Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Hierarchical Attention for Multimodal Depression Detection

Yongwei Liao

weijie0018@163.com

Shenzhen City Polytechnic

Yibin Zhang

Shenzhen City Polytechnic

Weiwei Chen

Shenzhen City Polytechnic

Xinyue Zhang

Shenzhen City Polytechnic

Bing Zhou

Shenzhen City Polytechnic

Research Article

Keywords: Depression Detection, Multimodal Learning, Feature Fusion, Large Model, Artificial Intelligence

Posted Date: May 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9188744/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cross-Hierarchical Attention for Multimodal Depression Detection

Yongwei Liao[1,2*], Yibin Zhang[1], Weiwei Chen[1], Xinyue Zhang[1], Bing Zhou[1]

[1*]School of Information and Communications Technology, Shenzhen City Polytechnic, Longgang, Shenzhen, 518000, Guangdong, China.
[2]Shenzhen Institute of Technology, Shenzhen, 518000, Guangdong, China.

*Corresponding author(s). E-mail(s): weijie0018@163.com;
Contributing authors: 37975964@qq.com; chenwwayne@gmail.com;
lizhenjun@szcp.edu.cnm; bingzhou6@gmail.com;

Abstract

Depression detection in video interview scenarios requires effective integration of multimodal information. This paper proposes Cross-Hierarchical Attention for Multimodal Depression Detection (CHAMDD), a Transformer-based multimodal depression recognition framework comprising four core parts: feature extraction, feature fusion, feature refinement, and depression detection. In feature extraction, AUFormer, Audio Spectrogram Transformer (AST), and Llama are respectively used to extract facial, audio, and text features. The feature fusion build a Cross-hierarchical Attention Fusion to realize gradual fusion from local representation to global alignment. The feature refinement designs a hierarchical temporal cross-modal refinement to model dynamic behavior and purify fused features. The depression detection uses multi-scale pooling to output binary classification and depression degree. Experiments on the LMVD dataset show that CHAMDD achieves an accuracy of 82.88%, outperforms the state-of-the-art method DepMamba by 4.55, and verifying its progressiveness and effectiveness.

Keywords: Depression Detection, Multimodal Learning, Feature Fusion, Large Model, Artificial Intelligence

1 Introduction

With the growing prominence of global mental health issues, depression–as one of the most prevalent mental disorders has made early detection and accurate assessment central research topics in clinical psychology and computational psychiatry. Traditional diagnostic approaches mainly rely on scale-based instruments (e.g., PHQ-9 [1, 2], HAMD [3]) and clinical interviews, which suffer from high subjectivity, inefficiency, and strong dependence on patients willingness to communicate and clinicians professional experience. In some regions, the misdiagnosis rate of depression can exceed 20% [4, 5]. In recent years, advances in artificial intelligence have driven significant progress in automated depression detection [6] using multimodal behavioral data?such as facial expressions, vocal prosody, and linguistic content [6, 7]. These approaches enable objective, non-invasive identification of behavioral biomarkers associated with depressive states in naturalistic settings like video interviews, offering a

promising technical pathway for computer-aided depression diagnosis [8–10].

Existing popular methods typically employ Convolutional Neural Networks (CNNs) [11], Recurrent Neural Networks (RNNs) [12], and their variants such as Long Short-Term Memory (LSTM) [13] networks for feature extraction and sequential modeling, along with tensor fusion, attention mechanisms, and other strategies for multimodal integration [4, 5]. Although recent advanced models represented by DepMamba [14] have achieved remarkable performance improvements, several critical challenges remain unresolved. Firstly, the efficiency of modeling complex long-sequence multimodal data is insufficient. Secondly, conventional feature fusion schemes fail to adequately capture fine-grained and asymmetric cross-modal interactions and dependencies. Thirdly, most models struggle to balance shared cross-modal features and modality-specific information, leading to limited discriminability and interpretability of fused representations. Accordingly, developing a unified framework capable of deep cross-modal understanding and fine-grained feature coordination is essential to further improve model performance and facilitate clinical translation.

To address these limitations, this paper introduces a novel Cross-Hierarchical Attention Multimodal Depression Detection framework (CHAMDD). Built on a pure Transformer architecture, the proposed framework establishes a unified four-stage pipeline covering feature extraction, feature fusion, feature refinement, and depression classification. The main contributions of this study are summarized as follows:

1. An innovative cross-hierarchical attention network is proposed, which realizes deep alignment and complementary enhancement of cross-modal information via a progressive fusion strategy ranging from local self-attention reorganization to global cross-attention interaction.
2. A hierarchical temporal-crossmodal decoder is designed, adopting a three-layer iterative structure of temporal decoding, cross-modal refinement, and interactive modeling to dynamically characterize behavioral evolution patterns and extract highly discriminative structured representations.
3. Extensive experiments conducted on the public LMVD benchmark dataset demonstrate that CHAMDD achieves 82.88% detection accuracy, remarkably outperforming the current state-of-the-art model DepMamba.

The rest of this paper is organized as follows: Section 2 reviews related work in depression detection and multimodal learning. Section 3 details the overall architecture and key components of the CHAMDD framework. Section 4 describes the experimental settings, presents quantitative and qualitative results, and provides in-depth analyses. Finally, Section 5 concludes the paper and discusses potential future research directions.

2 Related Work

Research on automatic depression detection has advanced from early single-modality analysis to contemporary deep learning and multimodal fusion. This section reviews related studies from single-modality depression detection, traditional multimodal fusion strategies, and Transformer-based multimodal detection.

