Mixture of Spectral Experts for Audio Deepfake Detection

Yaxuan Qiu, Zhe Li, Mieradilijiang Maimaiti, Zunwang Ke, Wushour Silamu



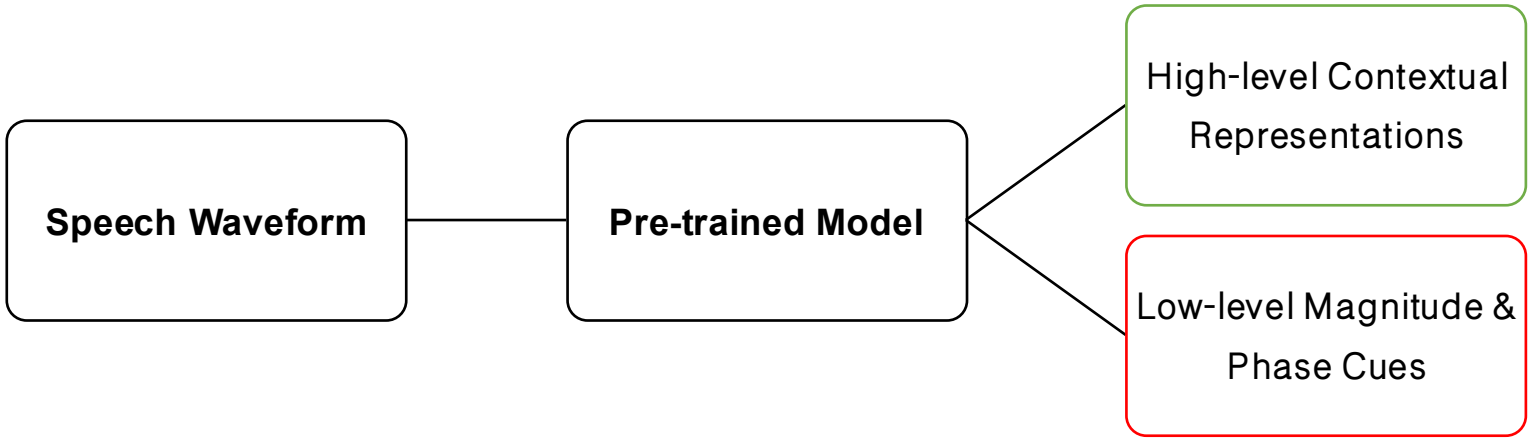


香 港 大 學
THE UNIVERSITY OF HONG KONG

Background

Current Status

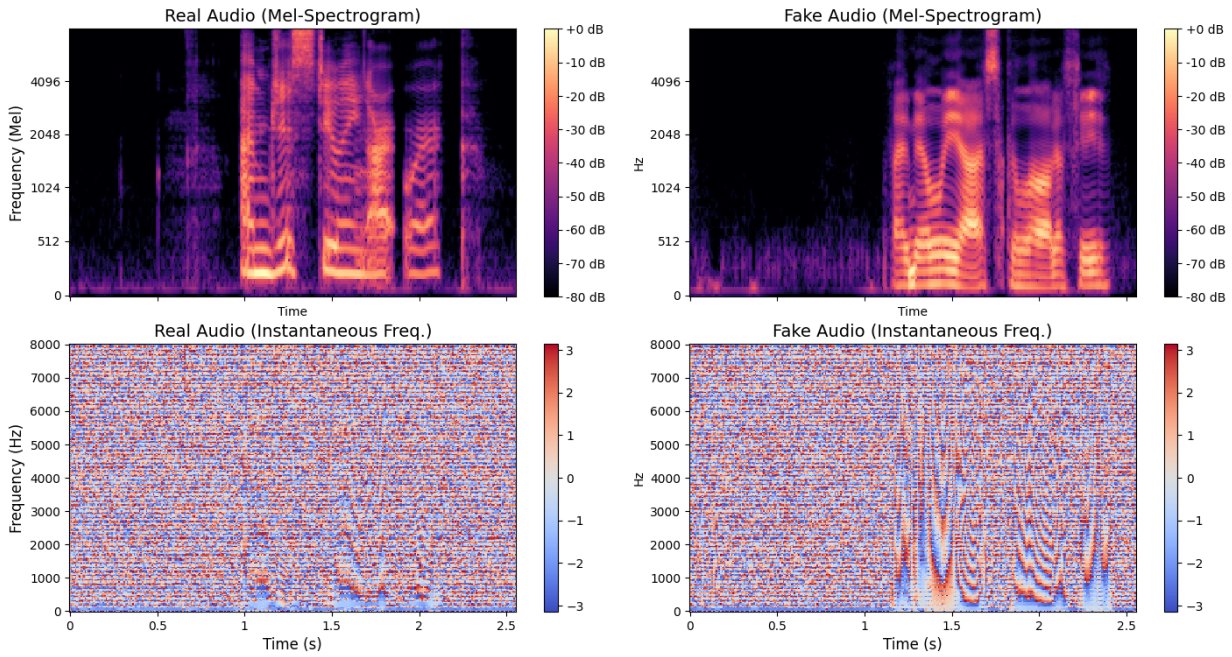
- Advancements in Speech Synthesis: Recent neural speech synthesis produces highly natural waveforms, making spoofing artifacts increasingly difficult to perceive.
- Effectiveness of PTMs: Self-supervised learning (SSL) pre-trained models (e.g., WavLM) provide robust representations and show strong performance in audio deepfake detec tion.
- Core Limitation: High-level contextual embeddings from PTMs tend to underrepresent crucial low-level physical artifacts.



Background

Motivation

- PTMs focus on high-level contexts but often ignore low-level magnitude and phase ano malies vital for distinguishing real from fake speech.
- Synthetic speech generation struggles with accurate phase reconstruction, leaving dete ctable spectral flaws.
- Full fine-tuning disrupts pre-trained acoustic priors, while current parameter-efficient me thods lack control over specific spectral dire ctions.



Methodology

Objective

To develop a robust audio deepfake detection framework that explicitly models physical cues and efficiently aligns pre-trained models to spectral artifacts.

