Big Data and Information Analytics

BDIA, 10: 96–129.
DOI: 10.3934/bdia.2026006
Received: 23 February 2026
Revised: 23 February 2026
Accepted: 05 April 2026
Published: 10 April 2026

https://www.aimspress.com/journal/BDIA

Research article

MEGAKANs: enhancing intermodal dependence with global channel-spatial attention and Kolmogorov-Arnold networks for multimodal sentiment analysis

Xinglong Shen and Xuesi Ma*

School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo 45400, China

* **Correspondence:** Email: maxuesi@hpu.edu.cn.

Abstract: Multimodal sentiment analysis (MSA), which integrates text, audio, and visual cues, plays a critical role in affective computing and human-computer interaction. However, existing fusion architectures—typically based on multilayer perceptrons (MLPs)—struggle to capture complex nonlinear dependencies across modalities, limiting their effectiveness in modeling subtle and implicit emotional expressions. To address this, we have created a novel framework named MEGAKANs, which introduces a highly expressive and interpretable fusion strategy. Unlike conventional fusion approaches that rely on static MLPs, MEGAKANs incorporates Kolmogorov-Arnold networks (KANs) into the mid-fusion stage, leveraging learnable functional decomposition to flexibly model high-order nonlinear interactions across modalities. Complementarily, embedding KANs into the global channel-spatial attention (GCSA) module can adaptively highlight salient emotional patterns across spatial and channel dimensions, thereby enhancing cross-modal alignment. MEGAKANs was rigorously evaluated on the benchmark multimodal sentiment dataset (CMU-MOSI) for binary, multi-class, and regression-based sentiment prediction tasks. Experimental results revealed that MEGAKANs surpasses state-of-the-art baselines, achieving a binary accuracy of 87.02% and reducing the mean absolute error (MAE) to 0.7265, thereby demonstrating superior robustness and generalization capabilities. Notably, the proposed model showed the greatest relative improvement in the underutilized visual modality, validating its ability to capture subtle affective cues. These results demonstrate not only the superior performance of MEGAKANs but also highlight the potential of KANs in multimodal learning, offering a scalable and interpretable solution for real-world affective computing applications.

Keywords: multimodal sentiment analysis; Kolmogorov-Arnold networks; global channel-spatial attention; multimodal fusion; modality contribution

1. Introduction

The rapid advancement of affective computing has led to growing interest in multimodal sentiment analysis (MSA), which seeks to infer human emotions by jointly modeling textual, auditory, and visual signals. From virtual assistants and autonomous driving to customer interaction systems, sentiment understanding has become a cornerstone for enhancing human-computer interaction [1–3]. While early sentiment analysis efforts focused predominantly on text, the rise of media-rich platforms such as YouTube and TikTok has catalyzed a shift toward multimodal frameworks that can capture a broader spectrum of affective cues [4].

MSA aims to resolve the ambiguities inherent in unimodal sentiment classification by leveraging the complementary nature of different modalities. Text provides semantic precision but often fails in contexts involving irony, sarcasm, or ambiguity. Audio contributes prosodic information (e.g., tone, pitch, rhythm), while visual input captures facial expressions and other non-verbal cues that are critical for emotional interpretation [5,6]. However, the integration of such heterogeneous signals poses several key challenges.

Existing fusion paradigms—typically categorized as early, late, or mid-fusion—struggle to fully exploit the potential of multimodal data. Early fusion often leads to redundancy and feature entanglement, while late fusion lacks the capacity to model cross-modal dependencies effectively [7,8]. Mid-fusion, though promising, tends to rely on conventional multilayer perceptrons (MLPs), which impose strong inductive biases due to their fixed activation structures. These models fall short in capturing nonlinear intermodal interactions and adapting to varying modal contributions across diverse sentiment contexts [9,10].

This manuscript argues that the limitation lies not in feature extraction but in the expressiveness of the fusion function itself. Modeling multimodal sentiment requires a fusion mechanism capable of adapting its functional structure to varying modality contributions and emotional contexts. Motivated by this observation, this work explores functional decomposition-based fusion as a principled alternative to MLP-centric designs.

To address these limitations, this paper proposes MEGAKANs: a novel multimodal sentiment analysis framework that integrates Kolmogorov-Arnold networks (KANs) and global channel-spatial attention (GCSA) into mid-fusion architecture. The acronym MEGAKANs reflects its design goals—multimodal sentiment analysis with enhanced intermodal dependence via global attention and KANs.

Unlike standard MLPs, KANs leverage learnable functional decomposition to model complex, high-order interactions across modalities. This allows the fusion module to dynamically adapt to feature distributions, providing a more flexible and expressive alternative to fixed activation pipelines [11,12]. To further improve cross-modal alignment, we introduce a lightweight GCSA module, which recalibrates the fused representations by selectively attending to salient spatial and channel-wise features. The synergy between KANs and GCSA enables MEGAKANs to model both structural dependencies and fine-grained attention, thus achieving interpretable and robust multimodal fusion with competitive efficiency.

We conducted extensive experiments on the CMU-MOSI dataset [13], evaluating MEGAKANs across binary classification, multi-class sentiment prediction, and regression tasks. The proposed model consistently outperforms state-of-the-art baselines across multiple metrics, including binary accuracy (Acc_2), five-class accuracy (Acc_5), seven-class accuracy (Acc_7), F1-score, and mean absolute error (MAE), demonstrating not only enhanced accuracy but also greater stability across modalities and tasks. Notably, MEGAKANs exhibits the most significant improvement in visual modality, validating its ability to extract and integrate subtle non-verbal affective cues.

The main contributions of this work are summarized as follows:

(1) This work introduces a functional-decomposition-based multimodal fusion paradigm by integrating KANs into the mid-fusion stage of multimodal sentiment analysis. The proposed approach models intermodal interaction as a composition of adaptive univariate functions, enabling more expressive and flexible nonlinear dependency modeling.

(2) A KANs-enhanced GCSA mechanism is proposed to jointly recalibrate modality-specific, channel-wise, and spatial features. By embedding KANs within the attention computation, the proposed design captures high-order cross-modal dependencies that conventional linear attention mechanisms fail to represent.

(3) The proposed framework provides improved interpretability of multimodal fusion, as the learned functional components and attention weights offer insights into the modality contribution and emotional salience under different contexts.

(4) Extensive experiments on the CMU-MOSI benchmark demonstrate consistent performance gains across binary, multi-class, and regression sentiment tasks, validating the effectiveness and robustness of functional-decomposition-based fusion in multimodal learning.

The remainder of this paper is structured as follows: Section 2 reviews related work on multimodal sentiment analysis fusion and KANs-based network structures. Section 3 details the architecture of MEGAKANs. Section 4 elaborates on the experiments and results, providing quantitative and qualitative evaluations. Finally, Section 5 concludes with a comprehensive discussion of findings and potential future research directions.

2. Related work

2.1. MSA and fusion strategies

MSA involves the integration of diverse data sources—typically text, audio, and visual modalities—to achieve a more accurate understanding of sentiment compared to unimodal approaches [14]. The key advantage of this approach lies in its ability to capture a more comprehensive range of emotional expressions, such as linguistic semantics from text, tonal cues from audio, and facial expressions from video. This multimodal integration enables models to handle complex emotional nuances that are often ambiguous when examined through one modality alone [15].

Traditional approaches to multimodal fusion typically fall into three categories [16]: early fusion, late fusion, and mid-fusion. Early fusion concatenates raw features from different modalities before learning joint representations but often suffers from modality imbalance and noise propagation [17]. Late fusion performs independent inference for each modality, followed by decision-level aggregation, limiting cross-modal interaction modeling [17]. In contrast, mid-level fusion—which integrates intermediate representations after individual feature encoding—has shown

superior performance by balancing flexibility and modality-specific learning [18].

Previously research has focused on leveraging deep learning models such as convolutional neural networks (CNNs) [19], recurrent neural networks (RNNs) [20], and transformers [21] to extract higher-level representations from each modality. In particular, the vision transformer (ViT) [22] has proven to be highly effective in capturing global dependencies in visual data, while models like BERT [23] and RoBERTa [24] have been central to text processing. In audio processing, architectures like YAMNet [25] utilize deep convolutional networks to capture abstract representations of speech tones and frequencies. Recent developments in hybrid fusion [17] and cross-modal learning [26] have allowed models to capture intricate modality interactions. State-of-the-art models now incorporate deep neural networks (DNNs) [27], transformer-based attention mechanisms, and contrastive learning to align multimodal representations dynamically.

Despite these advances, several challenges persist. First, feature fusion is often handled by shallow MLP-based networks, which are limited in their capacity to model nonlinear inter-modal interactions. Second, non-verbal cues—such as gaze dynamics, facial micro-expressions, and vocal intonation—are not always adequately emphasized or dynamically weighted. Finally, many methods lack interpretable mechanisms to analyze how modality contributions evolve across samples.

Given the limitations of existing multimodal fusion frameworks, we propose MEGAKANs, an innovative MSA framework that integrates KANs with GCSA to improve intermodal dependence modeling. This approach enhances nonlinear multimodal representation learning by leveraging KANs' ability to dynamically adapt to feature dependencies, while GCSA ensures effective global and local feature recalibration for cross-modal fusion. By combining these techniques, MEGAKANs overcomes the challenges of feature redundancy, modality imbalance, and nonlinear interdependence modeling, providing a scalable and robust solution for sentiment classification and regression tasks.

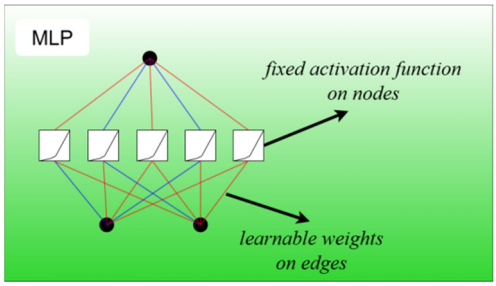
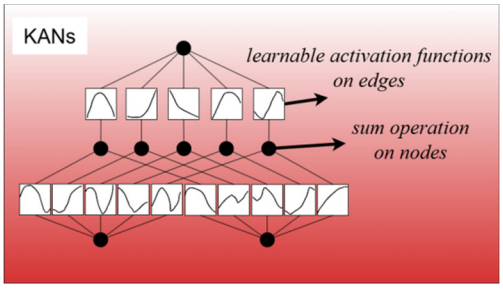
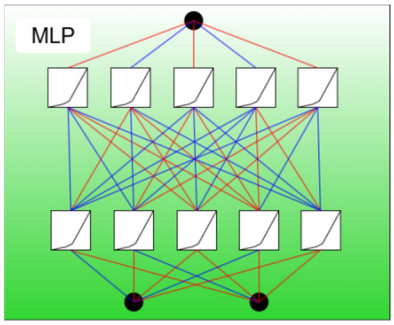
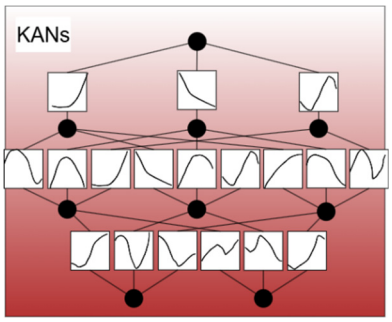
2.2. KANs and their emerging potential

To address these challenges, recent studies have turned to KANs [11], which are inspired by the Kolmogorov-Arnold representation theorem. The theorem states that any multivariate continuous function on a bounded domain can be expressed as a superposition of univariate functions. KANs implement this theoretical foundation by replacing the fixed activation functions in MLPs with learnable univariate spline functions and applying them to network edges instead of nodes. This structural reconfiguration enables KANs to dynamically approximate complex, high-dimensional nonlinear functions with greater expressive capacity and interpretability.

Unlike MLPs, where the nonlinearity is fixed and node-centered, KANs adaptively learn the structure of transformations based on data, offering fine-grained control over feature interactions. The advantages of this design include: (1) improved modeling of combinatorial structures, (2) higher approximation fidelity for nonlinear mappings, and (3) inherent interpretability due to the explicit form of learned transformations. Table 1 is a conceptual comparison of MLPs and KANs, showing the uniqueness of KANs compared to the traditional network structure.

Recent applications of KANs in deep learning have demonstrated their advantages in surrogate-assisted evolutionary algorithms, where they act as surrogate models to reduce computational costs in optimization problems [28]. KANs have also been explored for adversarial robustness, outperforming MLPs in resisting adversarial attacks in image classification tasks [29]. These findings suggest that KANs provide greater robustness and adaptability in complex decision-making scenarios [30].

Table 1. Comparison of KANs and multi-layer perceptron (MLP) models.

Model	MLP	KANs
Theorem	Universal approximation theorem	Kolmogorov-Arnold representation theorem
Formula (Shallow)	$f(x) \approx \sum_{i=1}^{N(\varepsilon)} a_i \sigma(w_i \cdot x + b_i)$	$f(x) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model structure (Shallow)		
Formula (Deep)	$MLP(x) = (w_3 \cdot \sigma_2 \cdot w_2 \cdot \sigma_1 \cdot w_1)(x)$	$KAN(x) = \phi_1 \cdot \phi_2 \cdot \phi_3(x)$
Model structure (Deep)		

In the context of sentiment analysis, KANs have been successfully applied to aspect-based sentiment analysis (ABSA), where they improve dependency modeling between words and their associated sentiments [31]. For instance, recent studies have combined KANs with graph convolutional networks (GCNs) to optimize the encoding of sentiment-specific dependencies, leading to more precise sentiment classification [32]. The MambaForGCN model, integrating KANs with syntactic GCNs, has demonstrated superior performance in modeling text and non-verbal features, further validating KANs' potential in MSA [33].

Despite these advancements, the application of KANs to multimodal fusion—especially at the mid-level representation stage—remains largely unexplored [34]. This work aims to bridge this gap by proposing a dedicated framework that integrates KANs into a multimodal fusion pipeline, hypothesizing that their superior nonlinear modeling capacity can lead to more effective and interpretable integration of textual, acoustic, and visual emotional cues.

3. Methodology

3.1. Overview of the MEGAKANs architecture

Multimodal sentiment analysis requires not only the effective extraction of modality-specific features but also a principled mechanism to model complex interactions among heterogeneous modalities. To address this challenge, this manuscript proposes MEGAKANs, a unified mid-fusion

framework that explicitly enhances intermodal dependency modeling through GCSA and KANs. As illustrated in Figure 1, the overall architecture follows a structured pipeline consisting of four main stages: (1) modality-specific feature extraction, (2) latent space alignment and concatenation, (3) KAN-enhanced GCSA-based fusion, and (4) task-specific prediction.

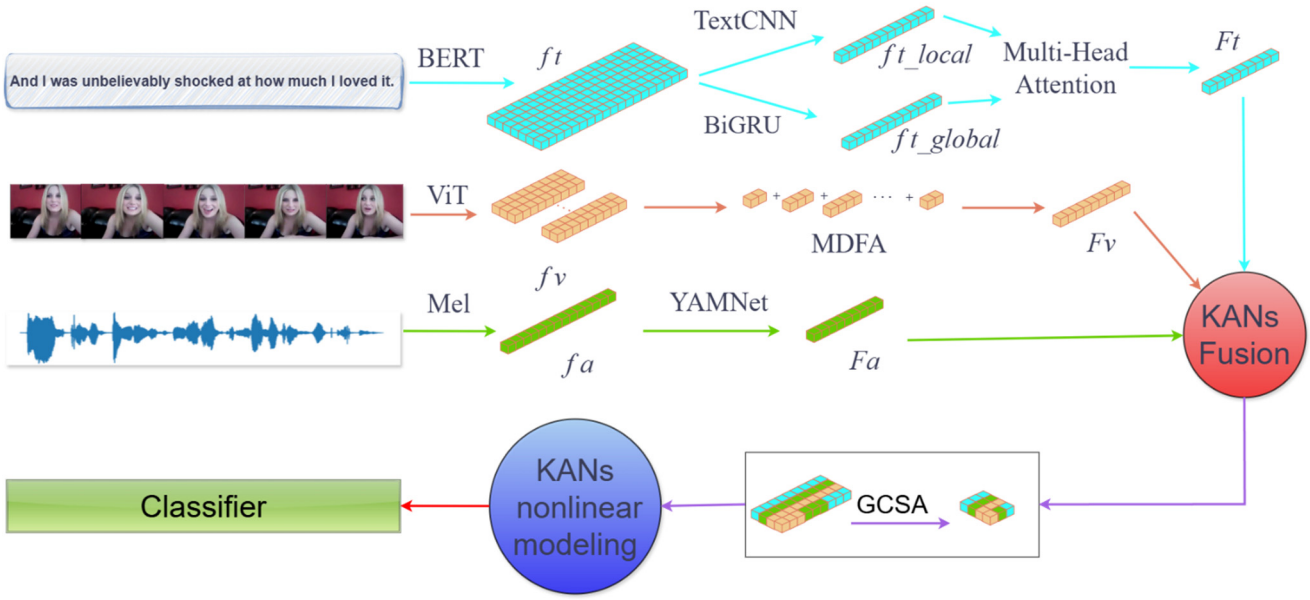


Figure 1. The overall architecture of the MEGAKANs framework.

In the first stage, each modality is encoded using a specialized backbone tailored to its signal characteristics. For the visual modality, raw video frames are processed through a ViT to extract frame-wise spatial features, denoted as $f_v \in \mathrm{R}^{T_v \times d_v}$, where T_v is the number of frames and d_v is the feature dimension. These features are further refined through a multi-dilated feature aggregation (MDFA) module to capture multi-scale temporal dependencies, yielding the aggregated visual representation F_v. For the audio modality, the features are represented by a Mel spectrogram to obtain f_a, which is then fed into the pre-trained YAMNet model. This model extracts advanced acoustic features and aggregates them through average pooling to generate a unified audio representation F_a. For the textual modality, input sentences are tokenized and encoded using bidirectional encoder representations from a transformers (BERT) model to generate contextual embeddings f_t. To further capture semantic granularity, two parallel branches are adopted: a TextCNN module for local n-gram features f_t^{local}, and a BiGRU module for global dependencies f_t^{global}. These are subsequently fused through a multi-head attention mechanism to obtain the final textual representation F_t.

In the second stage, heterogeneous modality features are projected into a shared latent space with consistent dimensionality and normalized distributions. This alignment step is crucial for multimodal fusion, as it mitigates scale discrepancies and prevents modality dominance caused by representational bias. After alignment, features from text, audio, and vision are concatenated to form a unified multimodal representation.

The third stage introduces the GCSA module, which serves as a key mechanism for intermodal interaction refinement. Rather than treating all fused features equally, GCSA adaptively reweights multimodal representations along both channel and spatial dimensions, allowing emotionally salient

cues to be emphasized while suppressing redundant or noisy signals. Importantly, feature importance is not manually defined, but is learned implicitly through end-to-end optimization driven by the sentiment prediction objective. The refined multimodal features are fed into a KANs-based fusion module, which replaces conventional MLPs.

By leveraging functional decomposition, KANs enable explicit, high-order nonlinear modeling of intermodal interactions, providing stronger expressive capacity and improved interpretability compared to static linear fusion layers. The same KAN-based architecture is employed as the final prediction head to support both classification and regression sentiment tasks. The details of module extraction are described in detail below.

3.2. Visual modality: multi-scale feature extraction from images

Visual information plays a critical yet challenging role in multimodal sentiment analysis, as emotional expressions conveyed through facial movements and gestures are often subtle, noisy, and highly subject-dependent. To robustly capture visual affective cues while mitigating sensitivity to image quality and individual expression variability, a two-stage visual encoding strategy is adopted. As shown in the visual branch of Figure 1, we propose a hybrid architecture that integrates a ViT, MDFA, and KANs to transcend the limitations of conventional CNN-based feature extractors.

First, video frames are processed using a ViT backbone to extract global spatial representations. Unlike convolution-based encoders that emphasize local patterns, ViT models long-range spatial dependencies through self-attention, enabling the extraction of holistic facial configurations and contextual visual semantics. This design choice enhances robustness to variations in pose, illumination, and background clutter, which are common sources of visual noise in real-world sentiment datasets. To better understand the principles of the ViT, one can refer to the model overview (see Figure 2).

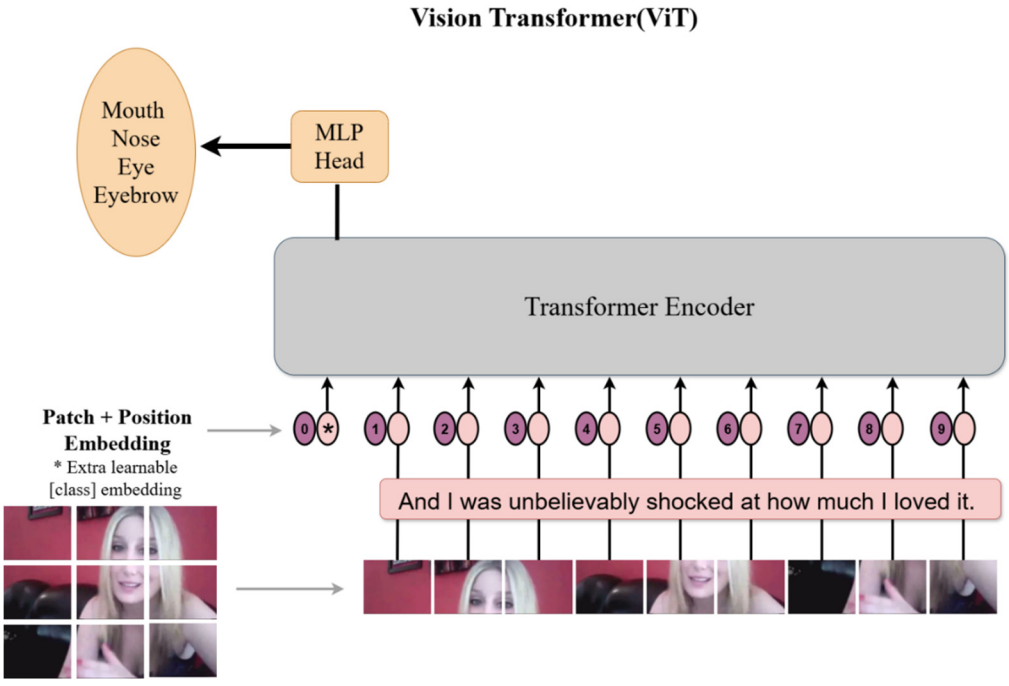


Figure 2. Model overview of the ViT.

However, frame-level representations alone are insufficient to capture the temporal dynamics of emotional expressions, which often evolve gradually rather than appearing instantaneously. To address this limitation, an MDFA module is applied to the sequence of ViT embeddings. By employing multiple dilation rates, MDFA aggregates temporal information at different receptive fields, allowing both short-term micro-expressions and longer-term expression trends to be modeled simultaneously. This multi-scale temporal design improves sensitivity to emotion intensity changes while reducing overfitting to transient visual artifacts.

The input feature map $F_v \in R^{H' \times W' \times D}$ undergoes parallel dilated convolutions with dilation rates $d \in \{1,3,6\}$, capturing both local and nonlocal dependencies. The dilated convolution operation is defined as:

$$F_v^{(d)} = \sum_{(i,j)} K(i,j) \cdot X(i \cdot d, j \cdot d), \tag{1}$$

where $K(i,j)$ denotes the convolution kernel, and d is the dilation rate, effectively enlarging the receptive field without additional computational overhead. The extracted feature maps are concatenated along the channel dimension:

$$F_v^{concat} = Concat[F_v^{(1)}; F_v^{(3)}; F_v^{(6)}]. \tag{2}$$

To avoid channel explosion and maintain computational efficiency, we employed a 1×1 convolutional bottleneck for dimensionality reduction:

$$F_v = \sigma(W_{(1\times1)}(F_v^{concat})), \tag{3}$$

where $W_{1\times1}$ is a learnable weight matrix and σ denotes a nonlinear activation function. This operation yields a refined, multi-scale feature map that integrates fine-grained and global details, complementing the ViT-derived representations.

Importantly, the proposed visual feature extraction strategy does not rely on explicit facial landmark detection or handcrafted expression rules. Instead, emotional saliency is learned implicitly through downstream supervision, allowing the model to adapt to diverse expression styles across individuals. This data-driven approach alleviates the subjectivity inherent in human emotional expression and reduces dependence on image-level perfection.

To finalize the feature extraction pipeline, we replace the conventional MLP classification head with KANs, leveraging the Kolmogorov-Arnold representation theorem to enable highly expressive feature modeling. Given an input feature representation F_v, KANs model the transformation as:

$$F_v^{KANs} = \sum_i g_i(h_i(F_v)), \tag{4}$$

where $h_i(\cdot)$ represents learnable nonlinear transformations (spline functions), capable of capturing complex feature interactions. $g_i(\cdot)$ represents a feature mixing function, dynamically weighing the contributions of different transformed representations. This final visual representation F_v^{KANs} is passed into the GCSA fusion module, where it is integrated with textual and acoustic features for downstream sentiment prediction.

3.3. Text modality: multi-granularity semantic encoding

Textual modality serves as the most explicit carrier of sentiment information in multimodal

sentiment analysis, as emotional states are often directly conveyed through lexical choice, syntactic structure, and contextual semantics. However, textual sentiment understanding is inherently complex due to phenomena such as sarcasm, contextual polarity shifts, and long-range semantic dependencies. To address these challenges, MEGAKANs adopts a hierarchical textual modeling strategy that jointly captures contextual semantics, local sentiment patterns, and long-term dependencies.

BERT is a powerful pre-trained language model that employs a bidirectional transformer architecture to capture complex contextual relationships in language. BERT extracts semantic and syntactic features from text sequences through multiple layers of transformers:

$$f_t = BERT(\mathrm{X}), \tag{5}$$

where $f_t \in R^{n \times d}$ represents the embeddings of each word in a sentence, n is the sentence length, and $d = 768$ is the embedding dimension.

While contextual embeddings encode global semantics, sentiment is often expressed through localized linguistic patterns, such as short phrases, intensifiers, or negations. To explicitly capture these patterns, a convolutional operation is applied over the contextual token representations. TextCNN extracts local features from text, such as n-gram patterns, using convolutional neural networks. The model applies convolution operations with kernels of different sizes (e.g., windows of 3, 4 or 5 words) to the input text, capturing local contextual information at different granularities. The structure of TextCNN is as follows:

$$c_i = ReLU(\mathrm{W}_i * f_t + b_i), \tag{6}$$

where W_i is the convolution kernel weight matrix, $*$ represents the convolution operation, f_t is the input feature matrix, and b_i is the bias term. Each convolutional kernel extracts features from different n-grams, and then the most salient features are selected through max-pooling:

$$f_t^{local} = MaxPooling(c_i). \tag{7}$$

This results in a fixed-length local feature vector representing the most informative n-gram patterns. The final TextCNN output is mapped into a fixed-size local feature vector for subsequent fusion.

In addition to local patterns, sentiment interpretation frequently depends on long-range semantic relations, especially in conversational or opinionated text. To model such dependencies, the contextual embeddings are further processed by a bidirectional recurrent mechanism:

$$\overrightarrow{h_t} = GRU(H_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = GRU(H_t, \overleftarrow{h_{t+1}}). \tag{8}$$

The final Bi-GRU output is:

$$f_t^{global} = h_t = [(\overrightarrow{h_t}); (\overleftarrow{h_t})]. \tag{9}$$

This bidirectional structure enables the model to aggregate sentiment-relevant information from both past and future contexts, enhancing robustness in cases where sentiment polarity emerges gradually or is revealed only at later stages of the utterance.

Finally, the local text features f_t^{local} obtained by TextCNN and the global text features f_t^{global} obtained by BiGRU are input into the multi-head attention mechanism to obtain the final output F_t:

$$F_t = Attention(Q,K,V) = softmax(\frac{f_t^{local} f_t^{global\boldsymbol{T}}}{\sqrt{d_k}})V. \quad (10)$$

This produces the final unified text representation F_t, incorporating both n-gram saliency and contextual dependencies.

Regarding multilingual applicability, the current study focuses on English-language benchmarks to ensure controlled and fair evaluation. Extending the framework to multilingual or cross-lingual sentiment analysis would require language-specific embedding adaptation or multilingual pretraining, which lies beyond the scope of this work and is identified as a promising direction for future research.

Similar to the image mode, we input the eigenvalues obtained from the text mode into F_t in the KAN structure as follows:

$$F_t^{KANs} = \sum_i g_i(h_i(F_t)), \quad (11)$$

where $h_i(\cdot)$ represents learnable nonlinear transformations (spline functions), capable of capturing complex feature interactions. $g_i(\cdot)$ represents a feature mixing function, dynamically weighing the contributions of different transformed representations.

3.4. Audio modality: temporal-frequency feature encoding and deep acoustic representation

Audio modality conveys rich paralinguistic information that is often unavailable in text or visual signals, such as tone, pitch variation, speaking rate, and prosodic emphasis. These cues play a critical role in revealing implicit emotions, especially in cases involving sarcasm, hesitation, or emotional intensity. However, audio sentiment modeling faces challenges arising from background noise, speaker variability, and temporal misalignment with other modalities. MEGAKANs addresses these challenges by emphasizing temporal dependency modeling and adaptive feature refinement, rather than relying solely on handcrafted acoustic descriptors.

First, we extract audio from video data using the MoviePy library and save it in .wav format. To ensure consistency and quality of the audio data, all extracted audio is resampled to a uniform sampling rate (e.g., 16 kHz). This process avoids frequency distortion caused by different sampling rates, ensuring more accurate feature extraction. The audio data is then converted into a Mel spectrum, which provides a two-dimensional time-frequency representation of the audio signal. The data is first subjected to a short-time Fourier transform (STFT):

$$X(n,k) = \sum_{t=0}^{L-1} x(t)w(t-nH)e^{-\frac{j2\pi kt}{L}}, \quad (12)$$

where $w(\cdot)$ is the window function, and H represents the frame shift. Then the calculation of the Mel spectrum is as follows:

$$M(m,n) = log\left(\sum_{k=f_{min}}^{f_{max}} |X(k)|^2 \cdot H_m(k)\right), \quad (13)$$

where $M(m,n)$ represents the value of the m-th Mel band at time frame n, $|X(k)|^2$ is the power spectrum of the Fourier transform, and $H_m(k)$ is the frequency response of the m-th Mel filter. The logarithmic amplitude transformation of the Mel spectrum makes it more suitable for neural network

processing, allowing learning over a wider energy range. The output size is 96 frames by 64 Mel bands (approximately 0.96 seconds by 64 dimensions).

YAMNet consists of convolution blocks, each of which includes a depthwise convolution (a 3 × 3 convolution for each channel) and a pointwise convolution (a 1 × 1 convolution for cross-channel mixing). The block computation is as follows:

$$DWConv: Z_d = K_d * X, PWConv: Y = K_p \otimes Z_d, \tag{14}$$

where $*$ denotes channel-by-channel convolution, and $\otimes$ denotes 1 × 1 convolution. Therefore, the final output F_a of the audio data can be obtained by inputting the audio features f_a from the Mel spectrum to YAMNet:

$$F_a = YAMNet(f_a). \tag{15}$$

YAMNet offers deep learning-based abstract feature representations. This ability to handle complex emotional scenarios is critical, as emotional expressions are often multi-layered and dynamic.

Finally, we leverage the Kolmogorov-Arnold representation theorem to enable highly expressive feature modeling. Given an input feature representation F_a, KANs model the transformation as:

$$F_a^{KANs} = \sum_i g_i(h_i(F_a)). \tag{16}$$

Here, $h_i(\cdot)$ represents learnable nonlinear transformations, capable of capturing complex feature interactions. $g_i(\cdot)$ represents a feature mixing function, dynamically weighting the contributions of different transformed representations. KANs enable adaptive, interpretable, and expressive modeling beyond traditional MLP layers, especially for capturing fine-grained emotional variations in auditory patterns.

3.5. Multimodal feature fusion via GCSAKANs

To enhance the representational quality of multimodal features and better guide the subsequent nonlinear fusion via KANs, we propose a GCSA mechanism. The GCSA module is designed to adaptively recalibrate multimodal features by emphasizing emotionally salient patterns while suppressing less informative signals. In this context, feature importance is not predefined or manually annotated but is implicitly learned through end-to-end optimization driven by the sentiment prediction objective. Specifically, GCSA decomposes attention modeling into two complementary dimensions: a channel attention module (CAM) and a spatial attention module (SAM), applied to refine the concatenated multimodal features.

3.5.1. Feature standardization and concatenation

Due to the inherent dimensional and distributional discrepancies across modalities, direct fusion of these features may lead to representation bias and unstable optimization. To enable effective cross-modal interaction, modality-specific features are first projected into a shared latent space with consistent dimensionality. This normalization step mitigates scale and distribution mismatches across modalities, ensuring that no single modality dominates the fusion process due to representational bias.

The features from different modalities may have different distributions and value ranges, so the standardization process eliminates these differences by mapping them to the same feature space, ensuring that they are treated fairly in the subsequent fusion process. In the previous section, we obtained the eigenvectors of each mode as F_t, F_a, F_v ($F_v^{KANs}, F_t^{KANs}, F_a^{KANs}$ for fusion KANs). We standardize each feature vector using the following formula:

$$F_m' = \frac{F_m - \mu_m}{\sigma_m}, m \in \{t, a, v\}, \tag{17}$$

where F_m' represents the standardized feature vector, and μ_m and σ_m are the mean and standard deviation of modality m (text, audio, or image). This operation aligns heterogeneous modality features into a unified latent space with consistent dimensionality and comparable statistical properties, thereby facilitating subsequent fusion operations.

After standardization, the multimodal feature vectors F_t', F_a', F_v' are concatenated to form a comprehensive feature vector:

$$F_{fusion} = [F_t'; F_a'; F_v'], \tag{18}$$

where F_{fusion} represents the concatenated comprehensive feature vector, containing all the information from text, audio, and image. To further optimize the fused features, we introduce a GCSA mechanism. GCSA dynamically adjusts the weights of the features, emphasizing important emotional features and de-emphasizing irrelevant ones. Applying GCSA to the fused feature matrix F_{fusion} enhances the model's attention and recognition performance. This unified representation serves as the input to the GCSA module, enabling joint modeling of inter-modal dependencies.

3.5.2. