LEARNING ALONG DEPTH: DEPTH-WEIGHTED LAYER SUPERVISION FOR SPEECH ANTI-SPOOFING

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in neural speech synthesis have made audio deepfakes increasingly hard to detect, and self-supervised Speech Foundation Models (SFMs) have become the standard front-end for anti-spoofing. Standard fine-tuning, however, supervises only the final representation. Intermediate layers therefore remain weakly adapted to the task, and the burden of discrimination shifts to ever more expressive back-end classifiers. In this work, we ask how supervision can instead reshape the internal representation hierarchy of the SFM encoder. Our layer-wise analysis of fine-tuned SFM shows that anti-spoofing discriminability emerges late in the network, and later still as model size grows, so most intermediate layers remain poorly aligned with the task. Motivated by this, we propose Depth-Weighted Layer Supervision (DWLS), an effective multi-layer supervision framework that progressively aligns representations across the encoder depth with the anti-spoofing objective. Across diverse deepfake benchmarks, DWLS paired with a lightweight classifier achieves competitive in-domain performance and strong out-of-domain generalization, suggesting that improving internal feature alignment can be a more efficient route to adaptation than increasing back-end complexity.

1 INTRODUCTION

Recent progress in neural speech synthesis has produced systems that generate natural, speaker-consistent audio (Tan et al., 2021). Models based on diffusion, flow matching, and large-scale neural vocoders now yield speech that is often hard to tell from genuine recordings (O'Shaughnessy, 2026). These advances benefit accessibility, dubbing, and conversational AI, but they also enable identity fraud, impersonation (Folorunsho & Boamah, 2025), and attacks on Automatic Speaker Verification (ASV) systems (Wang et al., 2024). In response, the anti-spoofing community has widely adopted self-supervised Speech Foundation Models (SFMs) as front-end feature extractors. Models such as `Wav2Vec2` (Baevski et al., 2020), pretrained on large unlabeled speech corpora, provide representations that clearly outperform handcrafted features such as CQCCs (Todisco et al., 2017), as shown by prior work (Kulkarni et al., 2024; Ge et al., 2025; Ali et al., 2026).

Yet how the layers of a SFM contribute to deepfake detection (El Kheir et al., 2025) still remains underexplored. Most work has instead focused on the back-end classifier (Tak et al., 2022b; Rosello et al., 2023; Xiao & Das, 2025; Liu et al., 2025a). Pretrained SFMs are known to be redundant across adjacent layers (Pasad et al., 2021; 2023; Dorszewski et al., 2025), a consequence of their task-agnostic objectives, and standard fine-tuning does not totally remove this redundancy. Because supervision reaches only the final layer, earlier layers receive weaker gradient signal and tend to stay close to their pretrained geometry (Papyan et al., 2020; Gromov et al., 2025; Súkeník et al., 2025). We argue that explicit layer-wise supervision is needed to restructure this internal feature space.

Accordingly, we prioritize front-end representation learning over back-end complexity. Our hypothesis is that *a well-aligned feature space reduces the need for heavy classifiers*. A preliminary analysis shows that, in post-trained SFMs, anti-spoofing discriminability appears only in the last layers, and increasingly late as models grow. We therefore introduce **Depth-Weighted Layer Supervision (DWLS)**, a technique designed to enforce layer-wise alignment throughout the encoder. With only lightweight classifier, DWLS achieves competitive in-domain results and generalizes well among the compared State-Of-The-Art (SOTA) systems, with strong gains on out-of-domain data.

To summarize, this paper makes the following main contributions:

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.

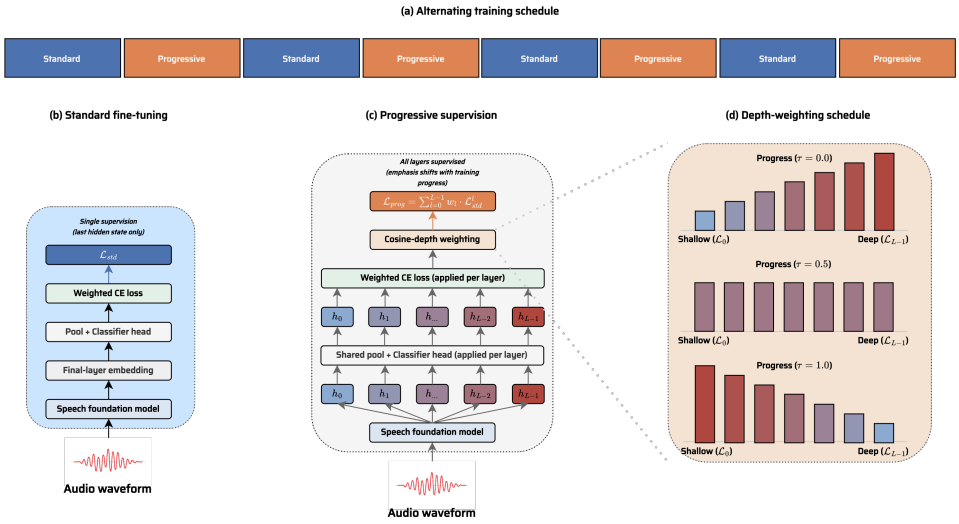


Figure 1: **Overview of the Depth-Weighted Layer Supervision.** Training alternates (**a**) between standard fine-tuning (**b**) and progressive supervision (**c**). During progressive phases, hidden states $h_0, \ldots, h_{L-1}$ pass through a shared pool and classifier, supervised by a depth-weighting schedule that shifts emphasis from deep to shallow layers (**d**).

1. We analyze layer-wise discriminability and representation in post-trained SFMs, showing that anti-spoofing capacity emerges late in the encoder and later as model size grows.
2. We propose DWLS, a multi-layer supervision scheme that explicitly realigns multi-layer encoder features toward anti-spoofing objectives.
3. We show that DWLS with a lightweight classifier is competitive with or better than larger back-ends on several in-domain and out-of-domain benchmarks, with the largest gains on out-of-domain data.

2 Related Work

Anti-Spoofing Countermeasures. Early detectors relied on handcrafted spectral features such as MFCCs, LFCCs, and CQCCs (Lei & Lopez, 2009; Todisco et al., 2017). The field has since moved to self-supervised SFMs, including `Wav2Vec2` (Baevski et al., 2020), `WavLM` (Chen et al., 2022), and `XLS-R` (Babu et al., 2022), which perform strongly across speech tasks (wen Yang et al., 2021) and now dominate anti-spoofing (Ali et al., 2026; Dowerah et al., 2026). Back-end classifiers on top of these features have evolved from recurrent networks (Guan et al., 2025) to graph neural networks (Tak et al., 2022b; Yang et al., 2025a; Wen et al., 2026) and Transformer variants (Rosello et al., 2023; Truong et al., 2024; 2025; Tran et al., 2026a). To reduce the cost of full self-attention, recent work uses state-space models (Tran et al., 2025b; Xiao & Das, 2025; Ng et al., 2026) and gated MLPs (Tran et al., 2025c), and replaces linear heads with Kolmogorov-Arnold Networks (KANs) (Phuong et al., 2025; Tran et al., 2025a; 2026b). Despite these advances, generalization to unseen attacks and domains remains an open problem (Müller et al., 2022; 2024).

Layer-Wise Feature Exploitation in SFMs. Information is distributed across SFM depth: shallow layers mostly capture acoustic and speaker-related traits, while deeper layers encode more abstract linguistic content (Pasad et al., 2021; 2023). Early anti-spoofing work exploited this with the weighted-layer-sum scheme popularized by SUPERB (wen Yang et al., 2021). Later methods include attentive layer merging (Pan et al., 2024), multi-fusion attentive pooling (Guo et al., 2024; Tran et al., 2024), dynamic weighting (Martín-Doñas & Álvarez, 2022; El Kheir et al., 2025), attention-based aggregation (Zhang et al., 2024; Xiao & Vu, 2025; Tran et al., 2025a), and gating mechanisms (Tran et al., 2025b;c). These methods all address how to *aggregate* intermediate features; none changes how the layers are trained. DWLS does the latter.

Representational Dynamics in Deep Encoders. Studies of self-supervised Transformer encoders report high redundancy across layers (Pasad et al., 2021; 2023; Dorszewski et al., 2025), and many layers can be pruned with little loss on downstream tasks (Dorszewski et al., 2025). Related work links this to layer-wise saturation (Gromov et al., 2025; Jiang et al., 2025; Súkeník et al., 2025) and to neural collapse near the network head (Papyan et al., 2020), both of which can leave intermediate features poorly aligned with the target task. Intermediate layers can also carry downstream-relevant cues that the final layer discards (Batra et al., 2026), and similar depth-wise saturation has been observed in Transformer back-ends for anti-spoofing (Tran et al., 2026a). Existing remedies mostly prune layers after training or force early layers to imitate the final output. We instead use a multi-stage, depth-weighted supervision schedule to align earlier representations directly with the task.

3 Analysis of Post-Trained Layer-Wise Representations

To empirically characterize how deepfake-discriminative information evolves across depth, we first evaluate post-trained countermeasures from Ge et al. (2025) on the In-The-Wild (`ITW`) dataset (Müller et al., 2022). Chosen for their robust generalization from large-scale post-training on tens of thousands of hours of diverse genuine and spoofed speech, we examine layer-wise detection performance using Equal Error Rate (EER %) alongside two representational geometry diagnostics such as angular distance and Centered Kernel Alignment (CKA) similarity - both measured relative to each model's last hidden state.

3.1 Problem Formulation

Speech Foundation Model as a Residual Encoder. A `Wav2Vec2`-style (`XLS-R`) SFM maps a raw waveform $\mathbf{a} \in \mathbb{R}^{T_a}$ to a sequence of latent speech units through a convolutional feature encoder $\mathcal{Z} = \mathrm{CNN}(\mathbf{a}) \in \mathbb{R}^{d \times T}$. This representation is contextualized by a transformer encoder consisting of $L-1$ blocks, yielding L total hidden states indexed by $\ell \in \{0, \ldots, L-1\}$, where $\ell = 0$ denotes the CNN feature output and $\ell = L-1$ denotes the last transformer layer. The frame-level hidden state $\boldsymbol{H}^{(\ell)} \in \mathbb{R}^{d \times T}$ evolves as

$$\boldsymbol{H}^{(\ell+1)} = \boldsymbol{H}^{(\ell)} + f_{\theta^{(\ell)}}\big(\boldsymbol{H}^{(\ell)}\big), \qquad \ell = 0, \ldots, L-2, \tag{1}$$

where $\boldsymbol{H}^{(0)} = \mathcal{Z}$ and $f_{\theta^{(\ell)}}(\cdot)$ denotes the composition of multi-head self-attention and feed-forward sublayers with parameters $\theta^{(\ell)}$. Unrolling equation 1 expresses the last representation as an additive sum over all block transformations:

$$\boldsymbol{H}^{(L-1)} = \boldsymbol{H}^{(0)} + \sum_{\ell=0}^{L-2} f_{\theta^{(\ell)}}\big(\boldsymbol{H}^{(\ell)}\big), \tag{2}$$

which motivates treating anti-spoofing discriminability as a property that is distributed across depth rather than concentrated solely at $\ell = L-1$.

For an utterance $\mathbf{a}_i$ with ground-truth label $y_i \in \{\text{bona fide}, \text{spoof}\}$, we obtain a fixed-dimensional layer representation via:

$$\boldsymbol{h}_i^{(\ell)} = \mathrm{Pool}\big(\boldsymbol{H}_i^{(\ell)}\big) \in \mathbb{R}^d, \qquad \ell = 0, \ldots, L-1, \tag{3}$$

where $\mathrm{Pool}(\cdot)$ denotes mean pooling over the frame dimension T. A lightweight fully connected linear back-end classifier $g(\cdot)$, parameterized by $\boldsymbol{W} \in \mathbb{R}^{2 \times d}$, is applied to the final representation to yield the predicted class $g(\boldsymbol{h}_i^{(L-1)}) = \arg\max_{k \in \{1,2\}} \big[\mathrm{SoftMax}(\boldsymbol{W}\boldsymbol{h}_i^{(L-1)})\big]_k$.

Layer-Wise Discriminability. To probe whether intermediate layers carry usable anti-spoofing information, we adopt a logit-lens-style diagnostic (Jiang et al., 2025): the terminal classifier $\boldsymbol{W} \in \mathbb{R}^{2 \times d}$ is applied directly to $\boldsymbol{h}_i^{(\ell)}$ at every depth ℓ. We compute the layer-wise Log-Likelihood Ratio (LLR) score for utterance i as:

$$s_i^{(\ell)} := [\boldsymbol{W}\boldsymbol{h}_i^{(\ell)}]_0 - [\boldsymbol{W}\boldsymbol{h}_i^{(\ell)}]_1, \qquad \ell = 0, \ldots, L-1, \tag{4}$$

where index 0 corresponds to the bona fide hypothesis $\mathcal{H}_0$ and index 1 to the spoof hypothesis $\mathcal{H}_1$. Given an evaluation dataset $\mathcal{D} = \{(\mathbf{a}_i, y_i)\}_{i=1}^N$, the layer-wise EER is defined as

$$\mathrm{EER}^{(\ell)}(\mathcal{D}_{\texttt{ITW}}) := \mathrm{EER}\Big(\{s_i^{(\ell)}\}_{i=1}^N, \{y_i\}_{i=1}^N\Big), \qquad \ell = 0, \ldots, L-1. \tag{5}$$

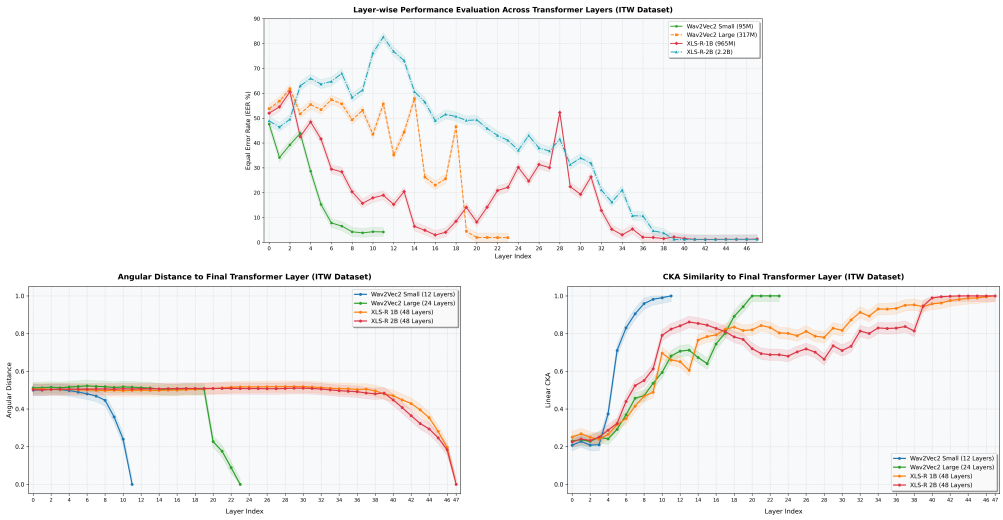


Figure 2: Layer-wise performance and geometric alignment analysis across speech foundation models on the In-The-Wild (ITW) benchmark. **(Top)** Layer-wise Equal Error Rate ($\mathrm{EER}^{(\ell)}$, %) across Transformer layers for Wav2Vec2 and XLS-R variants (lower is better). **(Bottom-Left)** Angular distance ($\mathrm{AngDist}^{(\ell)}$) relative to the final Transformer layer representation. **(Bottom-Right)** Linear CKA similarity ($\mathrm{CKA}^{(\ell)}$) relative to the final layer.

Representational Geometry Diagnostics. To disentangle *task-discriminative capacity* (Eq. 5) from *global geometric alignment*, we measure how far each intermediate representation sits from the terminal hidden state using two complementary diagnostics.

(i) Angular distance. We define the mean angular distance between layer ℓ and the final layer $L-1$, averaged over N evaluation samples, as

$$\mathrm{AngDist}^{(\ell)}(\mathcal{D}_{\mathtt{ITW}}) := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi}\arccos\left(\frac{\boldsymbol{h}_i^{(\ell)}\cdot\boldsymbol{h}_i^{(L-1)}}{\left\|\boldsymbol{h}_i^{(\ell)}\right\|_2\left\|\boldsymbol{h}_i^{(L-1)}\right\|_2}\right)\in[0,1]. \tag{6}$$

A small $\mathrm{AngDist}^{(\ell)}$ indicates that layer ℓ is directionally aligned with the final representation.

(ii) Centered Kernel Alignment. We complement the per-sample angular distance with linear CKA (Kornblith et al., 2019), which captures inter-example similarity structure. Let $\boldsymbol{Z}^{(\ell)} = [\boldsymbol{h}_1^{(\ell)},\ldots,\boldsymbol{h}_N^{(\ell)}]\in\mathbb{R}^{d\times N}$ collect the pooled representations across all N evaluation utterances at layer ℓ, centered such that $\sum_{i=1}^{N}\boldsymbol{h}_i^{(\ell)}=0$. The layer-wise linear CKA is defined as

$$\mathrm{CKA}^{(\ell)}(\mathcal{D}_{\mathtt{ITW}}) := \frac{\left\|(\boldsymbol{Z}^{(L-1)})^\top\boldsymbol{Z}^{(\ell)}\right\|_F^2}{\left\|\boldsymbol{Z}^{(\ell)}(\boldsymbol{Z}^{(\ell)})^\top\right\|_F\left\|\boldsymbol{Z}^{(L-1)}(\boldsymbol{Z}^{(L-1)})^\top\right\|_F}\in[0,1], \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. $\mathrm{CKA}^{(\ell)}\to 1$ indicates that the inter-example similarity structure at layer ℓ has converged to that of the last layer.

3.2 Empirical Observations

Discriminability emerges later in larger models. Fig. 2 shows that the depth at which anti-spoofing capacity ($\mathrm{EER}^{(\ell)}$, Eq. 5) appears grows with model size. For Wav2Vec2 Small (12 layers, 95M parameters), $\mathrm{EER}^{(\ell)}$ falls steadily from 47.6% at layer 0 to 3.86% at layer 9. Wav2Vec2 Large (24 layers, 317M) stays at 43–58% EER for about two-thirds of its depth and drops below 5% only after layer 19, reaching 1.92% at layer 23. In the 48-layer XLS-R models the pattern is more pronounced: XLS-R-1B fluctuates between 10% and 52% EER in its middle layers and stabilizes below 2% only after layer 38, while XLS-R-2B (2.2B) is uninformative for roughly 80% of its depth (up to 82.62% EER at layer 11) before dropping to about 1.2% in layers 39–47. Under standard fine-tuning, larger backbones therefore defer detection to a small final fraction of depth.

Geometric reorganization coincides with the EER drop. The geometry diagnostics point the same way. For most of the depth, intermediate features remain distant from the final-layer geometry ($\mathrm{AngDist}^{(\ell)} \approx 0.5$, $\mathrm{CKA}^{(\ell)} \approx 0.2$–$0.8$). Only in the last layers do they converge ($\mathrm{AngDist}^{(\ell)} \to 0$, $\mathrm{CKA}^{(\ell)} \to 1$), and the onset of this convergence lines up with the EER drop: around layer 8 for `Wav2Vec2 Small`, layer 19 for `Wav2Vec2 Large`, and layer 38 for both `XLS-R` models. This suggests that, under standard fine-tuning, discriminative structure is formed in a short late reorganization phase rather than gradually across depth.

4 Proposed Method

Motivated by our layer-wise analysis, we introduce Depth-Weighted Layer Supervision (DWLS), a training procedure that alternates between standard final-layer supervision and an auxiliary multi-layer objective, with the relative contribution of each layer supervised by a cosine depth-weighting schedule, described in Figure 1.

Alternating Training Schedule. Training is parameterized entirely by optimization steps. Let t denote the global optimization step index, W represent the number of warm-up steps (corresponding to an initial standard fine-tuning phase), and N denote the fixed step length for all subsequent alternating phases. For $t \geq W$, the active phase index $\phi(t) \in \{0, 1\}$ is determined by:

$$k(t) = \left\lfloor \frac{t - W}{N} \right\rfloor, \qquad \phi(t) = k(t) \bmod 2, \tag{8}$$

where $\phi(t) = 0$ corresponds to standard fine-tuning and $\phi(t) = 1$ corresponds to progressive multi-layer supervision. In phase $\phi = 0$ (and during warm-up $t < W$), only the pooled final-layer representation $\boldsymbol{H}^{(L-1)}$ is classified using the standard binary cross-entropy (CE) loss $\mathcal{L}_{\text{std}}$. In phase $\phi = 1$, intermediate representations are directly exposed to task supervision.

Multi-Layer Auxiliary Classification. During phase 1, we extract hidden representations $\boldsymbol{H}^{(\ell)} \in \mathbb{R}^{d \times T}$ across all depth levels $\ell = 0, \ldots, L-1$. Each representation is pooled using a shared function $\mathrm{Pool}(\cdot)$ and passed through a shared classification head $f_\theta(\cdot)$ to produce per-layer prediction logits:

$$\boldsymbol{h}_i^{(\ell)} = \mathrm{Pool}\big(\boldsymbol{H}_i^{(\ell)}\big) \in \mathbb{R}^d, \quad \hat{y}_i^{(\ell)} = g\big(\boldsymbol{h}_i^{(\ell)}\big). \tag{9}$$

For each representation, we evaluate the CE loss:

$$\mathcal{L}_{\text{std}}^{(\ell)} = \mathrm{CE}\left(\hat{y}^{(\ell)}, y; \mathbf{w}_c\right), \tag{10}$$

where $\mathbf{w}_c$ compensates for class imbalance.

Cosine Depth-Weighting Schedule. Let $\rho_\ell = \frac{\ell}{L-1} \in [0, 1]$ denote the normalized depth of layer ℓ, and let $r_t = (t - W) \bmod N \in \{0, \ldots, N-1\}$ denote the steps within the active progressive phase. The normalized progress through the phase is defined as $p_t = \frac{r_t}{N-1} \in [0, 1]$. We dynamically modulate the layer importance:

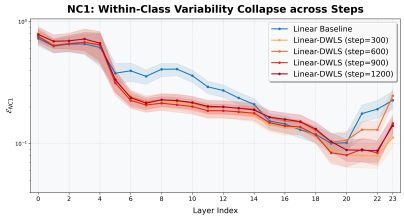
$$w_\ell(p_t) = \max\left(0, 1 + \cos(\pi p_t)(2\rho_\ell - 1)\right), \qquad \tilde{w}_\ell(p_t) = \frac{w_\ell(p_t)}{\sum_{\ell'} w_{\ell'}(p_t)}. \tag{11}$$

At the start of a phase ($p_t = 0$), $w_\ell \propto 2\rho_\ell$ and supervision concentrates on deeper layers; at the phase midpoint ($p_t = 0.5$), weighting becomes uniform ($\tilde{w}_\ell = \frac{1}{L}$); at the end ($p_t = 1$), $w_\ell \propto 2(1 - \rho_\ell)$ shifts focus toward shallow layers. The progressive objective is:

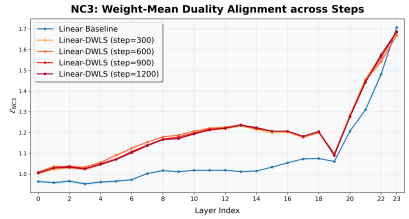
$$\mathcal{L}_{\text{prog}}(t) = \sum_{\ell=0}^{L-1} \tilde{w}_\ell(p_t) \mathcal{L}_{\text{std}}^{(\ell)}. \tag{12}$$

Overall Objective. The step-wise optimization target $\mathcal{L}(t)$ is summarized in Algorithm 1:

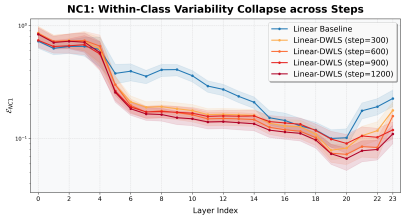
$$\mathcal{L}(t) = \begin{cases} \mathcal{L}_{\text{std}}(t) & \text{if } \phi(t) = 0 \text{ or } t < W, \\ \mathcal{L}_{\text{prog}}(t) & \text{if } \phi(t) = 1 \text{ and } t \geq W. \end{cases} \tag{13}$$



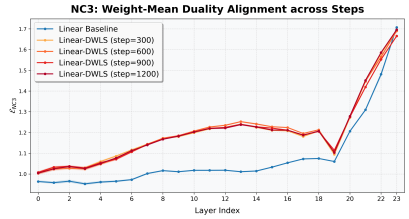
(a) NC1, deep-to-shallow



(b) NC3, deep-to-shallow



(c) NC1, shallow-to-deep



(d) NC3, shallow-to-deep

Figure 3: **Effect of layer-weighting order on Neural Collapse geometry.** Layer-wise NC1 (within-class variability) and NC3 (classifier–mean alignment) for the linear baseline and Linear-DWLS across training steps, under **(a, b)** deep-to-shallow and **(c, d)** shallow-to-deep layer weighting.

5 Experimental Setup

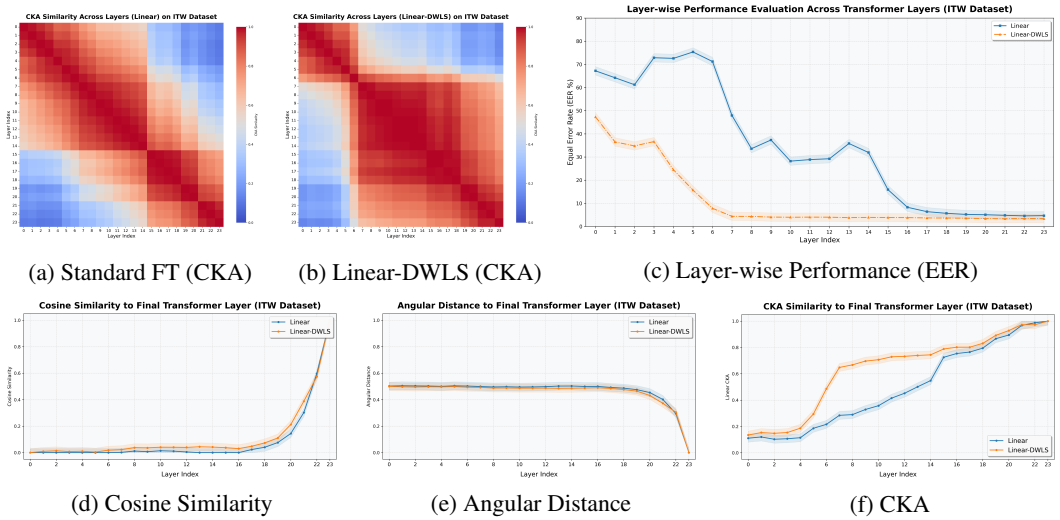
Datasets and Evaluation Metrics. We train all models on the ASVspoof 2019 (Wang et al., 2020) Logical Access (`19LA`) training partition and select the optimal checkpoint using the `19LA` development set, which features a disjoint set of speakers. To assess generalization and robustness under distribution shifts, we evaluate performance across a diverse suite of out-of-domain benchmarks, including in-the-wild recordings, diffusion- and vocoder-based synthetic speech, cross-lingual samples, and large-scale multilingual evaluation sets. Detailed dataset descriptions and statistics are provided in Appendix A.1. Following standard evaluation protocols, we report the EER as our primary metric where lower values indicate superior performance alongside the Area Under the ROC Curve (AUC). To assess statistical reliability, all metrics are reported with 95% confidence intervals (Metric $\pm$ margin) estimated via non-parametric percentile bootstrapping ($B = 1000$).

Implementation Details. We fine-tune the pre-trained `XLS-R-300M` backbone end-to-end jointly with a linear classification head, mapping the 1024-dimensional pooled representation directly to the two output classes. To stabilize training, we perform a warm-up phase during the first epoch using standard cross-entropy fine-tuning to adapt the backbone to deepfake detection, after which DWLS is applied using an alternating schedule. Complete training configurations and hyperparameters are detailed in Appendix A.2.

6 Main Result

6.1 Neural Collapse Analysis

To probe how fine-tuning organises the learned speech representations, we analyse them through the lens of *Neural Collapse* (NC), a geometric phenomenon observed in the terminal phase of classifier training (Papyan et al., 2020). NC describes a tendency for a network's last-layer features to collapse toward their class means, and for the classifier weights to align with those means, as training progresses. We use two complementary NC diagnostics adapted to our binary bona fide vs. spoof setting: *NC1*, which quantifies how tightly each class's embeddings cluster around its class mean, and *NC3*, which quantifies how closely the linear classifier's decision direction aligns with the direction separating the two class means. We use both metrics purely as descriptive diagnostics of the representation geometry, not as a claim that full collapse is reached; formal definitions, and their reduction to closed form for two classes are given in Appendix A.5.



(a) Standard FT (CKA) (b) Linear-DWLS (CKA) (c) Layer-wise Performance (EER)

(d) Cosine Similarity (e) Angular Distance (f) CKA

Figure 4: **Layer-wise representation geometry and performance evaluation on out-of-domain `ITW`. (a, b)** Pairwise CKA similarity matrices across 24 Transformer layers for Standard Fine-Tuning (FT) and `Linear-DWLS`. **(c)** Layer-wise EER (%) showing that `Linear-DWLS` achieves optimal performance by layer 7 (4.39%). **(d–f)** Representation alignment relative to the final layer measured via Cosine Similarity, Angular Distance, and Linear CKA.

Effect of Depth-Weighting Directionality. The progressive schedule $w_\ell(p_t)$ (Eq. 11) redistributes intermediate supervision across network depth. To evaluate how the directionality of this weighting affects representation geometry, we compare a shallow-to-deep schedule against a deep-to-shallow schedule across optimization steps ($N \in \{300, 600, 900, 1200\}$).

We observe a clear asymmetry between the two geometric properties:

- **NC1 Sensitivity:** Intra-class compactness ($\mathcal{E}_{\text{NC1}}$) is highly sensitive to weighting directionality. Under shallow-to-deep weighting, NC1 at the final layers displays monotonic convergence as training progresses, reaching its minimal variability at step 1200. Conversely, deep-to-shallow weighting exhibits performance saturation and slight feature degradation in deep layers (layers 20–23), where step 300 achieves lower final-layer variance than step 1200.

- **NC3 Invariance:** Weight-mean duality alignment ($\mathcal{E}_{\text{NC3}}$) is largely invariant to weighting directionality. Both schedules produce nearly identical layer-wise NC3 trajectories, characterized by an alignment dip near upper-middle layers and a characteristic rise toward the classification head.

This contrast suggests that linear classifier alignment (NC3) is rapidly established by the presence of intermediate-layer supervision regardless of scheduling direction, whereas sustained feature collapse (NC1) in deeper layers requires progressive shallow-to-deep optimization to avoid late-stage variance inflation.

Temporal Evolution Across Optimization Steps. Tracking metrics across optimization steps reveals distinct temporal dynamics between feature collapse and classifier alignment. NC3 stabilizes rapidly: its layer-wise profile at step 300 is virtually identical to that at step 1200, demonstrating that decision boundary duality is established early in training. In contrast, NC1 exhibits a slower, cumulative optimization dynamic. Under appropriate shallow-to-deep weighting, intermediate and deep layers continually compress within-class variability as DWLS accumulates alternating phases, confirming that nuisance variability suppression is a gradual effect driven by sustained intermediate supervision.

Table 1: Ablation study over depth-weighting directionality and optimization steps N, reported as AUC (%) (↑) and EER (%) (↓). **Bold**: best; underline: second best; D→S: deep to shallow; S→D: shallow to deep (Dir.: direction).

Dir.	N	$\mathcal{D}_{21LA}$		$\mathcal{D}_{21DF}$		$\mathcal{D}_{21DF}$		$\mathcal{D}_{ITW}$		$\mathcal{D}_{DFB}$		$\mathcal{D}_{IC}$		$\mathcal{D}_{M-EN}$		$\mathcal{D}_{LSV}$		$\mathcal{D}_{DFADD}$		$\mathcal{D}_{DIFFSSD}$	
		AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
D→S	300	99.61±0.07	1.94±0.20	99.84±0.01	1.88±0.09	**92.64**±0.10	**13.56**±0.14	**99.28**±0.08	**3.38**±0.20	99.29±0.16	3.84±0.54	**89.85**±0.34	**17.75**±0.45	95.74±0.18	9.25±0.26	99.67±0.08	1.86±0.32	92.10±0.84	14.44±1.17	90.22±0.24	15.42±0.31
	600	99.51±0.08	2.47±0.22	**99.86**±0.01	**1.76**±0.10	92.01±0.11	14.47±0.15	99.05±0.09	3.97±0.22	**99.62**±0.11	**2.87**±0.49	87.94±0.38	19.95±0.44	95.05±0.20	10.49±0.27	99.61±0.10	1.82±0.23	93.38±0.80	12.57±1.01	90.13±0.24	15.40±0.31
	900	**99.65**±0.06	**1.79**±0.20	99.82±0.01	1.95±0.11	92.39±0.10	14.03±0.14	99.07±0.09	3.87±0.22	99.48±0.15	3.13±0.53	89.45±0.35	18.56±0.45	97.00±0.14	**7.32**±0.23	**99.77**±0.08	**1.45**±0.25	93.13±0.81	13.10±1.06	**90.56**±0.24	**14.40**±0.30
	1200	99.56±0.07	2.11±0.18	99.84±0.01	1.80±0.08	92.31±0.11	13.85±0.14	99.16±0.09	3.90±0.23	99.16±0.21	3.75±0.60	89.68±0.33	18.17±0.42	**97.34**±0.13	7.74±0.22	99.57±0.10	1.86±0.25	**94.33**±0.73	**11.36**±1.05	88.93±0.26	16.79±0.31
S→D	300	99.51±0.08	2.55±0.20	**99.89**±0.01	**1.51**±0.08	91.36±0.11	15.66±0.14	99.22±0.08	3.78±0.23	**99.84**±0.07	**1.81**±0.40	89.65±0.35	18.45±0.46	97.11±0.12	8.55±0.22	**99.85**±0.08	**1.17**±0.27	**95.70**±0.63	**9.26**±1.05	**93.17**±0.22	**11.29**±0.26
	600	**99.63**±0.07	2.04±0.21	99.82±0.01	2.03±0.09	92.29±0.10	**14.18**±0.14	99.15±0.09	3.87±0.22	99.79±0.08	2.03±0.40	88.09±0.37	20.17±0.49	96.56±0.16	8.01±0.24	99.76±0.09	1.63±0.32	93.87±0.74	12.06±1.06	91.72±0.22	12.73±0.28
	900	99.59±0.07	2.49±0.20	**99.89**±0.01	1.69±0.11	91.74±0.15	14.43±0.21	**99.24**±0.08	3.86±0.23	99.83±0.06	1.94±0.41	88.34±0.40	18.88±0.47	96.47±0.15	9.43±0.25	99.77±0.08	1.74±0.30	94.43±0.71	11.53±1.01	90.41±0.24	15.22±0.30
	1200	**99.63**±0.06	**1.87**±0.18	99.85±0.01	1.73±0.10	**92.69**±0.10	15.11±0.14	**99.24**±0.07	**3.72**±0.22	99.62±0.10	3.09±0.53	**90.10**±0.34	**17.48**±0.43	**97.71**±0.11	**6.74**±0.20	99.73±0.07	1.63±0.23	92.56±0.86	13.50±1.07	88.58±0.27	18.24±0.32

6.2 Representation Alignment and Layer-wise Performance

Accelerated Feature Convergence Across Layers. The pairwise CKA matrices in Figure 4(a, b) reveal qualitatively different representational organisations between the two models. Standard fine-tuning produces a pronounced two-block structure: layers 0–14 form one internally coherent plateau, and layers 15–23 form a second, with near-zero cross-block similarity between them. This boundary coincides with the sharp EER transition visible in Figure 4(c), confirming that discriminative capacity in standard fine-tuning is confined to the top layers of the network. `Linear-DWLS` disrupts this block structure entirely, producing stronger off-diagonal similarity across the three overlapping regions (layers 0–6, 6–18, and 18–23) and a far more gradual representational evolution across depth. Quantitatively, Figure 4(f) shows that `Linear-DWLS` achieves layer-to-final-layer CKA above 0.65 by layer 7, whereas standard fine-tuning does not reach equivalent alignment until layer 17, a compression of approximately ten layers.

Early Generalisation and Sub-network Capability. This accelerated representational alignment translates directly into out-of-domain detection performance. As shown in Figure 4(c), standard fine-tuning requires 16 layers before EER drops below 10%, ultimately reaching 4.69% at the final layer (layer 23). `Linear-DWLS` achieves sub-10% EER by layer 6 (7.85%), reaches 4.39% EER at layer 7 – a 43.5% improvement over standard fine-tuning at the same depth – and continues to improve monotonically to 3.38% at the final layer. This confirms that the progressive depth-weighting schedule embeds class-discriminative capacity into shallow and intermediate layers, consistent with the early within-class feature collapse observed in the NC1 analysis: the layer at which NC1 begins to separate from the baseline ($\sim$layer 6) coincides precisely with the layer at which DWLS achieves near-optimal EER.

Directional Alignment Dynamics. The cosine similarity and angular distance curves in Figures 4(d, e) reveal a complementary picture. Both models maintain near-zero cosine similarity with the final layer (≤ 0.05) and near-constant angular distance (≈ 0.50, corresponding to approximately 90) across layers 0–19. Absolute directional alignment with the final-layer representation therefore remains a late, terminal-phase property in both models, concentrated in layers 20–23. `Linear-DWLS` shows a marginally steeper angular descent from layer 19 onward, reaching 0.37 at layer 22 compared to 0.46 for standard fine-tuning. Taken together, the three alignment metrics distinguish two complementary effects of DWLS: *subspace alignment*, captured by CKA, emerges much earlier under DWLS and accounts for the early EER improvement, while *directional alignment*, captured by cosine similarity and angular distance, remains a shared late-stage property driven by the transformer's residual structure rather than by the choice of training objective.

6.3 Ablation Study

Weighting Order and Generalisation. Table 1 reveals a trade-off between weighting directionality and out-of-domain robustness. Deep-to-shallow weighting at $N = 300$ steps achieves the best `ITW` performance (3.38% EER) despite exhibiting higher NC1 values than its shallow-to-deep counterpart, suggesting that less collapse toward the final-layer representation preserves information diversity useful for unseen attacks. Conversely, shallow-to-deep weighting at higher step counts drives NC1 lower but shows degradation on harder out-of-domain sets such as `DIFFSSD`, consistent with over-alignment to the in-domain `ASVspoof 2019` training distribution. Both schedules perform comparably on in-domain benchmarks (`21LA`, `21DF`), confirming that the generalisation trade-off is most consequential in the out-of-domain regime.

Impact of DWLS Across Back-End Architectures. Table 2 shows that DWLS consistently reduces EER across all three back-end heads compared to their respective baselines, confirming that the benefit of intermediate-layer supervision is architecture-agnostic. For the linear head, which has the least expressive capacity (2.05k parameters) and directly probes linear separability of the XLS-R space, DWLS yields meaningful gains on out-of-domain sets such as ITW (4.69% → 3.38%) and SC (19.56% → 17.75%), demonstrating that improved representation geometry alone — without any additional non-linear capacity — is sufficient to close much of the gap to more complex architectures. MLP-DWLS extends these gains further through its non linearity, achieving consistent improvements across most benchmarks. KAN-DWLS, leveraging learnable univariate functions without a dimensionality bottleneck, yields the strongest overall performance (5.61% average EER), with particularly large reductions on DFADD (7.41% → 4.00%) and LSV (1.51% → 1.33%), suggesting that the expressiveness of KAN complements richer intermediate representations with DWLS. More details about back-end heads are described in Appendix A.4.

Comparison to SOTA methods. Among all evaluated systems, KAN-DWLS achieves the best overall average EER (5.61%) and sets new best results on 19LA (0.07%), 21DF (1.10%), DFADD (4.00%), and LSV (1.33%). The linear head with DWLS is the most noteworthy result: despite having only ∼2.05k back-end parameters and no non-linearity, Linear-DWLS achieves 3.38% EER on ITW, outperforming the majority of prior systems including AASIST (11.20%), SLS (7.46%), Conformer (7.79%), and Mamba variants (5.12%–6.71%), KAN (5.31%) and MultiConv (4.44%). This result isolates the contribution of the training objective itself: the gains are not attributable to back-end capacity but to the intermediate-layer supervision provided by DWLS. MLP-DWLS and KAN-DWLS further extend this advantage, with KAN-DWLS remaining competitive with or superior to all compared systems on the majority of evaluation sets while operating at a comparable parameters.

Table 2: **Benchmark comparison on in-domain and out-of-domain evaluation sets (EER%, ↓). Bold**: best; underline: second best. All back-ends operate on top of XLS-R-300M.

Method	Params (M)	In-Domain			Out-of-Domain					Avg.
		$\mathcal{D}_{19LA}$	$\mathcal{D}_{21LA}$	$\mathcal{D}_{21DF}$	$\mathcal{D}_{ESDF}$	$\mathcal{D}_{ITW}$	$\mathcal{D}_{DFADD}$	$\mathcal{D}_{LSV}$	$\mathcal{D}_{SC}$	
AASIST (Tak et al., 2022b)	317.84	0.22	0.82	2.85	16.25	11.20	11.93	11.21	26.72	10.15
SLS (Zhang et al., 2024)	338.84	0.23	2.87	1.91	18.76	7.46	7.54	1.97	24.51	8.16
Conformer (Truong et al., 2024)	319.26	0.19	1.03	2.06	18.85	7.79	8.89	2.35	38.15	9.91
MultiConv (Tran et al., 2025c)	318.08	0.08	2.77	1.43	15.19	4.44	6.60	1.70	18.71	6.37
KAN (Tran et al., 2025a)	317.55	0.11	2.29	1.49	14.10	5.31	15.63	1.52	22.06	7.81
Mamba (Tran et al., 2025b)	317.52	0.11	1.78	1.51	13.58	5.12	8.62	1.82	17.87	6.30
Mamba (Xiao & Das, 2025)	317.38	0.42	0.93	1.88	14.40	6.71	10.70	2.23	27.58	8.10
Nes2NetX (Liu et al., 2025a)	315.95	0.12	1.88	1.49	22.06	5.52	11.15	2.88	54.28	12.42
RASA (Yang et al., 2025b)	–	0.09	0.89	1.24	–	4.74	–	–	–	–
Poin-HierNet (Yang et al., 2025a)	–	0.11	0.94	1.40	–	4.91	–	–	–	–
Mambo (Ng et al., 2026)	319.37	–	**0.81**	1.70	–	4.97	–	–	–	–
HyperPotter (Wen et al., 2026)	317.87	0.23	3.94	1.78	16.04	5.72	–	2.55	–	–
Linear (baseline)	315.44	0.10	2.38	1.49	**13.47**	4.69	17.51	1.74	19.56	7.62
Linear-DWLS	315.44	0.16 ↑0.06	1.93 ↓0.45	1.88 ↑0.39	13.56 ↑0.09	**3.38** ↓1.31	14.44 ↓3.07	1.86 ↑0.12	17.75 ↓1.81	6.87 ↓0.75
MLP (baseline)	316.50	0.15	1.40	2.18	15.16	5.41	12.06	2.42	19.48	7.28
MLP-DWLS	316.50	0.08 ↓0.07	2.21 ↑0.81	1.57 ↓0.61	14.37 ↓0.79	3.88 ↓1.53	7.16 ↓4.90	1.48 ↓0.94	18.13 ↓1.35	6.11 ↓1.17
KAN (baseline)	315.46	0.10	1.07	1.35	14.41	3.89	7.41	1.51	17.29	5.87
KAN-DWLS	**315.46**	**0.07** ↓0.03	2.48 ↑1.41	**1.10** ↓0.25	15.27 ↑0.86	3.57 ↓0.32	**4.00** ↓3.41	**1.33** ↓0.18	**17.09** ↓0.20	**5.61** ↓0.26

7 Discussion and Limitations

DWLS accelerates discriminative representation formation across encoder depth, with Linear-DWLS achieving near-optimal out-of-domain EER by layer 7 — consistent with early NC1 separation — while standard fine-tuning requires 16 layers to reach comparable performance. NC3 remains persistently high regardless of schedule, indicating that the linear classifier does not align with the class-mean axis, reflecting the high-dimensional covariance structure of the XLS-R space. Deep-to-shallow weighting better preserves within-class variability and transfers more robustly to unseen attacks, whereas shallow-to-deep weighting risks over-alignment to the 19LA distribution. Limitations include evaluation on a single backbone (XLS-R-300M), generalisation to adversarial attacks, codec artefacts, and investigation on layer-wise representation for early-exit mechanism.

8 Conclusion

Standard fine-tuning of speech foundation models supervises only the final layer, leaving intermediate representations weakly adapted to the anti-spoofing task and shifting the burden of discrimination to increasingly complex back-ends. We proposed Depth-Weighted Layer Supervision (DWLS), a multi-layer supervision framework that progressively aligns representations across encoder depth via a cosine depth-weighting schedule. DWLS consistently improves over all back-end baselines, and a lightweight linear classifier paired with DWLS already outperforms many prior systems on out-of-domain benchmarks, demonstrating that reshaping internal feature alignment is a more efficient route to robust speech anti-spoofing than increasing back-end complexity.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, generative AI tools were used strictly to polish the language and improve readability. Generative AI was not employed for any scientific claims, technical methodology, experimental setup, data analysis, or conclusions.

ETHICS STATEMENT

This work strictly adheres to the ICLR Code of Ethics. Our research focuses on improving the robustness and generalization of speech deepfake detection systems to safeguard media integrity and enhance the reliability of voice-based technologies. Robust spoofing detection is critical for countering the growing misuse of synthetic speech in fraud, impersonation, misinformation, and social engineering attacks. All experiments were conducted using standard, publicly available research datasets and pretrained models, fully respecting data privacy and consent protocols.

While evaluating detection systems can theoretically highlight weaknesses that inform adversarial synthesis strategies, we do not release new generation models, attack pipelines, or circumvention techniques. We firmly believe that transparent evaluation and rigorous benchmarking are essential for advancing trustworthy AI, and that the societal benefits of improving defensive deepfake detection far outweigh the potential risks.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed training procedure, architecture, layer-wise supervision strategy, depth-dependent weighting scheme, optimization settings, and evaluation protocol in the main paper in Sec. 5 and Appendix. The Appendix provides description about the datasets, preprocessing procedures, training configuration, hyperparameters, experimental protocols needed to reproduce the reported experiments. Where permitted by the corresponding dataset licenses, we use publicly available benchmark datasets and provide their sources and dataset splits.

REFERENCES

Hashim Ali, Nithin Sai Adupa, Surya Subramani, and Hafiz Malik. A superb-style benchmark of self-supervised speech models for audio deepfake detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16512–16516, 2026. doi: 10.1109/ICASSP55912.2026.11460421.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Arnesh Batra, Arush Gumber, Aniket Khandelwal, Jashn Khemani, and Anubha Gupta. Uncovering the latent potential of deep intermediate representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=6up1qGJwYZ.

Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J. Delp. Diffssd: A diffusion-based dataset for speech forensics. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889450.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Teresa Dorszewski, Albert Kjøller Jacobsen, Lenka Tětková, and Lars Kai Hansen. How redundant is the transformer stack in speech representation models? In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887866.

Sandipana Dowerah, Atharva Kulkarni, Ajinkya Kulkarni, Hoan My Tran, Joonas Kalda, Artem Fedorchenko, Benoit Fauve, Damien Lolive, Tanel Alumäe, and Mathew Magimai.-Doss. Speech df arena: A leaderboard for speech deepfake detection models. *IEEE Open Journal of Signal Processing*, 7:73–81, 2026. doi: 10.1109/OJSP.2026.3652496.

Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-Yi Lee, and Jyh-Shing Roger Jang. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 921–928, 2024. doi: 10.1109/SLT61566.2024.10832250.

Yassine El Kheir, Younes Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. Comprehensive layer-wise analysis of SSL models for audio deepfake detection. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4070–4082, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.227. URL https://aclanthology.org/2025.findings-naacl.227/.

Femi Folorunsho and Benedicta Frema Boamah. Deepfake technology and its impact: ethical considerations, societal disruptions, and security threats in ai-generated media. *International journal of information technology and management information systems*, 16(1):1060–1080, 2025.

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. Post-training for deepfake speech detection. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434709.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ngmEcEer8a.

Yu Guan, Wu Guo, Jie Zhang, and Zhijun Zhang. Fusing multi-layer features of the pre-trained model with grouped cross attention for spoofing speech detection. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 601–606, 2025. doi: 10.1109/APSIPAASC65261.2025.11249219.

Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, and Yuehai Wang. Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12702–12706, 2024. doi: 10.1109/ICASSP48485.2024.10447923.

Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vVxeFSR4fU.

Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, Wangyou Zhang, Seyun Um, Shinnosuke Takamichi, and Shinji Watanabe. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 6:68–77, 2025. doi: 10.1109/OJSP.2025.3529377.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.),

Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, and Mathew Maginai Doss. Exploring generalization to unseen audio data for spoofing: insights from SSL models. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pp. 86–93, 2024. doi: 10.21437/ASVspoof.2024-13.

Howard Lei and Eduardo Lopez. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *Interspeech 2009*, pp. 2323–2326, 2009. doi: 10.21437/Interspeech.2009-389.

Ziyao Li. Kolmogorov-arnold networks are radial basis function networks. *arXiv preprint arXiv:2405.06721*, 2024.

Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 20:12005–12018, 2025a. doi: 10.1109/TIFS.2025.3626963.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Ozo7qJ5vZi.

Juan M. Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9241–9245, 2022. doi: 10.1109/ICASSP43922.2022.9747768.

Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech 2022*, pp. 2783–2787, 2022. doi: 10.21437/Interspeech.2022-108.

Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2024. doi: 10.1109/IJCNN60899.2024.10650962.

Kwok-Ho Ng, Tingting Song, Yongdong WU, and Zhihua Xia. XLSR-MamBo: Scaling the hybrid mamba-attention backbone for audio deepfake detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31450–31462, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1573. URL https://aclanthology.org/2026.findings-acl.1573/.

Douglas O'Shaughnessy. Review of automatic text-to-speech synthesis methods. *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1–23, 2026. doi: 10.1109/TASLPRO.2026.3693972.

Zihan Pan, Tianchi Liu, Hardik B. Sailor, and Qiongqiong Wang. Attentive Merging of Hidden Embeddings from Pre-trained Speech Model for Anti-spoofing Detection. In *Interspeech 2024*, pp. 2090–2094, 2024. doi: 10.21437/Interspeech.2024-1472.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021. doi: 10.1109/ASRU51503.2021.9688093.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096149.

Tuan Dat Phuong, Long-Vu Hoang, and Huy Dat Tran. Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models. In *Interspeech 2025*, pp. 5633–5637, 2025. doi: 10.21437/Interspeech.2025-1411.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pp. 2757–2761, 2020. doi: 10.21437/Interspeech.2020-2826.

Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, 2019. doi: 10.1109/SPED.2019.8906599.

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech 2023*, pp. 5281–5285, 2023. doi: 10.21437/Interspeech.2023-1820.

Peter Súkeník, Christoph H. Lampert, and Marco Mondelli. Neural collapse is globally optimal in deep regularized resnets and transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8WKOk4U9R4.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 904–912, 2023. doi: 10.1109/CVPRW59228.2023.00097.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6382–6386. IEEE, 2022a.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*, pp. 112–119, 2022b. doi: 10.21437/Odyssey.2022-16.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45: 516–535, 2017. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2017.01.001. URL https://www.sciencedirect.com/science/article/pii/S0885230816303114.

Hoan My Tran, David Guennec, Philippe Martin, Aghilas Sini, Damien Lolive, Arnaud Delhay, and Pierre-François Marteau. Spoofed Speech Detection with a Focus on Speaker Embedding. In *Interspeech 2024*, pp. 2080–2084, 2024. doi: 10.21437/Interspeech.2024-481.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. Kan you hear the truth? audio deepfake detection with kolmogorov–arnold networks. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 561–565, 2025a. doi: 10.23919/EUSIPCO63237.2025.11226534.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection. In *Interspeech 2025*, pp. 5323–5327, 2025b. doi: 10.21437/Interspeech.2025-1703.

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. Multi-level ssl feature gating for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11766–11775, New York, NY, USA, 2025c. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754568. URL https://doi.org/10.1145/3746027.3754568.

Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-Francois Marteau, and David Guennec. When depth is redundant: Efficient transformer-based speech anti-spoofing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6380–6397, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.318. URL https://aclanthology.org/2026.findings-acl.318/.

Hoan My Tran, Aghilas Sini, David Guennec, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. Kan we make models simpler for audio deepfake detection with kolmogorov–arnold networks? In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 18057–18061, 2026b. doi: 10.1109/ICASSP55912.2026.11460320.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pp. 537–541, 2024. doi: 10.21437/Interspeech.2024-659.

Duc-Tuan Truong, Tianchi Liu, Ruijie Tao, Junjie Li, Kong Aik Lee, and Eng Siong Chng. Qamo: Quality-aware multi-centroid one-class learning for speech deepfake detection. *arXiv preprint arXiv:2509.20679*, 2025.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. Asvspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pp. 1–8, 2024. doi: 10.21437/ASVspoof.2024-1.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825, 2026. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2025.101825. URL https://www.sciencedirect.com/science/article/pii/S0885230825000506.

Qing Wen, Haohao Li, Zhongjie Ba, Peng Cheng, Miao He, Li Lu, and Kui Ren. Hyperpotter: Spell the charm of high-order interactions in audio deepfake detection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vIXuPi8zUM.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pp. 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Yang Xiao and Rohan Kumar Das. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*, 32:1276–1280, 2025. doi: 10.1109/LSP.2025.3547861.

Yixuan Xiao and Ngoc Thang Vu. Layer-Wise Decision Fusion for Fake Audio Detection Using XLS-R. In *Interspeech 2025*, pp. 5618–5622, 2025. doi: 10.21437/Interspeech.2025-1543.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm).*, 2019.

Mingru Yang, Yanmei Gu, Qianhua He, Yanxiong Li, Peirong Zhang, Yongqiang Chen, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. Generalizable Audio Deepfake Detection via Hierarchical Structure Learning and Feature Whitening in Poincaré sphere. In *Interspeech 2025*, pp. 2255–2259, 2025a. doi: 10.21437/Interspeech.2025-1594.

Mingru Yang, Yanmei Gu, Qianhua He, Peirong Zhang, Haolin He, Zhiming Wang, Huijia Zhu, Jian Liu, and Weiqiang Wang. Generalizable audio deepfake detection via risk-aware style alignment and structural empirical risk minimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11600–11609, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755593. URL https://doi.org/10.1145/3746027.3755593.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised XLS-r and SLS classifier. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=acJMIXJg2u.

Jim Zhao, Tin Sum Cheng, Wojciech Masarczyk, and Aurelien Lucchi. Optimizer choice matters for the emergence of neural collapse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9EPYWJrib1.

A Appendix

A.1 Dataset details

In this section, we provide exhaustive descriptions of the benchmark datasets utilized in this study (Table 3) to evaluate spoofing countermeasures.

ASVspoof 2019. The ASVspoof 2019 logical access (19LA[1]) scenario (Wang et al., 2020) serves as a foundational benchmark for assessing countermeasures against synthetic speech attacks directed at ASV systems. The 19LA track targets logical access threats, where spoofed audio is synthesized via TTS and VC systems. It integrates bona fide speech from the VCTK[2] corpus (Yamagishi et al., 2019) with spoofed samples created using 17 distinct generative mechanisms (6 known techniques in the training and development partitions, and 13 unknown techniques in the evaluation set). This configuration drives the development of robust countermeasures capable of identifying modern speech synthesis variations under controlled acoustic environments.

ASVspoof 2021. The ASVspoof 2021 (Liu et al., 2023) logical access (21LA[3]) scenario builds on the 19LA partition by subjecting the evaluation samples to realistic network transmission effects, including various telephony and VoIP lossy codecs, to mimic genuine communication channels. In

[1]https://doi.org/10.7488/ds/2555
[2]https://doi.org/10.7488/ds/2645
[3]https://doi.org/10.5281/zenodo.4817650

parallel, the ASVspoof 2021 deepfake (21DF[4]) track broadens the detection objective beyond ASV vulnerabilities to generic audio deepfake identification. This dataset comprises authentic recordings alongside spoofed utterances processed through diverse lossy compression algorithms and bitrates, introducing channel artifacts to challenge detector robustness.

ASVspoof 5. The ASVspoof 5 Wang et al. (2026) (25DF[5]) challenge introduces a highly intricate database for evaluating anti-spoofing and deepfake detection pipelines within realistic operational settings. Moving away from the pristine, studio-recorded environments of earlier challenge iterations, ASVspoof 5 leverages crowdsourced authentic speech spanning roughly 2,000 distinct speakers under heterogeneous acoustic backgrounds, drawn from the English portion of the Multilingual Librispeech (MLS) framework Pratap et al. (2020). The spoofed utterances are generated using advanced neural TTS/VC models and novel adversarial algorithms, with the evaluation sets featuring unseen codecs and acoustic compression to encourage real-world, in-the-wild generalization.

In-the-Wild. The In-the-Wild (ITW[6]) dataset (Müller et al., 2022) features roughly 37.9 hours of audio fragments from 58 English-speaking public figures and politicians. It contains 20.7 hours of authentic speech extracted from open-source podcasts and public addresses, paired with 17.2 hours of forged speech collected from 219 unconstrained internet sources. The audio clips are systematically balanced according to speech style, emotional delivery, environmental background noise, and length to represent real-world deployment. The ITW dataset contains unseen voice synthesis architectures, social media encoding artifacts, and varied transmission acoustics, acting as a rigorous out-of-domain evaluation where models trained on standard datasets frequently suffer dramatic performance degradation.

Fake-or-Real. The Fake-or-Real (FOR[7]) corpus (Reimao & Tzerpos, 2019) comprises more than 195,000 English utterances, containing approximately 87,000 synthetic and 111,000 authentic audio files created via prominent industrial and research TTS architectures. The genuine recordings are collected from heterogeneous open-source datasets and real-world media, incorporating vast diversity in speaker demographics, physiological age, regional accents, microphone hardware, and room impulse responses. The database is distributed in multiple variations, including FOR-original (untruncated utterances), FOR-norm (loudness normalized and balanced), FOR-2seconds (fixed 2-second windows), and FOR-rerecorded (re-played and recorded to capture realistic channel playback distortions). This study leverages the FOR-original variant to analyze out-of-domain transferability.

SpoofCeleb. The SpoofCeleb (SC[8]) benchmark Jung et al. (2025) provides a large-scale framework for voice deepfake detection, compiling real-world, in-the-wild authentic audio from 1,251 unique public figures derived from the VoxCeleb1 corpus via an automated processing chain encompassing transcription, diarization, noise attenuation, and quality thresholding. The corpus contains over 2.5 million files, where the spoofed counterparts are synthesized through 23 diverse TTS systems. The evaluation partition contains 40 speakers and 9 attack methodologies that emphasize high environmental reverberation, background noise, and wide speaker diversity to accurately mirror practical deepfake threats.

MLAAD. The Multi-Language Audio Anti-Spoofing Dataset (Müller et al., 2024) (MLAAD[9]) is an expansive multilingual evaluation suite engineered to combat the linguistic bias found in traditional benchmarks and analyze countermeasure generalization across wide cross-lingual contexts. It provides over 570 hours of synthetic speech ($\sim$243,000 utterances) generated via 119 diverse TTS pipelines. The underlying authentic speech is drawn from the M-AILABS collection in 8 native languages—English (M-EN), French, German, Italian, Polish, Russian, Spanish, and Ukrainian—and expands to 40 additional target languages via advanced machine translation. Audio samples

[4]https://doi.org/10.5281/zenodo.4835107
[5]https://doi.org/10.5281/zenodo.14498691
[6]https://deepfake-total.com/in_the_wild
[7]https://bil.eecs.yorku.ca/datasets
[8]https://huggingface.co/datasets/jungjee/spoofceleb
[9]https://deepfake-total.com/mlaad

are delivered at a 22,050 Hz sampling rate, with evaluation trials featuring synthetic noise injection, background music, and lossy codec compression.

LibriSeVoc. The LibriSeVoc (LSV[10]) benchmark (Sun et al., 2023) is an acoustic database engineered to assess voice deepfake systems on the footprints left by neural vocoders rather than focusing on end-to-end TTS/VC pipelines. Built upon the clean, multi-speaker LibriTTS corpus (Zen et al., 2019), it provides 92,407 audio files spanning roughly 244 hours at a 24 kHz sampling rate. The corpus includes 13,201 authentic recordings and 79,206 spoofed instances generated through self-vocoding mechanisms, where original speech is re-synthesized using six distinct neural vocoder paradigms across autoregressive, GAN-based, and diffusion-based architectures.

DFADD. The Diffusion and Flow-Matching Based Audio Deepfake Dataset (Du et al., 2024) (DFADD[11]) targets highly natural synthetic speech produced by modern diffusion and flow-matching TTS generators. Constructed using source files from the VCTK dataset, it features approximately 180 hours of total speech data, comprised of 44,455 authentic utterances and 163,500 spoofed audio instances, posing an empirical challenge to conventional classifiers.

DIFFSSD. The Diffusion-Based Synthetic Speech Dataset (DIFFSSD[12]) (Bhagtani et al., 2025) addresses generalization errors in countermeasures when subjected to high-fidelity generations from contemporary diffusion-based speech models. It contains approximately 200 hours of audio data, where real utterances are gathered from the LJ Speech (Ito & Johnson, 2017) and LibriSpeech (Panayotov et al., 2015) databases across 74 unique speakers. These are paired with 70,000 synthetic audio clips generated via 10 modern text-to-speech architectures, incorporating 8 open-source diffusion pipelines and 2 commercial platforms.

Dataset	Language	# Bona fide	# Spoofed	Attack
Training				
ASVspoof 2019 logical access (19LA) Wang et al. (2020)	English	2,580	22,800	TTS, VC
Development				
ASVspoof 2019 logical access (19LA) Wang et al. (2020)	English	2,548	22,296	TTS, VC
Evaluation				
ASVspoof 2019 logical access (19LA) Wang et al. (2020)	English	7,355	63,882	TTS, VC
ASVspoof 2021 logical access (21LA) / deepfake (21DF) Liu et al. (2023)	English	14,816 14,869	133,360 519,059	TTS, VC
ASVspoof5 (25DF) Wang et al. (2026)	English	138,688	542,086	TTS, VC, AT
Fake-or-Real (FoR) Reimao & Tzerpos (2019)	English	2,264	2,370	TTS
In-the-Wild (ITW) Müller et al. (2022)	English	19,963	11,816	Unknown
Diffusion-based Synthetic Speech Dataset (DFADD) Du et al. (2024)	English	755	3,000	TTS
Librisevoc (LSV) Sun et al. (2023)	English	2,641	15,846	Vocoded
SpoofCeleb (SC) Jung et al. (2025)	English	9,113	82,017	TTS
Multi-Language Audio Anti-Spoofing Dataset (MLAAD) MLAAD-ENMüller et al. (2024)	English	28,233	36,000	TTS
Diffusion-based Synthetic Speech Dataset (DIFFSSD) Bhagtani et al. (2025)	English	12,113	42,500	TTS

AT: Adversarial attack.

Table 3: Statistics of datasets used in the study.

A.2 Optimization and Training Configuration

As shown in Table 4, all models are implemented using PyTorch and fine-tuned using pretrained SFM sourced from HuggingFace[13]. Due to the variable durations of the input waveforms, dynamic batching is employed, where audio sequences within each batch are padded to match the maximum sample length of that specific batch. To counteract the severe class imbalance inherent in the 19LA training partition, we utilize a weighted cross-entropy loss function. This objective assigns a static weight of 0.9 to the minority bona fide class and 0.1 to the majority spoofed class, ensuring that the gradient updates remain sensitive to genuine speech distribution patterns. Optimization is

[10]https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts

[11]https://github.com/isjwdu/DFADD

[12]https://huggingface.co/datasets/purdueviperlab/diffssd

[13]https://huggingface.co/facebook/wav2vec2-xls-r-300m

performed using the Adam optimizer (Kingma, 2014) with a fixed learning rate of 2.5×10^{-6} and a weight decay coefficient of 10^{-4} to prevent overfitting. SFM is fine-tuned for a maximum of 10 epochs using a batch size of 5 samples on a NVIDIA H200 GPU. The random seed is fixed to 1234.

To improve out-of-domain generalization and simulate real-world channel variants, we also apply `RawBoost` data augmentation (Tak et al., 2022a)[14] to add various types of noise including linear and nonlinear convolutive, impulsive signal-dependent additive, stationary additive, and randomly colored noises during training.

Table 4: Hyperparameter details.

Configurations	Setting
Seed	1234
Batch size	5
Epochs	10
GPU	1 NVIDIA H200
Audio length	Dynamic batching
Optimizer	Adam
Learning rate	2.5×10^{-6}
Weight decay	10^{-4}
Weighted cross-entropy loss	0.9 real, 0.1 fake
Early-stop patience	3

A.3 Evaluation Metric

The Equal Error Rate (EER) corresponds to the operating point at which the False Acceptance (FA) rate equals the False Rejection (FR) rate. In the context of spoofing CMs, these rates are functions of a decision threshold τ_{CM} applied to the detection score:

$$P_{\text{fa}}^{\text{CM}}(\tau_{\text{CM}}) = \frac{\#\{\text{spoofed with score} > \tau_{\text{CM}}\}}{\#\{\text{spoofed trials}\}}, \tag{14}$$

$$P_{\text{miss}}^{\text{CM}}(\tau_{\text{CM}}) = \frac{\#\{\text{bona fide with score} \leq \tau_{\text{CM}}\}}{\#\{\text{bona fide trials}\}}. \tag{15}$$

The EER is obtained by sweeping τ_{CM} to the operating point where $P_{\text{fa}}^{\text{CM}} = P_{\text{miss}}^{\text{CM}}$. Lower EER values indicate stronger spoofing detection performance. In our framework, detection scores are computed as log-likelihood ratios (LLR):

$$\text{LLR}_t = \log p(X_t \mid \mathcal{H}_0) - \log p(X_t \mid \mathcal{H}_1), \tag{16}$$

where X_t denotes the input utterance, $\mathcal{H}_0$ the bona fide hypothesis, and $\mathcal{H}_1$ the spoofed hypothesis.

Area Under the ROC Curve (AUC). As a complementary threshold-independent metric, we additionally report the Area Under the Receiver Operating Characteristic curve (AUC), which summarises detection performance across all possible operating points τ_{CM}. The ROC curve traces the trade-off between the true positive rate $\text{TPR}(\tau_{\text{CM}}) = 1 - P_{\text{miss}}^{\text{CM}}(\tau_{\text{CM}})$ and the false positive rate $\text{FPR}(\tau_{\text{CM}}) = P_{\text{fa}}^{\text{CM}}(\tau_{\text{CM}})$ as τ_{CM} is swept over its full range, and the AUC is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(t)\big)\, dt \;\in [0, 1]. \tag{17}$$

An AUC of 1.0 corresponds to perfect discrimination between bona fide and spoofed utterances, while 0.5 corresponds to chance-level performance. Higher AUC values indicate stronger spoofing detection, and together with EER, AUC provides a fuller picture of system behaviour across the full score distribution rather than at a single operating point.

A.4 Baseline Details

Back-End Classification Baselines. To evaluate whether performance gains stem from improved representation quality rather than back-end parameter capacity, we attach three distinct classification architectures directly to the 1024-dimensional pooled representation $\boldsymbol{h}_i$:

(i) Linear Head. The minimal parametric baseline maps the pooled representation directly to class logits without intermediate projections or non-linearities:

$$\hat{\boldsymbol{y}}_i^{\text{linear}} = \mathbf{W}\boldsymbol{h}_i + \mathbf{b}, \tag{18}$$

[14] https://github.com/TakHemlata/RawBoost-antispoofing

where $\mathbf{W} \in \mathbb{R}^{2\times1024}$ and $\mathbf{b} \in \mathbb{R}^2$ comprise approximately 2.05k parameters. This head directly evaluates the linear separability of the underlying Transformer space.

(ii) Multi-Layer Perceptron (MLP) Head. To measure baseline performance under non-linear transformation with explicit dimensionality reduction, we employ a two-layer projection network:

$$\hat{\boldsymbol{y}}_i^{\text{mlp}} = \mathbf{W}_2\,\text{SiLU}\left(\mathbf{W}_1\boldsymbol{h}_i + \mathbf{b}_1\right) + \mathbf{b}_2, \tag{19}$$

where $\mathbf{W}_1 \in \mathbb{R}^{128\times1024}$, $\mathbf{b}_1 \in \mathbb{R}^{128}$, $\mathbf{W}_2 \in \mathbb{R}^{2\times128}$, and $\mathbf{b}_2 \in \mathbb{R}^2$. This adds non-linear capacity through the SiLU activation function while projecting the 1024-dimensional embedding into a 128-dimensional bottleneck before final prediction ($\sim$ 131.5k back-end parameters).

(iii) Kolmogorov-Arnold Network (KAN) Head. KANs replace fixed node activations with learnable univariate functions along network edges (Liu et al., 2025b). Following FastKAN (Li, 2024)[15], we parameterize edge functions using Gaussian Radial Basis Functions (RBFs) to preserve the intact 1024-dimensional feature dimension without a bottleneck projection, as proposed by Tran et al. (2026b):

$$\varphi_{m,n}(x) = w_b\,\text{SiLU}(x) + w_t s\sum_{k=1}^{G} c_k \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right), \tag{20}$$

where G is the grid size, μ_k and σ define the kernel centers and spread, and w_b, w_s, c_k are trainable weights. The output logit vector is formed by summing the non-linear univariate edge transformations:

$$\hat{\boldsymbol{y}}_i^{\text{kan}} = \sum_{j=1}^{1024} \boldsymbol{\phi}_j\left([\boldsymbol{h}_i]_j\right), \tag{21}$$

where each $\boldsymbol{\phi}_j : \mathbb{R} \to \mathbb{R}^2$ applies univariate non-linear mappings ($\sim$ 22.5k parameters) with $G = 10$.

A.5 Neural Collapse Analysis of the Fine-Tuned Speech Representations

Neural collapse (NC) describes a geometry that last-layer features and classifiers tend to adopt late in training (Papyan et al., 2020). We use it as a lightweight diagnostic of how fine-tuning shapes the utterance representation $\boldsymbol{h}_i \in \mathbb{R}^{1024}$ that is passed to the linear head (Eq. equation 18) for bona fide vs. spoof classification. Since our task has only two classes, we report two of the four NC properties: within-class variability collapse (NC1) and classifier–mean alignment (NC3). NC2 holds trivially for two classes, because the two centred class means are antipodal and equinorm by construction. We omit NC4, which is reported to hold whenever training accuracy is close to 100% (Zhao et al., 2026). We treat all metrics as descriptive diagnostics of the learned geometry rather than as evidence that full collapse is reached.

Setup. Let $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ be the mean embeddings of the bona fide and spoof classes, $\boldsymbol{\mu}_G = \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$, and $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Let $\boldsymbol{u} = \boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|$ be the unit vector from the bona fide mean to the spoof mean, and let $\boldsymbol{\Sigma}_W = \frac{1}{2}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)$ be the average of the two within-class covariances. Speech deepfake corpora are typically imbalanced, so we weight the two classes equally rather than by sample count. This adapts the balanced-class definitions of Papyan et al. (2020).

NC1: Within-class variability. NC1 measures how tightly the embeddings of each class cluster around their class mean, relative to the separation between the means. The standard definition is $\text{Tr}(\boldsymbol{\Sigma}_W\boldsymbol{\Sigma}_B^{\dagger})/C$ (Papyan et al., 2020). For two classes, $\boldsymbol{\Sigma}_B = \frac{1}{4}\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$, and it reduces exactly to

$$\text{NC1} = \frac{2\,\boldsymbol{u}^\top\boldsymbol{\Sigma}_W\,\boldsymbol{u}}{\|\boldsymbol{\Delta}\|^2}, \tag{22}$$

i.e., the within-class variance along the bona fide–spoof axis divided by the squared distance between the class means. Lower values indicate more compact and better separated classes. Because it only involves the direction $\boldsymbol{u}$, it can be computed without inverting a 1024×1024 matrix.

[15] https://github.com/ZiyaoLi/fast-kan

NC3: Classifier–mean alignment. NC3 measures whether the decision direction of the linear head matches the direction separating the class means. Following Papyan et al. (2020), it is the squared distance between the normalised classifier $\mathbf{W}$ and the normalised matrix of centred class means $\tilde{\mathbf{M}} = [\boldsymbol{\mu}_0 - \boldsymbol{\mu}_G,\ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_G]$:

$$\mathrm{NC3} = \left\| \frac{\mathbf{W}^\top}{\|\mathbf{W}\|_F} - \frac{\tilde{\mathbf{M}}}{\|\tilde{\mathbf{M}}\|_F} \right\|_F^2 . \tag{23}$$

For a two-class softmax, only the difference $\mathbf{w}_\Delta = \mathbf{w}_1 - \mathbf{w}_0$ between the two rows of $\mathbf{W}$ affects the prediction. We therefore apply Eq. equation 23 to the row-centred classifier, whose rows are $\mp\mathbf{w}_\Delta/2$. With $\cos\theta = \langle \mathbf{w}_\Delta, \boldsymbol{\Delta} \rangle / (\|\mathbf{w}_\Delta\| \|\boldsymbol{\Delta}\|)$, this gives

$$\mathrm{NC3} = 2\,(1 - \cos\theta), \tag{24}$$

so a value of 0 means the classifier points exactly along the line joining the class means.

Why this matters for speech deepfake detection. The pooled `XLS-R` embedding encodes speaker, language, content and channel information in addition to authenticity, and the spoof class mixes artefacts from many different generators and vocoders. A strongly collapsed representation (low NC1) means fine-tuning has suppressed much of this variation along the decision axis, which helps on the attacks seen in training. However, pulling all spoofed utterances toward a single mean may also discard attack-specific cues that would help on unseen generators. Likewise, a low NC3 means the classifier relies on the average difference between the two classes, which is dominated by the most frequent attack families in the training data. These properties therefore need not translate into better robustness, and we examine them empirically. For DWLS, we additionally compute NC1 at every layer using $\boldsymbol{h}_i^{(\ell)}$, following the layer-wise use of NC1 in Jiang et al. (2025), to show at which depth the two classes become separable.

A.6 Pseudocode in § 4

Algorithm 1 details the DWLS training step, alternating between the standard final-layer loss and a cosine depth-weighted combination of per-layer losses that shifts emphasis from deep to shallow layers across the progressive phase.

Algorithm 1 DWLS Training Procedure

Require: step t, phase length N, warm-up steps W, normalized depth ratios $\rho_\ell = \ell/(L-1)$
if $t < W$ **then** ▷ Initial warm-up
  $\phi \leftarrow 0$
else
  $\phi \leftarrow \left(\left\lfloor \frac{t-W}{N} \right\rfloor \bmod 2\right)$ ▷ Eq. 8
end if
if $\phi = 0$ **then** ▷ Standard phase
  $\hat{y}^{(L-1)} \leftarrow f_\theta\left(\mathrm{Pool}(\boldsymbol{H}^{(L-1)})\right)$
  return $\mathrm{CE}\left(\hat{y}^{(L-1)}, y; \mathbf{w}_c\right)$
else ▷ Progressive phase
  $r_t \leftarrow (t - W) \bmod N$ ▷ Step index in phase
  $p_t \leftarrow r_t/(N-1)$ ▷ Phase progress in $[0, 1]$
  for $\ell = 0, \ldots, L-1$ **do**
    $\hat{y}^{(\ell)} \leftarrow f_\theta\left(\mathrm{Pool}(\boldsymbol{H}^{(\ell)})\right)$
    $\mathcal{L}_{\text{std}}^{(\ell)} \leftarrow \mathrm{CE}\left(\hat{y}^{(\ell)}, y; \mathbf{w}_c\right)$
  end for
  $w_\ell \leftarrow \max\left(0,\ 1 + \cos(\pi p_t)(2\rho_\ell - 1)\right) \quad \forall \ell$ ▷ Eq. 11
  $\tilde{w}_\ell \leftarrow w_\ell / \sum_{\ell'} w_{\ell'}$
  return $\sum_\ell \tilde{w}_\ell \mathcal{L}_{\text{std}}^{(\ell)}$ ▷ Eq. 12
end if