emphasizes agreement between different views of a recording, through contrastive learning or non-contrastive self-distillation (Saeed et al., 2021; Niizumi et al., 2021; Huang et al., 2026). Another learns by predicting masked content, with the choice of target defining the supervision: spectrogram reconstruction operates on the acoustic signal (Huang et al., 2022a), while acoustic token prediction uses learned discrete targets (Chen et al., 2023). Continuous latent prediction instead uses teacher representations as targets, providing a common basis (Niizumi et al., 2024; Chen et al., 2024; Alex et al., 2025; Ghaffari et al., 2026). These objectives emphasize different aspects of the audio signal, making downstream transfer an important test of which acoustic distinctions their representations preserve.

Foundation models and cross-domain evaluation. Beyond audio, foundation-model research has pursued broadly transferable representations for visual understanding (Oquab et al., 2024), general time-series analysis (Goswami et al., 2024), and wearable and physiological applications (Narayanswamy et al., 2025; Zhang et al., 2025; Li et al., 2025; 2026; Lee et al., 2026; Zhou et al., 2026). Across these settings, a central question is how effectively pretrained representations support different

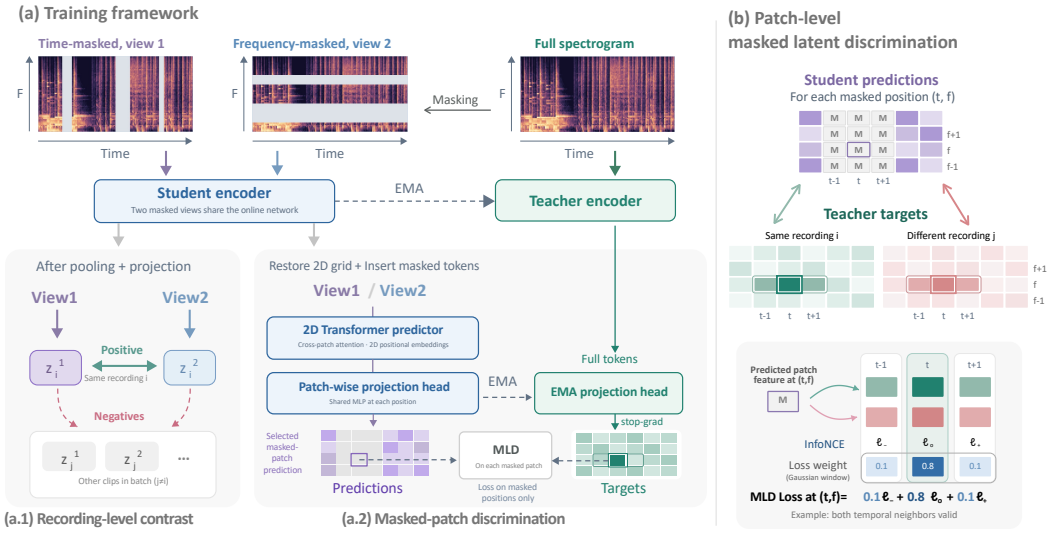

Figure 1: **Overview of AUDIOFM.** (a) A shared student encoder learns from complementary time- and frequency-masked views through recording-level contrast (a.1) and masked-patch prediction (a.2). An EMA teacher encodes the corresponding unmasked views to provide latent targets. (b) Detailed view of masked-patch discrimination in (a.2).

downstream tasks and data distributions. For audio, this question spans both general audio understanding and specialized acoustic applications. HEAR and HARES assess representations across speech, music, and environmental sound (Turian et al., 2022; Wang et al., 2021a), while OPERA and BioFoundation provide domain-focused evaluations in respiratory acoustics and bioacoustics (Zhang et al., 2024; Schwinger et al., 2025). However, coverage of general and specialized acoustic domains remains distributed across separate evaluations. Meanwhile, comparisons based on published fine-tuning results can reflect differences in downstream training recipes as well as pretrained representations. We therefore complement such results with a shared frozen-encoder evaluation across general and specialized acoustic tasks.

Global-local supervision and masked discrimination. Global and local supervision address complementary requirements: consistency across views of an instance and sensitivity to its constituent acoustic content. AudioMosaic studies recording-level contrast between complementary time- and frequency-masked views (Huang et al., 2026), while EAT and SSLAM explicitly couple utterance- and patch-level regression (Chen et al., 2024; Alex et al., 2025). Joint global-local contrast also has precedents in vision (Wang et al., 2021b; Zhou et al., 2022). Beyond supervision granularity, the prediction targets and candidate sets determine what a local contrastive objective distinguishes. SSAST and MAE-AST discriminate raw masked spectrogram patches using other masked positions in the same recording as candidates (Gong et al., 2022; Baade et al., 2022); SPIRAL and ConMIM instead contrast learned teacher representations, with negatives drawn from other positions in the same utterance or image (Huang et al., 2022b; Yi et al., 2023). AUDIOFM combines recording-level view contrast with contextual prediction of full-view teacher features at masked locations, drawing competing patch targets from other recordings at the same target coordinate. This candidate construction avoids requiring acoustically similar regions within a recording to be distinguished solely because they occupy different positions. We also give weaker supervision to nearby teacher targets along the time axis when those positions are masked in the student, encouraging local temporal consistency within this discriminative objective.

3 METHOD

Framework overview For each recording i in a minibatch of size B, we independently apply waveform augmentations to obtain two versions of the recording, following AudioMosaic (Huang et al., 2026). We convert them into log-Mel spectrograms X_i^1 and X_i^2, and embed each spectrogram as

a $T \times F$ grid of non-overlapping patch tokens, where T and F denote the numbers of patches along the time and frequency axes, respectively. Patch positions are indexed by $\mathcal{P} = \{1, \ldots, T\} \times \{1, \ldots, F\}$. We then apply complementary structured masking to these tokens: view 1 removes tokens at selected temporal positions across all frequency bands, while view 2 removes tokens at selected frequency bands across all time steps. Let $\mathcal{M}_i^v \subset \mathcal{P}$ denote the masked positions in view $v \in \{1, 2\}$. Only the remaining tokens are passed to the student Transformer, while the teacher processes all patch tokens of the corresponding spectrogram.

A shared Vision Transformer student encoder f_θ processes the visible patches with their positional embeddings and a class token. For global contrast, pooled encoder outputs pass through a shared projection head g_ψ and ℓ_2 normalization to produce recording embeddings $\mathbf{z}_i^v$. The positive pair consists of the two masked views of the same recording.

For local prediction, we remove the class token and restore the visible patch representations to their original grid coordinates. We fill masked positions with a shared learnable mask token, with encoder dimension of D. A shared 2D Transformer predictor d_η projects these features to a bottleneck dimension D_{pred}, adds 2D positional encodings and a view-specific masking-branch embedding, applies full time-frequency self-attention, and projects the outputs back to D. The positional encoding concatenates the time and frequency components, $E(t, f) = [E_{\text{time}}(t); E_{\text{freq}}(f)]$, each of dimension $D_{\text{pred}}/2$. A shared local projection head $h_\phi : \mathbb{R}^D \to \mathbb{R}^d$ maps the predictor outputs to a local contrastive embedding space of dimension d, yielding ℓ_2-normalized patch predictions $\mathbf{p}_{i,t,f}^v \in \mathbb{R}^d$. The predictor retains both grid axes; only predictions at masked positions contribute to the local loss.

3.1 Masked-Patch Prediction & Supervision

Our local objective uses contrastive self-supervision for masked predictions (Baevski et al., 2020; Yi et al., 2023), with full-view teacher targets and cross-recording negatives. We refer to this local objective as masked latent discrimination (MLD). We first define prediction-target discrimination at the masked patch's own position, then extend it to eligible temporal neighboring positions.

