Odyssey 2026
23-26 June 2026, Lisbon, Portugal

Can SSL Frontend Generalize to All-Type Audio Spoofing?

Arnab Das[1,4], *Yassine El Kheir*[1,4], *Fabian Ritter Guttierez*[2], *Tim Polzehl*[1,4], *Sebastian Möller*[1,3]

[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]Nanyang Technological University, Singapore
[3]Technical University of Berlin, Germany
[4]Gretchen AI, Germany

arnab.das@dfki.de

Abstract

Advances in generative AI have increased the exposure of automatic speaker verification systems to synthetic audio across diverse modalities. Recent spoofing detection approaches typically rely on self-supervised learning (SSL) frontends followed by neural classifiers. While such representations achieve strong performance within specific modalities, their ability to generalise across diverse deepfake types has not been systematically investigated. In this work, we benchmark SSL frontends for cross-modal audio deepfake detection spanning speech, environmental sounds, music, and singing voice. Our experiments reveal that no single SSL model generalises well across all modalities and that pre-training data distribution strongly influences detection performance. We show that a dual-frontend architecture combining a speech-specific and a general-audio SSL model yields the strongest cross-modal performance, reducing average equal error rate compared to the best single-frontend baseline.
Index Terms: deepfake detection, spoofing countermeasure, SSL frontend, cross-modal generalization

1. Introduction

The rapid advancements of generative AI are increasingly blurring the distinction between bonafide and synthetic audio, while also lowering the barrier to access such technologies, thereby exposing automatic speaker verification and other audio-dependent systems, whose reliability hinges on signal integrity, to significant vulnerabilities [1] posed by spoofing. Early research in developing a spoofing countermeasure has predominantly focused on synthetic speech. However, the horizon of audio deepfakes has quietly grown far broader in scope and expanded beyond speech to music[2], singing voice[3], and even to environmental sound deepfakes[4]. This poses risks to streaming platforms, amplifies the potential for disinformation campaigns, and creates unfair competition for artists. This broadening of scope introduces a critical challenge, as existing detection systems, which are often optimized for a single modality, may fail to generalise when confronted with heterogeneous and previously unseen attack types, especially in real-world scenarios.

Self-supervised learning has become the dominant paradigm for audio and speech representation, offering rich, transferable features learned from vast unlabeled corpora. Recent spoofing detection systems commonly adopt the recipe of extracting rich representations using SSL-based frontends followed by neural backends for classification. These approaches demonstrate strong performance within a specific modality; their cross-modal generalization capabilities remain underexplored. In particular, SSL models inherit strong inductive biases from their pre-training data, which can limit their effectiveness when applied to modalities outside their training distribution [5, 6].

In this work, we critically try to find the answer to a question: how well do current SSL frontends actually generalize across deepfake categories - spanning speech, music, environmental sound, and singing voice. More specifically, we aim to investigate how the pre-training regime of SSL models influences their ability to generalize across modalities for deepfake detection. In particular, we examine whether SSL models trained on speech-only or audio-only data can match the cross-modal performance of universal SSL frontends pre-trained on diverse, heterogeneous audio sources encompassing both speech and non-speech signals.

To answer these questions:

- We provide a systematic evaluation of SSL frontends for all-type audio deepfake detection across speech, music, sound, and singing modalities.
- We compare the SSL frontends pre-trained with unimodal data with universal SSL frontends, which are pre-trained with heterogeneous data.
- We analyze the generalization performance gap between models fine-tuned on single-modality deepfake data and those fine-tuned jointly on all modalities.
- We explore a series of architectural strategies that aim to improve cross-modal generalization. We investigate dual-frontend architectures that combine complementary SSL models to capture both speech-specific and general audio characteristics.
- We explore a parameter-efficient dual-frontend distillation framework that leverages knowledge transfer from a full audio SSL model to a lightweight counterpart.
- We further examine model performance under challenging mixed-audio conditions, where speech and environmental sounds are combined.

2. Background

2.1. SSL models

Over the years, self-supervised learning (SSL) has been widely used for speech deepfake detection due to SSL models ability for learning rich representations from large-scale unlabeled data. By leveraging pretext tasks such as contrastive prediction or masked reconstruction, models capture both local and global structures without reliance on manual annotations. These representations have shown strong transferability across downstream tasks, including speech deepfake detection. In this work, we employ multiple SSL models as front-end feature extractors, whose representations are subsequently provided as input to a neural detector backend. **XLS-R** [7], a large scale model based on wav2vec 2.0 [8] trained with publicly available multi-lingual speech data. Representation learned by this model demonstrates significant performance improvement for machine translation, speech recognition, and language identification tasks. **MERT** [9], a SSL model for music understanding, trained via masked acoustic modeling and pseudo-label prediction to encode tonal and pitch-related musical characteristics within its

