AGDT: AUDIO DEEPFAKE DETECTION VIA ASYMMETRIC GATED DUAL-STAGE TEMPORAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

In audio deepfake detection under unseen generators, channel distortions, and real-world recording conditions, discriminative evidence may be distributed across representation depths, temporal scales, and utterance-level statistics. To address these characteristics, we propose AGDT, a framework that aggregates evidence across layer depth, temporal scale, and distributional statistics, using XLS-R as its frontend. First, non-competitive sensitive layer gating is applied to the frame-level representations of the XLS-R Transformer layers, allowing multiple layers with complementary discriminative information to contribute to feature fusion. Next, a multi-scale temporal convolution module applies sequential channel and temporal gates to local residual updates before a Conformer backend, while a second module refines the contextual representations without additional gating. This asymmetric configuration places selective local processing before the backend encoder and ungated refinement after it. Finally, attentive statistics pooling combines weighted means and standard deviations to capture both the central tendency and temporal dispersion of frame-level features. For the downstream detection task, the model is trained only on the ASVspoof 2019 LA training set and evaluated on multiple benchmarks involving shifts across datasets, channels, and languages. It achieves equal error rates of 2.48%, 1.56%, 13.57%, and 5.83% on ASVspoof 2021 LA, ASVspoof 2021 DF, ASVspoof 5, and In-the-Wild, respectively. The experimental results demonstrate competitive performance across multiple evaluation domains.

1 INTRODUCTION

Recent advances in deep learning have substantially improved the naturalness and speaker similarity of synthetic speech generated by text-to-speech (TTS) and voice conversion (VC) systems. As shown in Figure 1, these technologies support beneficial applications, including accessibility tools and personalized voice assistants, but also create opportunities for impersonation, fraud, and misinformation. As high-quality speech synthesis becomes increasingly accessible, reliably distinguishing bona fide speech from synthetic or converted speech has become an important challenge for trustworthy audio processing.

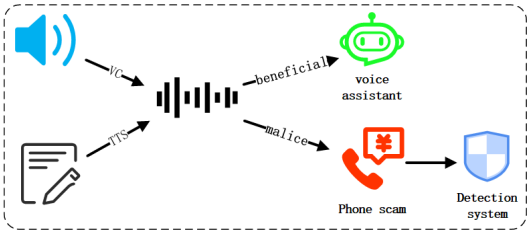

Figure 1: Beneficial applications and potential misuse of synthetic speech, motivating the development of audio deepfake detection systems.

Benchmark initiatives have established standardized evaluation protocols for speech spoofing and deepfake detection, but cross-domain generalization remains challenging. The ASVspoof (Todisco et al., 2019; Yamagishi et al., 2021; Wang et al., 2024) and Audio Deepfake Detection (ADD) challenge series (Yi et al., 2022) provide diverse evaluation datasets, while Speech DF Arena (Dowerah et al., 2026) offers an additional resource for comparing detectors across evaluation conditions. Despite low error rates on benchmarks such as ASVspoof 2019 LA (Todisco et al., 2019) , detector performance can deteriorate under channel compression, changes in language or recording conditions, and previously unseen generation

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.

methods. Such degradation suggests that models may exploit dataset-specific spectral patterns or acquisition conditions without learning sufficiently transferable spoofing cues. Cross-domain generalization should therefore be treated as a central objective of representation learning and temporal modeling.

Self-supervised speech models, such as XLS-R (Babu et al., 2022), provide a strong foundation for audio deepfake detection by learning transferable representations from large amounts of unlabeled speech. However, these representations are not explicitly pretrained to distinguish bona fide speech from synthetic speech, and different Transformer layers may encode speech information at different levels of abstraction, including acoustic and linguistic properties (Pasad et al., 2021). Consequently, using only the final layer may overlook complementary forgery cues preserved at intermediate depths. Moreover, informative cues may be distributed across several layers, making it useful to allow multiple layers to contribute simultaneously (Zhang et al., 2024a). These observations motivate adapting pretrained representations through task-dependent layer fusion and a backend that can integrate local evidence with its broader temporal context.

Conformer (Gulati et al., 2020; Rosello et al., 2023) is used as the detection backbone because it combines local acoustic modeling with global contextual interaction. Its convolution module captures neighboring-frame patterns, whereas self-attention models dependencies between distant positions. This combination is suitable for audio deepfake detection, where short acoustic discontinuities and inconsistencies spanning multiple phonemes require different temporal contexts. However, the standard Conformer uses a single convolution kernel within each block and does not explicitly construct parallel local receptive fields. This motivates supplementing its contextual modeling with dedicated multi-scale processing before and after the encoder. Unlike approaches that introduce multi-scale convolution inside each Conformer block, AGDT places multi-scale temporal convolution outside the Conformer encoder. We place two MSTCMs before and after the Conformer because local spoofing evidence has different roles before and after contextual encoding. Before the Conformer, the fused XLS-R representations retain relatively direct local acoustic cues, such as short-duration spectral distortions, discontinuities near phoneme boundaries, and vocoder-related artifacts. After Conformer encoding, the representations contain long-range contextual dependencies, but local artifacts may be mixed with or smoothed by global information exchange. The post-MSTCM performs ungated multi-scale residual refinement on these contextualized representations, enabling local evidence to be re-examined at multiple temporal scales. Using an ungated post-MSTCM avoids repeatedly imposing strong feature selection after contextual integration. Consequently, the dual-stage design forms a local enhancement–global interaction–local refinement pathway.

The main contributions of this work are summarized as follows:

- We incorporate sensitive layer fusion into the XLS-R frontend to exploit complementary representations across Transformer depths.
- We propose an asymmetric external multi-scale architecture in which a gated pre-MSTCM and an ungated post-MSTCM are placed around a standard Conformer encoder, forming a local enhancement–global interaction–local refinement pathway.
- We evaluate the resulting model across multiple datasets and analyze the effects of module placement, temporal receptive fields, and encoder architecture.

2 Related Work

Audio deepfake detection has progressed from handcrafted acoustic features to end-to-end neural architectures and self-supervised speech representations. As speech generation techniques and transmission conditions continue to evolve, recent work has increasingly focused on generalization to unseen attacks and robustness in realistic settings.

Early anti-spoofing systems extracted spectral, phase, and prosodic features and used conventional classifiers such as Gaussian mixture models and support vector machines (Sahidullah et al., 2015). Common features include MFCCs, LFCCs, and constant-Q cepstral coefficients (CQCCs) (Todisco et al., 2017). Although these methods were effective in early evaluations, their performance depends on the selected features and the characteristics of the spoofing attacks. This limitation motivated the use of neural networks to learn discriminative representations directly from data.

Convolutional neural networks have been widely used to capture local patterns associated with spoofed speech. Representative architectures include LCNN and Res2Net (Lavrentyeva et al., 2017; Li et al., 2021). LCNN uses max-feature-map activations for feature selection, whereas Res2Net improves multi-scale representation through interactions between channel groups. However, the temporal context of convolutional models is constrained by their depth and kernel configurations, which can limit the modeling of dependencies across distant speech segments.

Transformer-based architectures provide explicit mechanisms for modeling long-range dependencies. Conformer combines self-attention with convolutional modeling within each encoder block (Gulati et al., 2020), and has been adapted to anti-spoofing with pretrained speech representations (Rosello et al., 2023). However, the standard Conformer convolution module uses a single temporal kernel size per block and does not explicitly construct parallel local receptive fields. MultiConvformer addresses this limitation by replacing the single-kernel convolution inside each Conformer block with a gated multi-kernel convolution module (Prabhu et al., 2024). Its results on ASR and spoken language understanding demonstrate that multiple receptive fields can improve local context modeling. Nevertheless, this design performs multi-scale processing repeatedly within the encoder and was developed for speech recognition rather than audio deepfake detection.

Self-supervised learning provides transferable representations for speech processing and audio deepfake detection. Models such as wav2vec 2.0, HuBERT, WavLM, and XLS-R are pretrained on large amounts of unlabeled speech (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Babu et al., 2022). Layer-wise analysis further indicates that hidden representations encode speech information at different levels of abstraction (Pasad et al., 2021). For deepfake detection, SLS aggregates features from multiple XLS-R layers (Zhang et al., 2024a), while XLSR-Mamba uses a bidirectional state-space backend for temporal modeling (Xiao & Das, 2025).

Finally, utterance-level pooling aggregates frame-level features for classification. Attentive statistics pooling computes weighted means and standard deviations, thereby representing both the central tendency and temporal variation of an utterance (Okabe et al., 2018). Building on these studies, AGDT combines sensitive layer fusion, preserves the standard Conformer blocks, and places multi-scale temporal convolution outside the encoder. The pre-MSTCM selectively enhances local multi-scale evidence before global contextual interaction, while the post-MSTCM refines the contextualized representations after Conformer encoding. This asymmetric design explicitly separates local enhancement, global interaction, and local refinement, making it suitable for detecting spoofing artifacts that may play different roles before and after contextual encoding.

3 METHOD

As illustrated in Figure 2, AGDT comprises an XLS-R frontend with sensitive layer fusion, a gated pre-MSTCM, a Conformer encoder, an ungated post-MSTCM, attentive statistics pooling, and a linear classifier. Given a waveform $\mathbf{x} \in \mathbb{R}^N$, the model predicts whether the input speech is bona fide or spoofed by integrating information across representation layers and temporal scales.