Full-view latent targets. The teacher encoder $f_{\bar{\theta}}$ and local projection head $h_{\bar{\phi}}$ are updated as EMA of the student encoder f_θ and local projection head h_ϕ, respectively. For each student view, the teacher processes the corresponding full spectrogram X_i^v without structured masking (Figure 1). Its patch representations are projected into ℓ_2-normalized targets $\mathbf{q}_{i,t,f}^v \in \mathbb{R}^d$, with gradients stopped through the teacher branch.

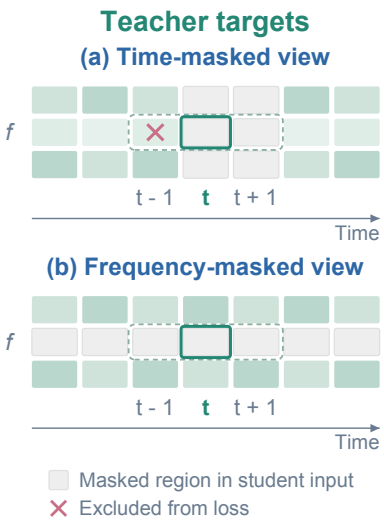

Figure 2: **Mask-aware target selection.** Temporal masking can exclude neighbors; frequency masking retains all in-bounds temporal neighbors.

Discriminating the corresponding teacher target. Let $\mathcal{N}_i$ be the recording indices in the effective contrastive pool for anchor i, including i itself. For a masked anchor $(t, f) \in \mathcal{M}_i^v$, same-position cross-recording contrast is defined as

$$\ell_i^v(\mathbf{p}_{i,t,f}^v; (t, f)) = -\log \frac{\exp\big((\mathbf{p}_{i,t,f}^v)^\top \mathbf{q}_{i,t,f}^v / \tau_p\big)}{\sum_{j \in \mathcal{N}_i} \exp\Big((\mathbf{p}_{i,t,f}^v)^\top \mathbf{q}_{j,t,f}^v / \tau_p\Big)}, \tag{1}$$

where $\tau_p > 0$ is the local temperature. Here, the positive is the teacher target $\mathbf{q}_{i,t,f}^v$ from recording i, and negatives come from other recordings $j \in \mathcal{N}_i$ at that same grid coordinate (t, f). Only targets at (t, f) enter the denominator; other positions within a recording are not negative candidates. A shared coordinate across recordings does not imply semantic alignment of their sound events.

Temporal neighborhood extension. We extend MLD with weaker supervision from nearby teacher targets, motivated by temporal-neighborhood learning (Tonekaboni et al., 2021) and neighborhood distribution smoothing (Yang et al., 2021). These works motivate a local-consistency prior; our prediction-target discrimination retains the cross-recording InfoNCE formulation above. We define

neighborhoods over temporal positions within a recording and keep the frequency coordinate fixed. For an integer radius $K \geq 0$, let $\mathcal{T} = \{-K, \ldots, 0, \ldots, K\}$, with offsets measured in temporal patch steps. The corresponding temporal window contains $W = 2K + 1$ positions. At offset δ, the same prediction $\mathbf{p}^v_{i,t,f}$ is contrasted with targets at $(t+\delta, f)$ through $\ell^v_i(\mathbf{p}^v_{i,t,f}; (t+\delta, f))$. Nearby temporal regions may share acoustic content, making teacher targets at neighboring positions within the same recording useful candidates for additional positive supervision. We use prior weights w_δ, assigning a positive weight to the reference target and smaller nonnegative weights to neighboring targets, preserving the reference position as the primary prediction target.

Mask-aware neighbor selection. As illustrated in Figure 2, an off-center target participates only if its location is within the grid and also masked in the anchor recording's student view. For each masked anchor, the valid offsets are

$$\mathcal{V}^v_{i,t,f} = \{\delta \in \mathcal{T} : (t + \delta, f) \in \mathcal{M}^v_i\}. \tag{2}$$

For each masked anchor, we retain only neighboring positions that are within the grid and masked in the same student view (Figure 2). The center position is always included because the anchor itself is masked. For each retained offset δ, teacher targets $\mathbf{q}^v_{j,t+\delta,f}$ from other recordings $j \neq i$ at the same coordinate serve as negatives.

Weighted loss aggregation. We define each anchor's neighborhood loss as a weighted average of the valid offset losses:

$$\ell^v_{i,t,f} = \sum_{\delta \in \mathcal{V}^v_{i,t,f}} \tilde{w}^v_{i,t,f,\delta}\, \ell^v_i(\mathbf{p}^v_{i,t,f}; (t + \delta, f)), \tag{3}$$

where $\tilde{w}^v_{i,t,f,\delta}$ denotes the normalized weight for each valid offset. After excluding invalid offsets, we rescale the remaining prior weights w_δ to sum to one for each anchor, preserving their relative proportions. For each valid offset δ, the contrastive denominator is computed over recordings at the target coordinate $(t + \delta, f)$. When $K = 0$, or when no off-center neighbor is valid, the neighborhood loss reduces to the same-position contrastive loss in Eq. 1. The local objective $\mathcal{L}_{\text{MLD}}$ first averages the anchor losses over all masked patches across recordings within each view, then averages the two view-level losses with equal weight.

3.2 Joint Training Objective

We use a symmetric NT-Xent objective for global contrast $\mathcal{L}_{\text{global}}$ (Huang et al., 2026; Chen et al., 2020). For anchor $\mathbf{z}^v_i$, the positive is the other masked view of the same recording, $\mathbf{z}^{3-v}_i$; both views of other recordings serve as negatives. Each denominator includes the positive and excludes only the anchor itself. We jointly optimize the student encoder, predictor, and projection heads with

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \lambda_{\text{local}} \mathcal{L}_{\text{MLD}}, \tag{4}$$

where λ_{local} weights the local loss relative to the global loss. The teacher encoder and local projection head are updated as exponential moving averages of their student. The momentum schedule is specified in Appendix §A.1. Only the pretrained student encoder is transferred to downstream tasks. The local predictor, both pretraining projection heads, and the teacher are discarded. Task-specific classifiers are trained during downstream probing or fine-tuning.

4 Experiments

4.1 Experimental Setup

Pretraining. We pretrain our model on the widely used AudioSet corpus (Gemmeke et al., 2017) without labels. Due to YouTube video deletions over time, our downloaded version consists of approximately 1.68M valid audio clips ($\sim$ 4,665 hours), representing a **smaller pretraining corpus** than the 1.91M–2.0M clips typically employed in prior SSL literature (Huang et al., 2022a; Chen et al., 2024; Huang et al., 2026; Alex et al., 2025; Chen et al., 2023).

Evaluation protocols. We assess downstream performance and transferability under three complementary settings. **Full fine-tuning:** We jointly fine-tune the encoder and task-specific classification heads on standard benchmarks: AudioSet-2M, AudioSet-20K, ESC-50, Speech Commands V1/V2, and FSD50K. Subject to the same attrition, our retained AudioSet evaluation split comprises 16,103 clips. **Linear probing and domain transfer:** We evaluate frozen representations across general audio, bioacoustics, medical acoustics, and speech emotion recognition. For baseline encoders, we obtain open-weight self-supervised pretrained checkpoints and evaluate them alongside AUDIOFM using a shared evaluation pipeline. Within each task, we use consistent data splits, label definitions, evaluation metrics, and probe-training protocols across encoders. Encoder parameters remain frozen, and only the downstream probe is optimized. We consider two configurations: a linear classifier applied directly to frozen features (Bare), and a linear classifier preceded by LayerNorm (LN). Each model–dataset–probe configuration is evaluated over three random seeds, and we report the mean performance across runs. Dataset protocols and optimization settings are detailed in Appendix §A.3 and Appendix §A.1. **Zero-shot transfer:** We evaluate zero-shot transfer through sound event tagging and emotion recognition, where shared label spaces allow source-trained predictors to be applied across datasets without target-domain training. We also examine medical-audio embeddings to explore whether general audio pretraining qualitatively captures structure relevant to physiological sounds.