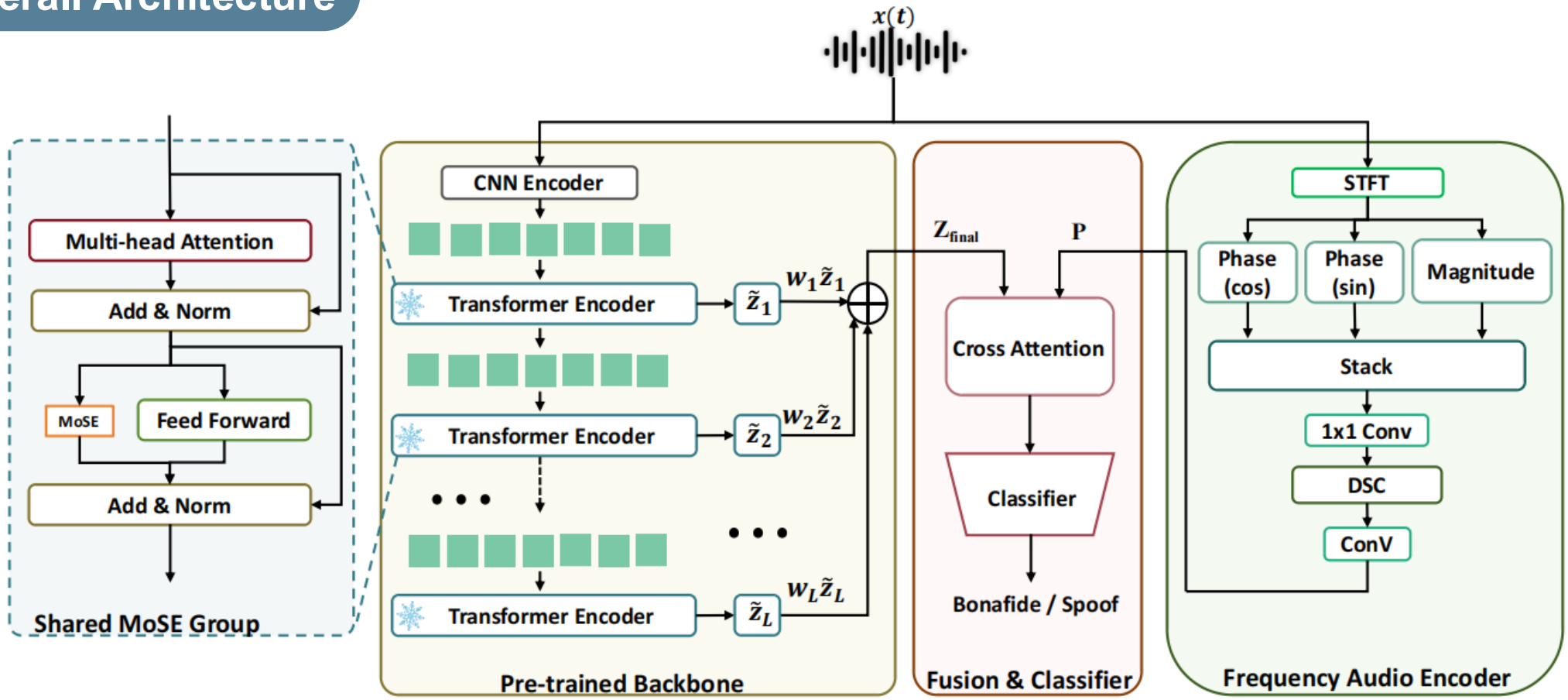
Contributions

1. Proposed a Frequency Audio Encoder (FAE) to explicitly capture magnitude and phase information.

2. Introduced Mixture of Spectral Experts (MoSE), a novel parameter-efficient fine-tuning (PEFT) method in the SVD domain.

Methodology

Overall Architecture



Pre-trained Backbone: Frozen WavLM extracts contextual features.