10.21437/Odyssey.2026-41

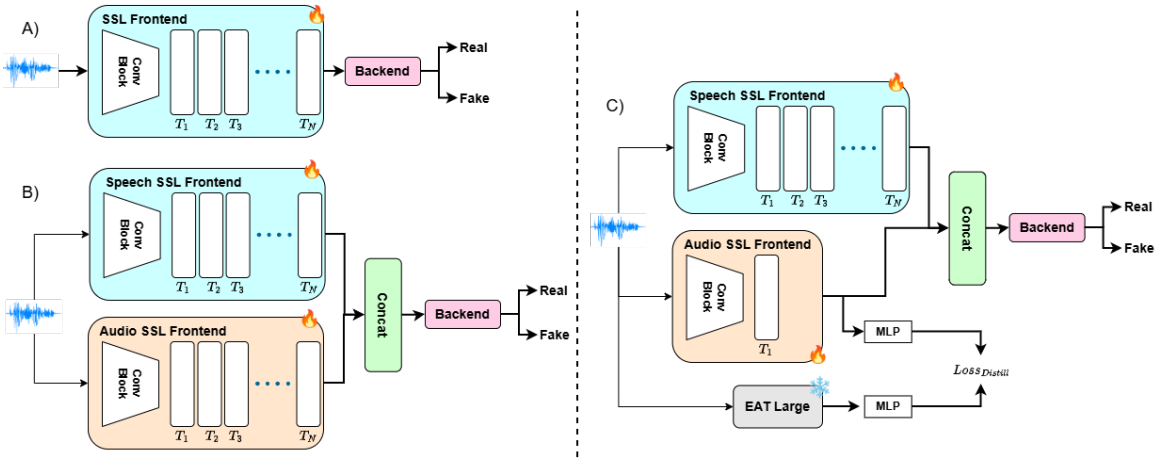

Figure 1: *High-level schematic of the deepfake detection model architectures with SSL: A) single-frontend, B) dual-frontend, and C) dual-frontend with distillation.*

learned representations. **EAT**[10], which extends the bootstrap SSL technique to generic audio representation learning with a novel utterance–frame objective and optimized masking strategy, achieving state-of-the-art (SOTA) performance on audio-related tasks such as audio classification while significantly improving pre-training efficiency. **USAD**[11], a unified audio representation framework for speech, sound, and music, that distills knowledge from domain-specific SSL models into a single encoder, achieving competitive performance across domains on benchmarks such as SUPERB and HEAR. **SPEAR**[12], another recent self-supervised framework that distills complementary knowledge from speech-specific and general-audio teachers into a unified model using vector-quantized tokens and masked prediction, achieving performance improvement on both domains. **Dasheng Tokenizer**[13], a continuous audio tokenizer that injects acoustic information into frozen semantic representations, enabling unified audio understanding and generation while outperforming prior codecs and encoders across diverse audio tasks, including audio classification and speech understanding.

2.2. Audio deepfake detection

The development of countermeasures for speech deepfakes has been inspired by benchmark challenges such as ASVspoof [14] and ADD [15]. Current speech deepfake detection systems predominantly follow a common architecture, leveraging SSL-based frontends coupled with neural backend classifiers, which have demonstrated notably strong performance in detecting synthetic speech [16, 17, 18]. Over the years, numerous speech deepfake datasets have been introduced, encompassing diverse spoofing scenarios such as multilingual attacks (MLAAD [19]), partial spoofing [20], and codec-based synthesis (CodecFake [21]). [22] leverages the foundation SSL model WavLM as a frontend and incorporates a mixture-of-experts strategy along with low-rank adaptation to achieve performance gains across multiple such speech deepfake datasets.

Beyond speech, rapid advances in text-to-audio (TTA) and text-to-music (TTM) synthesis have made synthetic music, general audio, and singing voice increasingly realistic, raising concerns about their potential for misuse [23]. The research community has responded with dedicated datasets and detection methods across these modalities. The work [24] provides an early investigation into music deepfake detection, examining both identification and interpretability of synthetic music. [25] builds on [24] findings and explores a hybrid approach combining MERT with the AASIST backend [26] for music deepfake detection. [18] proposes the architecture Nes2Net-X for singing voice deepfake detection based on WavLM representations. For environmental sounds, the recent ESDD challenge [27] saw that most participating systems relied on SSL-based frontends such as EAT, BEATs [28], and SSLAM [29].