3.1 XLS-R FEATURE EXTRACTION AND SENSITIVE LAYER FUSION

We employ XLS-R 300M (Babu et al., 2022) as the speech representation frontend. It consists of a convolutional feature encoder followed by $L = 24$ Transformer layers, each producing a frame-level representation $\mathbf{H}_l \in \mathbb{R}^{T \times D}$, where $D = 1024$. Because different layers retain speech information at different levels of abstraction (Pasad et al., 2021), all Transformer layer outputs are retained.

Following sensitive layer fusion (Zhang et al., 2024a), an input-dependent gate is estimated for each layer from its temporally averaged representation:

$$\mathbf{z}_l = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_{l,t}, \qquad g_l = \mathrm{sigmoid}\left(\mathbf{w}^\top \mathbf{z}_l + b\right), \qquad \mathbf{F} = \sum_{l=1}^{L} g_l \mathbf{H}_l.$$

The gates are not normalized across layers, allowing multiple informative layers to contribute simultaneously. The fused representation $\mathbf{F} \in \mathbb{R}^{T \times D}$ is projected to $d = 144$ dimensions, followed by two-dimensional batch normalization and SELU activation. The resulting sequence is denoted by $\mathbf{X}_0 \in \mathbb{R}^{T \times d}$.

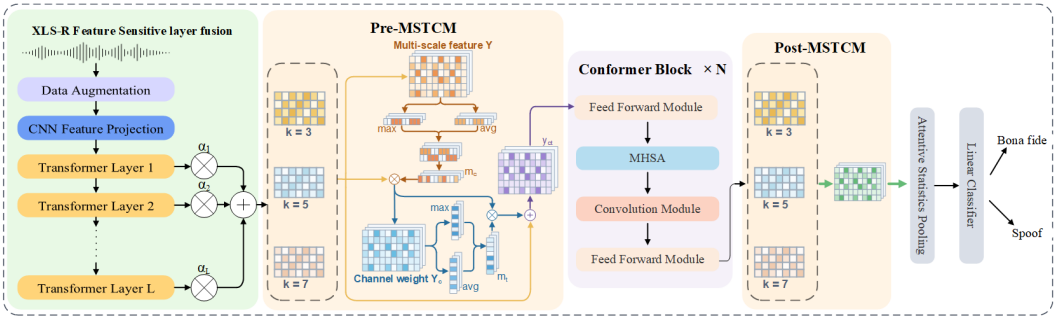

Figure 2: Overview of AGDT. XLS-R layer representations are adaptively fused and processed by a gated pre-MSTCM, a Conformer encoder, and an ungated post-MSTCM. Each MSTCM contains three parallel temporal convolution branches with kernel sizes $\{3, 5, 7\}$. Attentive statistics pooling and a linear classifier produce the final prediction.

3.2 Asymmetric Dual-Stage Multi-Scale Temporal Modeling

The temporal backend combines local multi-scale processing with global contextual modeling. A gated MSTCM is placed before the Conformer to selectively enhance local artifacts, whereas an ungated MSTCM is placed after the Conformer to refine the contextualized representation without imposing a second strong selection operation.

Each MSTCM contains three parallel branches with temporal kernel set $\mathcal{K} = \{3, 5, 7\}$. Each branch applies depthwise temporal convolution, batch normalization, GELU activation, pointwise convolution, and dropout. For stage $s \in \{\text{pre}, \text{post}\}$, the branch outputs are averaged as

$$\mathbf{Y}_s(\mathbf{X}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathbf{B}_k^{(s)}(\mathbf{X}),$$

where $\mathbf{B}_k^{(s)}(\mathbf{X}) \in \mathbb{R}^{T \times d}$ denotes the output of the branch with kernel size k. Depthwise convolution captures scale-specific temporal patterns, while pointwise convolution mixes information across channels. Unit stride and appropriate padding preserve the temporal length. The two MSTCMs use the same branch structure but have independent parameters.

The pre-MSTCM sequentially applies channel and temporal gates. The channel gate is estimated using temporal average and maximum pooling:

$$\mathbf{m}_c = \sigma\left(\phi\left(\text{Avg}_t(\mathbf{Y}_{\text{pre}})\right) + \phi\left(\text{Max}_t(\mathbf{Y}_{\text{pre}})\right)\right), \qquad \mathbf{Y}_c = \mathbf{Y}_{\text{pre}} \odot \mathbf{m}_c,$$

where ϕ is a shared two-layer pointwise mapping with GELU activation and hidden dimension $d/4$. The resulting channel gate $\mathbf{m}_c \in \mathbb{R}^{1 \times d}$ is broadcast along the temporal dimension.

