XLSR-EMA: A 3D JOINT FORGERY MANIFOLD FOR UNIVERSAL SPEECH DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained self-supervised learning (SSL) models have achieved remarkable success in speech deepfake detection. However, existing methods typically rely on the final hidden layer or simple layer aggregation, limiting their ability to exploit the hierarchical information encoded across SSL representations. This limitation becomes more pronounced for speech generated by modern Audio Language Models (ALMs), whose artifacts span multiple acoustic and semantic levels. We propose XLSR-EMA, a universal speech deepfake detection framework that organizes the hidden representations of a pre-trained XLS-R model into a unified **Time-Layer-Feature Joint Forgery Manifold**. An Efficient Multi-scale Attention (EMA) module models interactions among temporal, feature, and layer dimensions, while a Dual-Purpose Perceptual Augmentation (DPPA) strategy improves channel robustness and sensitivity to subtle synthetic artifacts. Experiments on multiple benchmarks demonstrate the effectiveness and cross-domain generalization of the proposed framework. Anonymous code is available online.

1 INTRODUCTION

Recent advances in generative audio models have enabled the creation of highly realistic synthetic speech through text-to-speech (TTS) Li et al. (2024) and voice conversion (VC) Yi et al. (2020) technologies. While these developments benefit human-computer interaction, they introduce serious security risks, including voice impersonation, misinformation, and fraud. The emergence of Audio Language Models (ALMs), which generate speech through codec-based token modeling, has further increased the realism of synthetic audio and challenged existing speech deepfake detection systems Yang et al. (2023); Xie et al. (2025); Li et al. (2024). To facilitate research in this area, the Codecfake benchmark was recently introduced to evaluate detection performance against modern codec-based generation methods Xie et al. (2025).

The ASVspoof challenges Wu et al. (2017) have significantly advanced speech anti-spoofing research. Early approaches mainly relied on handcrafted acoustic features, such as MFCC Davis & Mermelstein (1980), LFCC Todisco et al. (2018), and CQCC Todisco et al. (2016). Subsequently, deep neural architectures including SincNet Ravanelli & Bengio (2018) and FastAudio Fu et al. (2022) achieved notable improvements by learning discriminative representations directly from speech signals Chen et al. (2024). More recently, self-supervised learning (SSL) models, such as wav2vec 2.0 Baevski et al. (2020) and XLS-R Babu et al. (2022), have become the dominant feature extractors for speech deepfake detection. For example, combining SSL representations with AASIST Jung et al. (2022) achieves strong performance on the ASVspoof 2021 benchmarks Tak et al. (2022b).

Despite these advances, two challenges remain. First, most SSL-based methods utilize only the final hidden layer or perform simple aggregation across layers, overlooking the rich hierarchical information encoded throughout the network. Second, existing detectors often struggle to generalize to emerging codec-based ALM-generated speech, where forgery artifacts are distributed across multiple semantic and acoustic abstraction levels.

To address these limitations, we propose a universal speech deepfake detection framework based on a **Time-Layer-Feature Joint Forgery Manifold**. Instead of collapsing hidden representations, we preserve all 24 hidden layers of a pretrained XLS-R model and organize them into a unified three-dimensional representation. This design enables the model to jointly exploit low-level acoustic

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.

artifacts and high-level semantic inconsistencies. Furthermore, we introduce an Efficient Multi-scale Attention (EMA) module to model cross-layer, temporal, and feature-dimensional interactions within the manifold. To improve robustness under diverse transmission conditions, we additionally develop a Dual-Purpose Perceptual Augmentation (DPPA) strategy that balances channel robustness and forgery sensitivity.

The main contributions of this work are summarized as follows:

- We propose a novel Time-Layer-Feature Joint Forgery Manifold that preserves the complete hierarchical representations of a pretrained SSL model, enabling effective exploitation of multi-scale forensic cues for speech deepfake detection.
- We develop an EMA-based cross-dimensional interaction framework that models temporal, feature-dimensional, and layer-wise dependencies simultaneously, enhancing the detection of both low-level acoustic artifacts and high-level semantic inconsistencies.
- We introduce a Dual-Purpose Perceptual Augmentation strategy and conduct extensive evaluations on multiple benchmarks, including Codecfake and ASVspoof 2021 DF, demonstrating strong robustness and generalization against modern codec-based speech deepfakes.

2 Related Work

2.1 Speech Deepfake Detection and Artifact Evolution

Early speech deepfake detectors primarily targeted synthetic speech generated by traditional Text-to-Speech (TTS) Li et al. (2024) and Voice Conversion (VC) Yi et al. (2020) systems. These methods typically used handcrafted acoustic features, such as CQCC and LFCC, to expose spectral inconsistencies introduced by vocoder-based synthesis pipelines.

