Revealing the Truth with ConLLM for Detecting Multi-Modal Deepfakes

Gautam Siddharth Kashyap[1], Harsh Joshi[2], Niharika Jain[3], Ebad Shabbir[4]
Jiechao Gao[5*], Nipun Joshi[6], Usman Naseem[7*]

[1, 7]Macquarie University, Sydney, Australia
[2]Bharati Vidyapeeth's College Of Engineering, New Delhi, India
[3]Vivekananda Institute of Professional Studies (VIPS), New Delhi, India
[4]DSEU-Okhla, New Delhi, India
[5]Center for SDGC, Stanford University, California, USA
[6]Cornell University, New York, USA

Abstract

The rapid rise of deepfake technology poses a severe threat to social and political stability by enabling hyper-realistic synthetic media capable of manipulating public perception. However, existing detection methods struggle with two core limitations: (1) *modality fragmentation*, which leads to poor generalization across diverse and adversarial deepfake modalities; and (2) *shallow inter-modal reasoning*, resulting in limited detection of fine-grained semantic inconsistencies. To address these, we propose **ConLLM** (Contrastive Learning with Large Language Models), a hybrid framework for robust multimodal deepfake detection. **ConLLM** employs a two-stage architecture: stage 1 uses Pre-Trained Models (PTMs) to extract modality-specific embeddings; stage 2 aligns these embeddings via contrastive learning to mitigate *modality fragmentation*, and refines them using LLM-based reasoning to address *shallow inter-modal reasoning* by capturing semantic inconsistencies. **ConLLM** demonstrates strong performance across audio, video, and audio-visual modalities. It reduces audio deepfake EER by up to 50%, improves video accuracy by up to 8%, and achieves approximately 9% accuracy gains in audio-visual tasks. Ablation studies confirm that PTM-based embeddings contribute 9%–10% consistent improvements across modalities. Our code and data is available at: https://github.com/gskgautam/ConLLM/tree/main
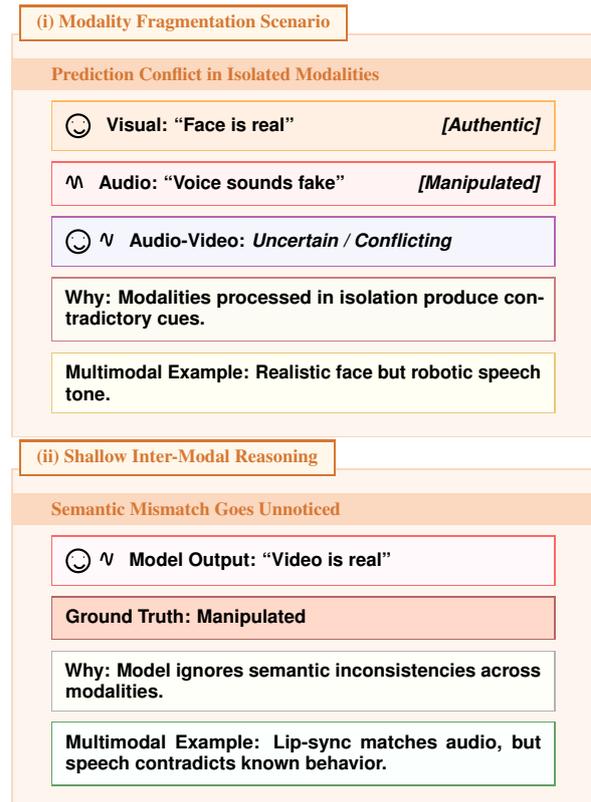
Figure 1: Illustration of common failure modes in multimodal deepfake detection. **Top:** Inconsistent unimodal predictions arise from poorly aligned audio-visual features (*modality fragmentation*). **Bottom:** Surface-level fusion leads to incorrect predictions due to the model's inability to reason over subtle cross-modal inconsistencies (*shallow inter-modal reasoning*).

1 Introduction

The rapid proliferation of deepfake technology, driven by advances in generative modeling and multimodal synthesis, poses growing societal and political risks (Kashyap et al., 2025b; Al-Khazraji et al., 2023). High-profile incidents—such as synthetic videos portraying political figures in fabricated scenarios—have underscored the need for effective and generalizable detection systems (Phukan et al., 2024a; Islam et al., 2024; Agarwal and Farid, 2021). As deepfake techniques evolve, they increasingly blend audio, video, and textual modalities using sophisticated architectures, including GANs, VAEs, and transformer-based TTS systems (Govindarajan et al., 2025; Masood et al., 2023; Yang et al., 2023; Cao et al., 2022). These developments have made it

---

*Corresponding Author: jiechao@stanford.edu, usman.naseem@mq.edu.au

significantly more difficult to distinguish between real and manipulated content.

Despite ongoing progress, existing detection methods predominantly focus on unimodal features, such as visual inconsistencies in facial movements (Tan and Le, 2019; Chollet, 2017) or artifacts in synthesized speech (Phukan et al., 2024b; Ranjan et al., 2022). However, multimodal deepfakes exploit cross-modal coherence to create more convincing forgeries, rendering unimodal detectors insufficient (Hao et al., 2022). While some multimodal approaches exist (Feng et al., 2023; Yu et al., 2023), they often suffer from two key limitations (see Figure 1): (1) *modality fragmentation*—where features extracted from different modalities remain isolated and poorly aligned, leading to poor generalization; and (2) *shallow inter-modal reasoning*—the inability to effectively model complex semantic relationships and subtle inconsistencies across modalities, limiting detection accuracy.

