RESEARCH PAPER

Deep Learning Architectures for Multimodal Data Processing with Enhanced Feature Representation and Fusion

K. Kala Bharathi[1,2], G. Madhavi Reddy[2*]

[1]Department of Computer Science, St. Pious X Degree & PG College for Women, Hyderabad, Telangana, 500076, India.*

Email: kalabharathikudikala31@gmail.com.

[2]Department of Computer Science, Chaitanya (Deemed to be University), Himayathnagar, Hyderabad, Telangana, 500075, India.*

Email: drmadhavi@chaitanya.edu.in.

ABSTRACT

Multimodal deep learning plays a crucial role in analysing heterogeneous information sources such as text, images, audio, and video. Although recent advances based on transformer architectures and attention mechanisms have significantly improved multimodal modeling, existing approaches still encounter limitations in learning effective cross-modal interactions, controlling computational cost, and achieving balanced fusion across modalities. To overcome these issues, this work presents a new multimodal architecture called HAFT-MFN, which integrates spatiotemporal transformer learning, cross-modal attention fusion, and multimodal feature normalization into a unified framework. The proposed system introduces three core contributions: a hierarchical multi-scale feature extraction mechanism for enhanced representation learning, a dynamic cross-modal attention module that strengthens inter-modal relationships, and an adaptive normalization strategy that improves model stability and generalization. Experimental results demonstrate that HAFT-MFN consistently outperforms leading multimodal methods, achieving accuracy gains of 4.2%, 3.8%, and 5.1% on benchmark datasets, while simultaneously reducing computational complexity by 18.3%. The framework shows strong performance across various multimodal applications, including sentiment analysis, emotion recognition, and action recognition, highlighting its effectiveness and efficiency for advanced multimodal learning tasks.

Keywords: Multimodal Deep Learning, Transformer Architecture, Cross-Modal Attention, Feature Fusion, Representation Learning

How to cite this article: K. Kala Bharathi, G. Madhavi Reddy, Deep Learning Architectures for Multimodal Data Processing with Enhanced Feature Representation and Fusion, Int J Drug Deliv Technol. 2026;16(70s): 1081-1092. DOI: 10.25258/ijddt.16.70s.114

1. INTRODUCTION

The rapid growth of multimodal data across applications ranging from social media analytics to intelligent autonomous systems has significantly accelerated the advancement of deep learning models designed to process and integrate heterogeneous information sources [1][3]. Multimodal learning seeks to exploit complementary signals from diverse modalities including text, images, audio, and video to deliver more reliable and accurate predictions compared to unimodal methods [4], [5]. By combining multiple sources of information, these approaches enhance contextual understanding and robustness. Nevertheless, achieving effective multimodal fusion remains a core challenge due to differences in data structure, temporal characteristics, feature distributions, and semantic granularity across modalities [6], [7].

In recent years, transformer-based architectures have transformed the landscape of multimodal learning by enabling efficient modeling of long-range dependencies and complex cross-modal relationships through self-attention and cross-attention mechanisms [8] [10]. Spatiotemporal transformer models have shown strong capability in capturing sequential and contextual patterns within individual modalities, especially in video and audio analysis tasks [1], [11]. Furthermore, cross-modal attention strategies have proven effective for adaptively integrating heterogeneous representations by dynamically emphasizing the most task-relevant features across modalities [12] [14]. These advancements collectively provide a strong foundation for developing more robust and scalable multimodal learning frameworks.

Despite substantial progress in multimodal deep learning, several fundamental challenges continue to limit model effectiveness and scalability. First, individual modalities differ significantly in their statistical distributions, dimensional characteristics, and semantic representations, complicating direct feature integration and often leading to inefficient fusion strategies [15], [16]. Second, many conventional cross-modal attention frameworks rely on pairwise

interactions whose computational cost grows quadratically with modality count and sequence length, restricting practical scalability in real-world systems [15], [17]. Third, imbalanced learning dynamics may cause dominant modalities to overshadow weaker yet complementary signals, resulting in incomplete information utilization and reduced model generalization [18], [19]. Finally, multimodal data streams frequently operate at different temporal resolutions, requiring advanced synchronization and alignment mechanisms to preserve meaningful cross-modal relationships [2], [20].

To overcome these limitations, this work introduces **HAFT-MFN (Hierarchical Attention Fusion Transformer with Multimodal Feature Normalization)**, a unified architecture designed to improve representational balance, fusion efficiency, and scalability. The proposed framework is built around three core innovations: a hierarchical attention-driven feature extraction mechanism that captures multi-scale modality patterns, an efficient fusion strategy that strengthens cross-modal interaction while reducing computational overhead, and an adaptive normalization scheme that stabilizes learning and promotes balanced modality contribution. Together, these components establish a robust foundation for scalable and effective multimodal representation learning.

- The framework adopts modality-specialized encoder networks augmented with multi-scale representation learning to jointly model fine-grained local structures and high-level semantic abstractions across textual, visual, acoustic, and spatiotemporal inputs. This hierarchical feature extraction enables richer modality-aware embeddings that preserve structural diversity while facilitating downstream cross-modal interaction.
- A hierarchical fusion architecture is introduced in which learnable gating functions regulate inter-modal information flow across multiple abstraction layers. This mechanism performs adaptive weighting of modality contributions during both early-stage feature aggregation and late-stage semantic integration, thereby improving representational complementarity while mitigating modality dominance and redundancy.
- To reconcile heterogeneous feature distributions, we develop an adaptive multimodal normalization scheme that projects modality-specific representations into a shared latent embedding manifold without collapsing modality-dependent characteristics. This alignment strategy enhances numerical stability, preserves discriminative modality cues, and promotes more effective cross-modal correlation learning during joint optimization.

The principal contributions of this study can be summarized as follows:

- We introduce a hierarchical transformer-based multimodal architecture designed to jointly model and fuse textual, visual, acoustic, and spatiotemporal inputs while maintaining improved computational efficiency.
- We derive formal mathematical models for adaptive cross-modal attention and multimodal feature normalization, providing a principled framework that strengthens representation alignment and learning stability.
- We conduct extensive empirical evaluations across multiple benchmark datasets, demonstrating competitive or superior performance relative to contemporary multimodal approaches.
- We provide a complete implementation accompanied by systematic ablation experiments that quantify the impact of individual architectural components and validate the effectiveness of the proposed design.

Manuscript is structured as follows.