Recent advances in generative audio have introduced Audio Language Models (ALMs), including UniAudio Yang et al. (2023), and related codec-based frameworks. Unlike conventional vocoders, ALMs synthesize speech by modeling discrete audio tokens, yielding artifacts across multiple abstraction levels, from local codec quantization distortions to high-level prosodic and semantic inconsistencies. Handcrafted features and single-scale detectors therefore generalize poorly to these emerging attacks. Broader AIGC forensics methods, such as VLForgery He et al. (2025) and OMAT Zhou et al. (2025), similarly underscore the need for robust, transferable forgery representations.

2.2 Self-Supervised Learning in Audio Forensics

Self-supervised learning (SSL) models pretrained on large-scale unlabeled speech corpora, including wav2vec 2.0 Baevski et al. (2020), HuBERT Hsu et al. (2021), XLS-R Babu et al. (2022), and WavLM Chen et al. (2022), now serve as dominant feature extractors for speech deepfake detection. Existing methods typically classify the final-layer representation Zhang et al. (2024), leveraging its strong semantic modeling capability while omitting complementary hierarchical evidence.

Different hidden layers encode complementary information, ranging from low-level acoustic characteristics to high-level linguistic representations. Multi-view collaborative learning networks Zhang et al. (2025), layer-wise decision fusion Xiao & Vu (2025), and multi-scale SSL detectors Wani & Amerini (2026) exploit complementary representations through aggregation. However, most existing methods pool or project hidden-layer representations before modeling their interactions, potentially discarding informative cross-layer correlations. This limitation motivates explicit joint modeling of hierarchical SSL representations for audio forensics.

2.3 Data Augmentation and Channel Robustness

Generalization remains challenging because recording devices, transmission channels, environmental noise, and lossy codecs can shift the input distribution. Representative augmentation methods, such as RawBoost Tak et al. (2022a), simulate realistic channel distortions by injecting additive

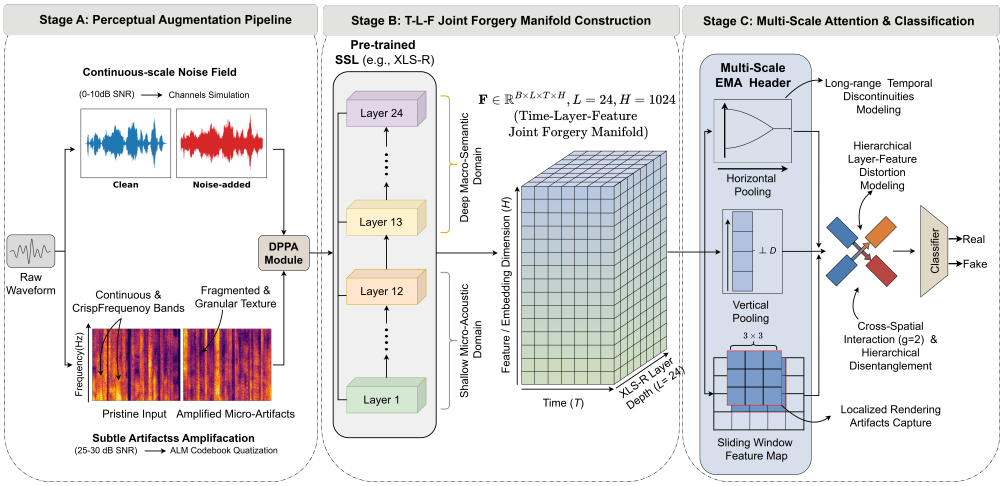


Figure 1: The overall framework of the proposed universal speech deepfake detection system. Stage A introduces Dual-Purpose Perceptual Augmentation (DPPA) to improve channel robustness while enhancing sensitivity to subtle forensic artifacts. Stage B stacks hierarchical hidden-layer representations from the pretrained XLS-R to construct a time-layer-feature joint forgery manifold $\mathbf{F} \in \mathbb{R}^{B\times L\times T\times H}$, where $L = 24$ denotes the layer depth and H denotes the hidden feature dimension. Stage C applies a multi-scale EMA-based attention classifier to capture temporal discontinuities, feature-dimensional distortions, and localized rendering artifacts for real/fake classification.

noise, nonlinear degradation, and reverberation into speech signals. Recent work also improves out-of-domain generalization with non-semantic universal audio representations Das et al. (2025), while replay attacks reveal that re-recording can make generated speech appear authentic to deepfake detectors Müller et al. (2025).