Explicit Extraction: FAE processes STFT to get magnitude and phase.
Adaptation & Fusion: MoSE adapts the backbone, and cross-attention fuses the features.

Methodology

Frequency Audio Encoder (FAE)

Goal: Explicitly capture low-level acoustic and spectral cues.

Magnitude Representation:

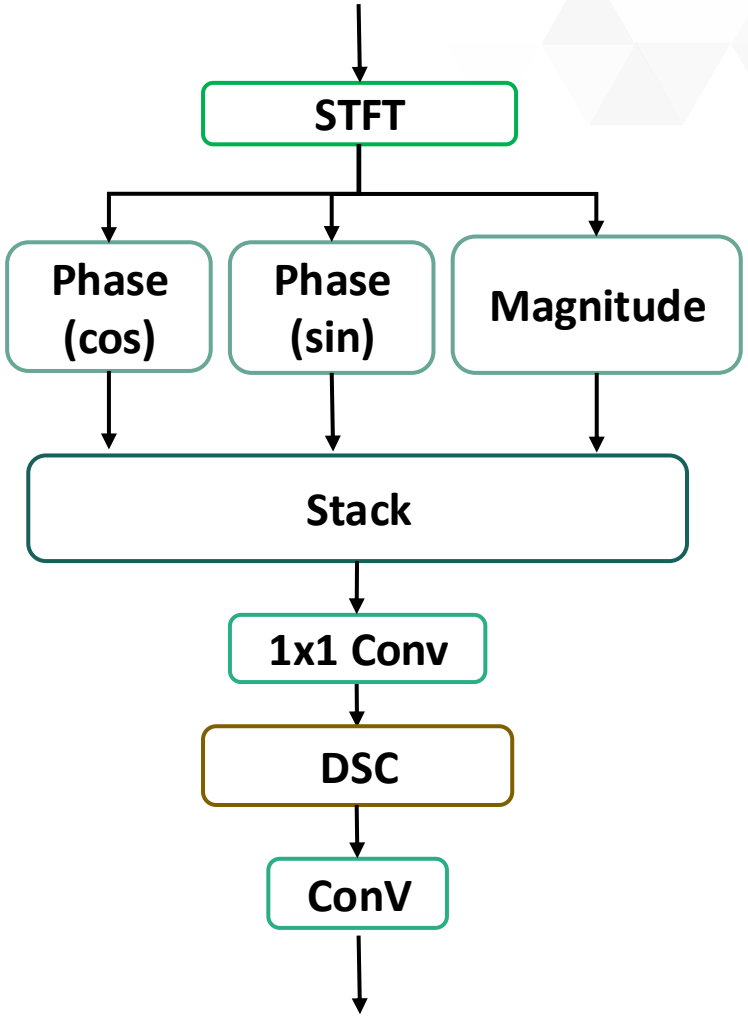
- Extracts spectral energy distribution to detect irregular harmonic structures