The temporal gate is subsequently computed from the channel-reweighted feature:

$$\mathbf{m}_t = \sigma\left(\text{Conv}_7\left([\text{Avg}_d(\mathbf{Y}_c); \text{Max}_d(\mathbf{Y}_c)]\right)\right), \qquad \mathbf{Y}_{c\tau} = \mathbf{Y}_c \odot \mathbf{m}_t.$$

Here, $\mathbf{m}_t \in \mathbb{R}^{T \times 1}$ assigns an input-dependent weight to each temporal position and is broadcast over the feature dimension. The convolution with kernel size 7 allows each weight to incorporate its neighboring temporal context.

The gated and ungated modules are respectively defined as

$$\mathcal{M}_{\text{pre}}(\mathbf{X}) = \text{LN}_{\text{pre}}\left(\mathbf{X} + \mathbf{Y}_{c\tau}\right), \qquad \mathcal{M}_{\text{post}}(\mathbf{X}) = \text{LN}_{\text{post}}\left(\mathbf{X} + \mathbf{Y}_{\text{post}}(\mathbf{X})\right).$$

This asymmetric design forms a local enhancement–global interaction–local refinement pathway. The pre-MSTCM selects informative channels and temporal regions before contextual encoding, whereas the post-MSTCM re-examines the contextualized features at multiple local scales without additional gating.

A fixed sinusoidal positional encoding is added to $\mathbf{X}_0$ before the pre-MSTCM. The output is then processed by four Conformer blocks (Gulati et al., 2020; Rosello et al., 2023), followed by the post-MSTCM. Each Conformer block contains two half-step feed-forward modules, multi-head self-attention, and a convolution module with residual connections. We use a feature dimension of

$d = 144$, four attention heads, a feed-forward expansion factor of 4, and a convolution kernel size of 31.

3.3 Attentive Statistics Pooling and Classification

Attentive statistics pooling summarizes the post-MSTCM output using its weighted mean and temporal dispersion (Okabe et al., 2018). For frame representation $\mathbf{h}_t \in \mathbb{R}^d$, an attention network f_{att} produces a scalar score, which is normalized over time:

$$a_t = f_{\text{att}}\left(\text{LN}(\mathbf{h}_t)\right), \qquad \alpha_t = \frac{\exp(a_t)}{\sum_{j=1}^{T} \exp(a_j)}.$$

The utterance-level representation is obtained from the attentive mean and standard deviation:

$$\boldsymbol{\mu} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t, \qquad \boldsymbol{\sigma} = \sqrt{\max\left(\sum_{t=1}^{T} \alpha_t (\mathbf{h}_t - \boldsymbol{\mu})^{\odot 2}, 10^{-5}\right)}, \qquad \mathbf{e} = [\boldsymbol{\mu}; \boldsymbol{\sigma}] \in \mathbb{R}^{2d}.$$

A linear classifier maps $\mathbf{e}$ to two logits corresponding to bona fide and spoofed speech. The model is optimized using cross-entropy loss, and the difference between the bona fide and spoof logits is used as the detection score during evaluation.

4 Experiments

We evaluate AGDT on multiple public benchmarks to examine its detection performance and cross-domain generalization. This section introduces the datasets, evaluation metric, and experimental settings, followed by comparisons with existing systems and component ablation studies.

4.1 Datasets

The downstream detector is trained on the ASVspoof 2019 LA (Todisco et al., 2019) training set, with its development set used for model selection. The LA dataset contains bona fide speech and spoofed speech generated using text-to-speech (TTS) and voice conversion (VC), with source recordings drawn from the VCTK corpus (Yamagishi et al., 2019) . Its relatively clean recording conditions provide limited coverage of the acoustic and transmission variations encountered in practical applications. Cross-domain performance is evaluated on ASVspoof 2021 LA, ASVspoof 2021 DF (Yamagishi et al., 2021) , and ASVspoof 5 (Wang et al., 2024). ASVspoof 2021 LA contains 148,176 trials and introduces variations in encoding and transmission conditions relative to ASVspoof 2019 LA. The ASVspoof 2021 DF evaluation set includes compressed bona fide and synthesized speech to reflect conditions associated with online media distribution. ASVspoof 5 incorporates crowdsourced speech and a broader range of generation and attack methods, providing a further test of generalization beyond the training domain.

Additional evaluations cover real-world recordings, Chinese speech, and other out-of-domain conditions. We use In-the-Wild (ITW) (Müller et al., 2022) , ADD 2022 Track 1 and Track 3 (Yi et al., 2022) , LibriSeVoc (Sun et al., 2023) , Fake or Real (FoR) (Reimao & Tzerpos, 2019) , and DFADD (Du et al., 2024) to examine performance under changes in recording conditions, language, and speech generation methods.