Although augmentation improves robustness to environmental variation, excessive perturbation may weaken subtle forgery traces introduced by modern neural codecs and ALMs. Practical detectors therefore require augmentation that enhances channel robustness while preserving discriminative forensic artifacts.

3 Methodology

This section describes the proposed universal speech deepfake detection framework. As illustrated in Fig. 1, the architecture comprises three stages: Dual-Purpose Perceptual Augmentation (DPPA), construction of the Time-Layer-Feature Joint Forgery Manifold using a pretrained XLS-R backbone, and cross-spatial feature extraction with an Efficient Multi-scale Attention (EMA) module.

3.1 Dual-Purpose Perceptual Augmentation (DPPA)

To mitigate real-world domain mismatch and improve generalization to deceptive generation methods, we introduce a customized Dual-Purpose Perceptual Augmentation (DPPA) strategy. DPPA models a dynamic, continuous-scale perceptual noise field with two complementary objectives: improving robustness under low-SNR conditions and amplifying micro-quantization distortions under high-SNR conditions.

Given a raw speech waveform $\mathbf{x}$, the augmented signal $\tilde{\mathbf{x}}$ is defined as:

$$\tilde{\mathbf{x}} = \mathbf{x} + \alpha \cdot \mathbf{n}_{perceptual}, \tag{1}$$

where $\mathbf{n}_{perceptual}$ denotes a continuous-scale Gaussian noise field generated dynamically for each batch, and α is a scaling factor sampled according to a uniform distribution that controls the Target Signal-to-Noise Ratio (SNR_{target}):

$$\text{SNR}_{target} \sim \mathcal{U}(\text{SNR}_{min}, \text{SNR}_{max}). \tag{2}$$

DPPA operates over two asymmetric acoustic regimes. In the low-SNR interval ($\text{SNR}_{target} \in [0, 10]$ dB), the added noise simulates channel transmission effects and background degradation. In the high-SNR interval ($\text{SNR}_{target} \in [25, 30]$ dB), the small perturbation acts as a perceptual carrier that magnifies discrete, localized codebook quantization irregularities in Audio Language Models (ALMs), thereby increasing the forensic sensitivity of the downstream network.

3.2 Time-Layer-Feature Forgery Manifold Construction

The augmented waveform $\tilde{\mathbf{x}}$ is then processed by a frozen, pretrained self-supervised learning (SSL) backbone, specifically an XLS-R model with 24 Transformer layers.

Prior approaches often collapse multi-layer feature maps through temporal pooling or linear projection before classification, potentially discarding useful layer-wise information. We instead stack the hidden representations from all XLS-R layers to construct a unified *Time-Layer-Feature (T-L-F) Joint Forgery Manifold*.

Let $f_l(\tilde{\mathbf{x}}) \in \mathbb{R}^{B\times T\times H}$ denote the representation extracted from the l-th hidden layer of XLS-R, where B denotes the batch size, T the temporal frame length, $H = 1024$ the hidden feature dimension, and $l \in \{1, \ldots, 24\}$ the layer index. We construct the joint manifold $\mathbf{F}$ by stacking these representations along the layer dimension:

$$\mathbf{F} = \text{Stack}_{l=1}^{24}\left(f_l(\tilde{\mathbf{x}})\right) \in \mathbb{R}^{B\times L\times T\times H}, \quad L = 24. \tag{3}$$

The resulting representation retains both shallow acoustic cues from lower XLS-R layers and higher-level semantic and prosodic cues from deeper layers.

3.3 Efficient Multi-Scale Attention on Forgery Manifold

To model cross-dimensional correlations in the manifold $\mathbf{F}$ while preserving its spatial structure, we adapt the Efficient Multi-Scale Attention (EMA) mechanism Ouyang et al. (2023) to audio forensics. EMA first partitions the joint manifold $\mathbf{F}$ along the depth (channel) dimension into g symmetrical sub-manifolds, denoted by $[\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_g]$, where each sub-manifold $\mathbf{F}_i \in \mathbb{R}^{B\times c\times T\times H}$ has a sub-channel depth of $c = L/g$.

Within each sub-manifold, a parallel tri-branch structure captures complementary synthetic speech artifacts:

1. **Temporal-Layer Path (Horizontal 1D Pooling):** The horizontal branch applies 1D global average pooling along the feature dimension H to model long-range temporal phase discontinuities and global rhythmic mismatches introduced by autoregressive ALMs. Its output represents the temporal evolution of artifacts across layers:

$$z_t^l = \frac{1}{H}\sum_{h=1}^{H} \mathbf{F}_i(l, t, h), \tag{4}$$

   where z_t^l represents localized acoustic energy fluctuations at frame t across the l-th sub-layer.

