Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077057

ARTICLE

SYMPHONIA–Enhanced Multimodal Emotion Recognition with Dual-Branch Dynamic Attention and Hierarchical Adaptive Fusion

Akmalbek Abdusalomov[1], Mukhriddin Mukhiddinov[2,3], Kamola Abdurashidova[2], Alpamis Kutlimuratov[4], Avazjon Marakhimov[5], Kuanishbay Seytnazarov[6] and Young-Im Cho[1,*]

[1]Department of Computer Engineering, Gachon University Sujeong-Gu, Seongnam-si, Gyeonggi-Do, Republic of Korea
[2]Department of Computer Systems, Tashkent University of Information Technologies Named after Muhammad Al-Khwarizmi, Tashkent, Uzbekistan
[3]Department of Industrial Management and Digital Technologies, Nordic International University, Tashkent, Uzbekistan
[4]Department of Applied Informatics, Kimyo International University in Tashkent, Tashkent, Uzbekistan
[5]Department of Information Processing and Management Systems, Tashkent State Technical University, Tashkent, Uzbekistan
[6]Department of General Education Disciplines and Distance Education, Nukus State Pedagogical Institute Named after Ajiniyaz, Nukus, Uzbekistan

*Corresponding Author: Young-Im Cho. Email: yicho@gachon.ac.kr

Received: 01 December 2025; Accepted: 03 March 2026; Published: 08 May 2026

ABSTRACT: Human emotions are intricate and difficult to decipher through various modalities. Current methodologies frequently employ inflexible fusion strategies that do not consider the dynamic and context-sensitive characteristics of emotional expressions in both visual and textual mediums. This paper presents SYMPHONIA (Synchronizing Facial and Textual Modalities for Emotion Understanding), an innovative architecture engineered to capture and amalgamate emotional signals from facial expressions and language, attuned to contextual and modality interactions. There are two parts to SYMPHONIA: a Facial Emotion Branch that uses Vision Transformers and facial landmarks, and a Textual Emotion Branch that uses RoBERTa embeddings and graph-based reasoning. A Dual-Branch Dynamic Attention Mechanism and a Hierarchical Adaptive Fusion Module are used to connect these branches. SYMPHONIA beat the best models on four datasets: IEMOCAP, MELD, CMU-MOSI, and CMU-MOSEI. It got 80.9% accuracy and 80.1% F1-score on IEMOCAP, which was better than Dualgats (74.8%) and EmoCLIP (75.3%). SYMPHONIA got 74.2% accuracy and 73.5% F1-score for MELD. It beat its competitors by getting a 0.86 Pearson correlation on MOSI and a 0.83 on MOSEI for predicting sentiment. Cross-dataset tests showed that SYMPHONIA could generalize, with 66.9% accuracy when trained on IEMOCAP and tested on MELD. This was better than all the baselines. These results show that SYMPHONIA is good at recognizing emotions and analyzing sentiment in different situations, which shows that it can adapt and do well in different settings.

KEYWORDS: Multimodal emotion recognition; RoBERTa; cross-modal attention; graph neural networks; contrastive learning; adaptive fusion; temporal modeling; affective computing; context-aware representation

1 Introduction

Emotion recognition systems represent a fundamental component of human–computer interaction (HCI) [1] and have significant applications in affective computing [2], virtual agents [3], behavioral analysis [4], and mental health diagnostics [5]. Earlier approaches primarily focused on facial expression recognition [6,7], textual analysis [8], and speech analysis [9], which collectively laid the foundation for emotion recognition research [10,11]. However, these unimodal systems often lacked robustness to noise,

ambiguity, and cross-modal biases [12]. In contrast, multimodal emotion recognition (MER) seeks to extract and integrate emotional cues from heterogeneous yet complementary modalities, thereby improving the reliability of emotion inference [13]. Despite the advances in MER, many existing models still rely on static fusion strategies or simplistic cross-modal interaction mechanisms [14]. Such approaches fail to capture the inherently dynamic and context-sensitive interplay between modalities. For instance, textual content may describe a smile as a joyful expression while simultaneously masking sarcasm, the interpretation of which may depend on preceding conversational context [15]. Furthermore, numerous multimodal feature integration methods treat features from different modalities as equally important regardless of contextual relevance. This assumption limits the model's ability to adaptively emphasize the most informative emotional indicators specific to a given scenario [16].

To deal with these constraint issues, we develop SYMPHONIA, a new multimodal framework capable of dynamically synthesizing facial expression and text emotion signals. The architecture consists of two modality-specific branches: A Facial Emotion Branch which employs Vision Transformers (ViT) with Landmark Guided Attention plus LSTM temporal modeling, and a Textual Emotion Branch derived from RoBERTa embeddings [16] with a TSG built with GAT. Both branches are strongly integrated via a Dual-Branch Dynamic Attention Mechanism which allows cross-modal, context-sensitive interactions to control influence between modalities. The hierarchical adaptive fusion module operates by aligning and integrating multimodal features across multiple layers. It labels features at different semantic levels, using adaptive control and contrastive self-supervised learning to ensure precise and efficient fusion.

By combining classification and regression methods, SYMPHONIA consistently surpasses current state-of-the-art models, achieving superior results across multiple datasets, including IEMOCAP, MELD, CMU-MOSI, and CMU-MOSEI. The redesigned framework has significantly enhanced both accuracy and long-term stability through a modular architecture and dynamic attention mechanisms. These improvements highlight the model's adaptability and its sensitivity to emotional context. SYMPHONIA's contribution reflects the evolution of affective computing, where emotion recognition has advanced toward a more complex, interdisciplinary approach supported by flexible and adaptive algorithmic solutions. By doing this, SYMPHONIA makes a strong starting point for more study and for growing technology that can understand feelings better in the future.

2 Related Works

Integrating information from various sources, such as video recordings that capture both speech patterns and facial expressions, greatly enhances emotion recognition models. This integration allows such systems to gain a deeper understanding of human emotions, contributing to the growing interest in this field [17]. However, traditional approaches that rely on a single modality, such as visual [18] or linguistic features, often struggle with reliability and generalization, particularly in noisy or uncertain environments [19,20]. These limitations have shifted research toward multimodal approaches, which offer a more accurate and contextually nuanced interpretation of emotions [21]. Facial expression recognition has traditionally been carried out using Convolutional Neural Networks (CNNs), which extract spatial features from static images or video sequences [22,23]. More recent studies, however, have shown that Vision Transformers (ViTs) [24] are significantly more effective. ViTs employ self-attention mechanisms to capture long-range relationships across different regions of the face [25]. Focusing attention on key landmarks that represent emotional expressions has also been shown to improve performance [26], allowing models to concentrate on critical areas such as the mouth, eyes, and eyebrows [27] for more precise emotional analysis. The growth of textual emotion recognition systems has paralleled the development of natural language processing, especially with the introduction of pre-trained transformer models like BERT and RoBERTa [28].

Barnet and his colleagues attest to the fact that these models provide rich contextual embeddings able to capture subtle semantic and syntactic relationships important for emotion classification tasks [29,30]. In spite of this, these models do not inter-relate the tokens flexibly nor adequately. This especially is important for understanding more complex emotional semantics in language. In response to these problems, higher order token interactions have been modeled by Graph Attention Networks (GATs) which use graph-based techniques and improve the explainability as well as the discrimination power of textual emotions [31]. MER, as with many other tasks, faces challenges with multimodal fusion even after the individual modalities have been sufficiently advanced [32]. As with many other tasks, MER faces challenges with multi-modal fusion even after the individual modalities have been sufficiently advanced. Strategies such as early fusion (feature-level concatenation) and late fusion (decision-level aggregation) often overlook the relational hierarchies and the intricate interdependencies of the modalities and their relative importance across different scenarios portraying emotions [33]. The problem of incorporating modality interactions has been addressed by TFN (tensor-based fusion networks) [34] and memory fusion networks (MFN) [35]. These approaches however still suffer from static representation bottlenecks and inefficient computations [36,37].

To address these issues, we introduce the SYMPHONIA framework (Facial-Textual Emotion Recognition), which incorporates the face and text modalities using: (i) dual branch dynamic attention mechanism with bidirectional context-aware modulation and (ii) hierarchical adaptive fusion based on contrastive self-supervised learning. As with all approaches, SYMPHONIA differs by emphasizing dynamically the most contextually relevant emotional interactions as intermodal contextual engagement and representation alignment at multiple semantics. All of these developments raise the bar for emotion recognition performance across complex systems by strengthening the model's interpretability, flexibility, and context understanding.

3 Proposed Model

We introduce the SYMPHONIA framework to address the limitations of current multimodal emotion recognition methods, such as their rigid fusion strategies and lack of context-aware adaptation. This innovative architecture provides a more flexible and responsive approach by dynamically and hierarchically merging textual and facial emotional information. Consequently, the final emotional embedding turns into a strong and evocative representation of the textual modality. The adaptive multimodal fusion process that follows is improved by the insightful information this description offers.