Phase Representation:

- Phase wrapping discontinuities.
- Decomposed into continuous sine ($\boldsymbol{\Phi}_{sin}$) and cosine ($\boldsymbol{\Phi}_{cos}$).

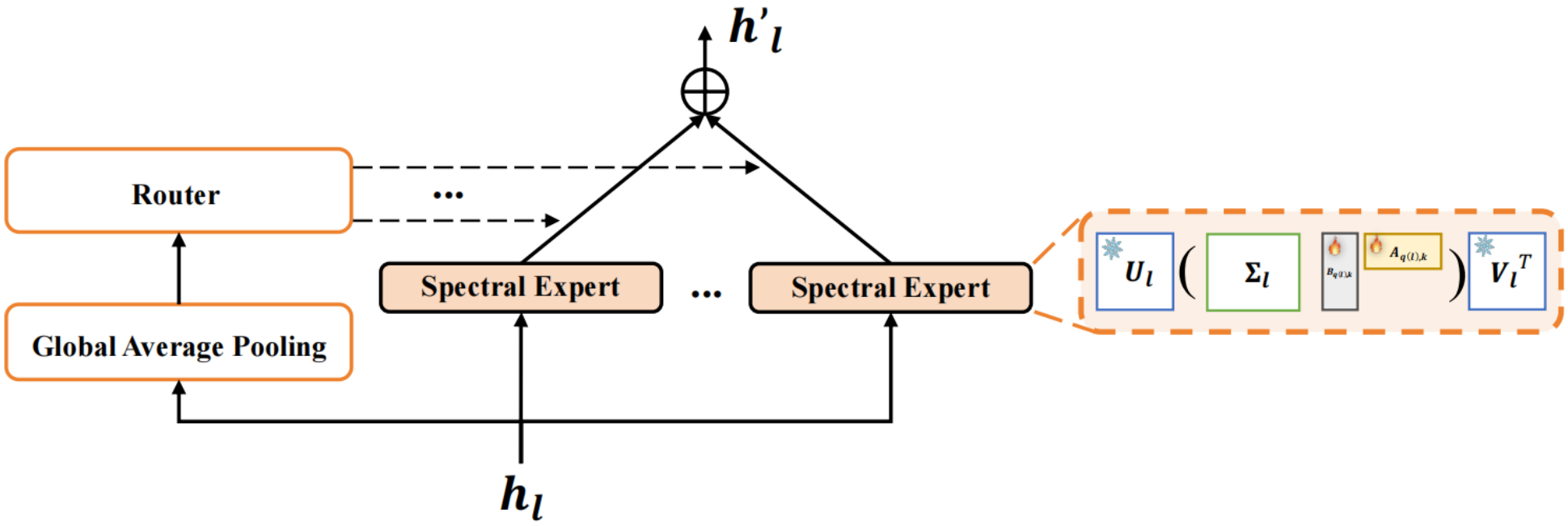
Stochastic Perturbation:

- Enhances robustness against unseen artifacts.