2. **Feature-Layer Path (Vertical 1D Pooling):** Codec-based generation may produce non-uniform distortions across hidden feature components. The vertical branch applies 1D global average pooling along the temporal dimension T to aggregate feature-wise statistics:

$$z_h^l = \frac{1}{T}\sum_{m=1}^{T} \mathbf{F}_i(l, m, h), \tag{5}$$

   where z_h^l represents feature-dimensional distortions within the h-th hidden component.

3. **Localized Artifact Path (2D Convolutional Branch):** To retain fine-grained local anomalies (e.g., localized neural rendering noise or high-frequency compression boundaries) that global pooling may smooth, the third branch applies a localized 3×3 2D convolution layer to the sub-manifold and extracts multi-scale structural details:

$$\mathbf{F}_{2d} = \text{Conv}_{3\times 3}(\mathbf{F}_i) \in \mathbb{R}^{B\times c\times T\times H}. \tag{6}$$

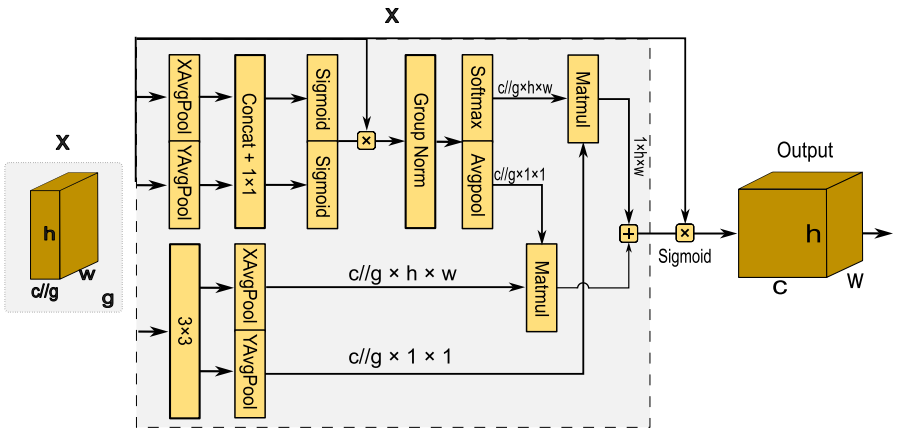


Figure 2: Internal structure diagram of EMA. Here, "g" represents the divided groups, "X Avg Pool" denotes the one-dimensional horizontal global pooling, and "Y Avg Pool" indicates the 1D vertical global pooling.

After multi-dimensional feature extraction, the two 1D directional pooling branches are concatenated along the spatial dimension. The directional attention map $\mathbf{A}_s$ is produced by a shared 1×1 2D convolution followed by a Sigmoid activation:

$$\mathbf{A}_s = \sigma\left(\text{Conv}_{1\times1}([\mathbf{z}_h, \mathbf{z}_v])\right), \tag{7}$$

where $[\cdot, \cdot]$ denotes concatenation, and $\mathbf{z}_h, \mathbf{z}_v$ are the reshaped horizontal and vertical pooling vectors.

Finally, the directional attention weights $\mathbf{A}_s$ modulate the original sub-manifold, and the result is combined with the enhanced localized 2D features. The final multi-scale representation $\mathbf{Y}_i$ for each sub-manifold group is defined as:

$$\mathbf{Y}_i = (\mathbf{F}_i \otimes \mathbf{A}_s) \oplus \sigma\left(\text{Conv}_{1\times1}(\mathbf{F}_{2d})\right), \tag{8}$$

where $\otimes$ denotes element-wise multiplication and $\oplus$ denotes element-wise addition.

4 Experiments

4.1 Datasets and Performance Metrics

For evaluation on codec-based speech deepfakes, we extract a 1% subset (7399 samples) from the Codecfake training set, denoted as Codecfake-mini, and combine it with the ASVspoof 2019 LA training set. We exclude the validation set because its distribution shift primarily concerns speakers and therefore does not mitigate overfitting to synthesis-specific artifacts Zhang et al. (2024). We evaluate the proposed method on ASVspoof 2021 LA & DF Yamagishi et al. (2021), the In-The-Wild benchmark Müller et al. (2022), and the Codecfake evaluation set Xie et al. (2025). Codecfake contains synthetic audio produced with audio codecs and thus evaluates generalization to codec-based generation methods. We report Equal Error Rate (EER) and minimum tandem detection cost function (min t-DCF) Kinnunen et al. (2020) as the primary metrics. Table 1 summarizes the dataset statistics:

4.2 Training Setup

Audio samples are cropped or concatenated into segments of approximately 4 seconds (64,600 samples). We use the Adam optimizer with a learning rate of 10^{-6}, a weight decay coefficient of 0, and a batch size of 10. Training runs for 100 epochs with early stopping (patience = 3). All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU using a fixed random seed.