4.2 Experimental Settings

Input preprocessing produces fixed-length waveforms for the XLS-R 300M frontend. Speech is resampled to 16 kHz, and the input length is fixed at 66,800 samples, corresponding to approximately 4.175 s. Longer utterances are truncated to their first 66,800 samples, whereas shorter utterances are periodically repeated and then trimmed to the required length. RawBoost (Tak et al., 2022) is applied online during training to introduce waveform perturbations associated with nonlinear distortion, impulsive interference, and additive noise. It is disabled for the development and evaluation sets to avoid augmentation-induced variation during scoring. The model is optimized using class-weighted cross-entropy. Spoofed and bona fide speech are assigned labels 0 and 1, with class weights of 0.1

Table 1: Cross-dataset detection performance in EER (%). Lower is better. Bold and underlined values indicate the best and second-best results in each column, respectively. Results marked with * are taken from Speech DF Arena (Dowerah et al., 2026), while all other results are obtained from our reproductions. All models are trained on the ASVspoof 2019 LA training set.

Publication	System	21LA	21DF	ASVspoof 5
ICASSP 2021	RawNet2*	9.48	22.38	40.67
ICASSP 2022	AASIST*	11.46	21.07	35.53
INTERSPEECH 2023	XLSR-Conformer	1.55	2.58	16.51
ICASSP 2024	Wav2Vec2-AASIST2*	**0.82**	6.63	16.24
ACM MM 2024	XLSR-SLS	2.86	2.24	18.76
INTERSPEECH 2024	XLSR-TCM	1.83	2.00	16.47
SPL 2025	XLSR-Mamba	0.93	1.88	<u>14.40</u>
TIFS 2025	Nes2NetX	2.49	<u>1.77</u>	17.94
Findings of ACL 2026	XLSR-MamBo	<u>0.88</u>	2.76	14.83
ours	AGDT	2.48	**1.56**	**13.57**

and 0.9, respectively. We use Adam with a learning rate of 10^{-6}, a weight decay of 10^{-4}, and a batch size of 20. Training is conducted on a single NVIDIA GeForce RTX 4090 GPU. The primary evaluation metric is the equal error rate (EER). It corresponds to the operating point at which the false acceptance and false rejection rates are equal. All EER values are reported as percentages, with lower values indicating better detection performance.

4.3 Comparison with State-of-the-Art Systems

The comparison includes RawNet2 (Tak et al., 2021) , AASIST (Jung et al., 2022) , Wav2Vec2-AASIST2 (Baevski et al., 2020; Zhang et al., 2024b) , XLSR-Conformer (Rosello et al., 2023) , XLSR-SLS (Zhang et al., 2024a) , XLSR-TCM (Truong et al., 2024) , XLSR-Mamba (Xiao & Das, 2025), Nes2NetX (Liu et al., 2025) , and XLSR-MamBo (Ng et al., 2026) . Table 1 presents their results alongside AGDT on ASVspoof 2021 LA, ASVspoof 2021 DF, and ASVspoof 5. AGDT achieves EERs of 2.48%, 1.56%, and 13.57%, respectively. Although it does not achieve the lowest EER on ASVspoof 2021 LA, it obtains the lowest reported EERs on ASVspoof 2021 DF and ASVspoof 5 among the systems listed. These results indicate competitive generalization across the evaluated conditions.

Evaluations on additional datasets reveal both the strengths and limitations of cross-domain transfer. As shown in Table 2, AGDT achieves EERs of 5.83%, 30.24%, 5.83%, and 3.54% on ITW, ADD 2022 Track 1, DFADD, and FoR, respectively, which are the lowest values among the listed systems. On ADD 2022 Track 3 and LibriSeVoc, it does not achieve the best results. These datasets differ from the training domain in language, recording conditions, or generation methods. The results therefore support the effectiveness of the proposed framework across several domain shifts, while the relatively high error rates on the ADD datasets indicate substantial room for improvement.

Feature visualization offers a qualitative view of the representations learned by the two models. As shown in Figure 3, the XLSR-Conformer baseline produces partially overlapping distributions of bona fide and spoofed samples, especially near the class boundary. In comparison, AGDT forms more compact and distinguishable class clusters, despite a small amount of residual overlap. This pattern qualitatively suggests that AGDT learns more discriminative representations on the visualized samples.

The condition-specific results show that detection difficulty varies across vocoder types and codec conditions. As illustrated in Figure 4, neural autoregressive vocoders consistently produce higher EERs than traditional, waveform-concatenation, and neural non-autoregressive systems across all nine codec conditions. In particular, the neural autoregressive category yields a pooled EER of 2.84%, whereas the overall pooled EER is 1.56%. This result indicates that autoregressive generation presents a greater challenge to AGDT, possibly because its sequential generation process produces fewer easily detectable local artifacts or introduces different temporal dependencies. Further analysis is therefore needed to determine which characteristics of autoregressive speech contribute to these errors.

Table 2: Performance on additional out-of-domain datasets in EER (%). Bold and underlined values indicate the best and second-best results in each column, respectively. Results marked with * are taken from Speech DF Arena (Dowerah et al., 2026), while all other results are obtained from our reproductions. All models are trained on the ASVspoof 2019 LA training set.

System	ITW	ADD22T1	ADD22T3	DFADD	LibriSeVoc	FoR
RawNet2*	49.19	50.35	32.29	39.07	48.17	65.68
AASIST*	43.01	47.91	33.18	41.86	38.02	21.64
XLSR-Conformer	8.42	31.40	34.88	18.42	5.23	7.90
Wav2Vec2-AASIST2*	11.20	<u>31.05</u>	<u>16.52</u>	11.93	11.21	7.47
XLSR-SLS	7.45	34.37	17.79	<u>7.54</u>	<u>2.00</u>	<u>5.08</u>
XLSR-TCM	6.40	33.39	21.76	8.89	5.83	15.73
XLSR-Mamba	6.71	34.23	19.37	10.70	2.23	6.71
Nes2NetX	<u>6.15</u>	34.59	25.64	11.15	2.81	6.01
XLSR-MamBo	9.16	31.82	**15.60**	11.16	**1.70**	9.75
AGDT (ours)	**5.83**	**30.24**	19.14	**5.83**	2.19	**3.54**



(a) XLSR-Conformer (b) AGDT

Figure 3: t-SNE visualization of features extracted before the final classification layer on ASVspoof 2021 DF. Both models use the same 20,000 sampled utterances, comprising 10,000 bona fide and 10,000 spoofed samples, with identical dimensionality-reduction settings. Blue and red points denote bona fide and spoofed speech, respectively.

4.4 Ablation Studies

We conduct ablation experiments to examine two aspects of AGDT: the contributions of its individual components and the placement of multi-scale temporal modeling. All variants use the same XLS-R front end, ASVspoof 2019 LA training set, and optimization and evaluation protocols. The results are presented in Table 3.

Contribution of individual components. Data augmentation provides consistent improvements across all four evaluation sets. In particular, removing it increases the EER on ITW from 5.83% to 9.77%, indicating that exposure to acoustic perturbations during training is important for robustness to unseen recording and channel conditions. Removing SLS improves the result on 21LA but degrades performance on 21DF, ITW, and ASVspoof 5. This result suggests that aggregating representations from different XLS-R depths may introduce redundant information under conditions closely related to LA, while preserving complementary cues that become useful under larger distribution shifts. A similar pattern is observed for ASP. The variant without ASP obtains the lowest EER on 21LA, whereas its performance degrades on 21DF and ITW. Modeling both the weighted

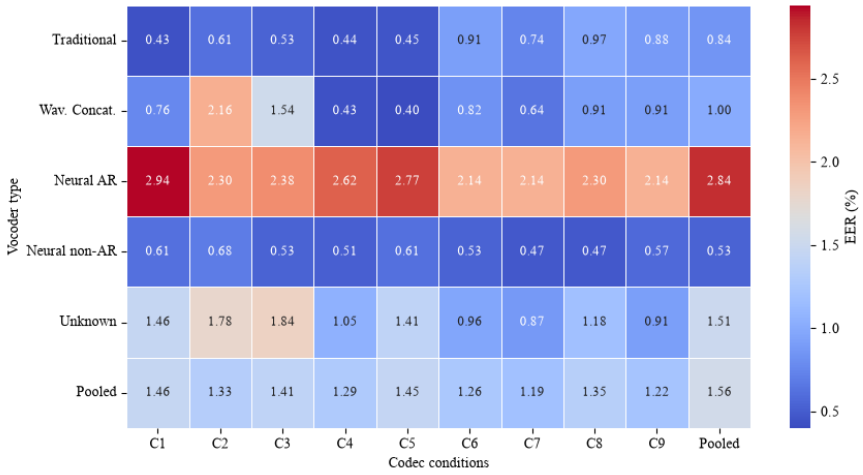

Figure 4: AGDT's EERs (%) across codec conditions and vocoder categories on ASVspoof 2021 DF. Wav. Concat. denotes waveform concatenation, AR denotes autoregressive, and Unknown indicates unavailable or incomplete vocoder metadata. C1–C9 represent different codec conditions. Pooled denotes the EER computed over all trials along the corresponding dimension. Colors range from blue to red as the EER increases.

Table 3: Ablation results for model components and temporal-modeling placement in terms of EER (%). Lower is better. Bold values indicate the lowest EER among all variants in each column.