Methodology

Mixture of Spectral Experts (MoSE)



Methodology

Mixture of Spectral Experts (MoSE)

SVD-Domain Adaptation:

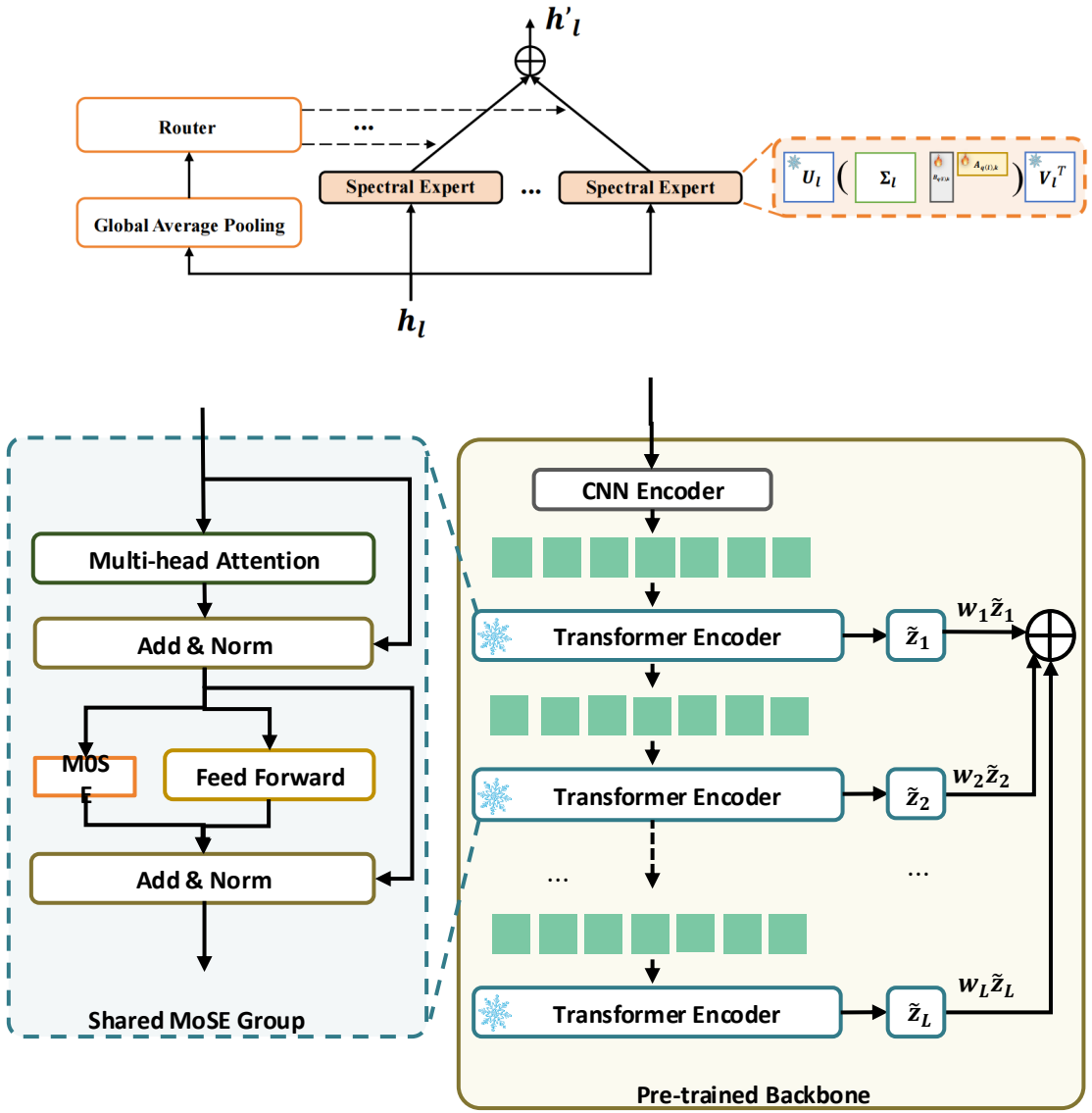
- Freeze: Pre-trained singular bases ($\boldsymbol{U}_l, \boldsymbol{V}_l^\top$) remain fixed to preserve acoustic priors.
- Update: Low-rank matrices modify only the singular values ($\boldsymbol{\Sigma}_l$).