Table 1: Detailed information about the training and evaluation sets.

Dataset	Real	Fake	Total
Train (Codecfake-mini)	1,057	6,342	7,399
Train (ASVspoof 2019 LA)	2,580	22,800	25,380
Eval (ASVspoof 2021 LA)	18,452	163,114	181,566
Eval (ASVspoof 2021 DF)	22,617	589,212	611,829
Eval (In-The-Wild)	19,963	11,816	31,779
Eval (Codecfake)	75,399	149,474	224,873

4.3 Experimental Results and Comparative Analysis

4.3.1 Performance on ALM-Generated and Real-World Deceptions

Table 2 compares the proposed framework with representative baselines on In-The-Wild (ITW), Codecfake, and ASVspoof 2021 DF in terms of EER and area under the receiver operating characteristic curve (AUC-ROC). The compared methods are RawNet3 (weon Jung et al., 2022), AASIST (Jung et al., 2022), SLS (Zhang et al., 2024), MCLN (Zhang et al., 2025), XLSR-Mamba (Xiao & Das, 2025), LWDF (Xiao & Vu, 2025), HISPASpoof (Risques et al., 2026), and PVP (Xue et al., 2026).

Table 2: Comparative analysis of EER (%) and AUC-ROC (%) on In-The-Wild, Codecfake, and ASVspoof 2021 DF datasets.

	In-The-Wild		**Codecfake**		**ASVspoof 2021 DF**	
Method	**EER (%)**	**AUC (%)**	**EER (%)**	**AUC (%)**	**EER (%)**	**AUC (%)**
LFCC-LCNN	19.98	86.42	41.25	62.15	5.43	97.10
RawNet3	15.42	91.10	38.67	65.48	4.21	98.02
AASIST	12.35	93.45	39.12	64.10	3.56	98.45
XLS-R + Linear	8.12	96.88	36.45	68.32	2.14	99.15
XLS-R + SLS	7.46	97.21	33.43	71.55	1.99	99.30
XLS-R + MCLN	6.54	97.90	32.10	72.84	1.76	99.45
XLS-R + Mamba	6.22	98.15	31.55	73.10	1.62	99.51
XLS-R + LWDF	5.98	98.32	28.45	76.50	1.48	99.63
XLS-R + HISPASpoof	5.82	98.42	27.10	77.34	1.42	99.68
PVP Classifier	6.14	98.10	25.65	79.20	1.35	99.72
Ours + Add. Data	**5.25**	**99.12**	**4.64**	**98.54**	**1.24**	**99.88**

Table 2 shows that the proposed framework performs favorably across the evaluated benchmarks. The largest margin is observed on **Codecfake**, which targets speech generated by next-generation Audio Language Models (ALMs). Compared with competitive baselines from 2025, including the multi-view collaborative learning framework (XLS-R + MCLN) and the selective state-space configuration (XLS-R + Mamba), which report EERs exceeding 31%, our model achieves lower EER and higher AUC-ROC.

This margin supports the use of explicit cross-layer modeling. Existing methods based on sequential modeling or feature alignment commonly aggregate hidden states through holistic projections, which may weaken the multi-layer structure and obscure non-uniform sub-band quantization artifacts introduced by neural audio codecs. The proposed 3D *Time-Layer-Feature Joint Forgery Manifold* instead retains this structure to model complementary forensic cues across layers.

We also evaluate on ASVspoof 2021 DF, which includes lossy codec conditions such as MP3, OGG, and AAC. As shown in Table 2, XLSR-EMA achieves the lowest EER and highest AUC, indicating robustness to codec-induced domain shifts. The tri-branch EMA design captures temporal and

Table 3: Cross-dataset generalization evaluation (EER%) under unseen synthesis paradigms.

Method	Train: ASVspoof2019 LA		Train: Codecfake-mini	
	Test: Codecfake	Test: ITW	Test: 21DF	Test: ITW
LFCC-LCNN	39.45	19.75	27.09	20.47
RawNet3	36.87	16.34	23.57	20.21
AASIST	36.32	15.56	20.05	17.98
XLS-R + Linear	32.71	12.38	18.67	15.09
XLS-R + SLS	33.58	12.07	17.66	13.54
XLS-R + MCLN	32.10	10.98	15.08	12.69
XLS-R + Mamba	30.59	8.44	12.87	10.24
XLS-R + LWDF	30.27	6.24	9.96	7.12
XLS-R + HISPASpoof	25.95	6.33	8.62	6.09
PVP Classifier	20.44	6.38	7.74	5.86
Ours (XLSR-EMA)	**17.76**	**5.52**	**4.43**	**4.80**

Table 4: Performance (EER%) under different EMA group configurations (g).