Ablation variant	21LA	21DF	ITW	ASVspoof 5
Component removal				
w/o DA	3.90	3.03	9.77	16.51
w/o SLS	1.70	3.08	7.34	15.84
w/o ASP	**1.36**	2.34	8.28	13.72
Temporal modeling and placement				
Standard Conformer	1.55	2.58	8.42	16.51
Internal multi-kernel	1.90	1.63	7.69	14.15
Pre-MSTCM only	1.66	3.26	12.15	15.43
Post-MSTCM only	2.20	2.23	6.19	15.97
Dual-stage AGDT	2.48	**1.56**	**5.83**	**13.57**

mean and temporal variation therefore appears more useful for heterogeneous and out-of-domain spoofing conditions than for the relatively constrained LA setting.

Effect of temporal modeling and module placement. To determine whether the improvement can be explained by multi-kernel convolution alone, we construct an internal multi-kernel variant. It removes the external pre- and post-MSTCM convolution branches, retains the pre-encoder channel–temporal gate, and replaces the single depthwise convolution in each Conformer block with three parallel branches using kernel sizes $\{3, 5, 7\}$. The branch outputs are fused by arithmetic averaging. Compared with the standard Conformer, the internal multi-kernel variant reduces the EER from 2.58% to 1.63% on 21DF and also improves ITW and ASVspoof 5. These results show that multiple local receptive fields are beneficial even when they are introduced inside the Conformer encoder. However, the internal design remains inferior to dual-stage AGDT on 21DF, ITW, and ASVspoof 5, suggesting that the location of multi-scale processing also affects cross-domain robustness. The single-stage variants further demonstrate that the two external MSTCMs play different roles. The post-MSTCM-only variant performs substantially better than the pre-MSTCM-only variant on ITW, suggesting that multi-scale refinement after contextual encoding helps recover local evidence that may have been mixed during global information exchange. Nevertheless, neither single-stage variant matches the complete model across the three more strongly shifted evaluation sets. The full design obtains the lowest EERs on 21DF, ITW, and ASVspoof 5, supporting the complementary use of early artifact enhancement and subsequent contextual refinement. The complete model does not achieve

Table 4: Temporal kernel configuration ablations in EER (%).

Kernel sizes	21LA	21DF	ITW
$\{3, 5\}$	**1.61**	1.76	7.30
$\{3, 5, 7\}$	2.48	**1.56**	**5.83**
$\{3, 7, 11\}$	1.82	1.77	6.80
$\{3, 7, 11, 15\}$	2.25	2.42	5.93

Table 5: Encoder architecture and depth ablations in EER (%).

Encoder configuration	21LA	21DF	ITW
Transformer ($N = 2$)	**1.53**	2.51	6.51
Transformer ($N = 4$)	2.45	4.47	10.15
Transformer ($N = 6$)	1.69	2.64	9.16
Conformer ($N = 2$)	1.66	2.35	6.60
Conformer ($N = 4$)	2.48	**1.56**	**5.83**
Conformer ($N = 6$)	1.69	2.08	7.08

the best result on 21LA, where several simplified variants perform better. The components therefore produce a domain-dependent trade-off rather than a uniform improvement on every dataset.

Effect of temporal kernel configurations. Table 4 indicates that increasing kernel size or spacing does not produce monotonic improvements. Small kernels are sensitive to short-duration artifacts, whereas moderately larger kernels capture patterns across neighboring frames. The adjacent $\{3, 5, 7\}$ configuration balances these receptive fields and performs best on 21DF and ITW. Wider or more sparsely distributed kernels may smooth subtle artifacts or duplicate the long-range modeling already provided by the Conformer. Although $\{3, 5\}$ performs best on 21LA, we select $\{3, 5, 7\}$ for its more consistent cross-domain performance.

Effect of encoder architecture and depth. The encoder ablation demonstrates the benefit of combining local and global modeling. Transformer self-attention captures long-range dependencies, while the Conformer additionally preserves neighboring acoustic patterns through convolution. At four layers, replacing the Transformer with the Conformer reduces the EER from 4.47% to 1.56% on 21DF and from 10.15% to 5.83% on ITW. Increasing the Conformer depth further does not improve cross-domain performance, possibly because repeated contextual interaction smooths weak local artifacts or encourages specialization to the training distribution. We therefore select the four-layer Conformer as a balance between local-detail preservation and global context modeling.

5 Conclusion

We presented AGDT, which aggregates evidence across network depth, temporal scale, and utterance-level statistics. Non-competitive sigmoid gates fuse multiple XLS-R layers, while asymmetric dual-stage temporal convolution provides gated local enhancement before Conformer encoding and ungated refinement afterward. Attentive statistics pooling captures feature means and temporal variation. Experiments across multiple datasets show competitive cross-domain performance, and ablations confirm the complementary, domain-dependent effects of these components.