Despite this progress in individual modalities, detection methods that generalise across all deepfake types remain underexplored, and the choice of SSL frontend for a unified framework is still unclear. Two recent studies take initial steps in this direction. [25] proposes a co-training approach with wavelet-based prompt tokens to improve the frequency sensitivity of SSL representations across multiple deepfake types, though their evaluation covers only three frontends, namely XLS-R, MERT, and WavLM. [30] evaluates audio large language models such as Qwen2-Audio-Chat and Qwen2.5-Omni for all-type deepfake detection with an emphasis on interpretability, however their evaluation is limited to in-domain settings and the results indicate poor generalisation for synthetic sound detection.

3. Methodologies

In this work, we explore three different architectural paradigms of using SSL as a frontend.

3.1. Single frontend

Under the single frontend paradigm, as illustrated in Figure 1A), we follow the de facto architecture commonly adopted in recent spoofing countermeasure systems. The input audio is first processed by a single SSL-based frontend, which produces a sequence of latent representations of size $Seq \times dim$, where Seq denotes the temporal sequence length and dim represents the embedding dimension. These embeddings are subsequently passed to a neural backend for classification. Importantly, the frontend is not kept frozen but is fully fine-tuned during training to better adapt to the task. For the backend, we employ BiCrossMamba-ST[17], a dual-branch spectro-temporal architecture that models spectral and temporal dynamics separately using BiMamba blocks, while enabling rich interaction between them through a mutual cross-attention mechanism.

3.2. Dual frontend

Under this setting, we employ two distinct frontends, as shown in Figure 1B). The first frontend is XLS-R, a speech-focused SSL model that operates directly on raw audio and produces frame-level

representations. The second frontend is a non-speech audio SSL model; in this work, we experiment with EAT Large[1], fine-tuned on the AudioSet-2M dataset [31], and Dasheng Tokenizer. Unlike XLS-R, EAT processes mel-spectrogram inputs and generates patch-wise representations, which are then flattened into a sequence of embeddings for a given audio sample. Again, both the frontends are fully finetuned as part of the training. Finally, the representations from both frontends are concatenated and provided as input to the backend model.

3.3. Dual frontend with distillation

In this configuration, we again adopt a dual-frontend design consistent with Sub-section 3.2 as depicted in Figure 1C). The speech-oriented SSL model XLS-R is retained unchanged as the first frontend. However, for the second (audio-focused) SSL frontend, we intentionally simplify the architecture by retaining only a single transformer layer and discarding the remaining deeper layers. This design choice reduces model complexity and the number of parameters while encouraging the network to learn efficient intermediate representations. To compensate for the reduced capacity, we employ a joint training strategy that combines the primary classification objective with a knowledge distillation objective. Specifically, the lightweight audio frontend is trained to align its representations with those of a fully pre-trained and frozen EAT Large, effectively transferring rich acoustic knowledge from the teacher model to the student. This setup enables the model to maintain strong representational power despite the architectural simplification. To compute the distillation loss, both the student features from the single-layer audio SSL model and the teacher features from the frozen full EAT Large are first projected into a shared embedding space. The alignment is then enforced using a combination of mean squared error (MSE) loss and cosine similarity loss, ensuring not only point-wise similarity but also preservation of the underlying geometric structure between student and teacher representations. The distillation loss objection is presented in Eqn.1, where $\mathbf{X}_s$ and $\mathbf{X}_t$ denote the normalized projected student and teacher representations, respectively, and λ_1,λ_2 are weighting coefficients controlling the contribution of each term.

$$\mathcal{L}_{\text{Distill}} = \lambda_1 ||\mathbf{X}_s - \mathbf{X}_t||_2^2 + \lambda_2 \left(1 - \frac{\mathbf{X}_s \cdot \mathbf{X}_t}{||\mathbf{X}_s||_2 ||\mathbf{X}_t||_2}\right) \quad (1)$$

For the binary classification objective, we adopt the A-Softmax [32] loss across all three experimental settings.

Table 1: *Datasets used for training, validation, and testing.*

Datasets	Train	Valid	Test
LA19	✓	✓	✓
Fakemusic	✓	✓	✓
EnvSDD	✓	✓	✓
CtrSVDD	✓	✓	✓
ItW	×	×	✓
Codecfake A3	×	×	✓
ESDD2 Mix	×	×	✓

[1] https://huggingface.co/worstchan/EAT-large_epoch20_finetune_AS2M

Table 2: *EER(%) comparison of detection models with different SSL frontends under single-dataset training (Train.: training set; Music, ESDD, SVDD denote FakeMusic, EnvSDD, and CtrSVDD).*