Groups (g)	21DF	ITW	C7	avg
$g = 1$	**1.35**	**6.21**	23.90	4.55
$g = 2$	1.55	6.69	**17.15**	**3.57**
$g = 3$	1.50	6.82	21.10	4.10
$g = 4$	1.55	7.01	21.45	4.45

21DF: ASVspoof 2021 DF; ITW: In-The-Wild; C7: Codecfake test (unseen codec); avg: average EER on Codecfake.

feature-wise artifacts through directional pooling, while DPPA improves tolerance to channel variation.

4.4 Cross-Dataset Generalization Evaluation

We conduct cross-dataset experiments to assess whether the proposed framework captures transferable forgery cues rather than dataset-specific artifacts under distribution shifts.

Specifically, we train all methods on ASVspoof2019 LA and evaluate them on Codecfake and ITW, which represents transfer from traditional vocoder-based attacks to speech generated by modern Audio Language Models (ALMs). We then train the models on Codecfake-mini and evaluate them on ASVspoof2021 DF and ITW to examine the reverse transfer direction.

Table 3 reports the results. XLSR-EMA achieves the lowest EER in all cross-domain settings, indicating robustness to unseen synthesis algorithms, codec configurations, speaker distributions, and recording environments.

The improvement over previous XLS-R based approaches is most pronounced in the challenging LA→Codecfake setting. This result suggests that preserving the hierarchical SSL topology in the proposed Time-layer-feature manifold helps capture forensic cues that transfer beyond dataset-specific generation traces.

4.5 Analysis of EMA Configurations

We examine how the internal structure of the EMA module affects detection performance by varying the group number g. Treating the XLS-R backbone's 24 layers as channels ($C = 24$), we evaluate $g \in \{1, 2, 3, 4\}$, which partitions the representation into subsets of 24, 12, 8, and 6 layers.

As shown in Table 4, $g = 1$ yields the lowest EER on vocoder-based datasets, where global feature interaction is beneficial. In contrast, $g = 2$ generalizes better to the ALM-based C7 attack. This difference is consistent with the hierarchy of SSL representations, in which lower layers encode discriminative forensic cues and higher layers capture semantic information Kheir et al. (2025). The $g = 2$ configuration balances interactions between these multi-scale features and better isolates ALM-generated micro-artifacts. Moreover, joint training on diverse datasets (e.g., Codecfake-mini and ASVspoof 2019 LA) yields the lowest average EER. These results indicate that the EMA module supports detection of both ALM-based general audio and traditional deepfake speech across multiple domains.

4.6 Ablation Study and Sensitivity Analysis

We conduct ablation experiments on the Codecfake and 21DF datasets to assess the contribution of each component. Table 5 summarizes the results.

Table 5: Ablation analysis of the proposed framework on Codecfake and 21DF datasets (EER, %).

Configuration	Description	Codecfake	21DF
Baseline Systems			
	XLS-R (Layer 24) + Classifier	33.43	5.12
EMA Group Variants			
$g = 1, c = 24$	Full manifold joint attention	10.74	1.88
$g = 2, c = 12$	Hierarchical grouped attention (Ours + Add. Data)	**4.64**	**1.24**
$g = 4, c = 6$	Over-fragmented sub-manifolds	6.12	1.45
Ablation Components			
w/o Horizontal Path	Disabling temporal-layer pooling	7.89	1.62
w/o Vertical Path	Disabling feature-layer pooling	8.34	1.71
w/o 2D Conv Path	Disabling localized artifact tracking	25.92	1.39
w/o DPPA	Dropping perceptual data augmentation	4.86	1.32

g: number of EMA sub-manifold groups; c: sub-channel depth per group.

4.6.1 Qualitative Visualization

Fig. 3 visualizes the DPPA spectral response and learned representation space. The residual exposes localized codec artifacts, transient phase distortion, and high-frequency energy loss while preserving the dominant speech structure. The t-SNE projection shows clearer separation between genuine speech and ALM-generated spoofed speech. These results serve as qualitative diagnostics rather than standalone evidence.