Baselines. We compare with recent open-source audio representation learning methods, including AudioMAE (Huang et al., 2022a), BEATs (Chen et al., 2023), EAT (Chen et al., 2024), SSLAM (Alex et al., 2025), BAT (Ghaffari et al., 2026), and AudioMosaic (Huang et al., 2026). For AudioSet-2M, we re-evaluate available author-released fine-tuned checkpoints on our retained evaluation subset without further training. For AudioSet-20K and FSD50K, we fine-tune released pretrained checkpoints on our training splits and evaluate them on the same evaluation splits we used. Other baseline fine-tuning results are taken from the original papers unless otherwise indicated.

Model and training configuration. Unless otherwise specified, AUDIOFM uses ViT-Base encoders (Dosovitskiy et al., 2021) for both the student and EMA teacher. Audio is resampled to 16 kHz mono and converted into 128-band log-Mel spectrograms with 1,024 time frames. We divide each spectrogram into non-overlapping 16×16 patches and construct two complementary views by masking 60% of temporal slices in one view and 40% of frequency-patch bands in the other (Huang et al., 2026). The default temporal-neighborhood radius is $K = 1$. Student parameters are optimized with AdamW (Loshchilov & Hutter, 2019) and cosine-annealed learning rate scheduler. We pretrain our models on a Cloud TPU v4-128 slice (64 chips) with a batch size of 6144. Detailed architecture specifications, training hyperparameters and downstream evaluation settings are provided in Appendix §A.1.

4.2 MAIN RESULTS

Supervised fine-tuning. We first assess AUDIOFM's adaptability through supervised fine-tuning on standard audio and speech benchmarks (Table 2). AUDIOFM achieves leading results among the compared methods across sound classification and keyword spotting. Its improvements over AudioMosaic extend to both AudioSet-20K and FSD50K, with a gain of 0.9 mAP on the latter. On AudioSet-2M, AUDIOFM remains competitive with SSLAM and AudioMosaic when evaluated on the same retained evaluation subset, despite using approximately 13% fewer pretraining and finetuning clips.

Linear probing and domain transfer. Beyond adaptation through fine-tuning, a general-purpose audio encoder can also serve as an off-the-shelf feature extractor for diverse downstream tasks. By keeping the encoder fixed and standardizing downstream training, linear probing helps isolate the contribution of pretrained representations from differences in fine-tuning recipes, including augmentation and optimization choices. We evaluate AUDIOFM and baseline encoders across diverse acoustic domains using a shared linear-probing protocol, assessing how readily task-relevant information can be extracted from their frozen features. As shown in Table 3, AUDIOFM matches or outperforms the evaluated baselines on the majority of tasks under both probing protocols. Its advantage is particularly evident under bare probing, where it achieves the best or joint-best performance on 10 of 12 tasks without downstream feature normalization. AUDIOFM leads all three general-audio

Table 2: **Comparison with representative audio self-supervised learning methods.** AS-20K, AS-2M, and FSD50K report mAP; the other tasks report accuracy (%). AS-2M baseline entries show published / re-evaluated mAP. Re-evaluation uses released fine-tuned checkpoints on the same evaluation subset as our model, without additional training. PT/FT: pre-training/fine-tuning; AS: AudioSet. AS*: our smaller available AudioSet subset (∼1.68M clips). ‡: Re-fine-tuned on our AudioSet-20K training split and evaluated on our retained evaluation subset.

Model	Backbone	Data (PT)	Params		Audio				Speech	
			PT	FT	AS-20K	AS-2M	FSD50K	ESC-50	SPC-2	SPC-1
BEATs$_{iter3}$ (Chen et al., 2023)	ViT-B/16	AS	182M	91M	38.0‡	48.0 / 46.7	57.6	95.6	98.3	97.7
Audio-MAE (Huang et al., 2022a)	ViT-B/16	AS	137M	86M	35.5‡	47.3 / 46.5	57.3	94.1	98.3	96.9
A-JEPA (Fei et al., 2023)	ViT-B/16	AS	354M	86M	38.4	48.6 / -	-	96.3	98.5	97.7
EAT (Chen et al., 2024)	ViT-B/16	AS	93M	88M	40.5‡	48.6 / 48.8	67.5	95.9	98.3	-
SSLAM (Alex et al., 2025)	ViT-B/16	AS	93M	88M	39.3‡	**50.2** / 50.0	66.6	96.2	98.1	98.8
BAT (Ghaffari et al., 2026)	ViT-B/16	AS	93M	88M	40.3‡	48.9 / -	68.0	96.0	99.1	-
AudioMosaic (Huang et al., 2026)	ViT-B/16	AS	86M	86M	41.8‡	**50.2** / 49.2	70.5	97.5	98.4	99.0
AUDIOFM	ViT-B/16	AS*	94M	86M	**42.3**	**50.2**	**71.4**	**98.2**	**99.8**	**99.8**

Table 3: **Linear Probing Performance**. We compare frozen representations evaluated under **Strict Bare Probing** (raw features without normalization) versus **Flexible Probing** (with downstream LayerNorm). Results are means over three random seeds, and all metrics are percentages (%). Best score is in **bold**, and the runner-up is underlined.

			Strict Bare Probing (Raw Features)						Flexible Probing (+LayerNorm)					
Domain	Dataset	Metric	A-MAE	BAT	EAT	SSLAM	AMos	AUDIOFM	A-MAE	BAT	EAT	SSLAM	AMos	AUDIOFM
General	AudioSet-20K	mAP	13.9	18.1	1.1	0.9	28.7	**31.2**	11.9	28.5	18.7	18.2	33.4	**35.0**
	FSD50K	mAP	32.7	53.9	31.4	30.3	53.7	**55.3**	33.8	53.3	39.3	37.0	55.4	**58.6**
	ESC-50	Acc	77.5	92.7	66.6	87.0	93.2	**94.5**	77.0	92.4	83.4	91.5	93.8	**94.7**
Bioacoustics	BEANS-Watkins	Acc	81.9	57.1	61.9	34.5	88.9	**89.1**	70.7	82.2	85.6	82.3	**89.3**	85.9
	BEANS-Dogs	Acc	86.1	59.5	60.9	37.4	**88.7**	85.9	74.6	82.5	**85.4**	78.7	85.1	83.5
	BirdSet-HSN	mAP	12.4	5.0	7.0	6.1	**13.1**	11.7	9.6	9.4	11.1	9.9	11.8	**11.9**
Medical	PSG-Audio (Binary)	Bal. Acc	71.8	75.1	70.2	67.6	77.2	**79.1**	69.7	77.5	75.6	74.2	75.6	**78.1**
	PSG-Audio (Subtype)	Macro-F1	30.4	26.7	21.3	22.9	34.7	**37.9**	26.4	**35.5**	31.7	32.7	31.4	33.1
	CirCor Murmur	Macro-F1	68.0	52.2	28.4	28.4	66.7	**72.7**	68.1	65.8	62.4	59.7	62.0	**68.9**
Speech Emotion	CREMA-D	Macro-F1	64.6	52.1	51.1	39.3	68.2	**68.7**	56.7	**69.8**	65.4	60.6	63.6	64.9
	RAVDESS	Macro-F1	57.5	40.5	41.5	34.9	53.5	**58.9**	47.0	**62.2**	54.0	49.4	53.8	56.8
	TESS	Macro-F1	99.0	84.1	86.4	78.6	**99.8**	**99.8**	91.2	99.1	99.2	98.3	**99.5**	**99.5**

benchmarks under both protocols, while its bare features also achieve the highest scores on all three medical-acoustic tasks. On PSG-Audio subtype classification and CirCor murmur classification, it improves macro-F1 over the strongest baseline by 3.18 and 4.70 percentage points, respectively. These superior results suggest that AudioFM is able to learn rich semantic audio representations.