2.1 Single-Modality Depression Detection

Early studies focused on extracting depressive information from a single behavioral modality. In the visual modality,early works mainly depended on facial expression and adopted manually defined facial action units (AUs) [7]. With the advancement of deep learning, models such as DepressNet [15] employed pre-trained convolutional neural networks to automatically learn depressive representations from facial expression. In the audio modality, studies focused on paralinguistic characteristics of speech. Early approach utilized handcrafted features such as MFCCs combined with classifiers such as support vector machines (SVMs). Deep learning models, such as DepAudioNet [16], directly learned complex patterns from raw audio or spectrograms. More refined architectures, e.g., TFCAV [17], incorporated time-frequency channel attention to enhance the extraction of depression-related speech features. In the textual modality, research analyzed linguistic content from interviews or questionnaires. Conventional methods directly used scale

scores including those from the Beck Depression Inventory (BDI) [18] and Patient Health Questionnaire [1, 2]. Recently, pre-trained language models (e.g., BERT-series models [19]) enabled deeper semantic and sentiment understanding, providing rich linguistic evidence for depression detection. Nevertheless, single-modality systems only capture partial manifestations of depression, suffer from limited information dimensions, are vulnerable to noise interference, and have bottleneck in performance.

2.2 Traditional Multimodal Fusion

To exploit complementary information across modalities, multimodal fusion has become the mainstream paradigm. Multimodal fusion [8, 20, 21] can be categorized as Decision-level fusion (late fusion), Feature-level fusion (mid-level fusion) and Model-level fusion. Decision-level fusion trains classifiers for each modality independently and integrates outputs via voting or weighted averaging. Although simple, this strategy neglects early-stage interactions among modalities. Feature-level fusion combines representations from different modalities before classification. Typical techniques include direct concatenation, tensor-based fusion (e.g., TFN [22]), and attention-based fusion (e.g., MFM-attention [23]). TAMFN [24] further integrated time-aware attention for dynamic temporal feature fusion. Feature-level fusion approaches capture partial cross-modal interactions but often rely on heuristic or complicated fusion designs. Model-level fusion constructs dedicated architectures for joint modeling. For instance, Bi-LSTM [25] is used to process temporally aligned audio and text features; DCGAN [26] has been explored to generate consistent multimodal representations for depression classification. CANAMRF [27] introduced cross-attention to adaptively weight modalities, representing an improvement over conventional deep fusion schemes. However, these methods are mostly built upon CNN or RNN backbones, which remain limited in modeling long-range global dependencies and complex non-linear cross-modal relationships.

2.3 Transformer-Based Multimodal Detection

Transformer has been introduced into depression detection due to its significant advantages in processing sequential data and cross modal alignment by using attention mechanisms [28]. As a backbone network, the Transformer has been used to replace CNNs/RNNs for single-modal feature extraction or as a fusion encoder. For example, Audio Spectrogram Transformer (AST) [29] achieved impressive performance in audio classification. For cross-modal fusion, the self-attention and cross-attention inherent in Transformers are naturally suited for multimodal interaction. DepMamba [14] model integrated state space model (SSM) for efficient sequence modeling and fusion, achieving state-of-the-art performance. Visual-language models CLIP [30] and ALBEF [31] learn cross-modal alignment via contrastive learning or multi-task training, and provide new paradigms for depression detection. However, the most existing works directly adopt generic cross-modal Transformer structures without fully accounting for the progressive temporal evolution of depressive behavioral information, the asymmetric and fine-grained dependencies between modalities (e.g., subtle facial expressions and speech pauses), and the trade-off between preserving modality specificity and enhancing inter-modal synergy during fusion.

To systematically address these challenges, this paper proposes CHAMDD. We design a cross-level attention network to achieve progressive, fine-grained fusion from local to global representations, and construct a hierarchical temporal-cross-modal decoder to explicitly model behavioral dynamics and asymmetric cross-modal interactions. Under a unified Transformer architecture, this framework can deepen understanding and more effective representation of multimodal information.

3 Methods

This section elaborates on the proposed Cross-Hierarchical Attention for Multimodal Depression Detection (CHAMDD).

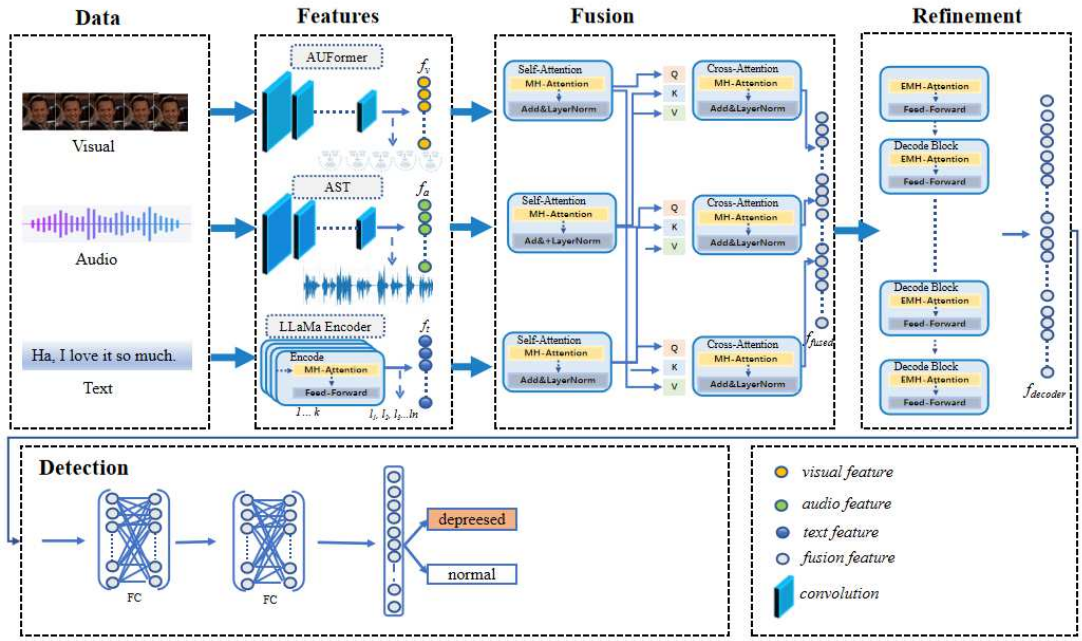


Fig. 1: The structure of CHAMDD.

3.1 The structure of CHAMDD

As illustrated in Figure 1, the framework is a unified architecture based on Transformer, consisting of four core modules in sequence: multimodal feature extraction, cross-hierarchical attention fusion, hierarchical temporal-crossmodal refinement, and depression detection. The entire process aims to conduct a progressive deep analysis of visual, audio, and text data from video interviews, moving from fine-grained local understanding to global context modeling. Firstly, modal-specific features are extracted from video, audio, and text inputs respectively. Subsequently, heterogeneous features are synchronized to a unified timeline through temporal alignment. Then, deep cross-modal interaction is performed via the Cross-hierarchical Attention Fusion Module (CHAF). Afterwards, the fused features undergo iterative purification through the Hierarchical Temporal-Cross-modal Refinement Module (HTCR) to learn highly discriminative depression-related patterns. Finally, the multi-task prediction head outputs classification and regression results.

Given that the LMVD [32] dataset used in this study only contains aligned video and audio streams (excluding textual modality), the fusion phase of CHAMDD primarily focuses on deep interaction and joint modeling of the audio-visual bimodal. Specifically, video and audio features are first temporally aligned, and then the aligned features are concatenated along the feature dimension to generate a unified 264-dimensional spatiotemporal feature vector. This concatenated sequence is input into a hierarchical and temporal cross-modal refinement module for further processing. CHAMDD is constructed based on the Transformer decoder and consists of four stacked decoder blocks, each of which includes a multi-head self-attention layer, a feedforward network layer, as well as residual connections and normalization. This design effectively suppresses noise and feature conflicts between modalities, thereby generating higher-level, contextually enhanced semantic representations and enhancing the model's sensitivity to key depressive behavioral patterns (such as micro-expressions, speech pauses, etc.)

3.2 Multimodal feature extraction

CHAMDD equips each modality with an optimized feature extractor based on Transformer to ensure the capture of the most discriminative information from the raw data.

3.2.1 Visual Feature Extraction via AUFormer

For the input video frame sequence, face detection and alignment are performed first. For each frame of the facial image $I_v \in \mathbb{R}^{H\times W\times 3}$ we employ the AUFormer model, which is a Vision Transformer optimized for Facial Action Unit (AU) detection. It captures both global and local dependencies of facial muscle movements through a multi-head self-attention mechanism. This process can be formalized as:

$$F_v = \text{AUFormer}(I_v) \in \mathbb{R}^{T_v\times D_v} \quad (1)$$

where T_u represents the number of time steps (frames), and D_u is the dimension of the visual features. These features encode both the geometric information of facial key points and the activation intensities of the AUs.

3.2.2 Audio Feature Extraction via AST

The original audio signal S_a is first converted into a log-Mel spectrogram $M \in \mathbb{R}^{T_a\times F}$, where T_a denotes the number of time frames and F denotes the number of frequency bins. The AST processes the spectrogram as a sequence of image patches:

$$F_a = \text{AST}(M) \in \mathbb{R}^{T_a\times D_a} \quad (2)$$

where the AST learns the time-frequency patterns within the spectrogram through its Transformer encoder, outputting high-level semantic audio features F_a.

3.2.3 Text Feature Extraction via LLaMA Encoder

The interview transcription text T is processed by the pre-trained Llama 2 model. We leverage its powerful language understanding capability to extract deep semantic and emotional features:

$$F_t = \text{Llama_Encoder}(T) \in \mathbb{R}^{L\times D_t} \quad (3)$$

where L denotes the sequence length, and D_t is the dimension of the text features. We typically take the representation of the [CLS] token or perform average pooling on the last hidden states to obtain a sequence-level representation $F_t \in \mathbb{R}^{D_t}$, which is then replicated along the time axis to align with the audio-visual features, resulting in $F_t \in \mathbb{R}^{T\times D_t}$.