Intuition (Neural Equalizer):

- Selectively scales spoofing-related spectral directions without altering the fundamental feature space.

Dynamic Routing & Sharing:

- Input-dependent gating adaptively mixes $\boldsymbol{K}$ experts based on context.
- Cross-layer sharing minimizes trainable parameters to just ~4.8M.



Experiments

Datasets:

Dataset	Real	Spoof
ASVspoof 2019 LA	12483	108978
ASVspoof 2021 LA	14816	133360
ASVspoof 2021 DF	14869	519059
In The Wild	19963	11816

Results

Results on ASVspoof 2019 LA

Model	EER (%)	min t-DCF
Wav+Spec-Res-TSSDNet [19]	3.39	-
LPS(F0) [20]	1.21	0.0358
AASIST [21]	0.83	0.0275
LFCC+ResNext [22]	0.61	0.0170
SE-Rawformer [23]	0.59	0.0184
DFSincNet [24]	0.52	0.0176
BiCrossMamba-ST [25]	1.08	0.0281
HuRawNet2 [26]	1.96	0.1393
wav2vec 2.0+LoRA [27]	1.30	-
wav2vec 2.0+MoE [4]	0.74	-
10L-WavLM-LSTM [28]	0.54	-
WavLM+MFA [3]	0.42	0.0126
wav2vec 2.0+VIB [29]	0.40	0.0107
XLSR-53+ASP [30]	0.31	-
AMFF+SSL+TDNN [31]	0.42	0.0190
Ours$_{(MoSE-WavLM-FAE)}$	**0.29**	**0.0081**

- Achieved 0.29% EER and 0.0081 min t-DCF on the evaluation set.
- Outperforms strong fine-tuned PTMs.
- Demonstrates the effectiveness of explicitly combining MoSE and FAE for detecting fine-grained artifacts

Results

Cross-Dataset Generalization

Model	EER (%)		
	21 LA	**21 DF**	**ITW**
BiCrossMamba-ST [25]	3.39	14.77	-
10L-WavLM-LSTM [28]	4.52	4.37	-
wav2vec 2.0-AASIST [32]	5.84	5.29	14.03
wav2vec 2.0-MoE-LoRA [32]	3.70	4.01	15.59
MoLEx [33]	4.31	**3.32**	9.60
WavLM+Ecapa [34]	6.68	15.94	34.64
WavLM+Ecapa+Glow [34]	8.54	26.26	32.07
Ours	**2.68**	3.89	**9.25**

Results

Ablation Study

Configuration	EER (%)	min t-DCF
Base Model	1.46	0.0377
w/o MoSE & FAE	0.72	0.0238
w/o FAE	0.51	0.0144
w/o MoSE	0.45	0.0140
Full Model	0.29	0.0081

- Removing FAE increases EER to 0.51%, confirming the necessity of explicit magnitude-phase cues.
- Removing MoSE increases EER to 0.45%, showing the importance of task-specific adaptation for frozen PTMs.

Results

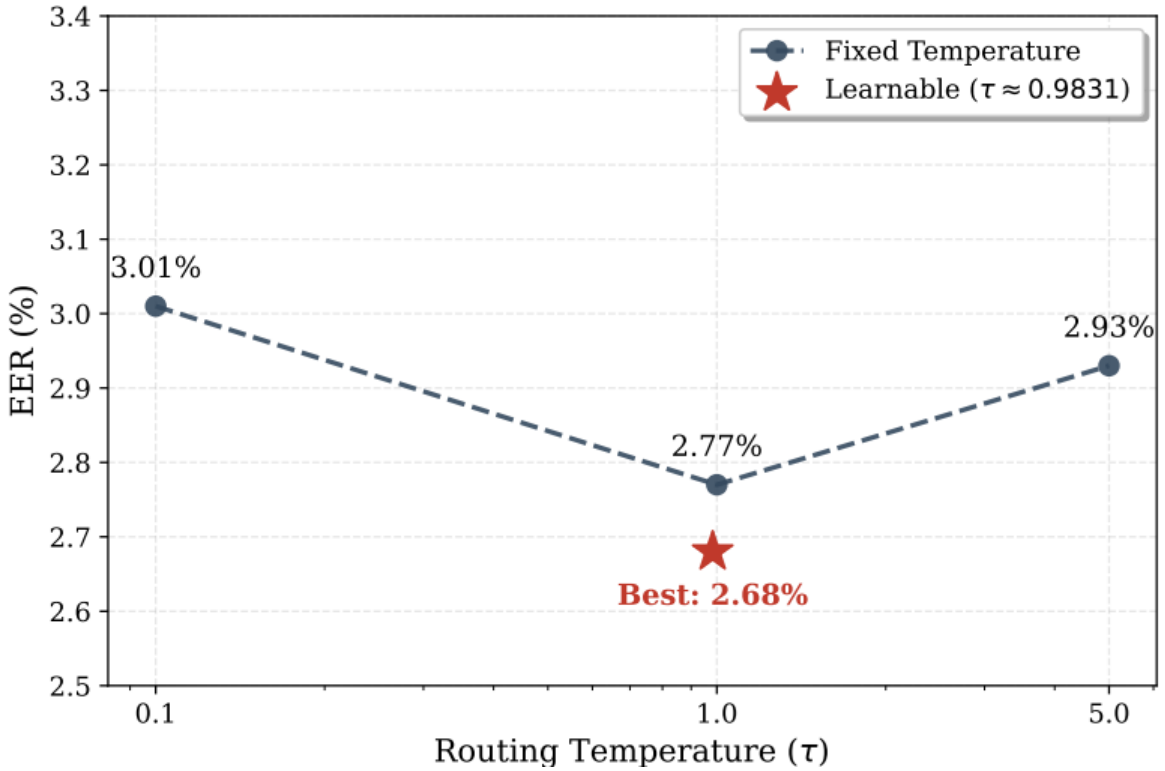
MoSE Configuration & Sharing Strategy

Experts K	Group Size G	EER (%)	
		19 LA	21 LA
1	2	0.48	4.17
2	2	0.35	3.21
4	**2**	**0.29**	**2.68**
8	2	0.31	2.89
4	4	0.32	3.24
4	6	0.38	3.47

- Number of Experts (K): $K = 4$ provides the optimal trade-off; further increasing to $K = 8$ slightly degrades performance.
- Layer-Group Size (G): Increasing group size from 2 to 4 or 6 degrades EER on both 19 LA and 21 LA.
- Trade-off: Highlights the balance between parameter sharing (efficiency) and adaptation capacity.

Results

Impact of Routing Temperature



- Evaluated fixed $\tau \in \{0.1, 1.0, 5.0\}$ versus a learnable τ.
- Exhibit a U-shaped performance trend, indicating sensitivity to manual tuning.
- Converges to $\tau \approx 0.9831$, achieving the lowest EER of 2.68% on 2021 LA.
- Dynamically learning τ better balances expert contributions compared to rigid fixed values.

Conclusion

- **Proposed Framework**: Integrated explicit frequency cues with pre-trained mo dels via a novel dual-branch architecture.
- **FAE (Frequency Audio Encoder)**: Successfully extracted crucial magnitude and phase anomalies overlooked by PTMs.
- **MoSE (Mixture of Spectral Experts)**: Achieved highly efficient SVD-domain adaptation (only ~4.8M parameters) acting as a neural equalizer.
- **Robust Performance**: Achieved highly competitive results on ASVspoof 2019 /2021 LA and demonstrated strong cross-dataset generalization.

Thanks for Your Attention