Frontends	Train.	LA19	Music	ESDD	SVDD
XLS-R	LA19	**0.43**	47.56	48.80	31.73
MERT		2.3	47.0	48.1	46.05
EAT		1.1	49.97	29.72	46.79
XLS-R	Music	39.45	28.07	52.20	50.27
MERT		60.26	3.19	39.82	52.57
EAT		47.06	**0.29**	29.50	53.23
XLS-R	ESDD	68.74	26.21	42.62	43.45
MERT		46.08	8.64	29.10	53.31
EAT		5.87	5.10	**2.9**	31.27
XLS-R	SVDD	4.2	33.52	49.0	**7.21**
MERT		38.44	33.9	49.7	27
EAT		14.18	31.98	36.10	19.4

Table 3: *Out-of-domain performance (EER(%)) comparison of detection models with different SSL frontends under single-dataset training.*

Frontends	Trained on	ItW	Codecfake A3
XLS-R	LA19	9.77	40.65
MERT		34.27	50.59
EAT		20.76	33.21
XLS-R	Fakemusic	45.86	44.07
MERT		33.46	38.77
EAT		12.66	52.23
XLS-R	EnvSDD	39.62	24.03
MERT		22.67	12.36
EAT		13.18	**3.59**
XLS-R	CtrSVDD	**5.17**	26.39
MERT		34.34	30.25
EAT		24.53	22.33

4. Experiment setup

4.1. Datasets

As our objective is to systematically evaluate the effectiveness of deepfake detection models with diverse SSL frontends across multiple deepfake modalities, careful selection of training and evaluation datasets is essential. The datasets used in this work are summarized in Table 1. **Speech:** We utilize the ASVspoof 2019 LA dataset [33], which includes spoofed samples generated by 17 distinct TTS and voice conversion systems, with disjoint attack conditions between training and evaluation splits. **Music:** For the music domain, we employ the FakeMusicCaps dataset [34], comprising both real and synthetic music generated by five state-of-the-art text-to-music (TTM) models conditioned on textual captions. **Sound:** We employ the EnvSDD corpus [35], a large-scale benchmark for environmental sound deepfake detection, comprising synthetic samples generated using both text-to-audio and audio-to-audio models. **Singing:** We use the CtrSVDD dataset [36], which comprises large-scale, controlled singing voice deepfakes generated by 14 different systems, created upon Mandarin and Japanese singing corpora.

In addition to these datasets, we further evaluate the models on the In-the-Wild (ItW) dataset [37] for real-world speech scenarios,

Table 4: *EER(%) comparison of detection models with different SSL frontends under combined dataset training for single- and dual-frontend architecture settings. Parameter counts are for the frontend only (backend excluded).*

Models	Param (M)	Trained on	LA19	Fakemusic	EnvSDD	SVDD	ItW	Codec A3	Avg.
XLS-R	317	LA19 + Fakemusic + EnvSDD + CtrSVDD	**0.78**	18.73	43.6	**8.33**	**3.47**	19.02	15.66
MERT	95		2.23	1.66	24.8	26.53	19.81	14.84	14.98
EAT	300		1.14	**0.17**	**2.8**	12.26	12.74	**1.28**	**5.07**
USAD	330		1.47	2.62	14.70	12.52	24.26	7.69	10.54
SPEAR Large	327		0.98	**0.18**	10.60	14.05	23.76	5.08	9.11
SPEAR XLarge	597		**0.86**	**0.18**	**10.5**	**8.24**	**11.42**	**4.91**	**6.02**
Dasheng Tok.	630		5.44	0.56	21.60	25.27	72.10	23.95	24.82
QWEN 3 Omni	650		5.02	1.31	18.42	19.62	8.18	15.37	11.32
XLS-R + Dasheng Tok.	947		**0.16**	0.02	17.2	8.44	7.5	14.29	7.94
XLS-R + EAT - 24 layers	617		0.52	**0**	**2.8**	10.21	6.68	**0.84**	3.51
XLS-R + EAT - 18 layers	463		0.74	0.02	3.32	**8.16**	4.64	1.26	**3.02**
XLS-R + EAT - 12 layers	307		0.9	**0**	4.0	8.74	8.41	1.4	3.91
XLS-R + EAT - 6 layers	154		0.67	0.01	6.28	12.6	16.46	4.08	6.68
XLS-R + EAT - 3 layers	78		1.86	0.38	14.02	22.93	61.89	7.68	18.13
XLS-R(24) + EAT(1) - Distill 24	338		0.37	0.35	24.7	9.9	4.34	14.42	9.01
XLS-R(24) + EAT(1) - Distill 18	338		0.37	0.75	21.12	8.91	**3.75**	8.78	7.28

which comprises noisy real and fake samples collected from the internet. We also assess out-of-domain generalization on the CodecFake A3 subset [21], focusing on fake sound detection under distribution shifts. Furthermore, to test the limitations of these models, we evaluate them on the highly challenging mixed-audio development subset of the ESDD2 [38], where speech and sound, either real or synthetic, are mixed. In this setting, all combinations (*spoof_bonafide, bonafide_spoof, spoof_spoof*) are labeled as fake except when both the speech and the sound components are bonafide.