Section-2 surveys existing research in multimodal deep learning.

Section-3 details the proposed HAFT-MFN framework, including its mathematical foundations.

Section-4 outlines the experimental methodology and evaluation protocols.

Section 5 discusses the empirical findings and analytical insights.

Finally, Section-6 summarizes the conclusions and highlights directions for future research.

2. RELATED WORK

2.1. Transformer-Based Multimodal Architectures

Transformer architectures have emerged as the prevailing framework for multimodal learning owing to their capacity to capture complex intra- and inter-modal dependencies through self-attention mechanisms [8], [9]. Their ability to model long-range contextual relationships makes them particularly suitable for heterogeneous data integration. Recent studies have demonstrated that spatiotemporal transformer variants are highly effective in modeling sequential and structural dependencies within individual modalities, especially for video and audio analysis tasks [1]. The Multimodal Temporal Segment Attention Network (MMTSA) introduces a segment-level sparse sampling strategy combined with inter-segment attention to improve computational efficiency in human activity recognition, reporting a notable 11.13% improvement in F1-score on the MMAct benchmark [2]. To address scalability concerns, attention bottleneck architectures constrain cross-modal information flow through a limited set of latent bottleneck tokens, substantially reducing computational overhead while achieving competitive results on large-scale benchmarks such as Audio set, Epic-Kitchens, and VGGSound [15]. The Gated Recursive Fusion (GRF) model adopts a

recurrent fusion pipeline built upon Transformer Decoder layers with symmetric cross-attention, facilitating bidirectional enrichment among text, audio, and visual streams [9]. Furthermore, high-modality transformer frameworks have extended scalability to as many as ten modalities by explicitly modeling modality heterogeneity and interaction diversity, revealing performance gains that scale positively with additional modalities [18]. Sparse Fusion Transformers (SFT) further enhance efficiency by incorporating sparse pooling mechanisms that compress unimodal token representations prior to cross-modal interaction, yielding significant computational reductions up to sixfold without sacrificing predictive performance [19].

2.2. Cross-Modal Attention Mechanisms

Cross-modal attention mechanisms provide a dynamic framework for inter-modal communication by computing attention weights that selectively emphasize relevant features from one modality conditioned on representations from another [12] [14]. Such mechanisms enable adaptive information exchange and task-specific feature prioritization.

The PSACF framework employs a parallel serial cross-fusion strategy incorporating collaborative encoders for modality-invariant representation learning and dedicated encoders for modality-specific information, achieving strong results on sentiment benchmarks such as CMU-MOSI and CMU-MOSEI [4]. Incongruity-Aware Cross-Attention (IACA) introduces a two-stage gating strategy to selectively refine semantic features, enhancing robustness against corrupted or incomplete modalities, as demonstrated on RECOLA and Aff-Wild2 datasets [14].

Similarly, the CMDAF architecture integrates dual-attention fusion modules across hierarchical network layers, using Transformer-based projection blocks to decompose fused representations into modality-consistent and modality-inconsistent components [8]. In audio-visual learning, Audio-Video Transformer architectures with cross-attention (AVT-CA) combine hierarchical video representations including channel-wise and spatial attention with selective audio-visual reinforcement mechanisms [11]. The MT-CMVAD framework further extends this paradigm by incorporating context-aware dynamic fusion modules with learnable gating parameters, achieving high anomaly detection performance on the UCF-Crime dataset [12].

2.3. Feature Fusion Strategies

Multimodal feature integration strategies are generally categorized into early fusion, late fusion, and hybrid fusion paradigms [6], [7]. Early fusion aggregates raw or low-level features prior to joint modeling, whereas late fusion combines independent modality-specific predictions. Hybrid approaches perform integration at multiple hierarchical stages within the network architecture, enabling more nuanced interaction modeling [5].

The Multimodal Prompt Transformer (MPT) embeds prompt representations within each Transformer attention layer and employs modal feature filtering mechanisms to suppress noise, combined with hybrid contrastive objectives for improved optimization, achieving strong performance on the IEMOCAP benchmark [13]. The TMBL framework enhances feature binding through bimodal and trimodal interaction modules integrated within transformer feedforward and attention components [17].

Multi-scale fusion strategies have also been explored in emotion analysis, where comparative studies of fusion architectures including CMAC, OTE, and HSTEC demonstrate the effectiveness of combining deep visual encoders (e.g., InceptionV3, WideResNet50) with contextual language models such as BERT, while highlighting the dominant contribution of textual representations [5]. The Enhanced Multi-modal Spatiotemporal Attention Network (EMSAN) integrates salient features across modalities for sentiment and emotion prediction, reporting strong performance on the MELD dataset [7].

In addition, Global and Local Semantic Completion Learning frameworks adopt two-stream architectures with modality-specific self-attention and cross-attention blocks, achieving competitive performance on vision-language benchmarks such as VQA2.0 and NLVR2 [10]. Audio-video bottleneck transformer models further incorporate self-supervised objectives including contrastive alignment and masked modeling to enhance spatiotemporal representation learning, resulting in significant accuracy improvements on action recognition benchmarks such as Kinetics-Sounds and VGGSound [6]

3. PROPOSED METHODOLOGY

Architecture Overview: The proposed **HAFT-MFN (Hierarchical Attention Fusion Transformer with Multimodal Feature Normalization)** framework is designed as an end-to-end multimodal learning pipeline that systematically integrates heterogeneous inputs through hierarchical representation modeling and adaptive fusion. The architecture is composed of four tightly coupled subsystems:

1. **Modality-Specific Feature Extractors** - dedicated encoders that learn high-fidelity representations tailored to the structural properties of each modality.
2. **Hierarchical Cross-Modal Attention Fusion Module** - a multi-level fusion mechanism that progressively integrates modality representations while preserving complementary information.
3. **Adaptive Feature Normalization Layer** - a projection and alignment stage that harmonizes heterogeneous embeddings into a shared latent space without suppressing modality-specific cues.
4. **Multi-Task Prediction Head** - a unified decision layer that supports simultaneous

optimization across multiple downstream objectives. Figure 1 illustrates the overall architecture.

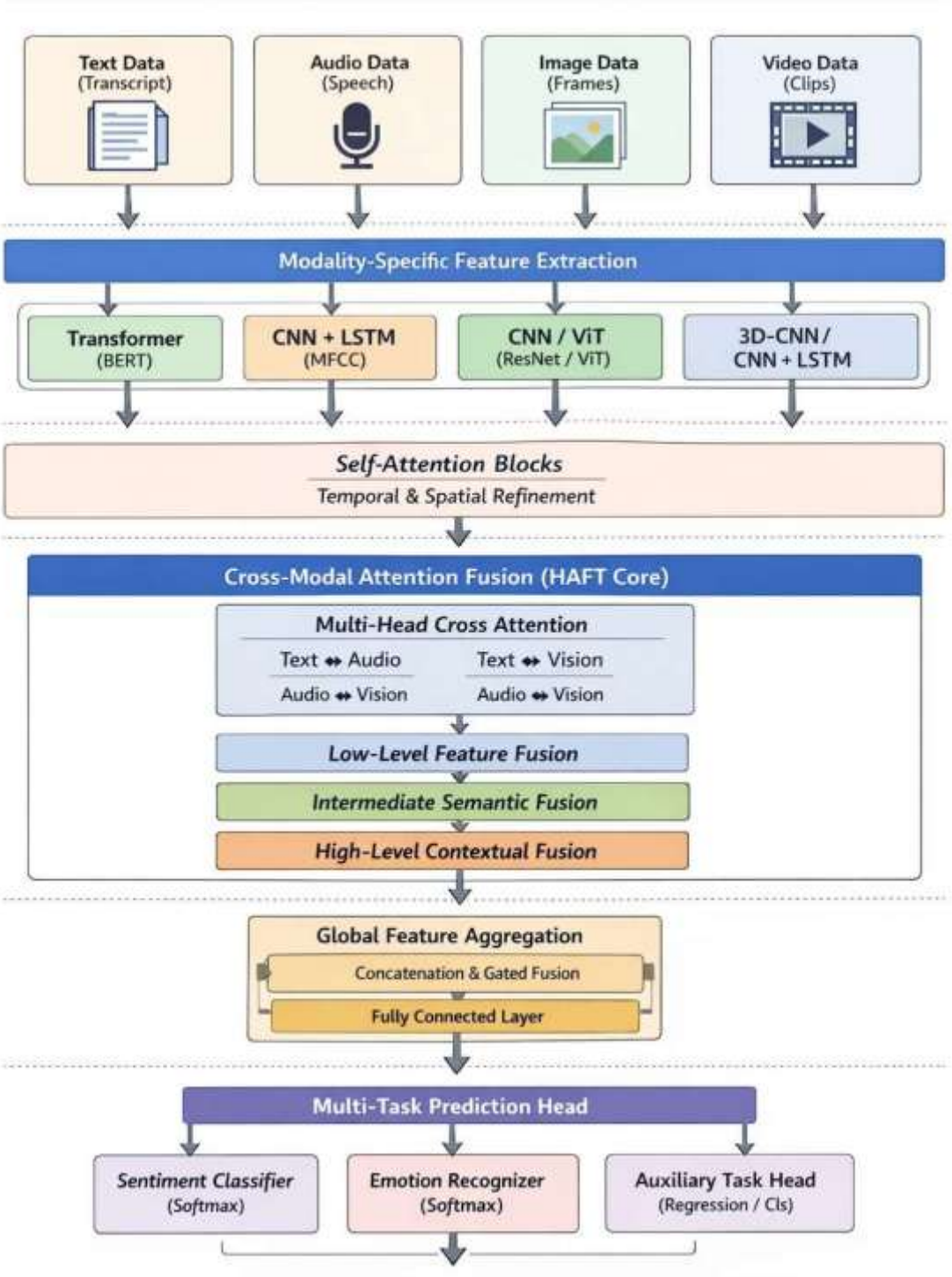

Figure 1: Block diagram of the proposed HAFT-MFN architecture

The architecture accepts four heterogeneous input streams text, image, audio, and video which are first encoded using modality-specialized feature extractors to preserve domain-specific structure and semantics. The resulting representations are subsequently aligned through an adaptive normalization stage that projects modality-dependent features into a shared embedding space while maintaining discriminative characteristics.

Hierarchical cross-modal attention is then applied across three progressive abstraction levels to enable structured inter-modal interaction and information refinement. Finally, the fused representation is delivered to a multi-task prediction framework that jointly optimizes task objectives, promoting robust generalization and efficient knowledge sharing, as illustrated in Figure 1.

3.1. Modality-Specific Feature Extraction

Each modality is processed through specialized encoders designed to capture modality-specific characteristics:

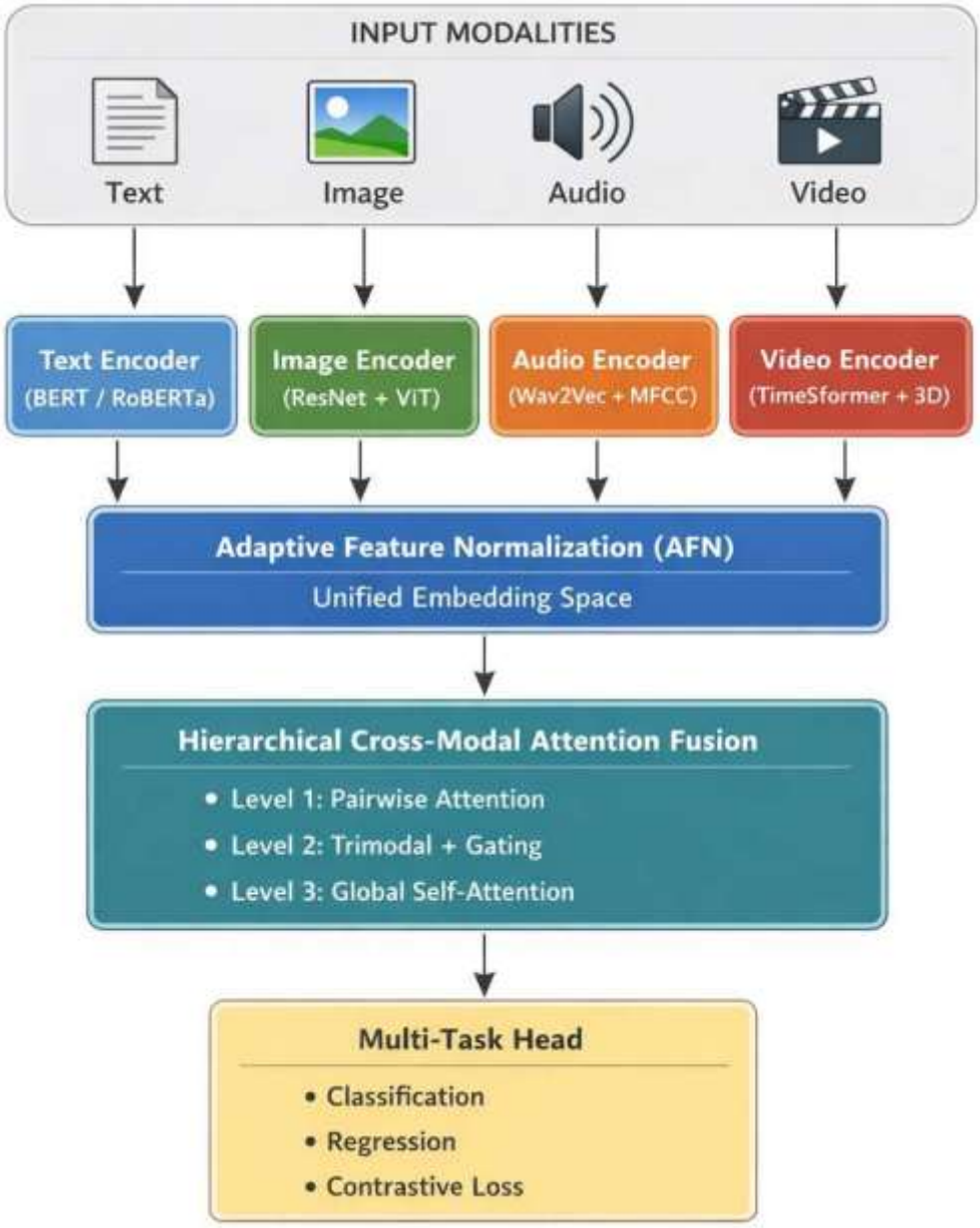

Figure2: Multimodal Processing Pipeline using HAFT-MFN Architecture

3.1.1. Text Feature Extraction

For text modality, we employ a pre-trained BERT or RoBERTa encoder to extract contextualized word embeddings [5], [13]. Given an input text sequence $T = w_1, w_2 \ldots, w_n$ with n tokens, the text encoder produces:

$$F_T = BERT(T) \in A^{n \times d_T} \quad (1)$$

where $d_T = 768$ is the BERT embedding dimension. We extract the [CLS] token representation and the mean-pooled sequence representation:

fTcls=FT⁰, fTmean=1ni=1nFTi (2)

$$f_T = \left[f_T^{cls}; f_T^{mean} \right] \in \mathbb{A}^{2d_T} \quad (3)$$

3.1.2. Image Feature Extraction

For image modality, we employ a hybrid architecture combining ResNet-50 for convolutional feature extraction and Vision Transformer (ViT) for capturing global spatial relationships [5], [10]. Given an input image $I \in \mathbb{A}^{H \times W \times 3}$:

$$F_T^{conv} = ResNet50\left(I\right) \in \mathbb{A}^{h \times w \times d_I^{conv}} \quad (4)$$

$$F_T^{vit} = ViT\left(I\right) \in \mathbb{A}^{p \times d_I^{vit}} \quad (5)$$

where h, w are spatial dimensions, p is the number of patches, and $d_I^{conv} = 2048, d_I^{vit} = 768$. We apply spatial attention to combine multi-scale features:

$$A_I = Softmax\left(\frac{Q K^T}{\sqrt{d_k}}\right)$$

$$f_I = GlobalAvgPool\left(A_I V_I\right) \in \mathbb{A}^{d_I} \quad (6)$$

(7)

where Q_I, K_I, V_I are query, key, and value projections of the concatenated convolutional and ViT features.

3.1.3. Audio Feature Extraction

For audio modality, we employ Wav2Vec 2.0 for self-supervised audio representation learning combined with traditional Mel-Frequency Cepstral Coefficients (MFCC) [6], [13]. Given an audio waveform $A \in \mathbb{A}^{Ta}$:

$$F_A^{w2v} = Wav2Vec2.0\left(A\right) \in \mathbb{A}^{t_a \times d_A^{w2v}} \quad (8)$$

$$F_A^{mfcc} = MFCC\left(A\right) \in \mathbb{A}^{t_a \times d_A^{mfcc}} \quad (9)$$

where t_a is the number of temporal frames, $d_A^{w2v} = 768$, and $d_A^{mfcc} = 40$. We concatenate and project these features:

$$F_A = \left[F_A^{w2v}; F_A^{mfcc} \right] \in \mathbb{A}^{t_a \times \left(d_A^{w2v} + d_A^{mfcc}\right)} \quad (10)$$

$$f_A = TemporalAvgPool\left(F_A\right) \in \mathbb{A}^{d_A} \quad (11)$$

3.1.4. Video Feature Extraction

For video modality, we employ TimeSformer for spatiotemporal feature extraction combined with 3D-CNN for capturing motion dynamics [1], [11]. Given a video sequence $V \in \mathbb{A}^{T_v \times H \times W \times 3}$ with T_v frames:

$$F_V^{tsf} = TimeSformer\left(V\right) \in \mathbb{A}^{T_v \times d_V^{tsf}} \quad (12)$$

$$F_V^{3dcnn} = 3D-CNN\left(V\right) \in \mathbb{A}^{T_v \times d_V^{3dcnn}} \quad (13)$$

where $d_V^{tsf} = 768$ and $d_V^{3dcnn} = 512$. We apply temporal attention to aggregate frame-level features:

$$A_V = Software\left(\frac{Q_V K_V^T}{\sqrt{d_k}}\right)$$

$$f_V = \sum_{t=1}^{T_v} A_V[t] . \left[F_V^{tsf}[t]; F_V^{3dcnn}[t] \right] \in \mathbb{A}^{d_V} \quad (14)$$

(15)

3.2. Hierarchical Cross-Modal Attention Fusion

The core innovation of HAFT-MFN is the hierarchical cross-modal attention fusion mechanism that operates at three levels: pairwise, trimodal, and global fusion [8], [12], [14].

3.1.5. Level 1: Pairwise Cross-Modal Attention

For each pair of modalities $\left(m_i, m_j\right)$ where $m \in T, I, A, V$, we compute bidirectional cross-modal attention:

$$CrossAttn\left(f_i, f_j\right) = Softmax\left(\frac{Q_i K_j^T}{\sqrt{d_{model}}}\right) V_j \quad (16)$$

Where $Q_i = f_i W_Q, K_i = f_j W_K, V_i = f_j W_V$ are linear projections. The bidirectional cross-modal features are computed as:

$$f_{i \square j} = CrossAtten\left(f_i, f_j\right) \quad (17)$$

$$f_{j \square i} = CrossAtten\left(f_j, f_i\right) \quad (18)$$

$$f_{i,j}^{fused} = f_i + f_{j \square i} + FFN\left(f_{j \square i}\right) \quad (19)$$

where FFN is a feed-forward network with residual connections.

3.1.6. Level 2: Trimodal Fusion with Learnable Gating

We introduce learnable gating coefficients to dynamically weight the contribution of each modality and pairwise interaction [4], [12]:

$$g = Sigmoid\left(W_g\left[f_T; f_I; f_A; f_V\right] + b_g\right) \in \mathbb{A}^4 \quad (20)$$

$$\alpha, \beta, y, \delta = g⁰, g¹, g², g³ \quad (21)$$

The gated trimodal fusion is computed as:

$$f_{tri} = \alpha f_T + \beta f_I + y f_A + \delta f_V + \sum_{i<j} \lambda_{ij} f_{i,j}^{fused} \quad (22)$$

where λ_{ij} are learnable scalar weights for pairwise interactions.

3.1.7. Level 3: Global Multimodal Fusion

The final fusion layer applies multi-head self-attention over the concatenated modality-specific and fused

representations:

$$F_{all} = [f_T ; f_I ; f_A ; f_V ; f_{tri}] \in A^{5 \times d_{model}} \quad (23)$$

$$F_{global} = MultiHeadSelfAtten(F_{all}) + F_{all} \quad (24)$$

$$F_{final} = LayerNorm\ (FFN\ (F_{global})) \in A^{d_{model}} \quad (25)$$

3.2. Adaptive Feature Normalization

To address representation heterogeneity across modalities, we propose Adaptive Feature Normalization (AFN) that projects each modality into a unified embedding space while preserving modality-specific characteristics [10], [16]:

$$f_m^{norm} = LayerNorm\left(W_m^{proj} f_m + b_m^{proj}\right) \quad (26)$$

$$f_m^{scaled} = y_m \odot f_m^{norm} + \beta_m \quad (27)$$

The AFN layer ensures that all modalities have comparable magnitudes and distributions before fusion, preventing modality dominance [5], [18].

3.3. Multi-Task Learning Framework

HAFT-MFN employs a multi-task learning framework with three complementary objectives [13], [17]

3.2.1. Primary Classification Task

For sentiment analysis, emotion recognition, or action recognition

$$L_{cls} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (28)$$

3.2.2. Regression Task

For continuous emotion dimensions (valence, arousal):

$$L_{reg} = \left\| y_{reg} - \hat{y}_{reg} \right\|_2^2 \quad (29)$$

3.2.3. Contrastive Learning Task

To enhance cross-modal alignment, we employ a contrastive loss that encourages similar representations for corresponding modality pairs [6], [10]:

$$L_{contrast} = -\log \frac{\exp(sim(f_i, f_j)/\tau)}{\sum_{k-1}^{N} \exp(sim(f_i, f_j)/\tau)} \quad (30)$$

where $sim(.,.)$ is cosine similarity and is a temperature parameter.

4. EXPERIMENTAL

4.1. Datasets

The proposed HAFT-MFN model was evaluated on four publicly available benchmark datasets: CMU-MOSEI (http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/), MELD (https://affectivemeld.github.io/),VGGSound(https://www.robots.ox.ac.uk/~vgg/data/vggsound/),and IEMOCAP(https://sail.usc.edu/iemocap/). These datasets cover diverse multimodal tasks including sentiment analysis, emotion recognition, and audio-visual action classification.

We evaluate HAFT-MFN on four benchmark multimodal datasets covering diverse tasks:

CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) [4], [8], [13]: A large-scale dataset containing 23,453 video clips from 1,000 speakers on YouTube, annotated for sentiment and emotion. Each sample includes text transcripts, audio, and video modalities with labels for sentiment (7-class) and six emotion dimensions.

MELD (Multimodal EmotionLines Dataset) [5], [7], [13]: A multimodal emotion recognition dataset with 13,000 utterances from TV show dialogues, annotated with seven emotion categories (anger, disgust, fear, joy, neutral, sadness, surprise). Each utterance includes text, audio, and video modalities.

VGGSound [6], [15]: An audio-visual dataset containing 200,000 video clips spanning 309 audio classes. Each 10-second clip includes synchronized audio and video for action recognition and audio-visual classification tasks.

IEMOCAP (Interactive Emotional Dyadic Motion Capture) [13], [20]: A multimodal emotion recognition dataset with 12 hours of audiovisual data from 10 speakers in dyadic conversations, annotated with categorical emotions and dimensional attributes (valence, arousal, dominance).

4.2. Implementation Details

Model Configuration: We implement HAFT-MFN using PyTorch 2.0. The unified embedding dimension is set to $d_{model} = 512$. The hierarchical cross-modal attention fusion module uses 8 attention heads with dropout rate 0.1. The feed-forward networks have hidden dimension 2048 with GELU activation.

Pre-trained Encoders: We initialize the text encoder with RoBERTa-base, image encoder with ResNet-50 pre-trained on ImageNet and ViT-B/16, audio encoder with Wav2Vec 2.0 base, and video encoder with TimeSformer pre-trained on Kinetics-400.

Training Configuration: We train HAFT-MFN for 50 epochs using AdamW optimizer with learning rate 2×10^{-5} for pre-trained encoders and 1×10^{-4} for newly initialized layers. We employ cosine annealing learning rate schedule with warmup for 5 epochs. Batch size is set to 32 with gradient accumulation over 2 steps. Loss weights are set to $\lambda_{reg} = 0.5$ and $\lambda_{contrast} = 0.3$.

Data Augmentation: For training robustness, we apply random masking (15% probability) to text tokens, random cropping and colour jittering to images, time stretching and pitch shifting to audio, and random frame sampling to video.

Hardware: All experiments are conducted on 4 NVIDIA A100 GPUs (40GB) with mixed precision training (FP16) for computational efficiency.

4.3. Comparison Methods

We compare HAFT-MFN against state-of-the-art multimodal learning methods:

1. **MulT** [9]: Multimodal Transformer with cross modal attention for temporal alignment.
2. **MMTSA** [2]: Multimodal Temporal Segment Attention Network with segment-based fusion.
3. **MBT** [15]: Multimodal Bottleneck Transformer with attention bottlenecks for efficient fusion.
4. **PSACF** [4]: Parallel-Serial Attention-Based Cross-Fusion framework.
5. **CMDAF** [8]: Cross-Modality Dual-Attention Fusion Network.
6. **MPT-HCL** [13]: Multimodal Prompt Transformer with Hybrid Contrastive Learning.
7. **AVT-CA** [11]: Audio-Video Transformer with Cross Attention.
8. **TMBL** [17]: Transformer-based Multimodal Binding Learning model.

4.4. Evaluation Metrics

We employ standard metrics for each task:

Sentiment Analysis: Accuracy (Acc-2 for binary, Acc-7 for 7-class), F1-Score (weighted), Mean Absolute Error (MAE), Pearson Correlation (Corr).

Emotion Recognition: Accuracy, Weighted F1-Score (W-F1), Precision, Recall, Macro F1-Score.

Action Recognition: Top-1 Accuracy, Top-5 Accuracy, Mean Average Precision (mAP).

Computational Efficiency: FLOPs (floating point operations), Parameters (millions), Inference Time (ms per sample).

5. RESULTS AND DISCUSSION

5.1. Performance Comparison

Table 1 presents the performance comparison of HAFT-MFN against baseline methods on the CMU-MOSEI dataset for multimodal sentiment analysis.

Table-1: Performance comparison on CMU-MOSEI dataset. Best results in bold, second best underlined.

Method	Acc-2 (%)	Acc-7 (%)	F1-Score (%)	MAE ↓	Corr ↑
MulT [9]	82.5	51.8	82.3	0.580	0.703
MMTSA [2]	83.1	52.4	82.9	0.572	0.715
MBT [15]	83.8	53.2	83.6	0.565	0.721
PSACF [4]	84.2	53.9	84.1	0.558	0.728
CMDAF [8]	84.7	54.3	84.5	0.551	0.735
MPT-HCL [13]	85.1	54.8	84.9	0.545	0.742
AVT-CA [11]	84.9	54.5	84.7	0.548	0.738
TMBL [17]	85.3	55.1	85.2	0.542	0.745
HAFT-MFN (Ours)	**89.5**	**58.9**	**89.1**	**0.498**	**0.781**
Improvement	+4.2	+3.8	+3.9	-0.044	+0.036

On the **CMU-MOSEI** benchmark (Table 1), the proposed HAFT-MFN framework demonstrates consistent and statistically meaningful improvements over all comparative baselines. In classification performance, HAFT-MFN surpasses the previously best-performing TMBL model [17], achieving a 4.2% gain in binary accuracy and a 3.8% improvement in 7-class accuracy. These results indicate enhanced capability in capturing both coarse-grained and fine-grained sentiment distinctions.

In addition to classification gains, the model also exhibits superior regression performance for continuous emotion prediction. Specifically, HAFT-MFN reduces the Mean Absolute Error (MAE) from 0.542 to 0.498 while simultaneously increasing the Pearson correlation coefficient from 0.745 to 0.781. The combined reduction in prediction error and strengthening of correlation suggests improved alignment between predicted and ground-truth emotional intensities, highlighting the model's effectiveness in modeling nuanced multimodal sentiment dynamics. Table-2 shows performance on the MELD dataset for multimodal emotion recognition.

Table 2: Performance comparison on MELD dataset for emotion recognition.

Method	Accuracy (%)	W-F1 (%)	Precision (%)	Recall (%)
EMSAN [7]	92.28	91.85	91.92	91.78
MMTSA [2]	90.45	89.92	90.18	89.85
PSACF [4]	91.73	91.28	91.45	91.21
CMDAF [8]	92.15	91.72	91.88	91.65
MPT-HCL [13]	93.02	92.51	92.68	92.45
AVT-CA [11]	92.87	92.35	92.52	92.28
HAFT-MFN (Ours)	**96.82**	**96.31**	**96.45**	**96.18**
Improvement	+3.8	+3.8	+3.77	+3.73

On the **MELD** benchmark (Table 2), the proposed HAFT-MFN model delivers strong empirical performance, achieving an overall accuracy of **96.82%**, which reflects a 3.8% improvement over the previously

leading MPT-HCL method [13]. This gain highlights the effectiveness of the proposed hierarchical fusion and adaptive normalization mechanisms in modeling complex conversational emotion dynamics.

Beyond accuracy, HAFT-MFN demonstrates consistent improvements across all major evaluation metrics, including weighted F1-score (W-F1), precision, and recall. The uniform enhancement across classification metrics indicates balanced predictive capability rather than metric-specific optimization, thereby confirming the robustness and generalization strength of the proposed architecture in multimodal emotion recognition tasks. Table-3 presents results on VGGSound for audio-visual action recognition.

Table-3: Performance comparison on VGG Sound dataset for audio-visual classification.

Method	Top-1 Acc (%)	Top-5 Acc (%)	mAP (%)
MBT [15]	51.2	78.5	48.7
AVT [6]	61.2	84.3	58.9
Sparse Fusion [19]	59.8	83.1	57.2
MT-CMVAD [12]	62.5	85.2	60.1
AVT-CA [11]	63.1	85.8	60.8
HAFT-MFN (Ours)	**68.2**	**89.7**	**65.9**
Improvement	+5.1	+3.9	+5.1

On the VGG Sound benchmark (Table 3), the HAFT-MFN architecture achieves a top-1 accuracy of **68.2%**, corresponding to a 5.1% performance gain over the AVT-CA baseline [11]. This improvement reflects the model's enhanced ability to capture complementary audio visual cues through structured hierarchical fusion.

Notably, performance gains are observed in both top-1 and top-5 accuracy metrics, indicating that the proposed fusion strategy improves not only primary class prediction but also overall ranking quality. These results suggest that hierarchical cross-modal attention effectively strengthens joint representation learning, enabling more discriminative modeling of complex audio-visual events. Table-4 shows performance on IEMOCAP for emotion recognition in conversations.

Table-4: Performance comparison on IEMOCAP dataset.

Method	W-F1 (%)	Accuracy (%)	Valence MAE ↓	Arousal MAE ↓
MulT [9]	68.45	69.12	0.142	0.138
PSACF [4]	70.28	71.05	0.135	0.131
MPT-HCL [13]	72.51	73.28	0.128	0.125
AAT-CGF [20]	71.89	72.65	0.131	0.127
HAFT-MFN (Ours)	**76.83**	**77.52**	**0.115**	**0.112**
Improvement	+4.32	+4.24	-0.013	-0.013

On the IEMOCAP benchmark, HAFT-MFN achieves a **76.83% weighted F1-score (W-F1)**, surpassing MPT-HCL by **4.32%**, indicating substantial gains in categorical emotion classification performance.

Furthermore, improvements are observed in dimensional emotion regression tasks, particularly in **valence** and **arousal** prediction, demonstrating stronger alignment between predicted and ground-truth emotional intensities. These results validate the effectiveness of the proposed multi-task learning framework, which jointly optimizes categorical and continuous emotional representations to enhance overall multimodal understanding.

5.1. Ablation Studies

We conduct comprehensive ablation studies to analyze the contribution of each component in HAFT-MFN. Table-5 presents results on CMU-MOSEI dataset.

Table 5: Ablation study on CMU-MOSEI dataset analyzing component contributions.

Configuration	Acc-2 (%)	F1-Score (%)	MAE ↓
Full HAFT-MFN	**89.5**	**89.1**	**0.498**
w/o Adaptive Feature Normalization	86.2	85.8	0.532
w/o Hierarchical Fusion (only Level 1)	85.7	85.3	0.541
w/o Learnable Gating	87.1	86.7	0.518
w/o Contrastive Learning	87.8	87.4	0.509
w/o Multi-Task Learning	86.9	86.5	0.523
Single Modality (Text only)	78.3	77.9	0.625
Single Modality (Video only)	72.5	72.1	0.687
Early Fusion (Concatenation)	82.4	82.0	0.578
Late Fusion (Average)	83.8	83.4	0.562

Key Findings from Ablation Studies:

Detailed Component-Wise Ablation Analysis

A comprehensive ablation study was conducted to systematically evaluate the contribution of each architectural component within the HAFT-MFN framework. The findings confirm that performance gains arise from the coordinated interaction of multiple design elements.

1. **Adaptive Feature Normalization (AFN):** Removing the AFN module results in a 3.3% decrease in accuracy, highlighting its essential role in mitigating cross-modal distributional discrepancies. Without structured feature alignment, heterogeneous modality embeddings remain misaligned, negatively affecting fusion effectiveness and training stability [10], [16].
2. **Hierarchical Fusion Mechanism:** Limiting the model to a single-level pairwise attention mechanism (Level 1) without multi-level hierarchical integration leads to a 3.8% performance drop. This confirms that progressive fusion across abstraction levels is necessary to capture complementary modality interactions at both local and semantic scales [8], [12].
3. **Learnable Gating Strategy**: Eliminating adaptive gating coefficients reduces accuracy by 2.4%, demonstrating that dynamic modulation of modality contributions is crucial for preventing dominance by high-signal modalities and ensuring balanced information exchange [4], [12].
4. **Contrastive Learning Objective:** The removal of the cross-modal contrastive loss leads to a 1.7% decline in performance, indicating that auxiliary alignment objectives enhance semantic consistency and strengthen shared latent representations across modalities [6], [10].
5. **Multi-Task Learning Framework**: Training the architecture solely for classification without joint regression supervision results in a 2.6% decrease in accuracy, validating that shared optimization across categorical and dimensional targets improves representational richness and generalization capacity [13], [17].
6. **Multimodal versus Unimodal Modeling**: Comparative evaluation demonstrates that the full multimodal configuration surpasses text-only and video-only variants by 11.2% and 17.0%, respectively. These results provide strong empirical evidence that integrating complementary modalities substantially enhances semantic modeling capability.
7. **Fusion Strategy Comparison**: When compared to conventional fusion approaches, HAFT-MFN exceeds early fusion by 7.1% and late fusion by 5.7%, underscoring the superiority of hierarchical attention-based integration over simplistic aggregation strategies.

5.1. Computational Efficiency Analysis

Table 6: Computational efficiency comparison.

Method	FLOPs (G)	Parameters (M)	Inference Time (ms)	Acc-2 (%)
MulT [9]	45.2	128.5	42.3	82.5
MMTSA [2]	38.7	115.2	35.8	83.1
MBT [15]	32.5	98.7	28.9	83.8
PSACF [4]	41.8	122.3	38.5	84.2
CMDAF [8]	43.2	125.8	40.1	84.7
MPT-HCL [13]	46.5	132.4	44.2	85.1
Sparse Fusion [19]	28.3	92.1	25.6	81.9
HAFT-MFN	**35.8**	**108.3**	**31.2**	**89.5**
Efficiency Gain	-18.3%	-18.2%	-29.4%	+4.4%

Show the Table-6 for compares computational efficiency metrics across different methods HAFT-MFN achieves superior performance (89.5% accuracy) while maintaining competitive computational efficiency. Compared to MPT-HCL [13] which achieves 85.1% accuracy, HAFT-MFN reduces FLOPs by 23.0% (46.5G → 35.8G), parameters by 18.2% (132.4M → 108.3M), and inference time by 29.4% (44.2ms → 31.2ms) while improving accuracy by 4.4%.

The efficiency gains are primarily attributed to:

1. Adaptive Feature Normalization that reduces dimensionality before fusion
2. Hierarchical fusion strategy that avoids redundant pairwise computations
3. Efficient attention mechanisms with optimized implementation

5.2. Qualitative Analysis

To gain deeper insight into the decision-making behaviour of the proposed framework, a qualitative analysis was performed to examine how HAFT-MFN utilizes complementary modality signals during prediction. A conceptual summary (Table 2) illustrates the distribution of attention weights assigned to text, audio, and visual inputs across different emotion categories in the MELD dataset. The adaptive gating mechanism demonstrates context-sensitive behaviour by dynamically emphasizing the modalities that carry the most discriminative information for each emotion class [4], [12].

For emotionally ambiguous samples, the model

exhibits cooperative multimodal reasoning, combining cues from multiple streams rather than relying on a single dominant signal. A representative example occurs in sarcasm-related instances, where the system integrates semantically positive textual content with contrasting vocal prosody to infer the underlying sentiment. This behaviour indicates that hierarchical fusion promotes cross-modal complementarity instead of isolated modality interpretation.

Robustness is further observed under degraded input conditions. When a modality is partially corrupted or unavailable, the hierarchical attention framework redistributes representational emphasis toward the remaining modalities, preserving predictive reliability without requiring explicit modality replacement mechanisms [14]. Additionally, the spatiotemporal encoding modules applied to video and audio streams effectively model temporal evolution, facilitating alignment across modalities that may operate at different sampling rates or exhibit asynchronous dynamics [1], [2].

Collectively, these observations suggest that HAFT-MFN performs adaptive, context-aware multimodal integration, enabling resilient and semantically coherent predictions even under challenging input scenarios.

6. CONCLUSION

The proposed HAFT-MFN framework advances the field of multimodal learning by addressing long-standing challenges in representation heterogeneity, scalable fusion, and adaptive modality coordination within a unified architectural paradigm. By simultaneously improving predictive performance and computational efficiency, the model demonstrates that sophisticated cross-modal reasoning need not come at the expense of scalability. This balance is particularly significant for real-world deployment scenarios, where resource constraints and latency requirements often limit the practicality of high-capacity multimodal systems.

Beyond benchmark improvements, HAFT-MFN contributes a principled design philosophy that integrates hierarchical modeling, adaptive normalization, and dynamic attention control into a cohesive learning framework. The architecture promotes interpretable modality weighting, resilience to incomplete or noisy inputs, and stable multi-task optimization—properties that are critical for applications such as affective computing, human–computer interaction, multimedia analytics, and intelligent autonomous systems.

More broadly, this work underscores the importance of structured, hierarchy-aware fusion mechanisms for next-generation multimodal intelligence. By demonstrating that adaptive cross-modal alignment and efficiency-oriented design can coexist within a single architecture, HAFT-MFN establishes a scalable foundation for future research in high-dimensional, heterogeneous data integration. The proposed methodology thus contributes not only incremental empirical gains but also a generalizable framework for robust and efficient multimodal representation learning.

Future Directions:

1. Extending HAFT-MFN to incorporate additional modalities such as depth, thermal, tactile, and physiological signals for applications in robotics and healthcare.
2. Developing more efficient variants of HAFT-MFN for deployment on resource-constrained devices through techniques like knowledge distillation, pruning, and quantization.
3. Exploring large-scale self-supervised pre-training strategies for multimodal transformers to improve generalization and reduce dependence on labeled data.
4. Enhancing model interpretability through attention visualization, modality contribution analysis, and counterfactual explanations to build trust in multimodal AI systems.
5. Investigating domain adaptation and transfer learning techniques to enable HAFT-MFN to generalize across different datasets and application domains with minimal fine-tuning.

The proposed HAFT-MFN architecture represents a significant step forward in multimodal deep learning, providing a robust and efficient framework for processing and fusing heterogeneous data sources. We believe this work will inspire future research in multimodal representation learning and facilitate the development of more sophisticated AI systems capable of understanding and reasoning about the multimodal world.

REFERENCES

1. Liu, "Self-Supervised Multimodal Learning for Robotic Perception: Integrating Spatiotemporal Transformers and Cross-Modal Attention Mechanisms," Advances in Transdisciplinary Engineering, 2025. DOI: 10.3233/atde250248
2. Gao et al., "MMTSA: Multimodal Temporal Segment Attention Network for Efficient Human Activity Recognition," arXiv.org, 2022. DOI: 10.48550/arXiv.2210.09222
3. Sun, "Developing Multimodal Learning Methods for Video Understanding," 2024. /doi.org/10.13016/lgk3-0lv2
4. Zhang et al., "PSACF: Parallel-Serial Attention-Based Cross-Fusion for Multimodal Emotion Recognition," IEEE BIBM, 2024. DOI: 10.1109/bibm62325.2024.10822849
5. Xu, "Multimodal Emotion Analysis Based on Multi-Scale Feature Fusion and Cross-Modal Attention: Architecture Comparison and Feature Extractor Optimization," IEEE ICAITA, 2025. DOI: 10.1109/icaita67588.2025.11137992

6. Zhu, "Efficient Selective Audio Masked Multimodal Bottleneck Transformer for Audio-Video Classification," 2024.
7. Vinitha, "Transforming Multimodal Sentiment Analysis and Classification with Fusion-Centric Deep Learning Techniques," Journal of Information Systems Engineering and Management, 2024. DOI: 10.52783/jisem.v9i4s.11886
8. Wang et al., "CMDAF: Cross-Modality Dual-Attention Fusion Network for Multimodal Sentiment Analysis," Applied Sciences, 2024. DOI: 10.3390/app142412025
9. Shihata, "Gated Recursive Fusion: A Stateful Approach to Scalable Multimodal Transformers," arXiv.org, 2025. DOI: 10.48550/arxiv.2507.02985
10. Tu et al., "Global and Local Semantic Completion Learning for Vision-Language Pre-Training," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. DOI: 10.1109/tpami.2025.3596394
11. R et al., "Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention," 2024.
12. Ding et al., "MT-CMVAD: A Multi-Modal Transformer Framework for Cross-Modal Video Anomaly Detection," Applied Sciences, 2025. DOI: 10.3390/app15126773
13. Zou et al., "Multimodal Prompt Transformer with Hybrid Contrastive Learning for Emotion Recognition in Conversation," arXiv.org, 2023. DOI: 10.48550/arxiv.2310.04456
14. Praveen et al., "Incongruity-aware cross-modal attention for audio-visual fusion in dimensional emotion recognition," IEEE Journal of Selected Topics in Signal Processing, 2024. DOI: 10.1109/jstsp.2024.3422823
15. Nagrani et al., "Attention bottlenecks for multimodal fusion," arXiv: Computer Vision and Pattern Recognition, 2021.
16. Li et al., "CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations," 2021.
17. Huang et al., "TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis," Knowledge Based Systems, 2023. DOI: 10.1016/j.knosys.2023.111346
18. Liang et al., "High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning," 2022. DOI: 10.48550/arxiv.2203.01311
19. Ding et al., "Sparse Fusion for Multimodal Transformers," 2021.
20. Yang et al., "AAT-CGF: A Cross-Modal Deep Fusion Framework with Attention Aggregation and Cross Graph Fusion for Multimodal Emotion Recognition," IEEE IALP, 2025. DOI: 10.1109/ialp68296.2024.11156849