3.3 Cross-hierarchical Attention Fusion (CHAF)

To overcome the limitations of simple concatenation or single-level attention fusion, we design a Cross-Modal Hierarchical Attention Fusion module, as illustrated in Figure 2. This module operates in two stages:

- Stage 1: Intra-modal Self-Attention, the features of each modality $F_m (m \in \{v, a, t\})$ independently pass through a self-attention layer to strengthen their internal structure and achieve deep understanding.

$$F_m^{self} = \text{LayerN}(F_m + \text{MHSelfAttn}(F_m)) \quad (4)$$

- Stage 2: Cross-modal Attention, we perform cross-attention computation by using one modality as the Query and the other modalities as the Key and Value. For example, fusion guided by the visual modality:

$$\begin{aligned} F_{v\to a} = \text{LayerN}(&F_v^{self} + \\ &\text{CrossAttn}(Q = F_v^{self}, \\ &K = F_a^{self}, V = F_a^{self})) \end{aligned} \quad (5)$$

Similar operations are applied for audio-guided and text-guided fusion. Finally, a gated summation mechanism dynamically aggregates all bimodal interaction results to generate the preliminary fused features $F_{fused} \in \mathbb{R}^{T\times D}$:

$$G = \sigma(W_g[F_{v\to a}; F_{a\to t}; F_{t\to v}]) \quad (6)$$

$$F_{fuse} = \sum_i g_i \cdot [F_{v\to a}; F_{a\to t}; F_{t\to v}] \quad (7)$$

Here σ denotes the sigmoid function, W_g represents learnable weights, and $[\cdot]$ denotes concatenation.

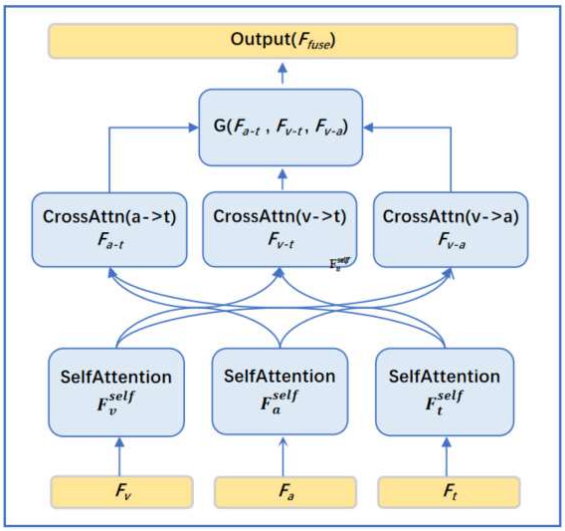


Fig. 2: The structure of cross-hierarchical attention fusion.

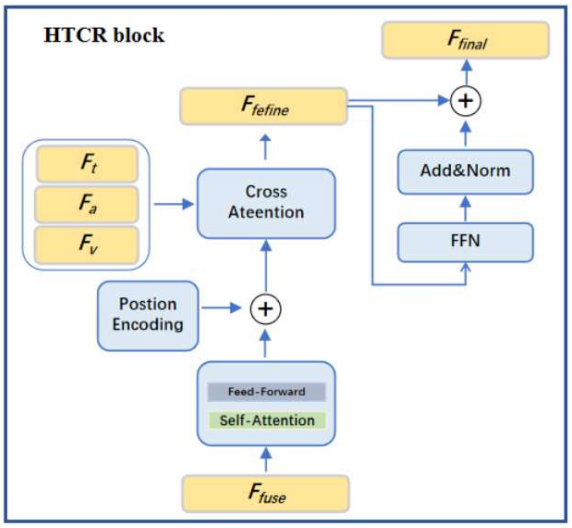


Fig. 3: The block of hierarchical temporal cross-modal refinement.

3.4 Hierarchical Temporal Cross-modal Refinement (HTCR)

As show in Figure 3, the initially fused features F_{fused} are fed into a hierarchical refinement module, derived from the Transformer decoder architecture, which has three iterative refinement.

1. Temporal Decoding Layer: A standard Transformer decoder layer (masked self-attention + feed-forward network) operates on F_{fused}, focusing on modeling the long-range temporal dynamics within the sequence.

$$F_{temp} = \text{T_Decoder}(F_{fused}) \tag{8}$$

2. Cross-modal Refinement Layer: This layer introduces a Cross-modal Residual Attention (CRA) mechanism. It takes F_{temp} as input and computes residual cross-attention with the original unimodal features F_v, F_a, F_t. This aims to supplement fine-grained modality-specific information, crucial for depression recognition,

that may have been lost during the initial fusion.

$$\text{Attn}_{re} = \text{S}\left(\frac{(F_{temp}W_Q)([F_v; F_a; F_t]W_K)^T}{\sqrt{d_k}}\right) \tag{9}$$

$$F_{refine} = F_{temp} + \text{Attn}_{re} \cdot ([F_v; F_a; F_t]W_V) \tag{10}$$

where S indicates Softmax.

3. Interaction Modeling Layer: A feed-forward network (FFN) performs a non-linear transformation and feature compression on $F_{refined}$, enhancing the model's expressive power and outputting the final refined features $F_{final} \in \mathbb{R}^{T\times D}$.

$$F_{final} = \text{LayerN}(F_{refine} + FFN(F_{refine})) \tag{11}$$

The entire HTRC module is formed by stacking N blocks composed of three layers like that, as show in Figure 3, achieving deep iterative refinement.

3.5 Depression Detection Head

The sparsified features F_{final} are fed into the detection head to perform depression binary classification and severity regression in parallel.

1)Multi-scale Temporal Pooling: We employ multi-scale 1D adaptive average pooling to extract global features at different temporal granularities from F_{final}, and concatenate them:

$$\begin{aligned}\mathbf{h} = [&\text{AdaptiveAvgPool1d}_{l_1}(F_{final});\\ &\ldots; \text{AdaptiveAvgPool1d}_{l_h}(F_{final})]\end{aligned} \tag{12}$$

where $l_1, \ldots, l_h$ are the target sequence lengths for different temporal granularities.