4.2. Implementation details

In this study, all models are trained on 4-second audio segments, where shorter inputs are handled using repetition padding. Optimization is performed using AdamW with a learning rate of 10^{-6}. Training is conducted on NVIDIA H100 GPUs with a batch size of 32 for up to 50 epochs, employing early stopping based on the validation set equal error rate (EER). We set the distillation loss coefficients λ_1 and λ_2 (cf. Sub-section 3.3) to 0.1 based on hyperparameter tuning.

5. Results & discussions

5.1. Single modality training

To evaluate cross-modality generalization to unseen deepfake types, we first train the models on single modalities and test them across all datasets and the results are presented in Table 2. When trained solely on the speech dataset LA19, the XLS-R frontend achieves the lowest EER of 0.43%, followed by EAT at 1.1%. However, both fail to generalize to other modalities. Conversely, when trained on the music deepfake dataset FakeMusic, the EAT model attains the best performance (0.29%), followed by MERT at 3.19%, while XLS-R performs poorly and achieves only 28.07% EER. Consistent with the previous setting, all models exhibit limited generalization to unseen modalities. For the training on the environmental sound dataset EnvSDD, the model with the EAT frontend achieves the lowest EER of 2.9%. Notably, this model also exhibits some cross-modal generalization, attaining EERs of 5.87% and 5.10% on LA19 and FakeMusic, respectively. In contrast, models with other frontends not only fail to generalize but also perform poorly even within their training modality. When trained on the singing voice deepfake dataset CtrSVDD, the XLS-R frontend achieves the best performance (7.21%) and demonstrates reasonable generalization to LA19 with an EER of 4.2%, while other models lag significantly behind. Overall, these results indicate that no single SSL frontend can effectively generalize across all deepfake modalities when trained on a single domain. Instead, XLS-R is better suited for speech and singing voice tasks, whereas EAT performs more effectively for music and environmental sound deepfakes — likely reflecting the strong dependency towards domain-specific knowledge acquired during pre-training of these SSL frontends. These results are consistent with the findings from [25, 30].

Additionally, we evaluate out-of-domain generalization performance using the ItW dataset for speech and the CodecFake A3 subset for sound, with results summarized in Table 3. The XLS-R model trained on CtrSVDD achieves the best EER on ItW (5.17%), while for the A3 subset, the EAT model trained on EnvSDD attains the lowest EER (3.59%).

5.2. Combined modality training

To evaluate how effectively models with different SSL frontends can capture deepfake artifacts across modalities, we adopt a unified training strategy by training the models jointly on all modalities by combining all four datasets and the results are summarized in the first 3 rows of Table 4. The results immediately reveal interesting insights. The model with the XLS-R frontend achieves the best performance on LA19, CtrSVDD, and ItW (0.78%, 8.33%, and 3.47%, respectively), whereas the EAT frontend performs best on FakeMusic, EnvSDD, and CodecFake A3 (0.17%, 2.8%, and 1.28%). These results highlight that the inductive biases acquired during SSL pre-training strongly influence deepfake detection performance. In particular, XLS-R, being trained on speech data, excels in speech-related modalities but struggles with non-speech domains, while EAT, trained on spectrogram-based audio data, performs significantly better on music and environmental sounds. This complementary behavior suggests that combining such SSL models could be beneficial for general-purpose deepfake detection, naturally motivat-

ing the question of whether SSL models pre-trained jointly on speech and general audio can achieve improved cross-modal performance.

5.3. Universal SSL frontends with combined training

We transition from unimodal SSL frontends to universal SSL models pre-trained on both speech and general audio data, with results summarized in the middle section of Table 4. Overall, SPEAR XLarge achieves the lowest average EER of 6.02% across all datasets. Both SPEAR Large and SPEAR XLarge demonstrate stronger generalization across speech, music, and sound modalities compared to XLS-R. For instance, SPEAR Large attains EERs of 0.98%, 0.18%, and 5.08% on LA19, FakeMusic, and CodecFake A3, respectively, while also reducing EnvSDD error to 10.6%—a fourfold improvement over XLS-R. Additionally, SPEAR XLarge achieves more than a 4% absolute EER reduction on CtrSVDD compared to EAT. However, it is important to note that SPEAR XLarge has roughly twice the number of parameters compared to XLS-R or EAT. In contrast, USAD and Dasheng Tokenizer do not exhibit comparable gains, though USAD achieves better average performance compared to XLS-R and MERT. Under this setting, we also evaluate the Audio Transformer (AuT) introduced in Qwen3-Omni [39]; however, it similarly fails to generalize across modalities, yielding consistently poor performance with an average EER of 11.32%. A plausible reason could be the low temporal resolution of AuT features (12.5 Hz), which may suffice for semantic understanding tasks but leads to loss of fine-grained temporal cues due to averaging, making them less suitable for deepfake detection.