Therefore, to address the aforementioned challenges, we propose **ConLLM** (Contrastive Learning with Large Language Models), a hybrid framework that integrates contrastive learning and large-scale pretrained LLMs (i.e., GPT-style transformer architectures) for robust deepfake detection across audio, visual, and audio-visual modalities. The framework follows a two-stage modular architecture. In Stage 1, Pre-Trained Models (PTMs) independently extract embeddings from each modality, preserving modality-specific characteristics without premature fusion. These embeddings are then refined in Stage 2, where contrastive learning aligns semantically consistent representations and separates forgeries from authentic content, directly addressing *modality fragmentation*. Simultaneously, LLM-based reasoning modules are applied to the aligned embeddings to capture subtle semantic inconsistencies and inter-modal dependencies, mitigating the limitations of *shallow inter-modal reasoning*. **ConLLM** guarantees contextual robustness and discriminative power by separating feature extraction and semantic integration. In summary, our main contributions are as follows:

- We propose **ConLLM**, a novel two-stage hybrid framework that first performs modality-specific embedding extraction via PTMs, followed by contrastive alignment and LLM-based reasoning for semantic refinement—addressing both *modality fragmentation* and *shallow inter-modal reasoning*.

- We comprehensively evaluate **ConLLM** across three modalities—audio, video, and audio-visual—to demonstrate its generalizability and robustness.

- **ConLLM** outperforms state-of-the-art baselines across all modalities, reducing audio deepfake EER by up to 50%, improving video accuracy by up to 8%, and achieving 9% accuracy gains in audio-visual tasks. PTM-based embeddings further contribute 9%–10% consistent performance improvements in ablations.

2 Related Works

Contrastive Learning. Contrastive learning has been widely used to improve representation learning by encouraging semantically similar instances to be closer in the embedding space while pushing dissimilar ones apart (Zhai et al., 2022). Recent works have explored contrastive objectives in unimodal settings such as voice spoofing (Kashyap et al., 2025c) and image forgery detection (Yu et al., 2024). However, these approaches often focus on single-modality embeddings and do not address cross-modal alignment, which is critical in multimodal deepfake detection.

Large Language Models. Large Language Models (LLMs), such as GPT-style transformers, have demonstrated impressive capabilities in semantic reasoning (Achiam et al., 2023) and few-shot learning (Wei et al., 2022). While LLMs have been used for vision-language tasks (Alayrac et al., 2022), their role in deepfake detection remains underexplored. Most existing detection models (Yan et al., 2023) rely on shallow fusion or early fusion strategies without leveraging LLMs' capability for high-level reasoning.

Multimodal Deepfake Detection. Multimodal approaches have emerged to address the limitations of unimodal detection by fusing features from audio and visual streams. Early works relied on hand-crafted fusion strategies (Chugh et al., 2020; Mittal et al., 2020), while recent methods use neural architectures for cross-modal learning (Kashyap et al., 2025a; Cheng et al., 2023; Feng et al., 2023; Yu et al., 2023; Cai et al., 2022). Despite performance gains, these models often suffer from *modality fragmentation* and *shallow inter-modal reasoning*.
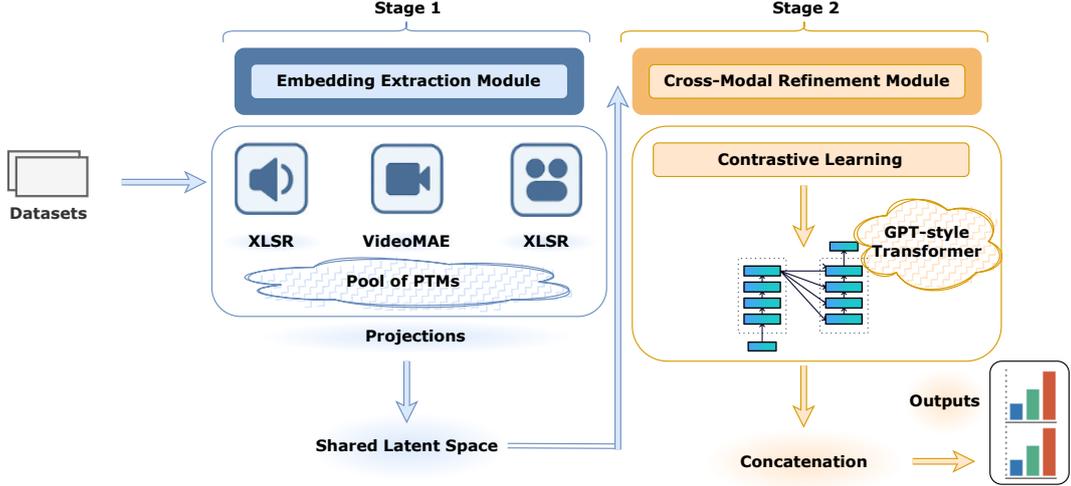
Figure 2: Overview of the **ConLLM** architecture. Modality-specific embeddings are extracted using PTMs (XLS-R, VideoMAE, VATLM) and projected into a shared latent space. A contrastive learning block aligns embeddings, followed by LLM-based refinement using a transformer. Refined embeddings are concatenated and passed through a classifier to detect deepfake content.

3 Model Architecture

The proposed **ConLLM** framework utilizes a novel two-stage architecture, as illustrated in Figure 2, to address the key challenges of *modality fragmentation* and *shallow inter-modal reasoning* in multimodal deepfake detection.