2)Parallel Prediction: The concatenated feature vector $\mathbf{h}$ is passed through two independent fully connected layers to generate the final predictions.

Classification Head: Outputs the probability of depression $p = \sigma(W_{cls}\mathbf{h} + b_{cls})$.

Regression Head: Outputs the predicted depression severity score $y_{reg} = W_{reg}\mathbf{h}+b_{reg}$ (e.g., a standardized PHQ-9 score).

The model is trained end-to-end using a multi-task loss:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{cls}(y, \hat{y}) + \beta \cdot \mathcal{L}_{reg}(s, \hat{s}) + \gamma \cdot \mathcal{L}_{aux} \tag{13}$$

Here, $\mathcal{L}_{cls}$ is the cross-entropy loss, $\mathcal{L}_{reg}$ is the smooth L1 loss, and $\mathcal{L}_{aux}$ is an auxiliary loss designed for auxiliary tasks (such as sentiment analysis).

4 Experiments

In this section, we evaluate the performance of CHAMDD on for depression detection. We will first introduce the dataset and evaluation metrics used in the experiments, followed by a detailed explanation of the experimental setup. Subsequently, we will present the comparative experimental results of CHAMDD against various state-of-the-art models on the LMVD dataset, and conduct analyses and discussions from multiple perspectives. Finally, through a series of ablation studies, we will deeply analyze the contribution of each key component within the model.

4.1 Datasets and Metrics

We employ the popular multimodal depression recognition benchmark dataset, LMVD [32]. This dataset comprises a large number of annotated clinical interview videos. Each participant provides synchronized facial video, audio recordings, and interview text transcriptions, accompanied by depression labels (binary classification: depressed/non-depressed) based on clinical assessments and severity scores (e.g., total PHQ-9 score). In our experiments, we follow the standard training, validation, and test sets.

To comprehensively evaluate model performance, we adopt the following metrics:

- Accuracy: The proportion of correctly predicted samples overall.
- F1 Score: The harmonic mean of precision and recall, which is more robust to class imbalance issues.
- AUC-ROC: Evaluates the model's comprehensive performance across all possible classification thresholds; values closer to 1 indicate better performance.
- RMSE: A commonly used measure of the difference between predicted severity scores and ground truth values; smaller values indicate better performance.
- PCC: Measures the linear correlation between predicted scores and ground truth scores; values closer to 1 indicate stronger correlation.

4.2 Experiment settings

Implementation Details as follows:

- Feature Extraction: Visual features are extracted using AUFormer, with input face images uniformly resized to 224×224 resolution. Audio features are extracted using AST, with input as 64-dimensional log Mel spectrograms. Text features are processed using Llama 2 tokenizer and pre-trained weights.
- Model Configuration: In the fusion and refinement modules of CHAMDD, the hidden dimension D of the Transformer layers is set to 768, with 12 attention heads. The HTCR module is stacked with $N = 3$ layers. Multi-scale pooling employs scales [1, 3, 5].
- Training: The AdamW optimizer is used with an initial learning rate of 5e-5, and a cosine annealing learning rate scheduler with warmup is adopted. The batch size is set to 16. Cross-entropy loss is used for classification tasks, and smooth L1 loss is used for regression tasks, with weights $\alpha = 1.0$ and $\beta = 0.5$. The model is trained on 4 NVIDIA A100 GPUs for 50 epochs, with early stopping based on validation F1 score.

We compare with the following representative state-of-the-art models, covering different fusion strategies and architectures:

- Late Fusion (LF): A baseline method that performs decision-level fusion after independent classification of each modality.
- TFN: A classic multimodal model based on tensor fusion.
- TAMFN: A multimodal fusion network incorporating time-aware attention.
- CANAMRF: A fusion model based on recurrent structures and adaptive weights.
- CFRDN: A text-centric model featuring feature decoupling and reconstruction.
- DepMamba: The current state-of-the-art model on this dataset, utilizing state space models for efficient sequential fusion.

4.3 Comparison between CHAMDD and the popular models

Table 1 presents a comparison between CHAMDD and other popular algorithms conducted on LMVD dataset, showing that CHAMDD achieves state-of-the-art performance. From Table 1 , we can make the following observations:

1) Performance Advantage: As shown in Table 1, CHAMDD significantly outperforms all baseline models across all five evaluation metrics. Particularly, compared to state-of-the-art model DepMamba, it achieves a 3.87 percentage point improvement in accuracy, a 0.039 increase in F1 score, and a 0.039 improvement in AUC. This fully validates the effectiveness of the cross-modal hierarchical attention fusion and hierarchical temporal-crossmodal refinement architecture.

2) Modality Fusion Advantages: The significant improvement of CHAMDD over DepMamba demonstrates that, when modeling cross-modal interactions, explicit, hierarchical attention mechanisms may be better at capturing the complex and asymmetric inter-modal dependencies required for depression recognition (such as subtle tonal changes inconsistent with brief facial expressions) compared to state space models.

3) Multi-task Benefits: In the severity regression task, CHAMDD also reduces RMSE by 0.32 and increases PCC by 0.054, indicating that the unified, refined feature representations generated by our framework are not only highly discriminative for classification tasks but also possess strong representational capabilities for continuous severity scores.

4.4 Comparison on the training process

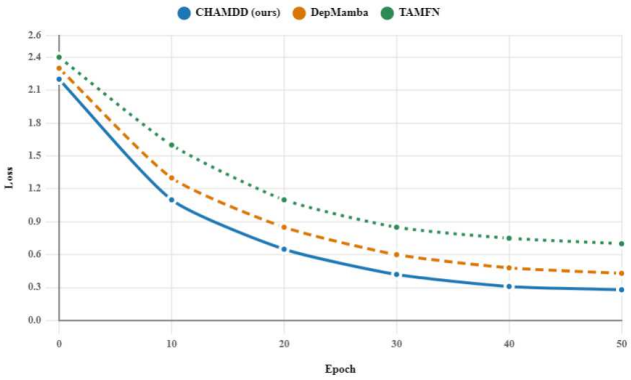


Fig. 4: The training loss.

We visualize the learning dynamics of the model during the training process, and plot the

Table 1: Comparison between CHAMDD and the baseline models on LMVD

Model	Accuracy (%)	F1	AUC	RMSE	PCC
Late Fusion	74.21	0.728	0.821	3.89	0.681
TFN	75.83	0.745	0.839	3.76	0.702
TAMFN	76.45	0.752	0.848	3.71	0.715
CANAMRF	77.12	0.761	0.857	3.65	0.728
CFRDN	77.90	0.770	0.863	3.59	0.736
DepMamba	79.01	0.783	0.875	**3.50**	0.751
CHAMDD (Ours)	**82.88**	**0.822**	**0.914**	3.18	**0.805**

Table 2: Comparison of Model Training Efficiency and Resource Consumption

Model	Params (M)	GPU Memory (GB)	Avg. Time per Epoch (s)	Final Loss
TAMFN	85.2	10.1	125	0.70
DepMamba	78.5	9.3	98	0.43
CHAMDD	**142.7**	**15.6**	**156**	**0.28**

curves of key indicators as shown in Figure 4. Figure 4 clearly illustrates the comparison of performance and resource consumption during the model training process. The loss curve graph depicts the loss reduction trend of three models, CHAMDD, DepMamba, and TAMFN, over 50 training epochs. By utilizing various colors and line types (solid/dashed), we can visually compare their convergence speeds and loss values. The peak video memory occupancy of each model is listed in the form of cards below the chart. The data shows that CHAMDD achieves faster loss reduction while also having lower video memory occupancy compared to the comparative models, which echoes the advantage described as "saving more than 12% of video memory".

Table 2 reports total parameter count (millions), peak GPU memory usage during training (GB), average time per training epoch (seconds), and final training loss for each model. CHAMDD achieves superior convergence (lowest final loss) at the cost of increased parameters and memory usage, demonstrating a favorable performance-efficiency trade-off.

From the figure 4 and table 2, we can make the following observations:

1) Convergence and performance: As shown in Figure 4, the loss curve of CHAMDD not only starts from a lower point (benefiting from better initialized features), but also decreases faster and ultimately converges to a lower loss value, which directly corresponds to its superior performance on the test set.

2) Trade-off between Efficiency and Effectiveness: Table 2 shows that due to its more complex hierarchical structure, CHAMDD has a higher parameter count and memory footprint than the comparison model, and its training time is also correspondingly longer. This reflects the "effectiveness-first" approach in model design. Nevertheless, its convergence speed (the number of epochs required to achieve similar performance) is faster, and the final performance improvement is significant. The increased resource overhead yields substantial performance benefits, which is acceptable in many practical applications

4.5 Ablation experiment

To verify the necessity of each core component in CHAMDD, we conducted a systematic ablation experiment, and the results are summarized in Table 3. From the table, we can observe that:

1) Fusion Module is Crucial: Variant B shows the most significant performance drop (-3.18%), strongly demonstrating that our proposed cross-modal hierarchical attention fusion is the primary contributor to model performance. Its two-stage "self-attention + cross-attention" design is indispensable for achieving effective cross-modal information interaction.

2) All Components of the Refinement Module Contribute: The performance degradation in variants C and D indicates that cross-modal residual

Table 3: Ablation Study of CHAMDD Model (Accuracy on Validation Set)

Exp. ID	Model Variant Description	Accuracy (%)	Relative Drop (pp)
A	**Full CHAMDD**	**81.23**	**-**
B	Remove **cross-modal hierarchical fusion**	78.05	-3.18
C	Remove **cross-modal residual attention**	79.41	-1.82
D	Remove **temporal decoding layers**	79.88	-1.35
E	Replace multi-scale pooling	80.17	-1.06
F	Remove **severity regression task**	80.65	-0.58

attention and temporal decoding layers enhance feature quality from two different but important perspectives: supplementing fine-grained information and modeling temporal dynamics, respectively.

3) Gains from Multi-task and Multi-scale Learning: The performance drops in variants E and F are relatively smaller but still observable, suggesting that multi-scale temporal pooling helps capture contextual information at different granularities, while the severity regression auxiliary task provides beneficial supervisory signals for regularizing the primary classification task through learning from continuous labels.

4.6 Analysis of Performance and Effectiveness

Figure 5 is visualization of attention map for key variants in ablation experiments. Cross-modal attention weights (video, audio) for the same interview segment (LMVD_test_0427, depressed) under three configurations. (A) Full CHAMDD with CHAF + HTCR: attention is sharply focused, highlighting strong associations between depressive AUs (AU4 brow lowerer, AU15 lip corner depressor) and audio prosodic features (pitch/energy) at critical moments. (B) w/o CHAF (simple concatenation + single attention): attention is diffuse and unstructured, failing to capture meaningful cross-modal interactions. (C) w/o CRA (no residual refinement): attention shows partial focus but lacks the fine-grained alignment seen in (A). Quantitative metrics confirm the qualitative observations: lower entropy and higher focus score correlate with better performance. Visual comparison shows that the attention distribution of the complete CHAMDD model is more concentrated and relevant across modalities (such as audio and video), indicating that it can better align emotional cues in different modalities. Variants that remove hierarchical fusion or CRA exhibit more diffuse attention maps or weaker inter-modality correlation.

4.7 Discussion and Limitations

Based on the comprehensive experimental results above, CHAMDD demonstrates exceptional multimodal depression detection capability. Its success can be attributed to:

1) A unified and powerful Transformer backbone, maintaining a consistently high-performance architecture from feature extraction to fusion and refinement;

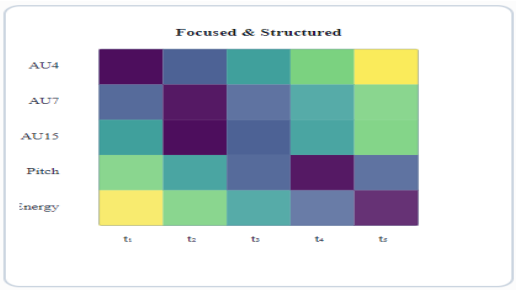
2) A hierarchical and fine-grained interaction design, simulating the cognitive process from local understanding to global reasoning;

3) A multi-task learning framework, effectively leveraging different supervisory signals within the data.

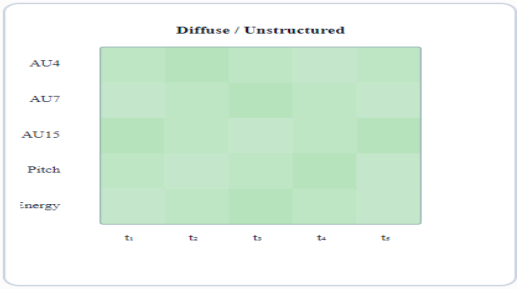
However, this study also has certain limitations: Firstly, the model complexity and computational cost are relatively high, posing challenges for deployment in resource-constrained environments. Secondly, experiments have currently only been conducted on the publicly available LMVD dataset; the model's generalizability needs to be validated on more diverse and larger-scale real clinical data. Future work will focus on model lightweighting, exploring more efficient novel attention mechanisms, and investigating generalizability across datasets and cross-cultural scenarios.

5 Conclusion

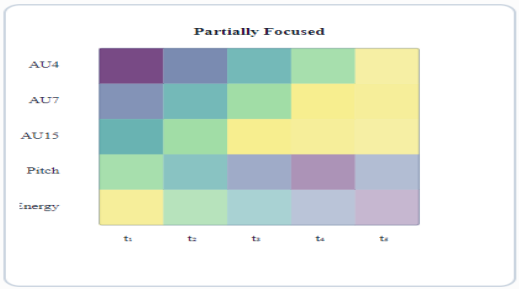
This paper proposes a novel Cross-modal Hierarchical Attention Multi-modal Depression Detection (CHAMDD) framework to address the



(a) Full CHAMDD Acc: 84.2%



(b) w/o CHAF Acc: 80.1%



(c) w/o CRA Acc: 81.2%

Fig. 5: Attention Map Visualization for Ablation Study Variants.

challenges of insufficient cross-modal interaction and inadequate fine-grained feature learning. Built entirely upon Transformer architecture, CHAMDD employs AUFormer, AST, and Llama 2 for visual, audio, and text modality encoding, ensuring consistent and advanced feature extraction. A two-stage hierarchical fusion strategy is introduced, where intra-modal self-attention first refines each modality independently, followed by cross-modal attention for deep alignment, effectively capturing complex cross-modal dependencies. Subsequently, a three-layer iterative refinement module with temporal decoding, cross-modal residual attention, and interaction modeling layers explicitly models temporal dynamics and supplements fine-grained information, yielding highly discriminative features. Extensive experiments on the LMVD dataset demonstrate that CHAMDD achieves state-of-the-art performance in both depression classification and severity regression tasks, outperforming existing models across all metrics. Ablation studies validate the necessity of each component, and visualization analyses confirm the model's superiority in capturing cross-modal correlations.

Declarations

Funding

This work was supported by (i)Natural Science Foundation of Shenzhen Municipality(Grant No.JCYJ20250604184310015) (ii)Shenzhen Institute of Technology School-level Project(Grant No.2611010).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Ethics declaration

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

This study is a secondary analysis of data collected in an earlier study. The dataset is publicly available at: https://github.com/helang818/LMVD/ [32].

References

[1] Kroenke, K., Spitzer, R.L.: The phq-9: A new depression diagnostic and severity measure. Psychiatric Annals **32**(9), 509–515 (2002)

[2] Williams, L.S., Brizendine, E.J., Plue, L., Bakas, T., W. Tu, H.H., Kroenke, K.: Performance of the phq-9 as a screening tool for depression after stroke. stroke **36**(3), 635–638 (2005)

[3] Wang, L., Chignell, M., Jiang, H., Lokuge, S., Mason, G., Fotinos, K., , Katzman, M.: Discovering the causal structure of the hamilton rating scale for depression using causal discovery. IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 1–4 (2021) https://doi.org/10.1109/BHI50953.2021.9508538

[4] He, L., Guo, C., Tiwari, P., Pandey, H.M., Dang, W.: Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. International Journal of Intelligent Systems **37**(12), 10140–10156 (2021) https://doi.org/10.1002/int.22426

[5] Suto, T., Fukuda, M., Ito, M., Uehara, T., Mikuni, M.: Multichannel near-infrared spectroscopy in depression and schizophrenia: cognitive brain activation study. Biological psychiatry **55**(5), 501–511 (2004)

[6] Tahir, W.B., Khalid, S., Almutairi, S.S., Abohashrh, M., Memon, S.A., Khan, J.: Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. IEEE Access **13**, 12789–12818 (2025) https://doi.org/10.1109/ACCESS.2025.3530862

[7] He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., Pan, X., Dang, W.: Deep learning for depression recognition with audiovisual cues: A review. Information Fusion **88**, 56–86 (2022) https://doi.org/10.1016/j.inffus.2021.10.012

[8] Nykoniuk, M., Basystiuk, O., Shakhovska, N., Melnykova, N.: Multimodal data fusion for depression detection approach. Computation **13**(9), 1–18 (2024) https://doi.org/10.3390/computation13010009

[9] Jiang, Y., Zhang, Z., Sun, X.: A multimodal dataset for depression and anxiety detection. ICPR Workshops **1**, 691–702 (2022) https://doi.org/10.1007/978-3-031-37660-3_49

[10] Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. CLPsych@NAACL-HTL, 88–97 (2018) https://doi.org/10.18653/v1/W18-0609

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[12] Cho, K., Merrieenboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734 (2014) https://doi.org/10.48550/arXiv.1406.1078

[13] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (2021) https://doi.org/10.1162/neco.1997.9.8.1735

[14] Ye, J., Zhang, J., Shan, H.: Depmamba: Progressive fusion mamba for multimodal depression detection. ICASSP, 1–5 (2025) https://doi.org/10.1109/ICASSP49660.2025.10889975

[15] Zhou, X., Jin, K., Shang, Y., Guo, G.: Visually interpretable representation learning for depression recognition from facial images. IEEE Transactions on Affective Computing **11**(3), 542–552 (2020) https://doi.org/10.1109/TAFFC.2018.2828819

[16] Ma, S., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: An efficient deep

model for audio based depression classification. AVEC@ACM Multimedia, 35–42 (2016) https://doi.org/10.1145/2988257.2988267

[17] Niu, J., Liu, B., Tao, J., Li, Q.: A time-frequency channel attention and vectorization network for automatic depression level prediction. Neurocomputing **450**, 208–218 (2021) https://doi.org/10.1016/j.neucom.2021.04.056

[18] Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F.: Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. Journal of personality assessment **67**(3), 588–597 (1996)

[19] Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies **1**, 4171–4186 (2019) https://doi.org/10.48550/arXiv.1810.04805

[20] Xiao, Z., Zhang, X., Sun, G., Shao, H.: Ftan: Feature transform and alignment network for cross-domain specific emitter identification. Signal Process **230**, 106800–109810 (2025) https://doi.org/10.1016/j.sigpro.2024.109800

[21] Wang, Y., Wang, Z., Li, C., Zhang, Y., Wang, H.: A multimodal feature fusion-based method for individual depression detection on sina weibo. IPCCC, 1–8 (2020) https://doi.org/10.1109/IPCCC50635.2020.9391501

[22] Zadeh, A., M. Chen, S.P., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. EMNLP, 1103–1114 (2017)

[23] Su, L., Hu, C., Li, G., Cao, D.: Msaf: Multimodal split attention fusion. CoRR abs/2012.07175 (2020)

[24] Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., L, Y., Hu, B.: Tamfn: Time-aware attention multimodal fusion network for depression detection. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 669–679 (2022)

[25] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics **2**, 207–212 (2016) https://doi.org/10.18653/v1/P16-2034

[26] Yang, L., Jiang, D., Sahli, H.: Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access, 24033–24045 (2020) https://doi.org/10.1109/ACCESS.2020.2970496

[27] Wei, Y., Zhang, Y., Zhang, S., Zhang, H.: Canamrf: An attention-based model for multimodal depression detection. PRICAI **2**, 111–116 (2023) https://doi.org/10.1007/978-981-99-7022-3_10

[28] Fan, H., Zhang, X., Xu, Y., Fang, J., Zhang, S., Zhao, X., Yu, J.: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion **104**, 102161–102169 (2024) https://doi.org/10.1016/j.inffus.2023.102161

[29] Gong, Y., Chung, Y., Glass, J.: AST: Audio Spectrogram Transformer. In: Interspeech 2021, pp. 571–575 (2021). https://doi.org/10.21437/Interspeech.2021-698

[30] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ICML, 8748–8763 (2021)

[31] Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS, 9694–9705 (2021)

[32] He, L., Chen, K., Zhao, J., Wang, Y., Pei, E., Chen, H., Jiang, J., Zhang, S., Zhang, J., Wang, Z., He, T., Tiwari, P.: Lmvd:

A large-scale multimodal vlog dataset for depression detection in the wild. Information Fusion, 103632–103642 (2026) https://doi.org/10.1016/j.inffus.2025.103632