These findings suggest that while jointly pre-trained speech–audio SSL models, such as SPEAR, capture more transferable representations required for cross-modal deepfake detection, they do not fully resolve the generalization challenge for all types of audio deepfake detection. It also indicates that strong performance of SSL models on semantic speech and audio tasks does not necessarily translate to robust cross-modality generalization in spoofing countermeasure tasks.

5.4. Dual frontends with combined training

We next extend the analysis to dual-frontend architectures as described in Sub-section 3.2, with results reported in the bottom section of Table 4. Models denoted as *XLS-R + EAT – N layers* combine XLS-R and EAT, where only the first N transformer layers from each frontend are retained. Since using all 24 layers from both models significantly increases the parameter count, we additionally explore truncated variants by removing higher layers, motivated by prior findings [40] that lower-layer representations are often more effective for deepfake detection.

Across this setting, all dual-frontend models exhibit strong performance on both LA19 and FakeMusic. Notably, the combination of XLS-R and Dasheng Tokenizer achieves the lowest EER of 0.16% on LA19, while the XLS-R + EAT configurations attain perfect detection (0% EER) when using either the full 24 layers or a reduced 12-layer setup from each frontend, indicating that complementary representations from speech and audio SSL models significantly enhance detection capability. Even the 6-layer configuration achieves an EER as low as 0.01% on FakeMusic, indicating that strong performance can be retained with significantly reduced depth. Overall, the combination of XLS-R and EAT proves highly effective for cross-modal generalization, with the 18-layer configuration yielding the best overall performance, achieving the lowest average EER of 3.02% among all frontend combinations, followed by 24-layer (3.51%) and 12-layer settings (3.91%). However, generalization degrades in the 6-layer and 3-layer configurations, particularly on EnvSDD, ItW, and CtrSVDD. This suggests that effective detection on these datasets requires a combination of low-level local and high-level global representations, whereas relying solely on shallow, local features is insufficient.

To further reduce the parameter footprint, we evaluate the architecture described in Sub-section 3.3, where all 24 layers of XLS-R are retained, while only the first layer of the EAT frontend is used. The output of this shallow audio branch is then distilled against a frozen EAT teacher, using representations from either the 24th or 18th layer; the results are reported in the final two rows of Table 4. Despite the extreme reduction in depth for the second frontend, both variants achieve performance comparable to the 24-, 18-, and 12-layer dual-frontend configurations on LA19, FakeMusic, and CtrSVDD. However, performance degrades substantially on EnvSDD and CodecFake A3. This observation indicates that, unlike speech and music, sound deepfake detection critically depends on deeper, high-level representations, and cannot be effectively captured using shallow features alone—even with distillation. In contrast, for Fakemusic, shallow spectral features are sufficient. Interestingly, the XLS-R(24) + EAT(1) – Distill 18 configuration demonstrates strong generalization on ItW, a real-world noisy speech dataset, achieving a low EER of 3.75% following the XLS-R only model.

Table 5: *EER(%) comparison of detection models with different SSL frontends on ESDD2 mixed speech and sound dataset.*

Frontends	Trained on	ESDD2
XLS-R	LA19 + Fakemusic + EnvSDD + CtrSVDD	33.41
EAT		**13.04**
SPEAR XLarge		17.46
XLS-R + EAT - 24 layers		16.59
XLS-R + EAT - 18 layers		<u>15.12</u>
XLS-R(24) + EAT(1) - Distill 18		23.32

5.5. Performance on mixture audio

When evaluated on the challenging mixed speech–sound corpus ESDD2, all frontend configurations exhibit a substantial degradation in performance (cf. Table 5). The best result is obtained by the EAT-only model with an EER of 13.04%, followed by the dual-frontend XLS-R + EAT (18-layer) configuration at 15.12%. This performance drop highlights the intrinsic difficulty of mixed-modality scenarios, where overlapping speech and environmental sound artifacts introduce complex interference patterns that are not well represented in unimodal training data. It suggests that existing SSL frontends, even when combined, struggle to disentangle and jointly model heterogeneous acoustic cues, particularly when both semantic (speech) and non-semantic (sound) artifacts coexist within the same signal.

5.6. Dataset specific insights

The above results provide several general insights into the characteristics of the datasets. For the LA19 dataset, shallow to mid-level speech representations are sufficient for effective detection; notably, the XLS-R + EAT - 6 layers model achieves a low EER of 0.67% despite its reduced depth.