The SYMPHONIA architecture comprises several vital components that function synergistically to extract, align, and integrate features on a modality level, thus providing robust, transparent, and contextually intelligent emotion recognition. More specifically, SYMPHONIA consists of: (i) Facial Emotion Branch that captures expressive visual dynamics through Vision Transformers with landmark-guided temporal attention and LSTM temporal modeling; (ii) Textual Emotion Branch using RoBERTa embeddings with temporal semantic interactions modeled by GAT to a temporal semantic graph; (iii) a bidirectional dynamic attention mechanism for context-sensitive modulation between modalities (Dual-Branch Dynamic Attention Mechanism); and (iv) a Hierarchical Adaptive Fusion Module that integrates multimodal features on several semantic levels with adaptive gating through contrastive self-supervised learning. Emotional categories are predicted based on the fused representation by the classification layer built on top of a transformer backbone. The subsequent subsections systematically describe each component of the SYMPHONIA framework, outlining the design rationale, architectural structure, and functional role within the complete emotion recognition pipeline (Fig. 1).

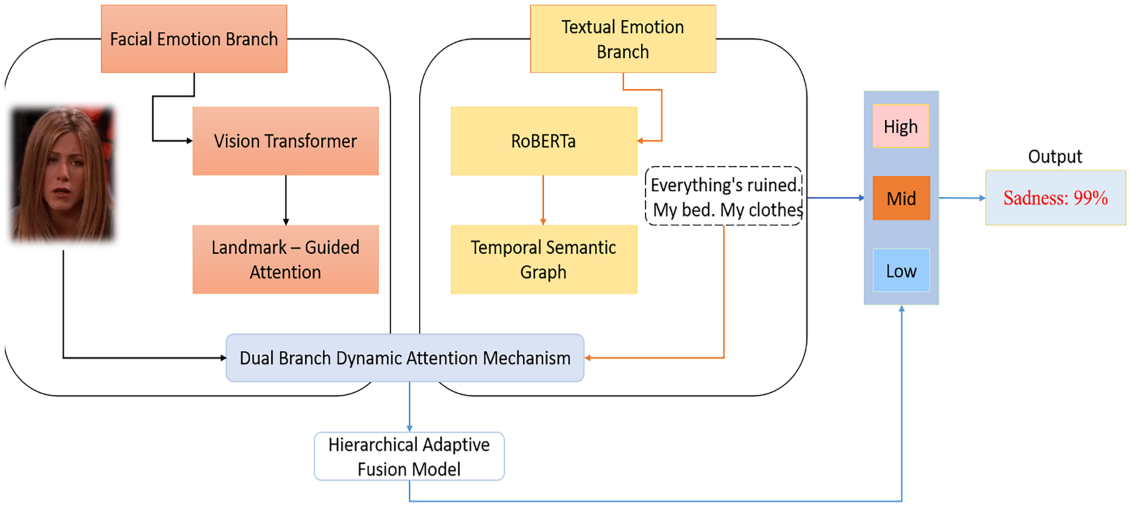

Figure 1: Overview of the SYMPHONIA framework. The architecture comprises two modality-specific branches: a facial emotion processing stream using ViT, facial landmark-guided attention, and LSTM-based temporal modeling; and a textual emotion branch leveraging RoBERTa embeddings enriched through a GAT. These branches are dynamically integrated via a Dual-Branch Dynamic Attention Mechanism and further aligned through a Hierarchical Adaptive Fusion Module operating at low, mid, and high semantic levels. The final output yields a precise emotion prediction, illustrated here with the classification result "Surprise: 96%".

3.1 Facial Emotion Branch

The main aim of the facial emotion branch is to capture expressive and discriminative facial features as well as model their temporal changes over time in order to interpret emotions accurately from visual data. We utilize a ViT for deep face feature extraction because of the impressive long-distance dependency capturing within regions of a face. ViT works by dividing the input face image into patches of a given size, embedding each patch, and then attending to them relation-wise with self-attention. As illustrated in Fig. 2, the attention weight matrix generated by a Vision Transformer head highlights how different facial patches attend to one another, thereby emphasizing salient regions that contribute most significantly to emotion inference. This attention-driven mechanism enhances the model's ability to capture subtle facial dynamics and contextual inter-patch relationships.

An input face image $I \in R^{H\times W\times C}$ where H, W, C represent height, width, and channels, respectively, partitioned into a sequence of N non-overlapping patches $\left\{I_p^i\right\}_{i=1}^{N}$ where each patch is of size P × P × C:$N = \frac{HW}{P^2}$. Each patch I_p^i is flattened and linearly projected into a D-dimensional embedding vector e_i:

$$e_i = LinearProj\left(I_p^i\right), e_i \in R^D \tag{1}$$

The ViT further prepends a learnable class embedding e_{cls} to the sequence of patch embeddings and adds positional embeddings e_{pos}^i to retain positional information:

$$Z_0 = \left[e_{cls}; e_1; e_2; \ldots; e_N\right] + e_{pos} \tag{2}$$

The Transformer encoder consists of multiple stacked layers comprising multi-head self-attention (*MHSA*) and multi-layer perceptron (*MLP*) blocks, calculated as follows:

$$Z_l' = MHSA\left(LN\left(Z_{l-1}\right)\right) + Z_{l-1} \tag{3}$$

$$Z_l = MLP\left(LN\left(Z_l'\right)\right) + Z_l' \tag{4}$$

here, Z_l is the output of layer l, LN represents layer normalization, and $l = 1, \ldots, L$ with L being the total Transformer layers. For highlighting pivotal emotional zones, we utilize Facial Landmark Attention (FLA).

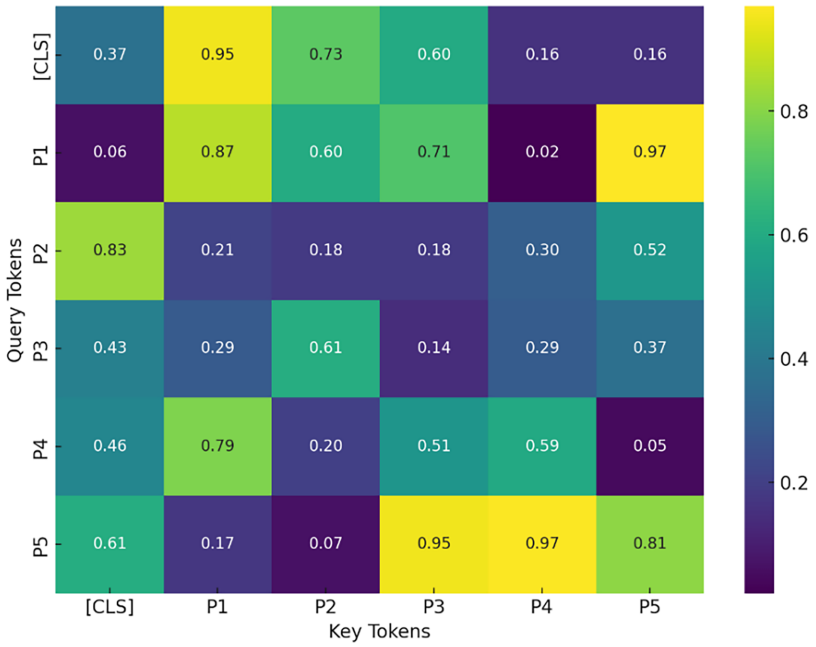

Figure 2: Attention weight matrix from vision transformer head in the facial emotion branch.

Utilizing facial landmark coordinates retrieved from a landmark detection model, we create attention masks which highlight landmarks located on the expression bearing regions which include eyes, eyebrows, and mouth:

$$A_{landmark} = \sigma\left(Conv\left(L\right)\right), A_{landmark} \in [0,1]^{H \times W} \tag{5}$$

where L represents a binary landmark heatmap, $Conv$ is a convolutional operation, and σ denotes the sigmoid activation function. The landmark attention mask is applied to each embedding vector e_i:

$$e_i' = e_i \odot Pooling\left(A_{landmark}\right) \tag{6}$$

This mechanism further enhances and reinforces the activation of emotionally expressive facial regions. To model temporal dependencies within facial embeddings over time, we employ an LSTM network. Considering the facial embedding sequence obtained from ViT for each frame t:

$$E_t = \left[e_{cls}', e_1', e_2', \ldots, e_N'\right], E_t \in R^{(N+1) \times D} \tag{7}$$

We utilize the class embedding e_{cls}' as it effectively summarizes global emotional context: $x_t = e_{cls}', x_t \in R^D$. LSTM processes this sequence temporally. Given an input sequence $X = [x_1, x_2, \ldots, x_T]$, the LSTM computes hidden states recursively as:

$$f_t = \sigma\left(W_f \times \left[h_{t-1}, x_t\right] + b_f\right) \tag{8}$$

$$i_t = \sigma\left(W_i \times \left[h_{t-1}, x_t\right] + b_i\right) \tag{9}$$

$$o_t = \sigma\left(W_o \times \left[h_{t-1,} x_t\right] + b_0\right) \tag{10}$$

$$\tilde{C}_t = tanh\left(W_c \times \left[h_{t-1}, x_t\right] + b_c\right) \tag{11}$$

$$C_t = f_t \odot C_{t-1} + i_t \bigodot \tilde{C}_t \tag{12}$$

$$h_t = o_t \odot tanh\left(C_t\right) \tag{13}$$

where f_t, i_t, o_t are forget, input, and output gates, respectively. C_t represents the cell state at time step t, and h_t is the hidden state. W_f, W_i, W_o, W_c, and b_f, b_i, b_o, b_c, are learnable parameters, σ represents sigmoid activation. The facial emotion branch generates temporally-aware and expression-focused embeddings through the integration of ViT feature extraction, landmark-based attention, and LSTM temporal modeling. These embeddings are then applied within the subsequent fusion and emotion classification of the SYMPHONIA model.

Fig. 3 displays the changes in emotion activation scores over time and across sequential facial frames as processed by the LSTM module within the Facial Emotion Branch. Every point reflects the internal estimation of an emotion by the model at a given time step. The LSTM's ability to model even the most minute changes over time is important for distinguishing dynamic from static emotion recognition.

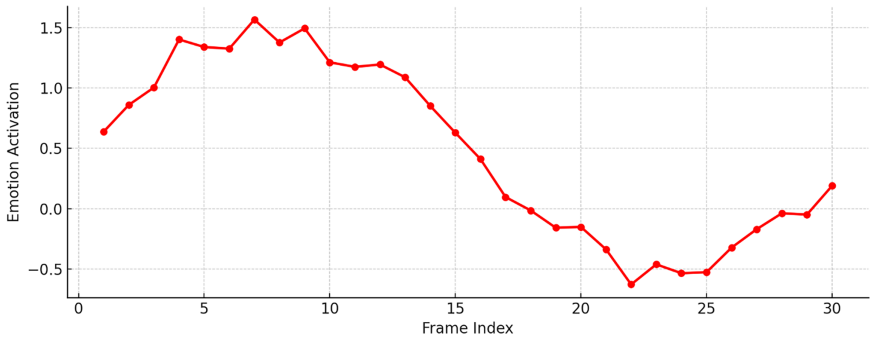

Figure 3: Temporal emotion activation curve from the facial LSTM module.

3.2 Textual Emotion Branch

The textual emotion branch is designed to extract semantically meaningful and context-aware embeddings that capture the subtle emotional nuances present in textual input. By leveraging advanced language modeling and graph-based relational learning, the framework achieves deep semantic understanding and explicitly models emotional dependencies among textual tokens. Textual modality embeddings are generated using RoBERTa, a Transformer-based encoder pretrained on large-scale corpora and widely recognized for its strong contextual representation capabilities. Given an input sentence S consisting of words (tokens) $w_1, w_2, \ldots, w_M$, where M is the total number of tokens, *RoBERTa* produces a sequence of contextual embeddings as:

$$E^{RoBERTa} = RoBERTa\left(\left[w_1, w_2, \ldots, w_M\right]\right), E^{RoBERTa} \in R^{M \times D_{text}} \tag{14}$$

here, D_{text} represents the dimensionality of RoBERTa embeddings. Each word embedding $e_m^{RoBERTa}$ captures semantic meaning and contextual nuances of the corresponding token w_M. To explicitly model relational and semantic interactions between words and enhance emotion-focused contextual representation, we introduce a TSG based on GAT. TSG effectively captures word-to-word emotional dependencies, enhancing the understanding of emotional nuances in textual sequences. The semantic graph $\mathcal{G} = (v, \varepsilon)$ is constructed

from the tokens of sentence S, where nodes $V = (v_1, v_2, \ldots, v_M)$ correspond to token embeddings from RoBERTa, and edges $\varepsilon \subseteq v \times v$ represent pairwise semantic interactions between tokens.

We define edge features between two nodes (words) as the scaled dot-product similarity between their corresponding RoBERTa embeddings, representing the degree of semantic relationship:

$$\alpha_{ij} = sortmax\left(\frac{\left(W_q e_i^{RoBERTa}\right) \times \left(W_k e_j^{RoBERTa}\right)}{\sqrt{d_k}}\right) \tag{15}$$

where W_q, W_k represent the parameter matrices that learn the embeddings and transform them into query and key vectors, respectively. d_k refers to the dimensionality of the transformed embeddings. α_{ij} denotes attention weight representing semantic connectivity from node i to node j.

The heatmap in Fig. 4 visualizes the attention weights calculated by the self-attention mechanism of RoBERTa, or by the TSG using a GAT, over a sample sentence. Each cell contains the attention strength from a query token (row) to a key token (column), thus showing which relationships among words the model focuses on while reasoning about emotions. Cumulatively, attention allocation for emotionally salient words like "happy" is higher, influencing the resultant emotion embedding.

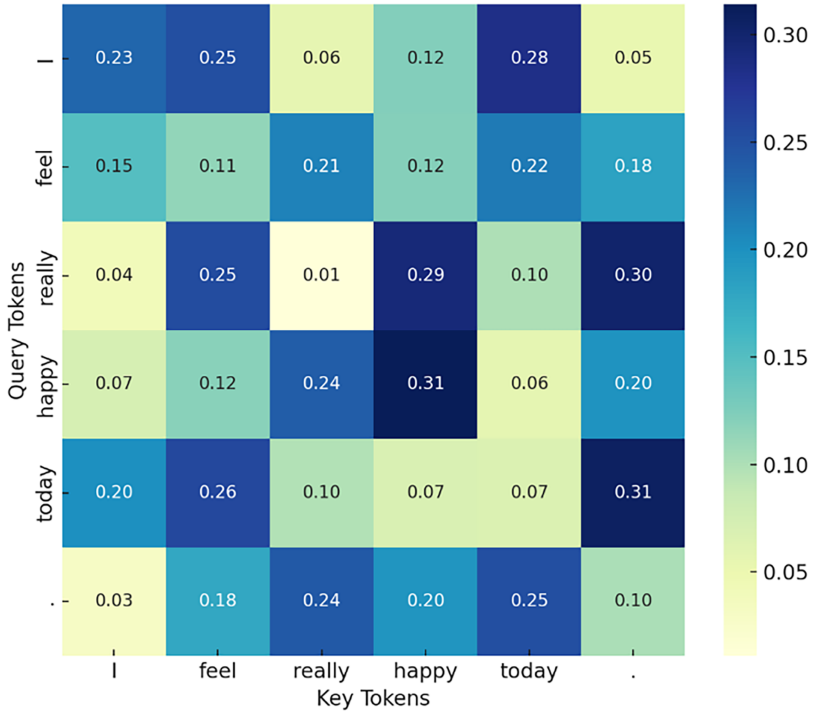

Figure 4: Token-level attention heatmap in the textual emotion branch.

We employ a multi-head Graph Attention Network to aggregate information from neighbor nodes, enhancing each token's embedding with its contextual semantic information. The GAT aggregation rule for updating node features is formulated as follows:

$$e_i' = ||_{k=1}^{K} \sigma\left(\sum_{j \in N_i} \alpha_{ij}^k W_v^k e_j^{RoBERTa}\right) \tag{16}$$

where e_i' denotes the updated embedding for token i, concatenated ($\|$) from K attention heads. N_i represents neighboring nodes (tokens) connected to node i. α_{ij}^k denotes attention weights computed by head k. W_v^k are learnable parameters (value matrices) for attention head k. σ denotes a non-linear activation function.

Fig. 5 presents the semantic graph illustrating the emotional states and contextual relationships associated with tokens in a sample sentence, as processed by the Textual Emotion Branch using a GAT. Each token is a node, and directed edges represent learned semantic relations endowed with weights of attention scores. Emotionally important words like "happy" can be extracted by the model through information aggregation from associated tokens with the help of contextual information. Thus, the representation of emotion is improved over the representation of words with the graph and context. For stability and efficiency, we typically average multi-head results:

$$e_i'' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k W_v^k e_j^{RoBERTa}\right) \tag{17}$$

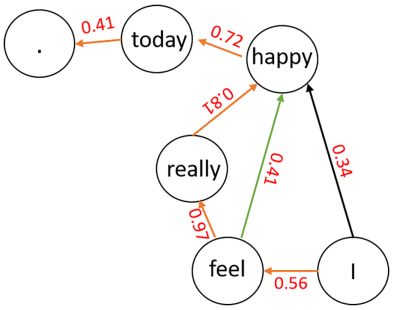

Figure 5: Semantic graph of token interactions from the TTSG.

To capture the entire sentence embedding, we utilize a weighted average pooling of the updated node embeddings e_i''. The graph attention mechanism pools express emotional salience with greater weight to certain words:

$$e_{final}^{text} = \sum_{i=1}^{M}\beta_i e_i'', \text{where } \beta_i = softmax\left(W_\beta e_i'' + b_\beta\right) \tag{18}$$

where $e_{final}^{text} \in R^{D_{final}}$ represents the final emotional embedding of text, where β_i are the learned emotional significance weights for each token embedding. W_β and b_β are learnable parameters. This emotional embedding integrates both the semantic and the relational emotional information contained in the textual modality. In this branch, RoBERTa performs the extraction of token embeddings deeply and contextually. These token embeddings are enriched by our TSG, which captures semantic emotional relationships between tokens through graph attention. Therefore, the final emotional embedding serves as a detailed and strong representation of the textual modality. This description offers valuable insights, improving the subsequent adaptive multimodal fusion process.

3.3 Dual-Branch Dynamic Attention Mechanism

The primary characteristic of SYMPHONIA is its Dual-Branch Dynamic Attention Mechanism, which effectively combines facial and textual emotional signals. Unlike traditional static or one-way attention systems, this method employs a bidirectional model in which the two modalities interact with each other. This enables the model to detect subtle emotional nuances by allowing each modality to dynamically influence the interpretation of the other. Traditional cross-modal attention methods frequently employ

fixed attention weights, which fail to account for the changing relevance of modalities in various emotional contexts. In contrast, SYMPHONIA's dynamic system uses facial features to guide text interpretation, emphasising emotional words that correspond to facial expressions. In a similar vein, textual characteristics impact facial image analysis, emphasising expressive facial cues that align with the text's emotional tone. A deeper, more nuanced understanding of emotions across various inputs is provided by this dynamic interaction between the two modalities, which also improves the model's adaptability.

The facial modality embedding $h_T \in R^{D_{face}}$ serves as a context vector to modulate textual embeddings $E''_{text} = [e''_1, e''_2, \ldots, e''_M] \in R^{M \times D_{text}}$. The attention scores between facial embedding h_T and textual embeddings e''_i are computed by:

$$u_i^{ft} = yanh\left(W_{ft}^{(1)} e''_i + W_{ft}^{(2)} h_T + b_{ft}\right) \tag{19}$$

$$\gamma_i^{ft} = \frac{exp\left(u_{ft}^{\top} u_i^{ft}\right)}{\sum_{j=1}^{M} exp\left(u_{ft}^{\top} u_j^{ft}\right)} \tag{20}$$

where $W_{ft}^{(1)}, W_{ft}^{(2)} \in R^{D_a \times D_{text}}, D_a \times D_{face}$ and $b_{ft} \in R^{D_a}$ all of which represent learnable model parameters. $u_{ft} \in R^{D_a}$ presents as a learnable vector responsible for transforming embeddings to attention scores. Attention weights from facial-to-textual modality is denoted as γ_i^{ft}. The facial-conditioned textual embedding c_{ft} is computed as a weighted sum of textual embeddings:

$$c_{ft} = \sum_{i=1}^{M} \gamma_i^{ft} e''_i, c_{ft} \in R^{D_{text}} \tag{21}$$

Similarly, the textual embedding $e_{final}^{text} \in R^{D_{text}}$ is utilized to dynamically guide the attention across facial embeddings $H = [h_1, h_2, \ldots, h_T] \in R^{T \times D_{face}}$. After obtaining dynamically attended cross-modal embeddings c_{ft} (facial-conditioned textual embedding) and c_{ft} (textual-conditioned facial embedding), we perform a dynamic integration step. The integration adaptively merges both modality embeddings using a learned gating mechanism, ensuring balanced contributions based on context:

$$g_{fusion} = \sigma\left(W_g\left[c_{ft}; c_{tf}\right] + b_g\right), g_{fusion} \in [0,1]^{D_{fusion}} \tag{22}$$

The final integrated embedding f_{fusion} is then computed as:

$$f_{fusion} = g_{fusion} \odot \left(W_f^{(1)} c_{ft}\right) + \left(1 - g_{fusion}\right) \odot \left(W_f^{(2)} c_{tf}\right) \tag{23}$$

where $W_g, W_f^{(1)}, W_f^{(2)}$ and b_g are learnable parameters. σ denotes sigmoid activation, providing dynamic control over each modality's influence. The resulting $f_{fusion} \in R^{D_{fusion}}$ represents the integrated multimodal emotional embedding, fully capturing the dynamic, bidirectional cross-modal interactions and emotional contexts between facial and textual modalities. By enabling both textual and facial embeddings to dynamically adapt and mutually inform one another, the Dual-Branch Dynamic Attention Mechanism greatly improves emotional recognition by capturing complex cross-modal emotional relationships.

In the SYMPHONIA framework, the adaptively merged embedding, f_{fusion}, offers a strong basis for later hierarchical fusion and emotion classification.

3.4 Hierarchical Adaptive Fusion Module

The main purpose of hope is to admit the possibility of a better future happening in the real world. This module cleverly interweaves human facial and textual data—at various levels—constructing prior comprehension toward emotional states. This multimodal fusion strategy patterns adaptive gating and self-supervised contrastive learning on aligned multimodal representations. Instead of forcing a snapshot collapse at some stage of fusion into a singular unified space, the fusion paradigm is staged to start from the early combination of modality-specific features, which capture primary emotional cues. At the intermediate stage, the model enforces semantic alignment across modalities with contrastive alignment, coherently clustering corresponding emotional representations while repelling unrelated ones. Attainment of the highest abstraction stage defines the use of attention-based adaptive gating to dynamically highlight the most critical components from each fusion stage. The module synthesizes a richly encoded and deeply hierarchical contextualized emotional embedding with robust alignment and high sensitivity, ensuring interpretability and structural coherence. In this stage, initial modality embeddings—facial h_T and textual e_{final}^{text} are combined. An adaptive gating mechanism integrates modality-specific features, preserving complementary emotional cues:

$$f_{low} = g_{low} \odot \left(W_l^{(1)} h_T \right) + (1 - g_{low}) \odot \left(W_l^{(2)} e_{final}^{text} \right) \tag{24}$$

where the gating factor g_{low} is computed as:

$$g_{low} = \sigma \left(W_g^{low} \left[h_T; e_{final}^{text} \right] + b_g^{low} \right), g_{low} \in [0,1]^{D_{low}} \tag{25}$$

$W_g^{low}, W_l^{(1)}, W_l^{(2)}, b_g^{low}$ are learnable parameters. σ represents sigmoid activation, ensuring adaptive gating. The low-level fused representation flow captures initial cross-modal interactions at a coarse semantic level.

The adaptive gating mechanism dynamically assigns weights to features derived from low-, mid-, and high-level fusion stages, with the high-level fusion contributing the most significantly, as illustrated in Fig. 6.

To enhance cross-modal semantic alignment, we propose a self-supervised contrastive learning-based fusion at the intermediate stage. The contrastive objective ensures that corresponding modality embeddings are closer in the joint embedding space, while mismatched modality pairs remain farther apart. We apply a contrastive loss ($L_{contrastive}$) inspired by InfoNCE loss formulation to align these representations:

$$L_{contrastive} = -\log \frac{exp\left(\frac{sim(c_{ft},\, c_{tf})}{\tau} \right)}{\sum_{n=1}^{N} exp\left(\frac{sim\left(c_{ft},\, c_{tf}^{(n)}\right)}{\tau} \right)} \tag{26}$$

where $sim\,(u,v) = \frac{u^{\top} v}{\|u\|\|v\|}$ denotes cosine similarity. τ is a temperature hyperparameter that scales the similarity. N represents the number of negative samples (Dualgatsligned embeddings). By minimizing this loss, the mid-level fusion learns modality embeddings with strong semantic correspondence, ensuring cross-modal representational consistency Fig. 3. The mid-level fused embedding f_{mid} is then obtained by averaging aligned embeddings, reflecting enhanced semantic coherence:

$$f_{mid} = \frac{1}{2} \left(c_{ft} + c_{tf} \right) \tag{27}$$

We combine the previously obtained low-level f_{low}, mid-level f_{mid}, and dynamically integrated embedding f_{fusion} into a unified high-level emotional embedding using attention-based adaptive gating.

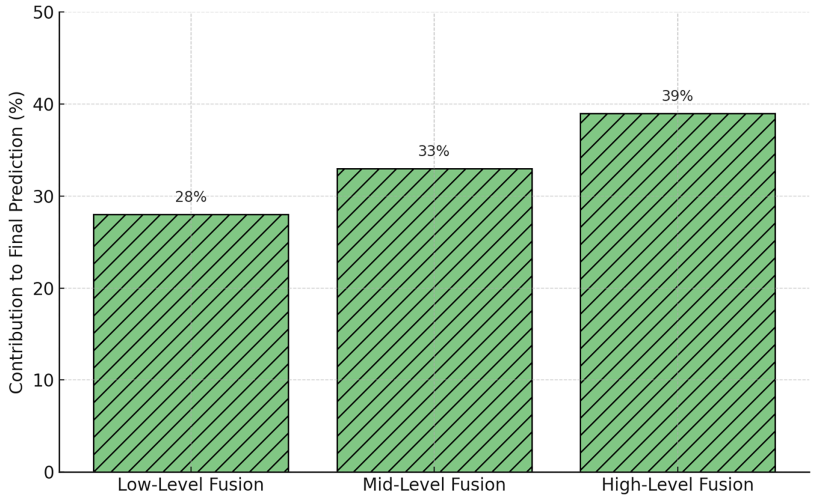

Figure 6: Contribution of each hierarchical fusion stage to final emotion prediction.

Fig. 7 demonstrates the cosine similarity between the token embeddings computed by RoBERTa (left) and those enhanced by GAT (right). The numbers show how semantically close each pair of tokens is. After the application of GAT, tokens that are semantically related, like "feel" and "happy", show greater similarity, suggesting proper alignment of words on emotions relevant to the model. Such shifts improve the model representation of text. Finally, we combine the previously obtained low-level f_{low}, mid-level f_{mid}, and dynamically integrated embedding f_{fusion} into a unified high-level emotional embedding using attention-based adaptive gating. We first stack the embeddings into a single set:

$$F = \left[f_{low}, f_{mid}, f_{fusion}\right], F \in R^{3\times D_{fusion}} \tag{28}$$

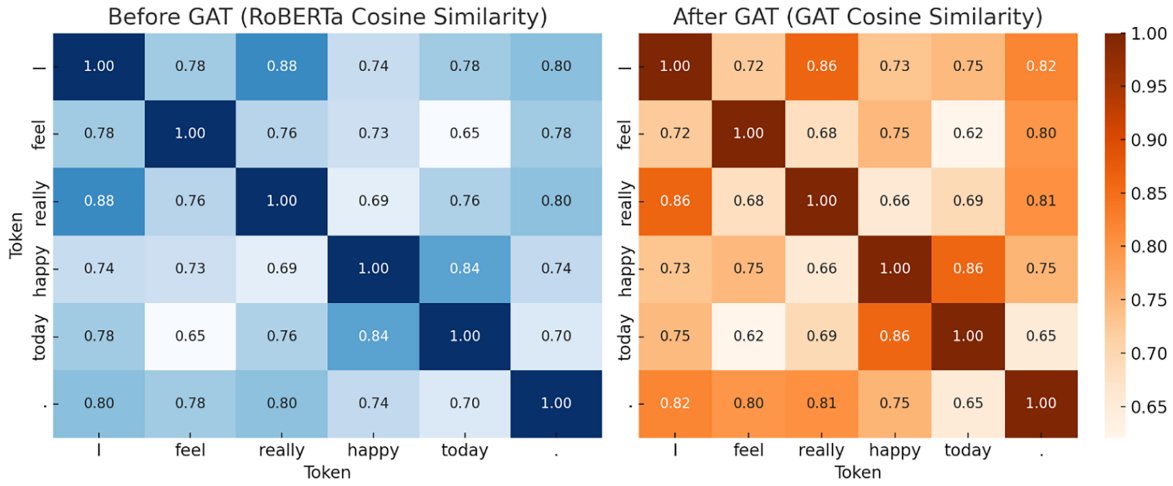

Figure 7: Cosine similarity matrices of token embeddings before and after GAT.

We then compute attention-based gating weights:

$$\alpha = softmax\left(W_a tanh\left(W_F F^{\top} + b_F\right) + b_a\right), a \in R^3 \tag{29}$$

W_a, W_F, b_a, b_F are learnable parameters.

After alignment, the embeddings from both modalities become semantically closer, demonstrating the effectiveness of the self-supervised contrastive learning framework in improving cross-modal representation coherence, as illustrated in Fig. 8. The final fused representation f_{high} is a weighted sum of these embeddings, adaptively emphasizing the most emotionally relevant features from each fusion stage:

$$f_{high} = \sum_{i=1}^{3} a_i F_i, f_{high} \in R^{D_{fusion}} \tag{30}$$

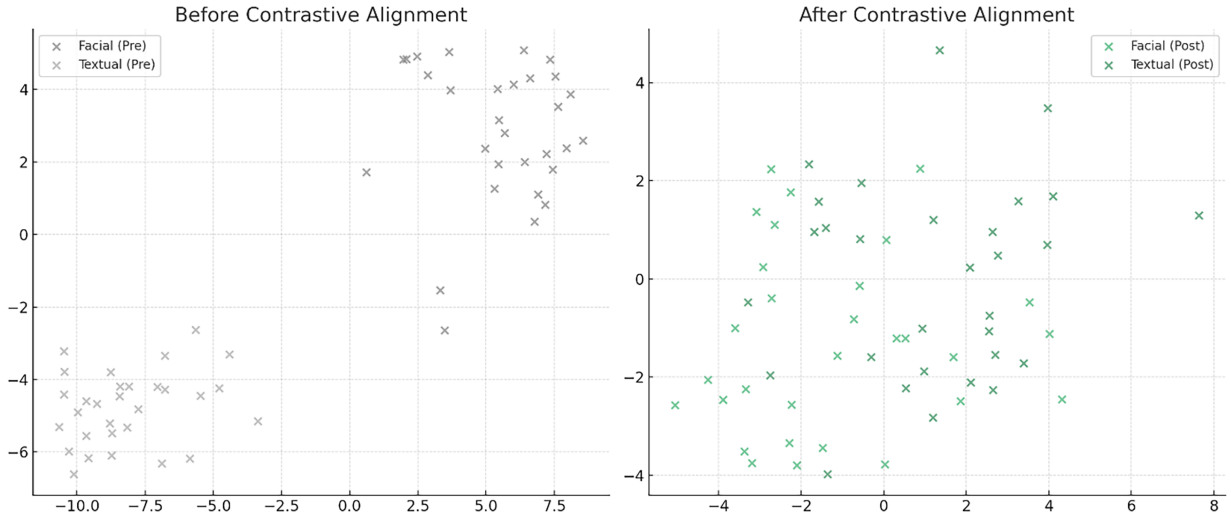

Figure 8: Visualization of facial and textual embeddings before and after contrastive alignment using t-SNE.

The described adaptive weighting structure enables each emotional layer to dynamically adjust its contribution based upon context so that each layer functions optimally during integration. The complete hierarchical adaptive fusion module is trained jointly under a single optimization criterion using cross-modal alignment (contrastive loss) in conjunction with a standard supervised classification loss, resulting in the following combined loss function:

$$L_{fusion} = L_{cls}\left(f_{high}, y\right) + \lambda L_{contrastive} \tag{31}$$

where y denotes ground-truth emotion labels. L_{cls} is typically implemented using cross-entropy or focal loss. λ is a hyperparameter balancing supervised classification and self-supervised contrastive learning objectives. The Hierarchical Adaptive Fusion Module successively merges the various modality embeddings, which are at different abstraction levels namely, low, mid, and high, thus creating a strong and semantically consistent multimodal emotional embedding f_{high}. The collaboration of adaptive gating and self-supervised contrastive learning guarantees the smooth integration and proper alignment of emotional signals from the face and the text. This mix increases the ability of the model to interpret emotions better and thus tends to robustness.

3.5 Final Classification Layer

The last step in the SYMPHONIA Framework involves the classification of the detailed multimodal embedding created through hierarchical adaptive fusion into specific emotion categories. The Final Classification Layer is designed as a Transformer-based module that can effectively leverage contextual embeddings to interpret complex multimodal emotional representations. Given the final multimodal embedding $f_{high} \in R^{D_{fusion}}$ we apply a Transformer encoder to further model the internal contextual dependencies, capturing

subtle emotional contexts that may span across different embedded dimensions. To apply the Transformer-based classification module, we first expand the embedding f_{high} into a sequence of vectors. We define a linear projection of the embedding into a sequence of L tokens, each with dimension d_{model}:

$$F_{seq} = [f_1, f_2, \ldots, f_L], f_i \in R^{d_{model}} \tag{32}$$

with each token embedding f_i defined as:

$$f_i = W_i^{proj} f_{high} + b_i^{proj}, i \in \{1, 2, \ldots, L\} \tag{33}$$

$W_i^{proj} \in R^{d_{model} \times D_{fusion}}$, $b_i^{proj} \in R^{d_{model}}$ are learnable parameters. Positional embeddings e_{pos} are added to each token to preserve ordering information:

$$F'_{seq} = \left[f_1 + e_{pos}^{(1)}, f_2 + e_{pos}^{(2)}, \ldots, f_L + e_{pos}^{(L)}\right] \tag{34}$$

We feed this sequence into a Transformer Encoder block comprising Multi-Head Self-Attention (MHSA) and feed-forward neural networks (FFN):

$$Z = TEncoder\left(F'_{seq}\right), Z \in R^{L \times d_{model}} \tag{35}$$

The Multi-Head Self-Attention module captures the contextual relationships across the different components of the multimodal embedding sequence. Given input X, self-attention operates as:

$$Att(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{36}$$

where $Q = XW_Q, K = XW_K, V = XW_V$, with learnable parameters W_Q, W_K, W_V. d_k is the dimensionality of key vectors. Multi-head attention further divides attention operations into h parallel heads, allowing parallel computation and diverse representation learning: $MHSA(X) = Concat(head_1, \ldots, head_h) W^o$. Each head is computed as follows: $head_i = Att(XW_{Q_i}, XW_{K_i}, XW_{V_i})$. Following the Transformer Encoder, we aggregate the output representations Z via mean pooling, summarizing emotional features comprehensively:

$$\mathcal{Z}_{final} = \frac{1}{L} \sum_{i=1}^{L} Z_i, \mathcal{Z}_{final} \in R^{d_{model}} \tag{37}$$

The pooled representation $\mathcal{Z}_{final}$ undergoes a linear transformation and softmax activation to predict emotion class probabilities:

$$\hat{y} = softmax\left(W_c \mathcal{Z}_{final} + b_c\right), \hat{y} \in R^C \tag{38}$$

where $W_c \in R^{C \times d_{model}}$ and $b_c \in R^C$ denote the learnable parameters of the classifier. C denotes the total number of emotion categories available. The predicted class is retrieved via the argmax operation: $y_{pred} = arg\max_{c \in [1, C]} (\hat{y}_c)$. The classification module is trained using a supervised cross-entropy loss, optimizing model parameters to minimize discrepancy between predicted and true emotion labels: $L_{cls} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$. where y_c denotes the ground truth emotion label in one-hot format. $\hat{y}_c$ denotes predicted emotion class probability from softmax.

The heatmap in Fig. 9 visualizes the attention scores computed over the projected token embeddings within the Transformer encoder of the final classification layer. Certain tokens like [CLS], facial embeddings

(F1, F2), textual embeddings (T1, T2), and even the [EOS] token are influenced differently. The model can contextually refine the unified representation through the attention mechanism with regard to the most emotionally predictive tokens before the finalized prediction of emotions.

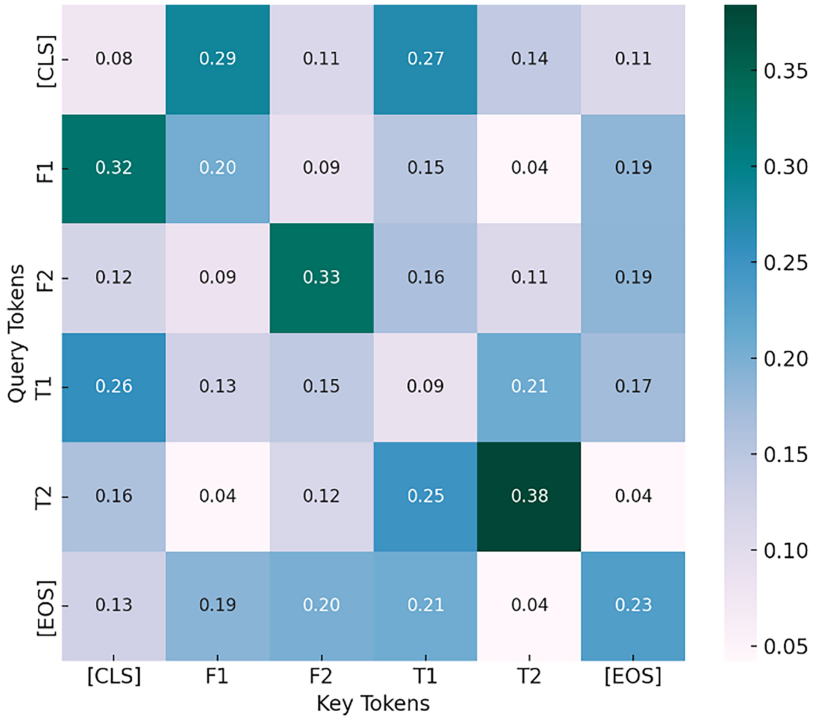

Figure 9: Transformer attention map within the final classification layer.

Combining with the previous fusion module losses, the overall training objective for the SYMPHONIA model is defined as: $L_T = L_{cls} + \lambda L_{contrastive}$ with hyperparameter λ balancing classification and contrastive alignment objectives. The last classification layer effectively performs the Transformer-based context modeling, drawing on complex cross-modal relationships captured by hierarchical adaptive fusion. Furthermore, by incorporating the usage of self-attention and robust pooling, the module for classification interprets multimodal emotional data accurately, guaranteeing the precision and reliability of emotion predictions.

4 Experimental Results

To systematically assess the impact of our proposed SYMPHONIA framework, we set out to methodically construct AWD-OMNIBUS and REPS-MSR, which, combined, offered the most comprehensive coverage of well-known multimodal datasets for emotion recognition alongside benchmark methods and evaluation protocols spanning all relevant implementation dimensions.

4.1 Datasets

To validate the accuracy and applicability of our proposed multimodal emotion recognition system, we performed a systematic evaluation on four publicly available benchmark datasets, IEMOCAP, MELD, CMU-MOSI, and CMU-MOSEI. We ensured that the datasets IEMOCAP and MELD, along with CMU-MOSI and CMU-MOSE, provided all relevant graphemic and phonemic systems, speaker heterogeneity, emotions within linguistics, and real-world complexities to make the results of the evaluation reliable and holistic Table 1.

Table 1: Dataset description.

Dataset	Modalities	Description	Emotions/Annotations	Samples
IEMOCAP	Video (face), Audio, Text	Interactive emotional dialogues performed by actors	Angry, Happy, Sad, Neutral, Excited, Frustrated	~12 h, 5 sessions
MELD	Video (face), Audio, Text	Clips from the 'Friends' TV show with natural emotional interactions	Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise	13,708 utterances, 1433 dialogues
CMU-MOSI	Video (face), Audio, Text	Opinion monologues with continuous sentiment intensity	Sentiment Intensity (Likert Scale), Positive/Negative	2199 opinion segments from 93 videos
CMU-MOSEI	Video (face), Audio, Text	Extended version of MOSI with diverse emotional annotations	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Sentiment Scores	~23,500 annotated utterances

The IEMOCAP dataset comprises approximately 12 h of richly annotated, audio-visual recordings of dyadic emotionally-charged dialogues performed by trained actors. It contains tri-modal video, audio, and text data, and is labeled with coarse-grained emotion categories including anger, happiness, sadness, neutral, excitement, and frustration. The MELD dataset is an extension of the FRIENDS television series corpus, so it contains a very large collection of naturalistic, spontaneous, multi-party conversations. It consists of more than 13,000 utterances in more than 1400 dialogues and has annotations of 7 emotions: anger, disgust, fear, joy, neutral, sadness, and surprise in video, audio, and text streams.

The CMU-MOSI dataset centers around the monologue video segments that have opinions, and it gives detailed layers of sentiment intensity alongside face and voice multimodal data. It contains 93 video clips from which 2199 opinion segments have been extracted. CMU-MOSEI augments MOSI by adding a collection of over 23,000 annotated utterances with diverse emotions such as anger, disgust, fear, happiness, sadness, and surprise alongside continuous sentiment scores and discrete values for each emotion. As with the other datasets, CMU-MOSEI preserves synchronized input from the facial video to the audio and text transcripts. The datasets serve as a solid benchmark for testing the model with regard to controlled and spontaneous interaction, layered and unlayered emotion, and a diverse range of human behavior.

4.2 Evaluation Metrics

To thoroughly assess our model performance, we employed standard evaluation metrics commonly used for multimodal emotion recognition:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{39}$$

$$Precision = \frac{TP}{TP + FP} \tag{40}$$

$$Recall = \frac{TP}{TP + FN} \tag{41}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{42}$$

Additionally, for sentiment intensity and continuous emotional dimensions (CMU-MOSI/CMU-MOSEI):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{43}$$

$$PearsonCorrelation = \frac{\sum_{i=1}^{N}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}} \tag{44}$$

4.3 Results

This section provides complete experimental findings capturing the application of our SYMPHONIA model in various datasets and evaluation frameworks. We include evaluation metrics of quantitative analysis against baseline models and ablation tests, generalization analysis across datasets, as well as some model interpretability evaluations. To demonstrate the comparative effectiveness of our model, we selected several state-of-the-art (SOTA) multimodal emotion recognition baseline methods for comparison in Table 2.

Table 2: Model's name and descriptions.

Baseline Method	Description
TFN (Tensor Fusion Network) [34]	Tensor-based fusion method.
MFN (Memory Fusion Network) [35]	Fusion using memory mechanisms.
Dualgats [32]	Uses invariant/specific features.
EmoCLIP [37]	CLIP-based zero-shot multimodal method.
MSER [38]	Cross-attention-based multimodal model.
EmoBERTa [16]	RoBERTa-based multimodal transformer.

Table 3 presents the summary statistics for class-based emotions and classification metrics (Accuracy, Precision, Recall, F1-score) for the IEMOCAP and MELD datasets. The results show that SYMPHONIA performed better than other methods in all the evaluation criteria.

Table 3: Emotion recognition performance (IEMOCAP & MELD).

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
TFN [34]	IEMOCAP	72.4	71.2	70.8	71.0
MFN [35]	IEMOCAP	73.1	72.5	71.8	72.2
Dualgats [32]	IEMOCAP	74.8	74.2	73.5	73.9
EmoCLIP [37]	IEMOCAP	75.3	74.5	74.0	74.2
SYMPHONIA (ours)	IEMOCAP	80.9	80.3	79.8	80.1
TFN [34]	MELD	64.2	63.8	63.4	63.6
MFN [35]	MELD	65.7	65.3	65.1	65.2
Dualgats [32]	MELD	67.4	67.0	66.7	66.8
EmoCLIP [37]	MELD	68.9	68.2	67.8	68.0
SYMPHONIA (ours)	MELD	74.2	73.7	73.4	73.5

Table 4 shows sentiment intensity prediction results for the CMU-MOSI and CMU-MOSEI datasets using Pearson correlation (Corr) and Mean Absolute Error (MAE).

Table 4: Sentiment intensity prediction (CMU-MOSI & CMU-MOSEI).

Model	Dataset	Corr ↑	MAE ↓
TFN [34]	MOSI	0.75	0.98
MFN [35]	MOSI	0.77	0.92
Dualgats [32]	MOSI	0.79	0.88
EmoCLIP [37]	MOSI	0.80	0.85
SYMPHONIA (ours)	MOSI	0.86	0.77
TFN [34]	MOSEI	0.72	0.91
MFN [35]	MOSEI	0.74	0.86
Dualgats [32]	MOSEI	0.76	0.84
EmoCLIP [37]	MOSEI	0.78	0.82
SYMPHONIA (ours)	MOSEI	0.83	0.74

The outcomes of this study show that SYMPHONIA performs better than other models at capturing nuanced emotional and sentiment representations and improves quantitative metrics across a wide range of emotional tasks. We performed thorough ablation studies to assess how particular features contribute to the model's overall performance (Table 5).

Table 5: Ablation study on IEMOCAP.

Ablation Condition	Accuracy (%)	F1-Score (%)
Facial Branch Only	67.1	66.8
Textual Branch Only	69.5	69.1
Without Dynamic Attention	74.3	74.0
Without Hierarchical Fusion	75.6	75.3
Full SYMPHONIA Model	80.9	80.1

The results demonstrate the combined and separate contributions to facial and textual elements, as dynamic attention and hierarchical fusion improve performance significantly (Fig. 10).

Omission of these components yields less effectiveness, thereby demonstrating their importance. To evaluate generalization ability, we trained SYMPHONIA on IEMOCAP and tested it on MELD in a direct transfer scenario, without any additional tuning (Table 6).

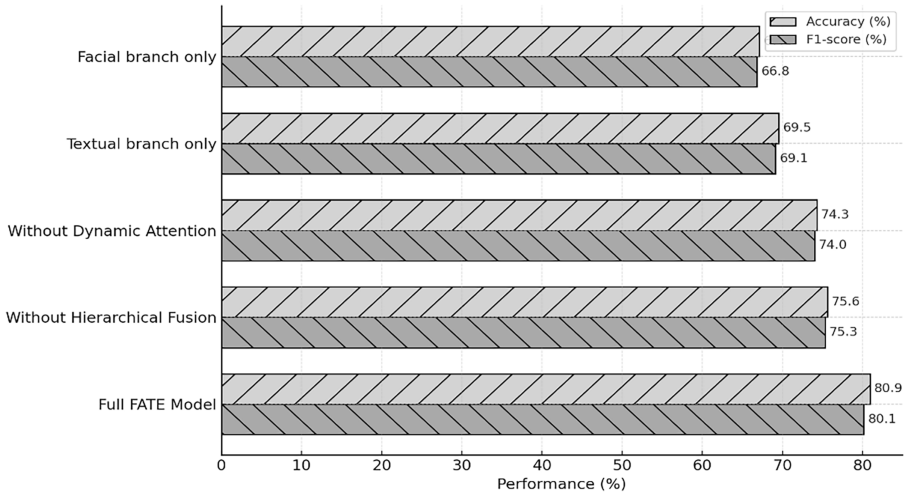

Figure 10: Ablation study visualization based on IEMOCAP dataset results.

Table 6: Cross-dataset generalization (Trained in IEMOCAP and tested on MELD).

Method	Accuracy (%)	F1-Score (%)
TFN [34]	50.8	50.3
MFN [35]	52.6	51.8
Dualgats [32]	55.1	54.4
EmoCLIP [37]	58.4	57.6
SYMPHONIA (ours)	66.9	65.8

Our results demonstrate the remarkable general performance of SYMPHONIA, especially on cross-dataset and cross-modal transfer tasks, in which it outperforms all baseline models consistently. A further qualitative study clearly shows how dynamic cross-modal attention resolves modality conflicts by paying attention to the most reliable emotional cues, which brings about performance improvement. Take for instance the text "I'm fine, everything's good"—which expresses a positive emotional tone—combined with subtle facial sadness as shown in Fig. 11.

While unimodal approaches using text let "happiness" slip through, SYMPHONIA, through facial-conditioned textual attention, identifies the true emotion as "sadness". On the contrary, another example features the phrase "That's amazing!" and a facial expression that is neutral or inconsistent. SYMPHONIA here adaptively attends to the verbal modality, classifying the facial expression as "surprise" or "happiness", which static mechanisms fail due to facial ambiguity. This demonstrates how dynamic attention of SYMPHONIA not only improves accuracy for target predictions but also extends fine-grained analysis for explaining the decisions made via contextually informative modalities using context-sensitive reasoning.

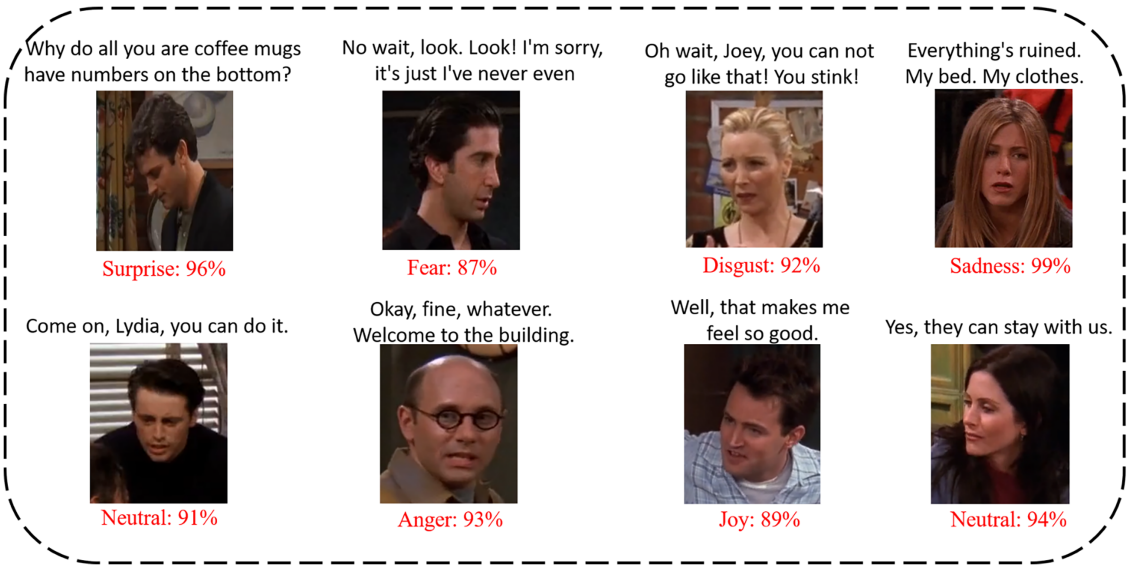

Figure 11: The SYMPHONIA model produces highly accurate predictions on multimodal input samples, effectively identifying a wide range of emotional expressions. These include Surprise (96%), Fear (87%), Disgust (92%), Sadness (99%), Neutral (91%, 94%), Anger (93%), and Joy (89%). These findings point to the model's effectiveness in capturing a wide range of emotional conditions for different subjects under various situations. The results further accentuate the capability of SYMPHONIA in distinguishing subtle facial expressions and closely matching them with textual cues via its dynamic attention mechanism and hierarchical fusion approach.

5 Conclusions

This paper proposes SYMPHONIA, a new model that effectively recognizes and integrates emotional signals from facial expressions and text. Instead of relying on a static approach, the signals are processed dynamically with the use of a dual-branch dynamic attention mechanism and adaptive hierarchical fusion. Extensive experiments were conducted to evaluate the performance of the model on IEMOCAP, MELD, CMU-MOSI, and CMU-MOSEI datasets, which were benchmarked against the state-of-the-art methods for multimodal emotion recognition. Experimental results manifest the proposed model outperforming current models and yielding higher accuracy while enhancing metrics such as F1-score, sentiment intensity correlation, and MAE. Concretely, SYMPHONIA achieved 80.9% in terms of accuracy and 80.1% in terms of F1-score on the IEMOCAP dataset, compared to Dualgats (74.8%) and EmoCLIP (75.3%). For the MELD dataset, SYMPHONIA achieved an accuracy of 74.2% and an F1-score of 73.5%, outperforming other baseline models. In the case of sentiment prediction, it resulted in a Pearson correlation of 0.86 on MOSI and 0.83 on MOSEI, outperforming all other baseline methods. Extensive ablation studies confirm that the model's success is largely due to its inclusion of the cross-modal dynamic attention and adaptive hierarchical fusion components for enhancing stability and explainability. This goes to support the hypothesis that context-aware, adaptive multimodal fusion improves model performance.

The generalization of SYMPHONIA across datasets further underlines its robustness toward domain shifts and variations of emotional contexts. For the IEMOCAP-trained model that was tested on MELD, SYMPHONIA achieved 66.9% accuracy and 65.8% F1-score, outperforming all baselines such as TFN, MFN, and Dualgats with a gap of more than 8%. Qualitative analyses further show that the model excels in the interpretation of complex signals through dynamic focusing on the most salient emotional cues, which is another proof of the strength of its framework.

Building on the promising results of SYMPHONIA, future work will explore the integration of additional modalities, such as audio and physiological signals (e.g., ECG, EEG, and GSR), to capture subtle and complementary emotional cues, thereby enhancing contextual understanding and predictive accuracy. Furthermore, adaptive multi-sensor fusion and space–frequency selective feature modeling strategies, as demonstrated in AMSO-SFS [39], will be investigated to improve robustness under challenging visual conditions. In addition, iterative refinement and scale-alignment mechanisms inspired by DI-MDE [40] may be incorporated to strengthen temporal consistency and dynamic scene representation in multimodal emotion recognition. Lastly, improvements in model reliability under low-resource conditions can be achieved with the help of advanced self-supervised learning, large-scale pretraining, and transfer learning; this will make a system robust, adaptive, and context-aware.

Acknowledgement: None.

Funding Statement: This study was funded by the Korea Agency for Technology and Standards in 2022, project numbers 1415181629 (20022340, Development of International Standard Technologies Based on AI Model Lightweighting Technologies).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Akmalbek Abdusalomov, Alpamis Kutlimuratov, Avazjon Marakhimov, Kuanishbay Seytnazarov and Young-Im Cho; data collection: Alpamis Kutlimuratov, Mukhriddin Mukhiddinov and Kamola Abdurashidova; software: Akmalbek Abdusalomov and Alpamis Kutlimuratov; analysis and interpretation of results: Akmalbek Abdusalomov, Mukhriddin Mukhiddinov, Alpamis Kutlimuratov, Avazjon Marakhimov and Kuanishbay Seytnazarov; draft manuscript preparation: Akmalbek Abdusalomov and Alpamis Kutlimuratov; supervision: Young-Im Cho. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository. "The data that support the findings of this study are openly available in IEMOCAP at https://sail.usc.edu/iemocap/index.html, MELD at https://affective-meld.github.io/ and CMU-MOSI at http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/".

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. Inf Fusion. 2024;102(3):102019. doi:10.1016/j.inffus.2023.102019.
2. Guo R, Guo H, Wang L, Chen M, Yang D, Li B. Development and application of emotion recognition technology—a systematic literature review. BMC Psychol. 2024;12(1):95. doi:10.1186/s40359-024-01581-4.
3. Zhang S, Yang Y, Chen C, Zhang X, Leng Q, Zhao X. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: a systematic review of recent advancements and future prospects. Expert Syst Appl. 2024;237:121692. doi:10.1016/j.eswa.2023.121692.
4. AV G, Mala T, Priyanka D, Uma E. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. Inf Fusion. 2024;105(2):102218. doi:10.1016/j.inffus.2023.102218.
5. Zhu X, Huang Y, Wang X, Wang R. Emotion recognition based on brain-like multimodal hierarchical perception. Multimed Tools Appl. 2024;83(18):56039–57. doi:10.1007/s11042-023-17347-w.
6. Meng T, Shou Y, Ai W, Yin N, Li K. Deep imbalanced learning for multimodal emotion recognition in conversations. arXiv:2312.06337. 2023.
7. Safarov F, Kutlimuratov A, Khojamuratova U, Abdusalomov A, Cho YI. Enhanced AlexNet with Gabor and local binary pattern features for improved facial emotion recognition. Sensors. 2025;25(12):3832. doi:10.3390/s25123832.

8. Richet N, Belharbi S, Aslam H, Schadt ME, González-González M, Cortal G, et al. Textualized and feature-based models for compound multimodal emotion recognition in the wild. arXiv:2407.12927. 2024.
9. Abdusalomov A, Kutlimuratov A, Nasimov R, Whangbo TK. Improved speech emotion recognition focusing on high-level data representations and swift feature extraction calculation. Comput Mater Contin. 2023;77(3):2915–33. doi:10.32604/cmc.2023.044466.
10. Rathi T, Tripathy M. Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: a review. Speech Commun. 2024;162:103102. doi:10.1016/j.specom.2024.103102.
11. Makhmudov F, Kutlimuratov A, Cho YI. Hybrid LSTM–attention and CNN model for enhanced speech emotion recognition. Appl Sci. 2024;14(23):11342. doi:10.3390/app142311342.
12. Alhussein G, Ziogas I, Saleem S, Hadjileontiadis LJ. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. Artif Intell Rev. 2025;58(7):198. doi:10.1007/s10462-025-11197-8.
13. Kalateh S, Estrada-Jimenez LA, Nikghadam-Hojjati S, Barata J. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. IEEE Access. 2024;12(4):103976–4019. doi:10.1109/ACCESS.2024.3430850.
14. Hazmoune S, Bougamouza F. Using transformers for multimodal emotion recognition: taxonomies and state of the art review. Eng Appl Artif Intell. 2024;133(3):108339. doi:10.1016/j.engappai.2024.108339.
15. Yin G, Liu Y, Liu T, Zhang H, Fang F, Tang C, et al. Token-disentangling mutual transformer for multimodal emotion recognition. Eng Appl Artif Intell. 2024;133(10):108348. doi:10.1016/j.engappai.2024.108348.
16. Kim T, Vossen P. EmoBERTa: speaker-aware emotion recognition in conversation with RoBERTa. arXiv:2108.12009. 2021.
17. Udahemuka G, Djouani K, Kurien AM. Multimodal emotion recognition using visual, vocal and physiological signals: a review. Appl Sci. 2024;14(17):8071. doi:10.3390/app14178071.
18. Gursesli MC, Lombardi S, Duradoni M, Bocchi L, Guazzini A, Lanata A. Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets. IEEE Access. 2024;12(15):45543–59. doi:10.1109/ACCESS.2024.3380847.
19. Deshmukh S, Gupta P. Application of probabilistic neural network for speech emotion recognition. Int J Speech Technol. 2024;27(1):19–28. doi:10.1007/s10772-023-10037-w.
20. Rakhimovich MA, Kadirbergenovich KK, Rakhmovich OU, Rustem J. A new type of architecture for neural networks with multi-connected weights in classification problems. In: Proceedings of the 12th World Conference "Intelligent System for Industrial Automation" (WCIS-2022); 2022 Nov 25–26; Tashkent, Uzbekistan. p. 105–12.
21. Shou Y, Meng T, Ai W, Zhang F, Yin N, Li K. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. Inf Fusion. 2024;112(9):102590. doi:10.1016/j.inffus.2024.102590.
22. Wang L, Kang X, Ding F, Nakagawa S, Ren F. A joint local spatial and global temporal CNN-transformer for dynamic facial expression recognition. Appl Soft Comput. 2024;161(2):111680. doi:10.1016/j.asoc.2024.111680.
23. Tagmatova Z, Umirzakova S, Kutlimuratov A, Abdusalomov A, Cho YI. A hyper-attentive multimodal transformer for real-time and robust facial expression recognition. Appl Sci. 2025;15(13):7100. doi:10.3390/app15137100.
24. Zhu A, Li K, Wu T, Zhao P, Hong B. Cross-task multi-branch vision transformer for facial expression and mask wearing classification. arXiv:2404.14606. 2024.
25. Zakieldin K, Khattab R, Ibrahim E, Arafat E, Ahmed N, Hemayed E. ViTCN: hybrid vision transformer with temporal convolution for multi-emotion recognition. Int J Comput Intell Syst. 2024;17(1):64. doi:10.1007/s44196-024-00436-5.
26. Tian Y, Zhu J, Yao H, Chen D. Facial expression recognition based on vision transformer with hybrid local attention. Appl Sci. 2024;14(15):6471. doi:10.3390/app14156471.
27. Nawaz U, Saeed Z, Atif K. A novel transformer-based approach for adult's facial emotion recognition. IEEE Access. 2025;13(3):56485–508. doi:10.1109/ACCESS.2025.3555510.
28. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health. 2024;11(2):e54369. doi:10.2196/54369.

29. Shelke N, Chaudhury S, Chakrabarti S, Bangare SL, Yogapriya G, Pandey P. An efficient way of text-based emotion analysis from social media using LRA-DNN. Neurosci Inform. 2022;2(3):100048. doi:10.1016/j.neuri.2022.100048.
30. Madrakhimov S, Makharov K, Khurramov A. On the transparency of decision-making in classification by precedents with fuzzy descriptions. IEEE Access. 2025;13:173656–64. doi:10.1109/ACCESS.2025.3616052.
31. Zhu P, Wang B, Tang K, Zhang H, Cui X, Wang Z. A knowledge-guided graph attention network for emotion-cause pair extraction. Knowl Based Syst. 2024;286(3):111342. doi:10.1016/j.knosys.2023.111342.
32. Zhang D, Chen F, Chen X. DualGATs: dual graph attention networks for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023 Jul 9–14; Toronto, ON, Canada. p. 7395–408.
33. Wang D, Guo X, Tian Y, Liu J, He L, Luo X. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 2023;136(2):109259. doi:10.1016/j.patcog.2022.109259.
34. Xiang A, Qi Z, Wang H, Yang Q, Ma D. A multimodal fusion network for student emotion recognition based on transformer and tensor product. In: Proceedings of the 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE); 2024 Aug 29–31; Jinzhou, China. p. 1–4.
35. Chudasama V, Kar P, Gudmalwar A, Shah N, Wasnik P, Onoe N. M2FNet: multi-modal fusion network for emotion recognition in conversation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2022 Jun 19–20; New Orleans, LA, USA. p. 4651–60.
36. Khujamatov EH, Abdullaev M, Umirzakova S. Analytical modeling of hybrid CNN-transformer dynamics for emotion classification. Mathematics. 2026;14(1):85. doi:10.3390/math14010085.
37. Foteinopoulou NM, Patras I. EmoCLIP: a vision-language method for zero-shot video facial expression recognition. In: Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG); 2024 May 27–31; Istanbul, Turkiye. p. 1–10.
38. Khan M, Gueaieb W, El Saddik A, Kwon S. MSER: multimodal speech emotion recognition using cross-attention with deep fusion. Expert Syst Appl. 2024;245(22):122946. doi:10.1016/j.eswa.2023.122946.
39. Abdusalomov A, Umirzakova S, Bakhtiyor Shukhratovich M, Mukhiddinov M, Kakhorov A, Buriboev A, et al. Drone-based wildfire detection with multi-sensor integration. Remote Sens. 2024;16(24):4651. doi:10.3390/rs16244651.
40. Abdusalomov A, Umirzakova S, Shukhratovich MB, Kakhorov A, Cho YI. Breaking new ground in monocular depth estimation with dynamic iterative refinement and scale consistency. Appl Sci. 2025;15(2):674. doi:10.3390/app15020674.