Several limitations remain. Training uses only ASVspoof 2019 LA, and performance on ASVspoof 2021 LA and ADD 2022 shows room for improvement under channel, language, and generation shifts. The roles of the two temporal stages also require analysis of intermediate representations, gate responses, and frame-level saliency. Future work will examine domain-invariant layer selection, calibration for unseen generators, and local artifacts to improve cross-language and acoustic robustness.

We use public datasets and official evaluation protocols. To support reproducibility, we plan to release the training code, configurations, preprocessing and evaluation scripts, random seeds, checkpoint-selection criteria, and software dependencies.

AI USE STATEMENT

In this work, we used generative AI tools for translation and language polishing. We have not used generative AI tools for the generation of synthetic data sets, formulation of mathematical claims, assistance in the writing of proofs, or the design of research methodology.

Specifically, we used AI to assist with:

Translation: The original manuscript was translated into English with the assistance of AI.

Language polishing: The tool was used to improve grammar, sentence clarity, and overall readability of the author-written text.

Literature synthesis : The tool was used to summarize and analyze existing literature to help the authors understand the current research landscape.

REFERENCES

Arun Babu et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

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*, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

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*, pp. 1–9, 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*, pp. 921–928, 2024.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html.

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. *arXiv preprint arXiv:2106.07447*, 2021. URL https://arxiv.org/abs/2106.07447.

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6367–6371, 2022. doi: 10.1109/ICASSP43922.2022.9747766.

Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Interspeech 2017*, pp. 82–86, 2017. doi: 10.21437/Interspeech.2017-360. URL https://www.isca-archive.org/interspeech_2017/lavrentyeva17_interspeech.html.

Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Replay and synthetic speech detection with Res2Net architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. URL https://arxiv.org/abs/2010.15006.

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, 2025. doi: 10.1109/TIFS.2025.3626963.

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

Kwok-Ho Ng, Tingting Song, Yongdong Wu, and Zhihua Xia. XLSR-MamBo: Scaling the hybrid mamba-attention backbone for audio deepfake detection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31450–31462, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1573. URL https://aclanthology.org/2026.findings-acl.1573/.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. Attentive statistics pooling for deep speaker embedding. In *Interspeech 2018*, 2018. URL https://www.isca-archive.org/interspeech_2018/okabe18_interspeech.html.

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

Darshan Prabhu, Yifan Peng, Preethi Jyothi, and Shinji Watanabe. MultiConvformer: Extending conformer with multiple convolution kernels. In *Proceedings of Interspeech*, pp. 232–236, 2024. doi: 10.21437/Interspeech.2024-2384.

Ricardo Reimao and Vassilios Tzerpos. FoR: A dataset for synthetic speech detection. In *Proceedings of the International Conference on Speech Technology and Human-Computer Dialogue*, 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. URL https://www.isca-archive.org/interspeech_2023/rosello23_interspeech.html.

Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. In *Interspeech 2015*, 2015. URL https://www.isca-archive.org/interspeech_2015/sahidullah15_interspeech.html.

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

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6369–6373, 2021. doi: 10.1109/ICASSP39728.2021.9414234.

Hemlata Tak, Madhu R. Kamble, Jose Patino, Massimiliano Todisco, and Nicholas W. D. Evans. RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6382–6386, 2022. doi: 10.1109/ICASSP43922.2022.9746213.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 2017. doi: 10.1016/j.csl.2017.01.001. URL https://www.eurecom.fr/en/publication/5146.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of Interspeech*, pp. 1008–1012, 2019. doi: 10.21437/Interspeech.2019-2249.

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. URL https://www.isca-archive.org/interspeech_2024/truong24b_interspeech.html.

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 *Proceedings of the Automatic Speaker Verification Spoofing Countermeasures Workshop*, pp. 1–8, 2024. doi: 10.21437/ASVSPOOF.2024-1.

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.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019. URL https://doi.org/10.7488/ds/2645.

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 *Proceedings of the ASVspoof Challenge Workshop*, pp. 47–54, 2021. doi: 10.21437/ASVSPOOF.2021-8.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, Haizhou Li, et al. ADD 2022: The first audio deep synthesis detection challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 9216–9220, 2022.

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*, 2024a. doi: 10.1145/3664647.3681345. URL https://openreview.net/pdf?id=acJMIXJg2u.

Yuxiang Zhang, Jingze Lu, Zengqiang Shang, Wenchao Wang, and Pengyuan Zhang. Improving short utterance anti-spoofing with AASIST2. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11636–11640. IEEE, 2024b. doi: 10.1109/ICASSP48485.2024.10448049.