In the case of FakeMusic, shallow spectral features alone appear to be both necessary and sufficient. While the XLS-R-only model performs poorly, incorporating even a minimal audio branch—as in the XLS-R + EAT - 3 layers model configuration—yields a strong EER of 0.38%. This suggests that, with appropriate SSL feature selection, FakeMusic detection is comparatively less challenging. However, it remains unclear whether this behavior stems from the specific characteristics of the TTM models used in this particular

dataset or reflects a broader property of music deepfakes, remains a topic of further investigation.

In contrast, for the environmental sound deepfake detection dataset EnvSDD, deep SSL representations are essential. Performance degrades progressively as the depth of the audio branch is reduced—from 24 layers (2.8% EER) to 3 layers (14.02% EER) — indicating that shallow or mid-level features alone are insufficient, although they still contribute to some extent.

Joint training across all modalities substantially improves generalization to unseen speech data in the ItW dataset. The XLS-R frontend only model achieves 9.77% EER when trained only on LA19 and 5.17% when trained on CtrSVDD; however, joint training on all four datasets reduces the error to 3.47%, indicating enhanced robustness. Notably, performance on the CodecFake A3 subset follows a similar trend to EnvSDD across all training settings, reflecting consistent behavior for sound-based deepfake detection.

6. Conclusion & outlook

The scope of synthetic audio has rapidly expanded beyond speech to include modalities such as environmental sounds, music, and singing voice, necessitating the development of robust countermeasure systems capable of detecting deepfakes irrespective of their domain. As SSL-based frontends have become the de facto standard in modern spoofing detection countermeasures, their effectiveness across diverse modalities has not been thoroughly investigated. Our findings reveal a clear limitation of unimodal SSL representations: the XLS-R frontend, despite strong performance on speech and singing voice, fails to generalize to music and environmental sound deepfakes—even under combined multi-modal training. In contrast, EAT, trained on spectrogram-based audio data, effectively captures enough local and global spectral characteristics that are crucial for detecting non-speech deepfakes. However, universal SSL approaches such as USAD and Dasheng Tokenizer, though pre-trained on heterogeneous data, do not consistently resolve this limitation. Latest universal SSL frontends like SPEAR demonstrate improved cross-modal generalization, but still fall short of a perfect unified solution. Notably, the audio encoder of recent LLM-based systems (e.g., QWEN) also fails to deliver competitive performance in this task, suggesting that strong semantic modeling alone is insufficient for spoofing detection.

The most effective configuration arises from combining complementary inductive biases, specifically through a dual-frontend setup using XLS-R and EAT. Interestingly, reducing the depth of each frontend improves performance, with the optimal configuration retaining only 18 layers from each model, highlighting that both low-level and mid-level representations are critical for deepfake detection. However, this improvement comes at the cost of increased overall model complexity. Furthermore, evaluation under realistic mixed-audio conditions reveals a substantial performance degradation across all models, underscoring the difficulty of jointly modeling heterogeneous and overlapping acoustic artifacts. Overall, this study demonstrates that robust all-type deepfake detection cannot be achieved with current unimodal or even universal SSL frontends alone. Instead, it requires architectures that effectively integrate complementary representations or the development of stronger SSL models trained on diverse audio distributions. As future work, we plan to investigate the potential of full audio-LLM models of different sizes, leveraging both encoder and decoder components, to address the challenges of generalized deepfake detection.

7. Acknowledgements

This research has been partly funded by the Federal Ministry of Research, Technology and Space (BMFTR), Project: Fake-O-Meter (16KIS2657K).

8. References

[1] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio deepfake detection: What has been achieved and what lies ahead," *Sensors*, vol. 25, no. 7, p. 1989, 2025.

[2] Y. Li, M. Milling, L. Specia, and B. W. Schuller, "From audio deepfake detection to ai-generated music detection–a pathway and overview," *arXiv preprint arXiv:2412.00571*, 2024.

[3] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, T. Toda, and Z. Duan, "Svdd 2024: The inaugural singing voice deepfake detection challenge," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 782–787.

[4] H. Yin, Y. Xiao, R. K. Das, J. Bai, and T. Dang, "Environmental sound deepfake detection challenge: An overview," *arXiv preprint arXiv:2512.24140*, 2025.

[5] V. Cabannes, B. Kiani, R. Balestriero, Y. LeCun, and A. Bietti, "The ssl interplay: Augmentations, inductive bias, and generalization," in *International conference on machine learning*. PMLR, 2023, pp. 3252–3298.

[6] N. Anguita, F. Locatello, A. M. Saxe, M. Mondelli, F. Mancini, S. Lippl, and C. Domine, "A theory of how pretraining shapes inductive bias in fine-tuning," *arXiv preprint arXiv:2602.20062*, 2026.

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[10] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[11] H.-J. Chang, S. Bhati, J. Glass, and A. H. Liu, "Usad: Universal speech and audio representation via distillation," *arXiv preprint arXiv:2506.18843*, 2025.

[12] X. Yang, Y. Yang, Z. Jin, Z. Cui, W. Wu, B. Li, C. Zhang, and P. Woodland, "Spear: A unified ssl framework for learning speech and audio representations," *arXiv preprint arXiv:2510.25955*, 2025.

[13] H. Dinkel, X. Sun, G. Li, J. Mei, Y. Niu, J. Liu, X. Li, Y. Liao, J. Zhou, J. Zhang *et al.*, "Dashengtokenizer: One layer is enough for unified audio understanding and generation," *arXiv preprint arXiv:2602.23765*, 2026.

[14] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[15] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[16] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[17] Y. E. Kheir, T. Polzehl, and S. Möller, "Bicrossmamba-st: speech deepfake detection with bidirectional mamba spectro-temporal cross-attention," *arXiv preprint arXiv:2505.13930*, 2025.

[18] T. Liu, D.-T. Truong, R. K. Das, K. A. Lee, and H. Li, "Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 12 005–12 018, 2025.

[19] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller, P. Syga, P. Sperl, and K. Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.

[20] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino, and N. Evans, "An initial investigation for detecting partially spoofed audio," *arXiv preprint arXiv:2104.02518*, 2021.

[21] H. Wu, Y. Tseng, and H.-y. Lee, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," *arXiv preprint arXiv:2406.07237*, 2024.

[22] Z. Pan, S. H. Bhupendra, and J. Wu, "Molex: Mixture of lora experts in speech self-supervised models for audio deepfake detection," *arXiv preprint arXiv:2509.09175*, 2025.

[23] W. Guo, Y. Zhang, C. Pan, R. Huang, L. Tang, R. Li, Z. Hong, Y. Wang, and Z. Zhao, "Techsinger: Technique controllable multilingual singing voice synthesis via flow matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 978–23 986.

[24] D. Afchar, G. Meseguer-Brocal, and R. Hennequin, "Detecting music deepfakes is easy but actually hard," *arXiv preprint arXiv:2405.04181*, 2024.

[25] Y. Xie, R. Fu, Z. Wang, X. Wang, S. Cao, L. Ma, H. Cheng, and L. Ye, "Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception," *arXiv preprint arXiv:2504.06753*, 2025.

[26] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. 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)*. IEEE, 2022, pp. 6367–6371.

[27] H. Yin, Y. Xiao, R. K. Das, J. Bai, and T. Dang, "The first environmental sound deepfake detection challenge: Benchmarking robustness, evaluation, and insights," *arXiv preprint arXiv:2603.04865*, 2026.

[28] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[29] T. Alex, S. Ahmed, A. Mustafa, M. Awais, and P. J. Jackson, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," *arXiv preprint arXiv:2506.12222*, 2025.

[30] Y. Xie, X. Guo, J. Zhou, T. Wang, J. Liu, R. Fu, X. Wang, H. Cheng, and L. Ye, "Interpretable all-type audio deepfake detection with audio llms via frequency-time reinforcement learning," *arXiv preprint arXiv:2601.02983*, 2026.

[31] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[32] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[33] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[34] L. Comanducci, P. Bestagini, and S. Tubaro, "Fakemusiccaps: A dataset for detection and attribution of synthetic music generated via text-to-music models," *Journal of Imaging*, vol. 11, no. 7, p. 242, 2025.

[35] H. Yin, Y. Xiao, R. K. Das, J. Bai, H. Liu, W. Wang, and M. D. Plumbley, "Envsdd: Benchmarking environmental sound deepfake detection," *arXiv preprint arXiv:2505.19203*, 2025.

[36] Y. Zang, J. Shi, Y. Zhang, R. Yamamoto, J. Han, Y. Tang, S. Xu, W. Zhao, J. Guo, T. Toda *et al.*, "Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection," *arXiv preprint arXiv:2406.02438*, 2024.

[37] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" *arXiv preprint arXiv:2203.16263*, 2022.

[38] X. Zhang, H. Yin, Y. Xiao, L. Zhang, T. Dang, R. K. Das, and M. Li, "Esdd2: Environment-aware speech and sound deepfake detection challenge evaluation plan," *arXiv preprint arXiv:2601.07303*, 2026.

[39] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[40] Y. El Kheir, Y. Samih, S. Maharjan, T. Polzehl, and S. Möller, "Comprehensive layer-wise analysis of ssl models for audio deepfake detection," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 4070–4082.