4.6.2 Effectiveness of the Tri-Branch EMA Architecture

Table 5 shows that each EMA branch contributes to performance. Removing the horizontal 1D pooling path (w/o Horizontal Path) increases the Codecfake EER from 4.64% to 7.89%, indicating that global temporal modeling captures frame-wise phase inconsistencies from autoregressive ALMs. Removing the vertical path (w/o Vertical Path) raises the EER to 8.34%, while the local 2D convolutional path preserves fine-grained acoustic rendering artifacts that global pooling may smooth.

4.6.3 Impact of Sub-Manifold Grouping (g)

The group parameter g controls cross-spatial interaction within the Time-layer-feature manifold. With $g = 1$ ($c = 24$), holistic attention across all 24 layers yields 10.74% EER on Codecfake. Setting $g = 2$ ($c = 12$) reduces the EER to 4.64%. Dividing the manifold into two symmetrical groups separates interacting lower-layer acoustic cues from layers 1–12 and higher-layer semantic information from layers 13–24, while EMA models associations between these two distinct forensic

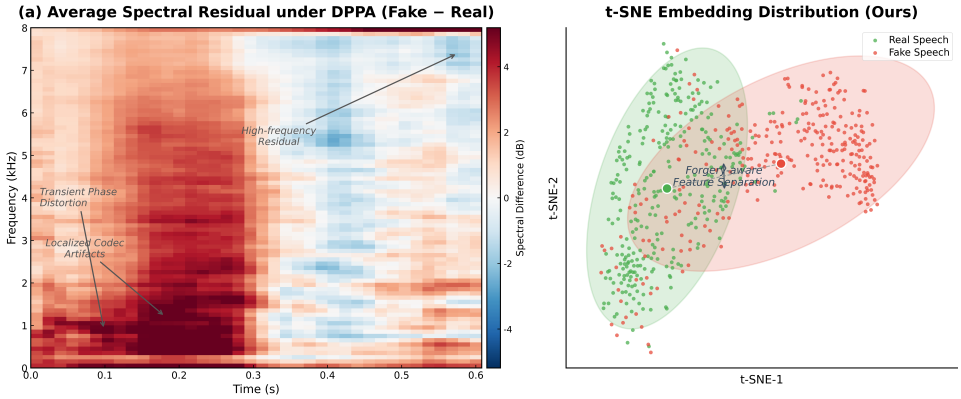


Figure 3: Qualitative visualization of the proposed framework. (a) Average spectral residual (Fake − Real) under DPPA, illustrating the amplification of codec-related artifacts, transient distortion, and high-frequency energy discrepancy. (b) t-SNE projection of the learned temporal layer-wise feature representations, showing clear separation between genuine speech and ALM-generated spoofed speech.

domains. More groups (e.g., $g = 4, c = 6$) restrict each sub-attention cell and degrade performance (6.12% EER).

4.6.4 Contribution of the DPPA Strategy

Finally, removing the customized Dual-Purpose Perceptual Augmentation (DPPA) strategy (w/o DPPA) increases the average EER to 4.86% and 1.32% on Codecfake and 21DF, respectively. DPPA provides regularization by injecting continuous-scale Gaussian noise, simulating real-world transmission channels under low SNR and amplifying deterministic quantization artifacts under high SNR to improve out-of-distribution generalization.

5 Conclusion

In this paper, we propose a universal speech deepfake detection framework to counter the emerging threats of next-generation Audio Language Models (ALMs). Instead of collapsing multi-layer features into static embeddings, our framework preserves the full hidden states of a pretrained XLS-R model as a 3D *Time-Layer-Feature Joint Forgery Manifold*, maintaining the critical layer-wise topological hierarchy. To capture cross-dimensional artifacts, we integrate an adapted Efficient Multi-scale Attention (EMA) module with a parallel tri-branch structure, which simultaneously models long-range temporal discontinuities, spectral distortions, and localized micro-rendering noise. Furthermore, the introduced Dual-Purpose Perceptual Augmentation (DPPA) enhances robustness against channel fluctuations and codebook quantization anomalies. Extensive evaluations on In-The-Wild, Codecfake, and ASVspoof 2021 DF benchmarks demonstrate that our method achieves state-of-the-art performance.

In the future, we plan to extend this work in three directions: (1) **Multi-Modal Extension:** Remapping the 3D joint manifold into higher-dimensional tensors to defend against synchronized audio-visual deepfakes. (2) **Lightweight Deployment:** Investigating structured network pruning and quantization for the XLS-R-EMA block to lower the real-time factor (RTF) for edge deployment. (3) **Interpretability:** Leveraging the multi-scale attention maps from the EMA module to localize manipulated temporal-spectral regions, advancing the system into an explainable forensic tool.

Generative AI Use Disclosure

Generative AI tools assisted with language polishing, formatting, code implementation, and results analysis; the authors verified all outputs and retain full responsibility.

REFERENCES

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, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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. doi: 10.1109/JSTSP.2022.3188113.

Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Zhao Lv, and Cunhang Fan. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. In *Interspeech 2024*, pp. 2720–2724, 2024. doi: 10.21437/Interspeech.2024-698.

Arnab Das, Yassine El Kheir, Carlos Franzreb, Tim Herzig, Tim Polzehl, and Sebastian Möller. Generalizable Audio Spoofing Detection using Non-Semantic Representations. In *Interspeech 2025*, pp. 4553–4557, 2025. doi: 10.21437/Interspeech.2025-1555.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980. doi: 10.1109/TASSP.1980.1163420.

Quchen Fu, Zhongwei Teng, Jules White, Maria E. Powell, and Douglas C. Schmidt. Fastaudio: A learnable audio front-end for spoof speech detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3693–3697, 2022. doi: 10.1109/ICASSP43922.2022.9746722.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. *arXiv preprint arXiv:2503.06142*, 2025.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371, 2022. doi: 10.1109/ICASSP43922.2022.9747766.

Yassine El Kheir, Youness Samih, Suraj Maharjan, Tim Polzehl, and Sebastian Möller. Comprehensive layer-wise analysis of ssl models for audio deepfake detection. *arXiv preprint arXiv:2502.03559*, 2025.

Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, and Douglas A. Reynolds. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2195–2210, 2020. doi: 10.1109/TASLP.2020.3009494.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. Audio anti-spoofing detection: A survey. *arXiv preprint arXiv:2404.13914*, 2024.

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 Müller, Piotr Kawa, Wei-Herng Choong, Adriana Stan, Aditya Tirumala Bukkapatnam, Karla Pizzi, Alexander Wagner, and Philip Sperl. Replay Attacks Against Audio Deepfake Detection. In *Interspeech 2025*, pp. 2245–2249, 2025. doi: 10.21437/Interspeech.2025-20.

Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096516.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018. doi: 10.1109/SLT.2018.8639585.

Maria Risques, Kratika Bhagtani, Amit Kumar Singh Yadav, and Edward J. Delp. Detecting and attributing synthetic spanish speech: The hispaspoof dataset. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13572–13576, 2026. doi: 10.1109/ICASSP55912.2026.11464323.

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, 2022a. doi: 10.1109/ICASSP43922.2022.9746213.

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.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *The Speaker and Language Recognition Workshop (Odyssey 2016)*, pp. 283–290, 2016. doi: 10.21437/Odyssey.2016-41.

Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion. In *Interspeech 2018*, pp. 77–81, 2018. doi: 10.21437/Interspeech.2018-2289.

Taiba Majid Wani and Irene Amerini. Multi-Scale Self-Supervised Learning for Efficient Audio Deepfake Detection. *IEEE Signal Processing Letters*, 33:46–50, 2026. doi: 10.1109/LSP.2025.3634032.

Jee weon Jung, Youjin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. In *Interspeech 2022*, pp. 2228–2232, 2022. doi: 10.21437/Interspeech.2022-126.

Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Héctor Delgado. Asvspoof: The automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):588–604, 2017. doi: 10.1109/JSTSP.2017.2671435.

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. *ISCA Interspeech, Rotterdam, The Netherlands*, pp. 5618–5622, 2025.

Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, Long Ye, and Yi Sun. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*, 33:386–400, 2025. doi: 10.1109/TASLPRO.2025.3525966.

Jun Xue, Tong Zhang, Zhuolin Yi, Yihuan Huang, Yi Chai, Yiyang Zhang, and Yanzhen Ren. Profiling the voice: Speaker-specific phoneme fingerprinting for speech deepfake detection. *arXiv preprint arXiv:2605.17737*, 2026. URL https://arxiv.org/abs/2605.17737. Accepted by IJCAI 2026.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 47–54, 2021. doi: 10.21437/ASVSPOOF.2021-8.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda. Voice conversion challenge 2020 — intra-lingual semi-parallel and cross-lingual voice conversion—. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98, 2020. doi: 10.21437/VCCBC.2020-14.

Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yifang Guo, Yushu Zhang, and Guoai Xu. Multi-view collaborative learning network for speech deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1075–1083, 2025.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 6765–6773, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681345. URL https://doi.org/10.1145/3664647.3681345.

Yue Zhou, Xinan He, KaiQing Lin, Bin Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *arXiv preprint arXiv:2506.00874*, 2025.