3.1 Embedding Extraction Module

This module extracts modality-specific embeddings from *audio*, *video*, and *audio-visual* inputs using PTMs. For audio, XLS-R converts speech signals into robust representations; for video, VideoMAE captures spatiotemporal features; and for audio-visual data, VATLM jointly models both modalities to encode cross-modal interactions. Each PTM generates high-dimensional embeddings, which are then mapped to a shared latent space via modality-specific projection functions. These functions align heterogeneous embeddings into a unified space, enabling joint learning across modalities in later stages.

Formally, let $\mathcal{D} = \{\mathcal{D}_a, \mathcal{D}_v, \mathcal{D}_{av}\}$ denote the input dataset, where $\mathcal{D}_m$ corresponds to each modality $m \in \{a, v, av\}$. Each $\mathcal{D}_m$ is passed through a PTM ϕ_m to yield embeddings $\mathbf{z}_m = \phi_m(\mathcal{D}_m)$, where $\mathbf{z}_m \in \mathbb{R}^{d_m}$. These are then projected to a shared latent space $\mathbb{R}^{d_s}$ using ψ_m, yielding $\mathbf{h}_m = \psi_m(\mathbf{z}_m)$ with $\mathbf{h}_m \in \mathbb{R}^{d_s}$. This ensures embedding compatibility for subsequent alignment and reasoning.

3.2 Cross-Modal Refinement Module

The second module applies contrastive learning to align embeddings across modalities and utilizes a GPT-style transformer architecture to refine the learned multimodal representations. The objective is to minimize the distance between embeddings of authentic samples while maximizing the distance between manipulated ones.

Let $\mathcal{H} = \{\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_{av}\}$ represent the set of embeddings for a given data sample, where $\mathbf{h}_a$, $\mathbf{h}_v$, and $\mathbf{h}_{av}$ denote the audio, visual, and fused audio-visual embeddings, respectively. The contrastive loss $\mathcal{L}_{\text{contrast}}$ is computed as $\mathcal{L}_{\text{contrast}} = -\sum_{(i,j) \in \mathcal{P}} \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \in \mathcal{N}} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$ where $\mathcal{P}$ denotes the set of positive pairs (e.g., authentic samples across modalities), $\mathcal{N}$ denotes negative pairs, $\text{sim}(\cdot, \cdot)$ computes cosine similarity, and τ is the temperature parameter.

After contrastive alignment, the embeddings $\mathbf{h}_m^{\text{input}} = \{\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_{av}\}$ are input to a GPT-style transformer module denoted as $\phi_{\text{Transformer}}$, which models inter-modal semantic dependencies and contextualizes the embeddings: $\mathbf{h}_m^{\text{refined}} = \phi_{\text{Transformer}}(\mathbf{h}_m^{\text{input}}) = \mathcal{F}_{\text{Transformer}}(\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_{av})$. The core of $\mathcal{F}_{\text{Transformer}}$ is the multi-head self-attention mechanism defined as $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$ where the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices are computed from the input embeddings using learned linear projections: $\mathbf{Q} = \mathbf{W}_Q \mathbf{h}_m$,

$\mathbf{K} = \mathbf{W}_K \mathbf{h}_m$, $\mathbf{V} = \mathbf{W}_V \mathbf{h}_m$ with $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ being trainable parameters, and d_k the dimensionality of the keys.

The refined embeddings $\mathbf{h}_m^{\text{refined}}$ are concatenated into a single vector: $\mathbf{h}_{\text{concat}} = [\mathbf{h}_a^{\text{refined}}, \mathbf{h}_v^{\text{refined}}, \mathbf{h}_{av}^{\text{refined}}]$ which is then passed through a classification head[1] f_{cls} to produce the final prediction: $\hat{y} = f_{\text{cls}}(\mathbf{h}_{\text{concat}})$, where $\hat{y} \in \{0, 1\}$.

Our cross-modal refinement module is inspired by the transformer architecture popularized by GPT models, which use self-attention to capture complex dependencies. Unlike the official GPT models—which publicly support primarily text (and in some variants audio or video)—our architecture is customized to jointly process and reason over audio, visual, and audio-video embeddings.

4 Experimental Setup

We select modality-specific PTMs based on the SUPERB benchmark (Yang et al., 2021) to ensure high-quality feature extraction. For audio, we use **XLS-R**[2] (Babu et al., 2021), which achieves strong performance in speech tasks (Hsu et al., 2021) and has shown promise in audio deepfake detection (Hsu et al., 2021). For video, **VideoMAE**[3] (Tong et al., 2022) is used for its efficient spatiotemporal feature learning and success in video understanding (Arnab et al., 2021). For audio-visual inputs, we employ **VATLM**[4] (Zhu et al., 2023), a unified transformer effective in multimodal representation learning across speech and vision (Gong et al., 2022).

4.1 Datasets

We evaluate **ConLLM** on six widely-used benchmarks across three modalities. For audio-based detection, we use the **ASVSpoof 2019 (LA)**[5] dataset (ASV) (Wang et al., 2020), which provides both bona fide and spoofed speech samples. We follow the official LA protocol for training and evaluation. We also use **DECRO**[6] (Ba et al., 2023), a multilingual dataset with English (**D-E**) and Chinese (**D-C**) audio samples to assess cross-lingual performance.

For video-based deepfake detection, we use **Celeb-DF (CDF)**[7] (Li et al., 2020), containing 590 real and 5,639 manipulated videos with enhanced synthesis quality, and **WildDeepfake (WD)**[8] (Zi et al., 2020), which includes 707 deepfake videos from uncontrolled environments to evaluate real-world generalizability.

For audio-visual detection, we use **FakeAVCeleb (FAFC)**[9] (Khalid et al., 2021), which includes 500 real and 19,500 deepfake samples, and the **DeepFake Detection Challenge (DFDC)**[10] (Dolhansky et al., 2020), the largest available benchmark with 128,154 manipulated videos across diverse subjects and environments.

Note: We adhere to the official training and evaluation splits provided by each dataset's authors.

4.2 Evaluation Metrics

We evaluate the performance of **ConLLM** using three standard metrics: Equal Error Rate (EER), Area Under the Curve (AUC), and Accuracy (ACC). These metrics are selected based on the modality and nature of the task.

For audio-based deepfake detection, we report **Equal Error Rate (EER)**, which is the point at which the False Positive Rate (FPR)[11] equals the False Negative Rate (FNR)[12]. EER provides a concise and meaningful indicator of verification accuracy, especially suited for binary spoof detection.

For video and audio-visual deepfake detection, we report both **Area Under the Curve (AUC)** and **Accuracy (ACC)**. AUC reflects the model's ability to distinguish between authentic and manipulated content across different thresholds, while Accuracy measures the overall proportion of correctly classified samples.

Note: In all result tables, ↑ indicates that higher values are preferable, whereas ↓ indicates that lower values are preferable.

4.3 Hyperparameters

For training and evaluation, **ConLLM** was configured with the following hyperparameters: a learning rate of 10^{-3}, batch size of 32, and a total of

---

[1]The final classification is performed using a sigmoid activation function.

[2]https://huggingface.co/facebook/wav2vec2-xls-r-1b

[3]https://huggingface.co/docs/transformers/en/model_doc/videomae

[4]https://github.com/microsoft/SpeechT5/tree/main/VATLM

[5]https://www.asvspoof.org/index2019.html

[6]https://zenodo.org/records/7603208

[7]https://github.com/yuezunli/celeb-deepfakeforensics

[8]https://github.com/OpenTAI/wild-deepfake

[9]https://github.com/DASH-Lab/FakeAVCeleb

[10]https://ai.meta.com/datasets/dfdc/

[11]The proportion of negative samples incorrectly classified as positive.

[12]The proportion of positive samples incorrectly classified as negative.

50 epochs. Momentum was set to 0.9, and weight decay to 10^{-4}. A dropout rate of 0.5 was used to mitigate overfitting, and ReLU was selected as the activation function. The model was optimized using the Adam optimizer. A step decay learning rate schedule was employed, with a decay factor of 0.1 applied every 10 epochs. *Note:* All hyperparameters were held constant across datasets to ensure fair and consistent performance evaluation.

5 Result Analysis

5.1 Comparison to State-of-the-Art

Tables 1, 2, and 3 demonstrate the strong performance of **ConLLM** on deepfake detection across multiple modalities.

In audio deepfake detection (see Table 1), **ConLLM** achieves exceptionally low Equal Error Rates (EER) of 0.21% on the ASV dataset and 0.01% on the D-E dataset. These results outperform prior works such as CL+GL (Kashyap et al., 2025c), MiO (Phukan et al., 2024b) and Res-TSSDNet (Ba et al., 2023). Note that since D-C was used primarily as a test set for cross-lingual evaluation in (Ba et al., 2023), we exclude models trained and tested on D-C from this comparison. This highlights **ConLLM**'s capability to detect subtle audio manipulations reliably.

For video deepfake detection (see Table 2), **ConLLM** achieves accuracy rates of 98.75% and 85.00% on the CDF and WD datasets, respectively, with Area Under Curve (AUC) values of 99.98% and 94.50%. These results surpass existing approaches such as RECCE (Cao et al., 2022) and Multi-Attention (Zhao et al., 2021), demonstrating strong visual feature analysis.

In audio-visual tasks (see Table 3), **ConLLM** achieves accuracies of 98.75% on FAVC and 96.50% on DFDC, along with AUC scores of 99.98% and 98.71%, respectively. This outperforms state-of-the-art models including PVASS (Yu et al., 2023) and AVAD (Feng et al., 2023), showcasing effective multimodal fusion.

To evaluate the impact of large-scale pretrained language model architectures, we compare two versions of **ConLLM**–one initialized with pretrained transformer weights inspired by GPT-style architectures, and another trained from scratch without such initialization. The pretrained variant consistently outperforms the non-pretrained model, demonstrating the benefits of leveraging pretrained knowledge embedded in these transformer archi-

Model	ASV ↓	D-E ↓
CQT-DCT-LCNN (Lavrentyeva et al., 2019)	1.84	–
STATNet (Ranjan et al., 2022)	–	0.20
Res-TSSDNet (Ba et al., 2023)	–	0.02
MiO (Phukan et al., 2024b)	0.41	0.04
CL+GL (Kashyap et al., 2025c)	0.33	0.01
ConLLM w/ GPT Weights	**0.21**	**0.01**
ConLLM w/o GPT Weights	0.50	0.05

Table 1: Performance comparison of EER (%) for audio deepfake detection on ASVSpoof 2019 (LA) and DE-CRO datasets.

Method	CDF ACC (%) ↑	CDF AUC (%) ↑	WD ACC (%) ↑	WD AUC (%) ↑
Xception (Chollet, 2017)	97.90	99.73	77.25	86.76
EfficientNet-B4 (Tan and Le, 2019)	97.63	99.20	81.63	90.36
F3-Net (Qian et al., 2020)	95.95	98.93	80.66	87.53
Mutil-Attention (Zhao et al., 2021)	97.92	99.94	82.86	90.71
RFM (Wang and Deng, 2021)	97.96	99.94	77.38	83.92
RECCE (Cao et al., 2022)	98.59	99.94	83.25	92.02
ConLLM w/ GPT Weights	**98.75**	**99.98**	**85.00**	**94.50**
ConLLM w/o GPT Weights	96.50	99.20	81.00	90.00

Table 2: Performance Comparison of Video Deepfake Detection Models on Celeb-DF and WildDeepfake Datasets

Method	FAVC ACC (%) ↑	FAVC AUC (%) ↑	DFDC ACC (%) ↑	DFDC AUC (%) ↑
MDS (Chugh et al., 2020)	82.8	86.5	89.8	91.6
EmoForen (Mittal et al., 2020)	78.1	79.8	80.6	84.4
JointAV (Zhou and Lim, 2021)	82.5	83.3	90.2	91.9
BA-TFD (Cai et al., 2022)	80.8	84.9	79.1	84.6
AVFakeNet (Ilyas et al., 2023)	78.4	83.4	82.8	86.2
VFD (Cheng et al., 2023)	81.5	86.1	80.9	85.1
AVoiD-DF (Yang et al., 2023)	83.7	89.2	91.4	94.8
AVAD (Feng et al., 2023)	94.2	94.5	93.2	96.7
PVASS (Yu et al., 2023)	95.7	97.3	96.3	98.9
ConLLM w/ GPT Weights	**98.75**	**99.98**	**96.50**	**98.92**
ConLLM w/o GPT Weights	95.00	98.00	90.00	94.00

Table 3: Performance Comparison of Audio-Visual Deepfake Detection on FakeAVCeleb and DeepFake Detection Challenge Datasets

tectures, beyond merely increasing model capacity.

5.2 Computational Experiments

In this section, we evaluate the computational efficiency of **ConLLM** across multiple benchmark datasets using three key metrics: Memory Usage[13] (MU), Floating Point Operations[14] (FLOPs), and Inference Time[15] (IT). These metrics are essential for understanding the feasibility of deploying deep-

---

[13]The amount of RAM consumed by a model during inference.

[14]The total number of arithmetic computations performed by a model.

[15]The time taken by a model to process an input and generate an output.

Dataset	Method	IT (ms) ↓	MU (GB) ↓	FLOPs ↓
Audio				
ASV	Audio Flamingo	280	11.5	190
	Audio Flamingo 2	170	5.2	105
	Audio Flamingo 3	130	4.1	85
	ConLLM	**55**	**1.6**	**45**
D-E	Audio Flamingo	285	11.7	195
	Audio Flamingo 2	175	5.3	108
	Audio Flamingo 3	135	4.3	88
	ConLLM	**60**	**1.8**	**48**
Video				
CDF	Video-LLM	320	12.8	215
	MiniGPT4-Video	180	5.8	115
	ConLLM	**65**	**2.0**	**55**
WD	Video-LLM	330	13.1	220
	MiniGPT4-Video	185	6.0	118
	ConLLM	**70**	**2.2**	**60**
Audio-Visual				
FAFC	Video-LLaMA	360	13.8	235
	video-SALMONN-o1	190	6.3	125
	ConLLM	**75**	**2.5**	**70**
DFDC	Video-LLaMA	365	14.0	240
	video-SALMONN-o1	195	6.6	130
	ConLLM	**80**	**2.8**	**75**

Table 4: Computational efficiency of **ConLLM** compared to modality-specific large models. ConLLM achieves significantly reduced inference time (IT), memory usage (MU), and FLOPs across all modalities.

fake detection models in real-world scenarios.

Table 4 compares **ConLLM** with representative multimodal models, including AudioFlamingo (Kong et al., 2024), AudioFlamingo 2 (Ghosh et al., 2025), and AudioFlamingo 3 (Ghosh et al., 2025) for audio datasets, Video-LLM (Cheng et al., 2024) and MiniGPT4-Video (Ataallah et al., 2024a) for video datasets, and Video-LLaMA (Zhang et al., 2023) and video-SALMONN-o1 (Sun et al., 2025) for audio-visual datasets. The results demonstrate that **ConLLM** consistently achieves superior efficiency. For example, on the ASV dataset, **ConLLM** processes inputs in approximately 55 ms using 1.6 GB of memory and 45 GFLOPs—significantly outperforming AudioFlamingo 3 (Ghosh et al., 2025), which requires around 130 ms, 4.1 GB, and 85 GFLOPs, respectively.

This efficiency advantage persists across all datasets and modalities. On video datasets like CDF and WD, **ConLLM** reduces inference time by more than 2x and uses less than half the memory compared to larger models like MiniGPT4-Video (Ataallah et al., 2024b). Similarly, in audio-visual tasks such as FAFC and DFDC, **ConLLM** demonstrates substantial reductions in computational costs while maintaining competitive performance.
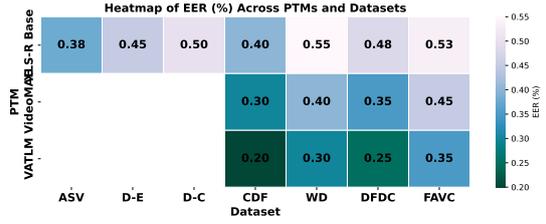
Note: We focus our comparison on domain-relevant multimodal models with native capabilities to handle audio, video, or audio-visual embeddings, deliberately excluding unimodal or purely text-based LLMs (e.g., GPT-3.5 or LLaMA) which are not directly comparable in this context.

Figure 3: Heatmap of Equal Error Rate (EER %) across different PTMs and datasets

| Model	D-C (Train) → D-E (Test)	D-E (Train) → D-C (Test)
XLS-R + x-vector	16.14	35.72
Whisper + Unispeech-SAT	15.74	44.26
ConLLM	**1.5**	**2.0**

Table 5: EER (%) ↓ for for Cross-Lingual Generalization against (Phukan et al., 2024b).

5.3 Additional Analysis

To assess the performance of various PTMs in detecting deepfakes across multiple datasets, we evaluate the EER for each model, as shown in Figure 3. The XLS-R Base model exhibits the most consistent performance, with EER values ranging from 0.38% to 0.55% across all datasets. Video-MAE performs well on the CDF, WD, DFDC, and FAFC datasets, achieving EERs between 0.30% and 0.45%, though its results are not available for the ASV, D-E, and D-C datasets. VATLM shows the lowest performance across datasets, with EER values between 0.20% and 0.35%, indicating comparatively limited effectiveness in these tasks.

Furthermore, Table 5 compares the cross-lingual generalization of **ConLLM** on the DECRO dataset against (Phukan et al., 2024b), demonstrating superior generalization. Specifically, **ConLLM** achieves relatively low EERs when trained on the D-C dataset and tested on the D-E dataset (1.5%), and vice versa, when trained on D-E and tested on D-C (2.0%). These results indicate that **ConLLM** possesses strong cross-lingual generalization capabilities, effectively adapting across different language domains. The small variation in performance between the two transfer tasks highlights the model's robustness in handling new linguistic contexts while maintaining low error rates in both
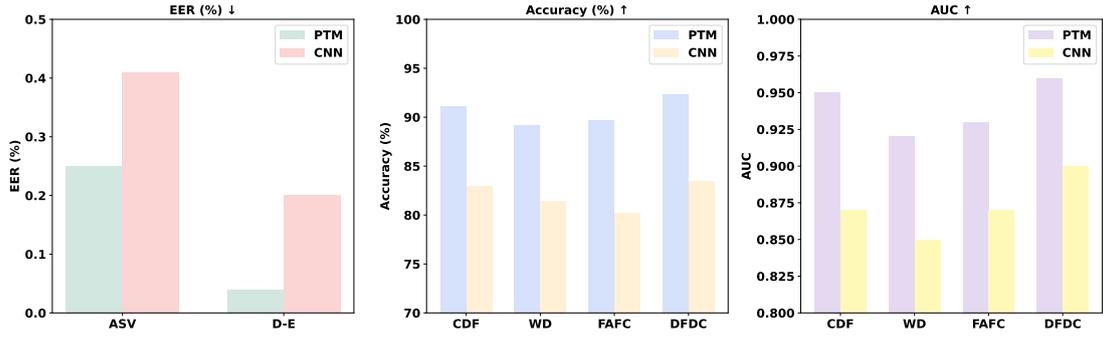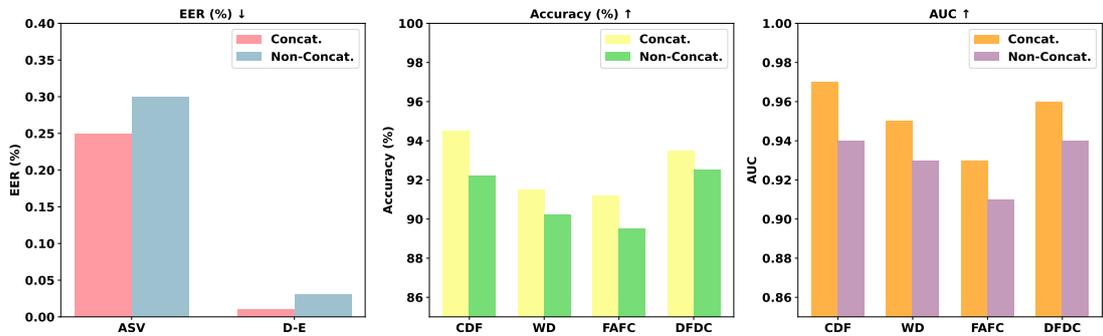
(a) Performance Comparison of PTM-based and CNN-based Features
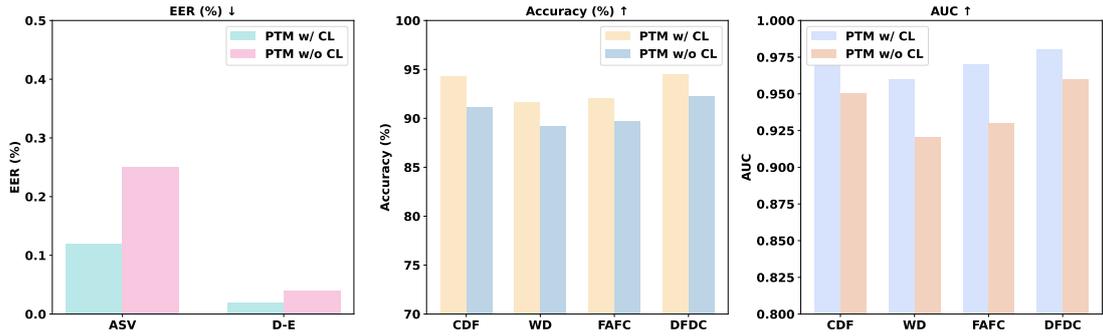
(b) Performance Comparison of PTM with and without Contrastive Learning
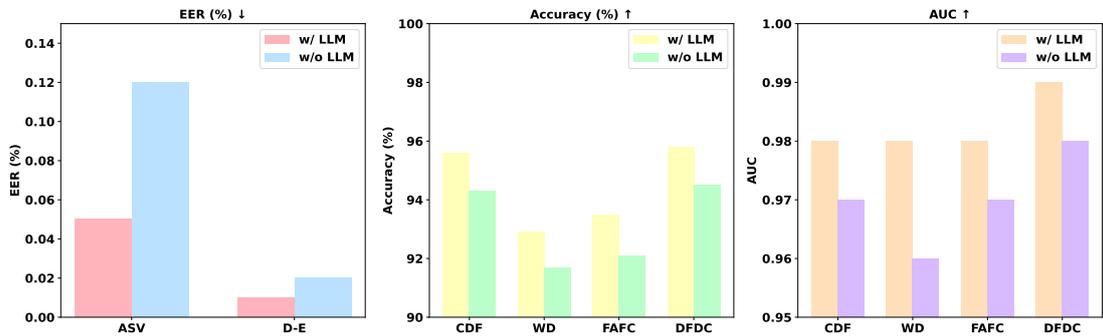
(c) Performance Comparison of PTM with Contrastive Learning and LLM-Based Embedding Refinement vs. PTM with Contrastive Learning

(d) Performance Comparison of Fusion Strategies with PTM, Contrastive Learning, and LLM-Based Embedding Refinement

Figure 4: Ablations to evaluate the impact of (i) modality-specific embedding extraction using PTMs, (ii) contrastive learning for cross-modal alignment, (iii) embedding refinement via LLMs, and (iv) comparison to alternative fusion strategies

directions.

6 Ablation Studies

We perform comprehensive ablation studies to analyze the contribution of key components in the **ConLLM** framework: (i) modality-specific embedding extraction using PTMs, (ii) contrastive learning for cross-modal alignment, (iii) LLM-based embedding refinement, and (iv) fusion strategies.

Modality-Specific Embedding Extraction. Replacing PTM-based embeddings with traditional CNN features significantly degrades performance across datasets (see Fig. 4a). On the ASV dataset, focused on voice spoof detection, EER increases from 0.21% to 0.41%, indicating CNNs' limited capacity to capture subtle speech inconsistencies. Similarly, for D-E, EER rises from 0.04% to 0.20%. Video datasets also show performance drops; on CDF, accuracy decreases from 91.1% to 82.9% and AUC from 0.95 to 0.87, highlighting PTMs' superior ability to capture fine-grained audio-visual cues. WD experiences a similar decline, with accuracy dropping from 89.2% to 81.4% and AUC from 0.92 to 0.85. For multimodal datasets FAFC and DFDC, PTMs yield notable gains in accuracy (80.2% to 89.7% on FAFC, 83.5% to 92.3% on DFDC) and AUC, underscoring their effectiveness in extracting rich multimodal representations. These results arise because PTMs are pretrained on large-scale diverse data and thus encode richer, more abstract, and context-aware features tailored for speech and visual modalities. Traditional CNNs, while effective for generic image tasks, lack the temporal and semantic understanding needed to detect subtle deepfake artifacts, leading to weaker discriminative power.

Contrastive Learning for Cross-Modal Alignment. Introducing contrastive learning improves alignment of audio-visual embeddings (see Fig. 4b). EER on ASV halves from 0.25% to 0.12%, while on D-E it improves from 0.04% to 0.02%. Video datasets benefit notably–CDF accuracy rises from 91.1% to 94.3% and AUC from 0.95 to 0.97; WD accuracy improves from 89.2% to 91.7%, AUC from 0.92 to 0.96. Multimodal datasets also show marked gains, with FAFC accuracy increasing from 89.7% to 92.1%, AUC from 0.93 to 0.97, and DFDC accuracy from 92.3% to 94.5%, AUC from 0.96 to 0.98. The observed improvements are due to contrastive learning's ability to explicitly align embeddings from different modalities by bringing matching pairs closer and pushing mismatched pairs apart in the latent space. This alignment enhances the model's sensitivity to cross-modal inconsistencies, which are critical cues for deepfake detection.

LLM-Based Embedding Refinement. Further refinement with LLMs yields consistent improvements by incorporating semantic context (see Fig. 4c). On ASV, EER improves from 0.12% to 0.05%, and on D-E from 0.02% to 0.01%. Video datasets like CDF see accuracy increase from 94.3% to 95.6%, AUC from 0.97 to 0.98, while WD accuracy rises from 91.7% to 92.9% and AUC from 0.96 to 0.98. For FAFC and DFDC, LLM refinement enhances accuracy to 93.5% and 95.8%, and AUC to 0.98 and 0.99 respectively. These gains result from the LLM's capacity to incorporate high-level semantic understanding and contextual reasoning into the embeddings. Leveraging LLMs enables the model to capture subtle linguistic and contextual nuances that signal sophisticated forgeries, thereby enhancing detection robustness.

Fusion Strategies. Comparing concatenation-based and non-concatenation fusion methods (see Fig. 4d), concatenation consistently outperforms across datasets. On ASV, EER is 0.21% versus 0.30%; CDF accuracy is 94.5% versus 92.2%. WD accuracy is 91.5% compared to 90.2%, and FAFC and DFDC also show superior accuracy and AUC with concatenation fusion. This advantage arises because concatenation preserves modality-specific information by retaining full feature representations from each modality before joint processing, allowing richer joint reasoning. Non-concatenation fusion methods such as addition or averaging may dilute or lose important modality-specific signals early, leading to less discriminative combined embeddings.

7 Conclusion and Future Works

In this study, we proposed **ConLLM**, a state-of-the-art framework for deepfake detection that effectively leverages multimodal features and advanced learning techniques to deliver exceptional performance across diverse datasets. Our extensive experiments demonstrate that **ConLLM** consistently outperforms existing deepfake detection models, excelling in both unimodal and multimodal scenarios. This adaptability enables robust detection of

emerging deepfake generation methods, highlighting its practical applicability in real-world settings.

Limitations

Despite the promising performance of **ConLLM**, several limitations should be acknowledged. First, the model's performance depends heavily on the quality and diversity of the training data, which may limit its generalizability to novel or unseen deepfake generation techniques. Moreover, the computational demands for training and inference in multimodal deepfake detection pose challenges for deployment in real-time and resource-constrained environments. Finally, although **ConLLM** shows strong results on several popular benchmarks, further evaluation on a wider variety of real-world datasets is necessary to comprehensively assess its robustness, scalability, and interpretability.

Ethics Statement

The development and deployment of **ConLLM** are guided by ethical considerations to ensure responsible use in deepfake detection. The model aims to mitigate the risks associated with the proliferation of manipulated content, supporting efforts to combat misinformation and protect individuals from potential harm. However, it is important to acknowledge that the application of deepfake detection technologies raises concerns about privacy, consent, and potential misuse. To address these, we advocate for the transparent and ethical use of **ConLLM** while ensuring that appropriate safeguards are in place to protect personal and sensitive information.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shruti Agarwal and Hany Farid. 2021. Detecting deepfake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 981–989.

Samer Hussain Al-Khazraji, Hassan Hadi Saleh, Adil Ibrahim KHALID, and Israa Adnan MISHKHAL. 2023. Impact of deepfake technology on social media: Detection, misinformation and societal implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics*, 23:429–441.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024a. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. In *arXiv preprint arXiv:2404.03413*.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. 2024b. Goldfish: Vision-language understanding of arbitrarily long videos. In *European Conference on Computer Vision*, pages 251–267. Springer.

Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, pages 2033–2044.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. 2022. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10. IEEE.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122.

Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. 2023. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang,

Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. In *arXiv preprint arXiv:2406.07476*.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 2020. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Chao Feng, Ziyang Chen, and Andrew Owens. 2023. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. 2022. Contrastive audio-visual masked autoencoder. In *arXiv preprint arXiv:2210.07839*.

Vijay Govindarajan, Pratik Patel, Sahil Tripathi, Md Azizul Hoque, and Gautam Siddharth Kashyap. 2025. Magic-enhanced keyword prompting for zero-shot audio captioning with clip models. *arXiv preprint arXiv:2509.12591*.

Hanxiang Hao, Emily R Bartusiak, David Güera, Daniel Mas Montserrat, Sriram Baireddy, Ziyue Xiang, Sri Kalyan Yarlagadda, Ruiting Shao, János Horváth, Justin Yang, et al. 2022. Deepfake detection using multiple data modalities. In *Handbook of digital face manipulation and Detection: From DeepFakes to morphing attacks*, pages 235–254. Springer International Publishing Cham.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. volume 29, pages 3451–3460. IEEE.

Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. 2023. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124.

Masabah Bint E Islam, Muhammad Haseeb, Hina Batool, Nasir Ahtasham, and Zia Muhammad. 2024. Ai threats to politics, elections, and democracy: A blockchain-based deepfake authenticity verification framework. *Blockchains*, 2(4):458–481.

Gautam Siddharth Kashyap, Mohammad Anas Azeez, Rafiq Ali, Zohaib Hasan Siddiqui, Jiechao Gao, and Usman Naseem. 2025a. Childguard: A specialized dataset for combatting child-targeted hate speech. *arXiv preprint arXiv:2506.21613*.

Gautam Siddharth Kashyap, Niharika Jain, Ebad Shabbir, Harsh Joshi, Usman Naseem, and Jiechao Gao. 2025b. Clarity: A lightweight multimodal transformer for harmful content detection. *IEEE Transactions on Artificial Intelligence*, (99):1–11.

Gautam Siddharth Kashyap, Zohaib Hasan Siddiqui, Mohammad Anas Azeez, Rafiq Ali, Shantanu Kumar, Navin Kamuni, and Jiechao Gao. 2025c. Fooling the forgers: A multi-stage framework for audio deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 25125–25148.

Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2023. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832.

Orchid Chetia Phukan, Gautam Siddharth Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024a. Are paralinguistic representations all that is needed for speech emotion recognition? In *Proc. Interspeech 2024*, pages 4698–4702.

Orchid Chetia Phukan, Gautam Siddharth Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024b. Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake. *arXiv preprint arXiv:2404.00809*.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer.

Rishabh Ranjan, Mayank Vatsa, and Richa Singh. 2022. Statnet: Spectral and temporal features based multitask network for audio spoofing detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. In *arXiv preprint arXiv:2502.11775*.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093.

Chengrui Wang and Weihong Deng. 2021. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. 2023. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029.

Yang Yu, Xiaolong Liu, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C Kot. 2023. Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6926–6936.

Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. 2024. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12765–12774.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *arXiv preprint arXiv:2306.02858*.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194.

Yipin Zhou and Ser-Nam Lim. 2021. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809.

Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2023. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390.