Fine-grained sound event detection.
Beyond clip-level recognition, we evaluate whether the frozen representations retain the temporal detail needed to detect sound events in finer temporal resolution on AudioSet-Strong. AUDIOFM outperforms the evaluated baselines across frame mAP, PSDS1, and PSDS2 (Figure 3). These gains extend the benefits observed in clip-level classification to both frame-level prediction and event-based detection. Dataset and metric details are provided in Appendix §A.3.2.

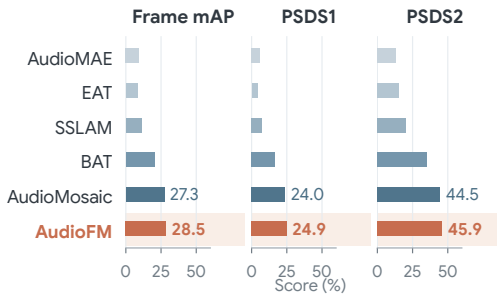

Figure 3: **Frozen-encoder SED on AS-Strong.**

4.3 ABLATION STUDIES AND ANALYSIS

We examine local supervision through frozen-encoder probing, and assess model and data scaling through AS-20K fine-tuning (Figure 4). Full neighborhood configurations and masking-ratio sensitivity results are provided in Appendix §A.1 and Appendix §A.2, respectively.

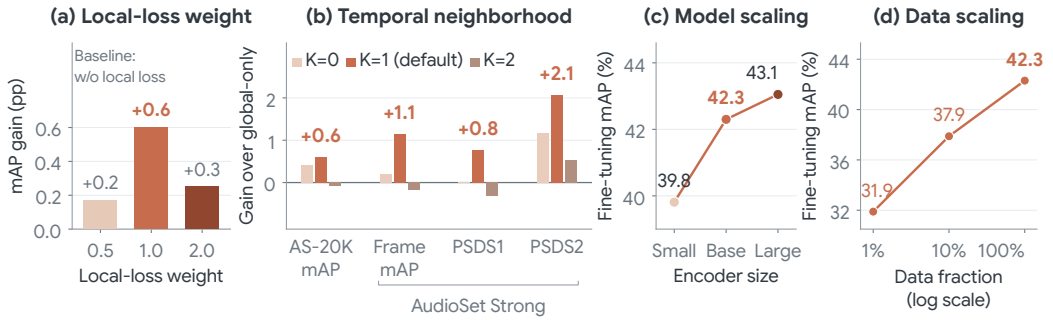

Figure 4: **Ablations and scaling.** (a) Local-loss weight λ_{local} with the default temporal neighborhood ($K = 1$). (b) Temporal-neighborhood radius. Panels (a,b) report frozen-encoder probing gains over training without local loss; (c,d) AS-20K fine-tuning performance as encoder size and pretraining data increase.

Model design. Adding masked-patch supervision improves AS-20K probing over global-only training across all tested positive local-loss weights, with $\lambda_{\text{local}} = 1.0$ performing best (Figure 4(a)). Within the local objective, the compact neighborhood ($K = 1$) outperforms center-only supervision on both AS-20K classification and AudioSet Strong detection, including frame mAP and both PSDS metrics (Figure 4(b)). This supports using immediate temporal neighbors as additional local supervision. The wider $K = 2$ setting performs worse, although it also changes the target weights and therefore does not isolate the effect of temporal span.

Scaling with model size and data. We observe consistent scaling gains in AS-20K fine-tuning with increasing encoder capacity and pretraining data (Figure 4(c,d)). Performance increases from Small to Base to Large, while increasing the pretraining data yields consistent gains across the two orders of magnitude evaluated. These results show that both model capacity and pretraining data contribute to downstream performance within the evaluated range.

4.4 Cross-Domain Generalization

We evaluate whether source-trained predictors transfer to new datasets and speakers without target-domain training, and further examine medical-audio embeddings extracted directly from pretrained encoders to provide a qualitative view of their representations in a specialized acoustic domain (Figure 5).

Sound event predictors transfer across datasets. Here we apply AudioSet-trained classifiers directly to RealDESED (Schmid et al., 2026), mapping their outputs to 12 shared event categories while keeping both the encoder and classification head fixed. AudioFM achieves 79.9% mAP, outperforming all evaluated baselines (Figure 5a). This result demonstrates that the source-trained predictor remains effective on domestic soundscapes without target-specific adaptation. The class mapping is provided in Appendix §A.5.

Emotion recognition across corpora and speakers. We next examine whether emotion classifiers trained on one corpus or speaker remain effective on another. We train linear emotion classifiers on frozen source-domain representations and evaluate them directly on unseen corpora or speakers. The evaluation covers CREMA-D → RAVDESS / TESS and bidirectional transfer between the two TESS speakers. AudioFM ranks first on three of the four transfer tasks and second on the remaining task, with particularly strong results on CREMA-D → RAVDESS and both cross-speaker directions (Figure 5b,c). These results provide evidence that its representations capture emotion-relevant information that transfers beyond the source corpus and speaker. Detailed protocols are provided in Appendix §A.4.

Frozen representation visualization reflects domain-relavent structure Figure 5(d-e) examines how AudioFM's pretrianing shape the zero-shot representations of out-of-domain medcal acoustics (PSG-Audio test cohort). AudioFM distinguishes normal breathing from apnea. Notably, when

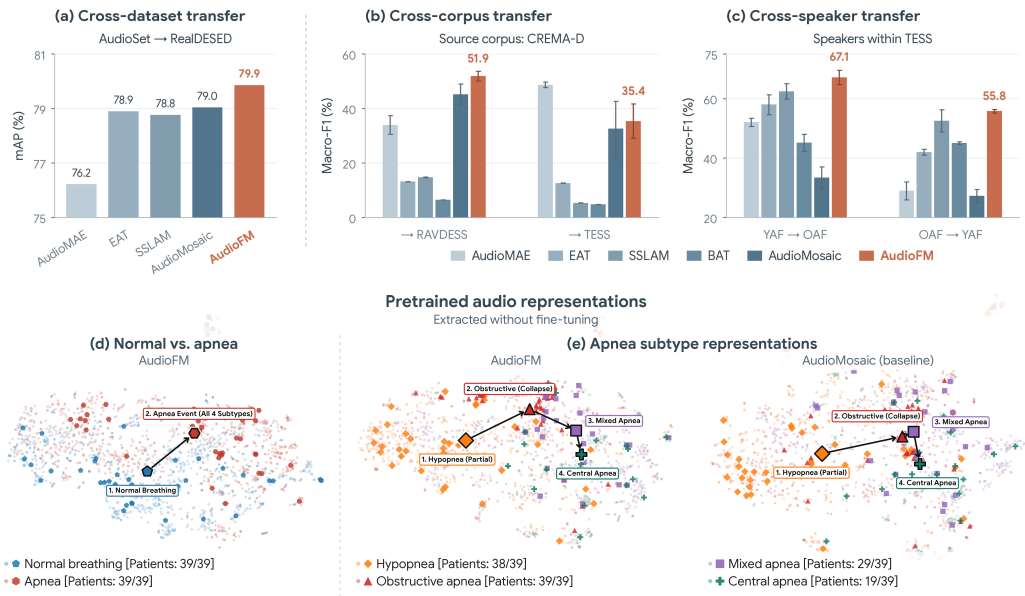

Figure 5: **Cross-domain transfer and pretrained audio representations.** (a–c) Direct transfer without target-domain training: (a) clip-level sound event tagging on RealDESED, (b) cross-corpus and (c) cross-speaker emotion recognition. Error bars indicate standard deviations over three seeds. (d–e) t-SNE visualizations of PSG-Audio embeddings extracted from pretrained encoders without fine-tuning: (d) normal breathing versus apnea using AudioFM, and (e) four apnea subtypes using AudioFM and AudioMosaic.

zooming into the four apnea subtypes, AUDIOFM exhibits a clearer arrangement of subtype centers in the projected embedding space compared with the baseline, spanning hypopnea, obstructive, mixed, and central apnea, motivating further exploration of its use in medical acoustic analysis.

5 DISCUSSION

Limitations. Our pretraining and scaling analyses focus on AudioSet and a limited range of model capacities, with all recordings converted to single-channel audio. Extending this framework to larger, more diverse corpora and multichannel inputs that preserve spatial cues remains an open direction. While our evaluation spans four acoustic domains, robustness to systematic shifts in recording devices and acquisition conditions remains unexplored. Future work could investigate transfer under a broader range of recording conditions and extend the evaluation to additional specialized tasks. In particular, our medical-audio findings highlight the potential of pretrained audio representations for physiological sound analysis, motivating future studies across broader tasks and clinical settings.

Conclusion. We introduced AUDIOFM, a self-supervised audio framework that combines recording-level cross-view contrast with masked latent discrimination on a two-dimensional time–frequency grid. Predicted masked-patch representations identify their corresponding full-view teacher targets among candidates from other recordings. The temporal-neighborhood extension retains this reference-position supervision while incorporating weaker supervision from eligible neighbors. AUDIOFM delivers leading performance across a broad range of audio tasks, matching or outperforming the evaluated state-of-the-art audio SSL baselines on the majority of frozen-encoder classification tasks and surpassing them across all reported temporal localization metrics. Beyond standard fine-tuning benchmarks, our study provides a systematic cross-domain evaluation through 12 frozen-encoder classification tasks spanning general audio, bioacoustics, medical acoustics, and speech emotion, complemented by frame-level sound event detection and cross-corpus and cross-speaker emotion transfer. This combined coverage examines representation quality across acoustic domains, temporal resolutions, and adaptation settings. Together with strong fine-tuning and cross-domain transfer results, these findings demonstrate the effectiveness of combining recording-level alignment with discriminative prediction of masked local representations.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing, grammatical corrections, and stylistic revisions of author-drafted text. They were not involved in the development of the core methodology. The authors made all research decisions and take full responsibility for the manuscript, experimental findings, and claims.

ETHICS STATEMENT

This study evaluates audio representations using publicly available research datasets spanning general audio, bioacoustics, medical acoustics, and speech emotion recognition. The medical-audio experiments are representation-learning benchmarks and do not establish clinical diagnostic validity. Likewise, results on acted speech emotion datasets should not be interpreted as validation for inferring individuals' emotional states in real-world settings. Potential applications raise privacy and misuse concerns, including non-consensual profiling. Deployment would require application-specific validation and appropriate ethical oversight.

REFERENCES

Tony Alex, Sara Atito, Armin Mustafa, Muhammad Awais, and Philip J. B. Jackson. SSLAM: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2506.12222.

Alan Baade, Puyuan Peng, and David Harwath. MAE-AST: Masked autoencoding audio spectrogram transformer. In *Interspeech*, 2022. URL https://www.isca-archive.org/interspeech_2022/baade22_interspeech.html.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020.

Cagdas Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović. A framework for the robust evaluation of sound event detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. EAT: Self-supervised pre-training with efficient audio transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3807–3815, 2024. doi: 10.24963/ijcai.2024/421. URL https://www.ijcai.org/proceedings/2024/421.

Bryan Darquea. PSG-Audio – apnea audios. Kaggle, n.d. URL https://www.kaggle.com/datasets/bryandarquea/psg-audio-apnea-audios. Preprocessed dataset. Accessed September 19, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Houtan Ghaffari, Lukas Rauch, Christoph Scholz, and Paul Devos. BAT: Better audio transformer guided by convex gated probing. *arXiv preprint arXiv:2602.16305*, 2026.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022. doi: 10.1609/aaai.v36i10.21315. URL https://ojs.aaai.org/index.php/AAAI/article/view/21315.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. BEANS: The benchmark of animal sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096686.

Shawn Hershey, Daniel P. W. Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R. Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370, 2021.

Hanxun Huang, Qizhou Wang, Xingjun Ma, Cihang Xie, Christopher Leckie, and Sarah Erfani. AudioMosaic: Contrastive masked audio representation learning. *arXiv preprint arXiv:2605.14231*, 2026. URL https://arxiv.org/abs/2605.14231.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. URL https://arxiv.org/abs/2207.06405.

Wenyong Huang, Zhenhe Zhang, Yu Ting Yeung, Xin Jiang, and Qun Liu. SPIRAL: Self-supervised perturbation-invariant representation learning for speech pre-training. In *International Conference on Learning Representations*, 2022b. URL https://arxiv.org/abs/2201.10207.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Georgia Korompili, Anastasia Amfilochiou, Lampros Kokkalas, Stelios A. Mitilineos, Nicolas-Alexander Tatlas, Marios Kouvaras, Emmanouil Kastanakis, Chrysoula Maniou, and Stelios M. Potirakis. PSG-Audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Scientific Data*, 8:197, 2021. doi: 10.1038/s41597-021-00977-w.

Simon Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md Sazzad Hissain Khan, Keum San Chun, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, et al. Himae: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *International Conference on Learning Representations*, volume 2026, pp. 64859–64895, 2026.

Xian Li, Nian Shao, and Xiaofei Li. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1336–1351, 2024.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 354–379. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.19/.

Zechen Li, Keerthana Natarajan, Weizhi Zhang, Menglian Zhou, Simon A. Lee, Yuwei Zhang, Maxwell A. Xu, Zeinab Esmaeilpour, Flora D. Salim, Mark Malhotra, Lindsey Sunden, Shwetak Patel, Yuzhe Yang, and Ahmed A. Metwally. GlucoFM: A dual-stream foundation model for continuous glucose monitoring. *arXiv preprint arXiv:2605.30865*, 2026. URL https://arxiv.org/abs/2605.30865.

Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13(5):e0196391, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yb4QE6b22f.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021. doi: 10.1109/IJCNN52387.2021.9534474. URL https://doi.org/10.1109/IJCNN52387.2021.9534474.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked modeling duo: Towards a universal audio pre-training framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2391–2406, 2024. doi: 10.1109/TASLP.2024.3389636. URL https://ieeexplore.ieee.org/document/10502167.

Jorge Oliveira, Francesco Renna, Paulo Costa, Marcelo Nogueira, Ana Cristina Oliveira, Andoni Elola, Carlos Ferreira, Alipio Jorge, Ali Bahrami Rad, Matthew Reyna, Reza Sameni, Gari Clifford, and Miguel Coimbra. The CirCor DigiScope phonocardiogram dataset. PhysioNet, Version 1.0.3, 2022a.

Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alipio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, Andoni Elola, Ali Bahrami Rad, Reza Sameni, Gari D. Clifford, and Miguel T. Coimbra. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535, 2022b. doi: 10.1109/JBHI.2021.3137048.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS). Borealis, Version 1, 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. BirdSet: A large-scale dataset for audio classification in avian bioacoustics. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=hJVdwBpWjt.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. URL https://arxiv.org/abs/2010.10915.

Florian Schmid, Paul Primus, Alexander Fichtinger, Tara Jadidi, Tobias Morocutti, and Gerhard Widmer. Realdesed: A real-world domestic sound event detection benchmark. *arXiv preprint arXiv:2607.16736*, 2026.

Raphael Schwinger, Paria Vali Zadeh, Lukas Rauch, Mats Kurz, Tom Hauschild, Sam Lapp, and Sven Tomforde. Foundation models for bioacoustics – a comparative review. *arXiv preprint arXiv:2508.01277*, 2025. URL https://github.com/DBD-research-group/BioFoundation.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic evaluation of audio representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pp. 125–145. PMLR, 2022. URL https://proceedings.mlr.press/v176/turian22a.html.

Luyu Wang, Pauline Luc, Yan Wu, Adrià Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, João Carreira, and Aaron van den Oord. Towards learning universal audio representations. *arXiv preprint arXiv:2111.12124*, 2021a. URL https://arxiv.org/abs/2111.12124.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3023–3032. IEEE, 2021b.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11842–11851, 2021. URL https://proceedings.mlr.press/v139/yang21m.html.

Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.09616.

Yuwei Zhang, Tong Xia, Jing Han, Yu Yvonne Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f803abdcad9de35b45d5a656dade45c-Abstract-Datasets_and_Benchmarks_Track.html.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025. URL https://arxiv.org/abs/2506.09108.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. In *Forty-third International Conference on Machine Learning*, 2026.

Qiang Zhou, Chaohui Yu, Hao Luo, Zhibin Wang, and Hao Li. MimCo: Masked image modeling pre-training with contrastive teacher. *arXiv preprint arXiv:2209.03063*, 2022. URL https://arxiv.org/abs/2209.03063.

A APPENDIX

A.1 MODEL AND TRAINING DETAILS

Table 4 summarizes optimization settings by dataset. Input, model, and pretraining-objective configurations follow the table.

Table 4: **Pre-training (PT) and fine-tuning (FT) hyperparameters** for AUDIOFM.

	Pre-training	Fine-tuning					
Configuration	AS-2M PT	AS-2M	AS-20K	FSD50K	ESC	SPC-2	SPC-1
Optimizer	AdamW						
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$						
Weight decay	0.05	0.01					
LR layer decay	1.0	0.75	0.75	0.75	0.75	0.75	0.75
Base learning rate	6e-4	1.5e-4	5e-5	5e-5	5e-5	5e-4	5e-4
Learning rate schedule	Half-cycle cosine decay						
Minimum learning rate	1e-6	3e-7	1e-7	1e-7	1e-7	1e-6	1e-6
Gradient clipping	1.0	None					
Warm-up steps	2,500	7,812	2,000	2,000	2,000	3,300	2,000
Training steps	125,000	78,125	100,000	200,000	50,000	132,600	40,000
Batch size	6144	256	64	128	32	128	128
Weighted sampling	False	True	False	False	False	False	False
Weighted sampling size	-	200,000	-	-	-	-	-
Augmentation	R	R	R	R	R	R+N	R+N
SpecAug (time/frequency)	-	192/48	192/48	192/48	96/24	48/48	48/48
Drop path	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Dropout	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mixup	0.0	0.8	0.8	0.8	0.0	0.8	0.8
Multilabel	N/A	True	True	True	False	False	False
Loss Function	NT-Xent + MLD	BCE	BCE	BCE	CE	BCE	BCE
Normalization mean	-4.268	-4.268	-4.268	-6.846	-6.627	-6.846	-6.846
Normalization std.	4.569	4.569	4.569	5.565	5.359	5.565	5.565

R: random temporal start with cyclic rolling; N: spectrogram noise at 20 dB SNR. BCE: binary cross entropy; CE: cross entropy; NT-Xent: normalized temperature-scaled cross entropy; MLD: masked latent discrimination.

Model architecture. The encoder is a ViT-Base with 12 layers, hidden dimension $D = 768$, 12 attention heads, and learnable 2D positional embeddings. The shared predictor has four Pre-LN Transformer blocks at bottleneck dimension $D_{\text{pred}} = 384$ with six attention heads, input/output projections between D and D_{pred}, a final LayerNorm, and masking-branch embeddings. Its axis-separated 2D SinCos position encoding concatenates 192-dimensional temporal and frequency components. The global projector outputs 128-dimensional normalized embeddings. The student and EMA teacher local projection heads share the same architecture: a two-layer MLP with LayerNorm, dimensions $768 \to 512 \to 128$, and output ℓ_2 normalization ($d = 128$).

Pretraining objective and updates. We combine $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{MLD}}$ with $\lambda_{\text{local}} = 1.0$, using temperatures $\tau_g = 0.1$ and $\tau_p = 0.2$. The reference neighborhood radius is $K = 1$ (window width $W = 3$), with prior weights $(w_{-1}, w_0, w_{+1}) = (0.1, 0.8, 0.1)$. Each temporal patch step corresponds to 160 ms, so side targets are at offsets ± 160 ms with frequency held fixed. These offsets specify target positions, rather than the encoder's or predictor's receptive field. Weights are renormalized over mask-aware valid offsets using Eq. 3: if only one side neighbor is valid, the center and neighbor receive weights $8/9$ and $1/9$; if neither is valid, the center receives weight 1. Teacher parameters use EMA momentum increasing from 0.990 to 0.9995. Student parameters use AdamW with linear warm-up followed by cosine decay; The remaining optimization settings are given in Table 4.

Neighborhood ablation configurations. The center-only MLD setting corresponds to $K = 0$ ($W = 1$). The default $K = 1$ setting uses prior weights $(0.1, 0.8, 0.1)$, equivalent to a Gaussian sampled at offsets $\{-1, 0, 1\}$ and normalized on this support, with $\sigma = (2 \ln 8)^{-1/2} \approx 0.49$ temporal patch steps; we describe this setting as center-weighted. The wider $K = 2$ ($W = 5$) variant uses prior

weights $(0.03, 0.20, 0.54, 0.20, 0.03)$ at temporal offsets $(-2, -1, 0, 1, 2)$, keeping frequency fixed. All temporal variants apply the same mask-aware selection and per-anchor weight renormalization (Eqs. 2–3). The listed values are prior weights before this selection. The wider variant reduces the center's prior weight from 0.80 to 0.54, so this comparison varies both temporal span and weight distribution.

Linear probing. For the utterance-level probing benchmark (Table 3), we evaluate both a strict linear classifier on frozen encoder features (Bare) and a linear classifier preceded by LayerNorm (Flexible). All probes are trained with AdamW ($\beta_1 = 0.95$, $\beta_2 = 0.999$, weight decay 5×10^{-7}) using a peak learning rate of 1×10^{-3} with linear warm-up followed by decay to 1×10^{-5}. Training steps and batch sizes are scaled to corpus size: 200,000 steps (batch 128) for FSD50K; 50,000 steps (batch 32) for ESC-50; 15,000 steps (batch 64) for BirdSet-HSN; 10,000 steps for AudioSet-20K (batch 128) and PSG-Audio Binary/Subtype (batch 128/64); 5,000 steps for BEANS-Watkins, BEANS-Dogs, CirCor Murmur (batch 32), and CREMA-D (batch 64); and 3,000 steps (batch 32) for RAVDESS and TESS. Short-clip tasks (ESC-50, BEANS-Watkins, and BirdSet-HSN) use a 5.12-second input window, whereas all remaining tasks use a 10.24-second window. No data augmentations (e.g., Mixup or SpecAugment) are applied during linear probing.

A.2 Masking-Ratio Sensitivity

We vary the temporal and frequency masking ratios separately around the default configuration $(r_t, r_f) = (0.60, 0.40)$ and evaluate the resulting representations using AudioSet-20K linear probing (Figure 6). Increasing r_t from 0.60 to 0.75 yields nearly identical mAP (34.98 and 34.97), whereas reducing it to 0.45 lowers mAP to 32.37. For frequency masking, the default $r_f = 0.40$ achieves 34.98 mAP, compared with 34.01 at $r_f = 0.25$ and 32.44 at $r_f = 0.75$. Thus, the tested higher temporal masking ratio retains performance, while both tested alternatives to the default frequency masking ratio reduce it.

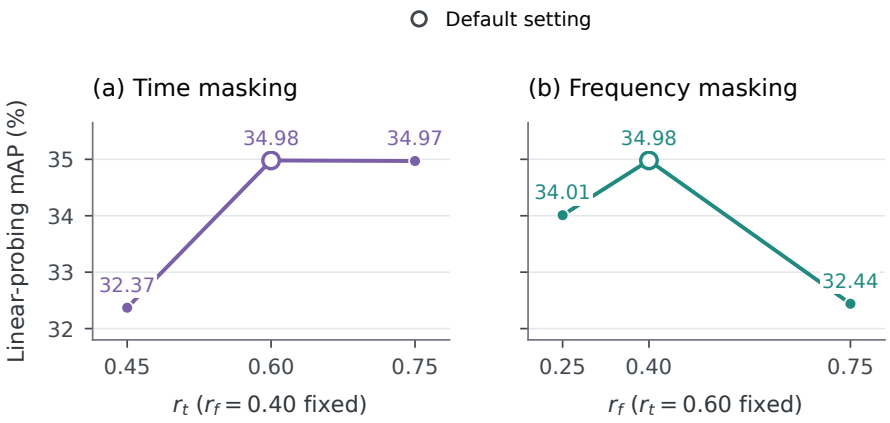

Figure 6: **Sensitivity to structured masking ratios.** AudioSet-20K linear-probing mAP when varying (a) r_t with $r_f = 0.40$ fixed, and (b) r_f with $r_t = 0.60$ fixed. Both panels use the same vertical scale. Hollow markers identify the shared default configuration $(r_t, r_f) = (0.60, 0.40)$.

A.3 Linear Probing: Datasets and Evaluation Protocols

We assess frozen audio representations at two temporal resolutions: utterance-level classification and frame-level sound event detection. Throughout this section, *utterance-level* refers to predictions for an entire audio sample, including both speech and non-speech recordings. The encoder remains frozen while the downstream classifier is trained on labeled examples. Table 5 summarizes the task coverage; optimization settings are given in §A.1.

Table 5: **Linear probing datasets and prediction targets.** Utterance-level tasks span four acoustic domains; AudioSet Strong additionally supports temporal event detection. Entries describe the task coverage, rather than the number of recordings retained in our experiments.

Domain	Dataset	Prediction target
Utterance-level classification		
General audio	AudioSet-20K	Sound event categories
	ESC-50	Environmental sound class
	FSD50K	Sound event categories
Bioacoustics	BEANS-Watkins	Marine mammal species
	BEANS-Dogs	Individual dog identity
	BirdSet-HSN	Bird species present in a soundscape
Medical acoustics	PSG-Audio (Binary / Subtype)	Sleep apnea screening and subtype classification
	CirCor Murmur	Patient-level heart murmur classification
Speech emotion	CREMA-D / RAVDESS / TESS	Expressed emotion
Frame-level sound event detection		
General audio	AudioSet Strong	Event class and temporal extent

A.3.1 Utterance-Level Linear Probing

For each utterance-level task, the classifier maps a frozen audio representation to a single class or a set of labels. Probe training settings are summarized in §A.1. Dataset descriptions below specify the prediction problem, retained sample counts, and evaluation protocol.

General Audio.

- **AudioSet-20K.** AudioSet (Gemmeke et al., 2017) comprises approximately 10-second YouTube excerpts annotated across 527 sound event classes. For linear probing, we train on the balanced AudioSet-20K subset (16,608 retained clips) and evaluate multi-label recognition on the AudioSet evaluation split (16,103 retained clips) using class-averaged mean average precision (macro mAP).
- **ESC-50.** ESC-50 (Piczak, 2015) contains 2,000 five-second environmental recordings organized into 50 balanced classes (40 clips per class). Following the official five-fold partition (1,600 training and 400 test clips per fold), which keeps excerpts from the same source recording within a single fold, we report top-1 classification accuracy averaged across the five folds.
- **FSD50K.** FSD50K (Fonseca et al., 2021) contains variable-duration Freesound recordings with multi-label annotations across 200 sound event categories. The official development set is subdivided into training and validation portions, alongside a separate evaluation set (10,231 clips). For our FSD50K linear-probing and fine-tuning experiments, we combine the official training and validation portions together (40,965 retained clips) for training with fixed hyperparameters, and evaluate on the official evaluation split (10,231 clips) using macro mAP.

Bioacoustics.

We use the BEANS-Watkins, BEANS-Dogs, and BirdSet-HSN benchmarks as integrated in BioFoundation (Hagiwara et al., 2023; Rauch et al., 2025; Schwinger et al., 2025).

- **BEANS-Watkins.** Derived from the Watkins Marine Mammal Sound Database, this benchmark defines a 31-class marine mammal species classification task over underwater vocalizations (1,017 training and 339 test clips), evaluated by top-1 accuracy.
- **BEANS-Dogs.** This task contains domestic dog barks annotated by individual identity across ten dogs (415 training and 139 test clips). We treat it as 10-class individual identification and report top-1 accuracy.
- **BirdSet-HSN.** The High Sierra Nevada (HSN) benchmark covers 21 bird species, pairing focal Xeno-Canto training recordings (5,460 five-second segments) with annotated 5-second soundscape

evaluation windows (12,000 test segments). Because multiple species can co-occur within a window, we train with binary cross-entropy and report macro mAP.

Medical Acoustics.

These tasks evaluate transfer to clinical respiratory and cardiac recordings under strict patient-disjoint partitions.

- **PSG-Audio (Binary and Subtype).** PSG-Audio (Korompili et al., 2021) pairs overnight polysomnography with synchronized microphone recordings for sleep apnea diagnosis. Using the 16 kHz preprocessed release (Darquea, n.d.), which segments the microphone channel into 10-second windows (103,210 segments across 192 patients: 64,585 respiratory-event segments and 38,625 normal-breathing segments), we construct a patient-disjoint split stratified by per-patient apnea burden: 134 training patients (70,825 segments), 19 validation patients (11,036 segments), and 39 test patients (21,349 segments). Because baseline recording gain varies by up to 22 dB across subjects—whereas per-segment normalization would erase the relative amplitude drop between quiet apneic phases and recovery breaths—we normalize all segments of each patient by that patient's global full-night root-mean-square (RMS) energy. We evaluate two tasks: (i) **PSG-Audio (Binary)**, which distinguishes normal breathing from any apneic event across all retained segments (70,825 train / 21,349 test) using balanced accuracy; and (ii) **PSG-Audio (Subtype)**, which performs 4-class differential diagnosis among pathological events (*hypopnea*, *mixed apnea*, *obstructive apnea*, and *central apnea*; 45,120 train / 13,765 test segments) using macro-averaged F1 (Macro-F1).
- **CirCor Murmur.** The CirCor DigiScope Phonocardiogram dataset (Oliveira et al., 2022b;a) contains 3,163 heart-sound recordings from 942 pediatric patients, captured across up to four standard auscultation locations per patient and annotated for murmur presence (*absent*, *present*, or *unknown*). We use a stratified patient-disjoint 80/20 split comprising 755 training patients (2,534 recordings) and 187 test patients (629 recordings). Because CirCor recordings are natively sampled at 4 kHz with diagnostic cardiac energy concentrated in 20–600 Hz, standard resampling to 16 kHz would compress the signal into the lowest bins of an 8 kHz Mel filterbank while leaving frequencies above 2 kHz empty. Instead, we feed the native 4 kHz samples directly to the 16 kHz frontend without interpolation ($4\times$ rate shift), mapping the 20–600 Hz cardiac band into 80–2,400 Hz where pretrained Mel filterbanks have high resolution. To cover full recordings (median duration 21.5 s), we extract overlapping 10.24-second windows (5.12-second stride) and mean-pool their frozen embeddings into a single recording-level feature. Probes are trained on recording-level features using patient murmur labels; at test time, predictions across all auscultation locations of a patient are aggregated via the official PhysioNet clinical escalation rule (*present* $>$ *unknown* $>$ *absent*) and evaluated by patient-level Macro-F1.

Speech Emotion Recognition.

We evaluate frozen representations on three acted-speech corpora to examine whether self-supervised acoustic representations retain vocal affect and paralinguistic cues. CREMA-D and RAVDESS use **actor-disjoint** partitions to assess generalization to unseen speakers, while TESS uses a **word-disjoint** partition to assess generalization to unseen lexical content. Table 3 reports mean performance across three random initialization seeds ($N = 3$) under both Bare probing (without downstream LayerNorm) and LN probing (with downstream LayerNorm).

- **CREMA-D.** CREMA-D (Cao et al., 2014) comprises audio-visual recordings of twelve sentences spoken by 91 actors (48 male, 43 female) of diverse ages and ethnic backgrounds across six core emotion categories: anger, disgust, fear, happiness, neutrality, and sadness. We exclude a single corrupted zero-energy recording (`1076_MTI_SAD_XX`), yielding 7,441 valid clips. We establish an actor-disjoint partition: 73 actors (5,965 clips) form the training set, while 18 unseen actors (1,476 clips) form the test set. The test cohort is strictly stratified to balance sex (9 male, 9 female) and demographic diversity. The prediction target is 6-class classification, evaluated by macro-averaged F1 (Macro-F1).
- **RAVDESS.** The Ryerson Audio-Visual Database of Emotional Speech and Song (Livingstone & Russo, 2018) contains 1,440 speech recordings produced by 24 professional actors (12 male, 12 female) across eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

We employ the audio speech subset and construct an actor-disjoint split: Actors 01–20 (10 male, 10 female; 1,200 clips) serve as training data, while Actors 21–24 (2 male, 2 female; 240 clips) are held out exclusively for testing. The prediction target is 8-class classification, evaluated by Macro-F1.

- **TESS.** The Toronto Emotional Speech Set (Pichora-Fuller & Dupuis, 2020) contains 2,800 recordings from two actresses: a young adult female (YAF, 26 years old) and an older adult female (OAF, 64 years old). Each speaker utters 200 target words embedded in the carrier phrase *"Say the word [target]"* across seven emotional states: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. For utterance-level probing, we use a **word-disjoint** partition over the 200 NU-6 target words: 160 words (2,240 clips, 1,120 per speaker) form the training set, while 40 unseen words (560 clips, 280 per speaker) form the test set, fully balanced across both speakers and seven emotions. Both speakers appear in training and test data, so this protocol evaluates generalization to unseen words. Cross-speaker transfer is evaluated separately in §A.4. We evaluate 7-class classification using Macro-F1.

A.3.2 Frame-Level Linear Probing for Sound Event Detection

AudioSet Strong and temporal targets. We use AudioSet Strong (Hershey et al., 2021) to assess whether frozen representations support temporally localized event predictions. Unlike utterance-level classification, this task retains the onset and offset annotations and predicts class activity along the recording. A linear classification layer operates on the frozen time-indexed representations; this preserves the distinction between evaluating the encoder's temporal information and adapting the encoder through full fine-tuning.

Frame-wise and event-based metrics. Frame-wise mAP evaluates class-specific prediction rankings over annotated time steps, with the averaging convention stated alongside the results. PSDS (Bilen et al., 2020) evaluates detected event intervals across operating points using temporal-intersection criteria. It therefore complements frame-wise mAP rather than serving as another name for it. We report PSDS1 and PSDS2 using the Polyphonic Sound Detection Score framework (Bilen et al., 2020), which evaluates event detection across operating points using temporal-intersection criteria.

Neighborhood ablation protocol. The neighborhood variants use the same frozen-encoder SED protocol as the baseline comparison in Figure 3. Their classification and SED results are summarized in Figure 4(b) and analyzed in §4.3. The prior weights and mask-aware target selection are specified in Appendix §A.1.

A.4 Zero-Shot Emotion Transfer: Protocol Details

This section details the source and target splits for the zero-shot emotion transfer evaluation in §4.4. The pretrained encoder remains frozen during source-domain probe training. The source-trained classifier is then applied to the target corpus or speaker without target-domain gradient updates. Figure 5(b,c) reports the results under Bare probing, without downstream LayerNorm or feature normalization.

Cross-corpus transfer (CREMA-D → RAVDESS / TESS). A linear probe is trained on the training partition of CREMA-D (73 actors, 5,965 clips) using its six emotion classes (*anger, disgust, fear, happiness, neutrality, sadness*) and evaluated directly on the corresponding six-class subsets of the target corpora:

- **RAVDESS:** Excluding *calm* and *surprised* leaves 1,056 speech clips across the six shared classes.
- **TESS:** Excluding *pleasant surprise* leaves 2,400 clips across the six shared classes.

These settings assess transfer across corpora with different speakers, lexical content, and recording conditions.

Cross-speaker transfer (TESS YAF ↔ OAF). We train a 7-class linear probe on all 1,400 clips of the younger speaker (YAF, 26 years old) and evaluate it directly on all 1,400 clips of the older

speaker (OAF, 64 years old), then reverse the direction. This protocol isolates generalization across speaker age and vocal timbre under matched lexical content.

Audio preprocessing, probe settings, and reporting. All emotion audio recordings are resampled to 16 kHz mono. Because speech emotion utterances are typically 1.5–3.5 seconds long, each clip undergoes deterministic symmetric zero-padding to match the 10.24-second input window (163,840 samples) expected by the pretrained audio encoders, centering the utterance without altering speech tempo or pitch. A single pooled representation is extracted per clip from the frozen encoder. Probes are trained directly on these clip-level representations using strict linear classification without LayerNorm or downstream feature normalization. Probe parameters are optimized on source-domain training data using cross-entropy loss, and we report Macro-F1 (%) (mean $\pm$ standard deviation across three random initialization seeds, $N = 3$) evaluated on the target domain.

A.5 Zero-Shot RealDESED Evaluation Protocol and Class Mapping

Table 6 specifies the ontology-based class mapping used for zero-shot cross-dataset sound event tagging on RealDESED (§4.4). We evaluate 12 of its 15 event classes, each mapped to one or more corresponding AudioSet categories, while excluding the remaining three classes (*coffee machine*, *light switch*, and *window open/close*) that lack direct AudioSet counterparts.

Inference and aggregation protocol. We evaluate all 1,008 recordings in the RealDESED test split (resampled from 48 kHz to 16 kHz mono), collapsing the onset/offset event annotations into clip-level multi-hot ground-truth targets indicating whether each class occurs anywhere within a recording (retaining recordings without target events as all-negative examples). Because RealDESED recordings extend up to $\sim$31 seconds and exceed the 10.24-second input context of the pretrained encoders, each recording is partitioned into consecutive non-overlapping 10.24-second windows (with the final tail window zero-padded to 10.24 seconds). Each window is scored independently by the frozen encoder and its AudioSet-2M fine-tuned 527-way classification head, and the resulting 527-dimensional predictions are aggregated across all windows of a recording via element-wise max-pooling. Finally, for RealDESED classes mapped to multiple AudioSet categories in Table 6, the target class score is obtained by taking the maximum score over its mapped AudioSet nodes, and we report macro-averaged mAP across the 12 classes.

Table 6: **Ontological Mapping Between RealDESED Classes and AudioSet Categories** for zero-shot sound event tagging.

RealDESED Target Class	AudioSet Ontology ID	AudioSet Label Description
bell_ringing	/m/0395lw	Bell
cutlery_dishes	/m/04brg2, /m/023pjk	Dishes, pots, and pans; Cutlery, silverware
door_open_close	/m/02dgv	Door
footsteps	/m/07pbtc8	Walk, footsteps
keyboard_typing	/m/01m2v, /m/0316dw	Computer keyboard; Typing
keychain	/m/03v3yw	Keys jangling
microwave	/m/0fx9l	Microwave oven
phone_ringing	/m/07pp8cl, /m/01hnzm	Telephone bell ringing; Ringtone
running_water	/m/02jz0l	Water tap, faucet
toilet_flushing	/m/01jt3m	Toilet flush
vacuum_cleaner	/m/0d31p	Vacuum cleaner
wardrobe_drawer_open_close	/m/0642b4, /m/0fqfqc	Cupboard open or close; Drawer open or close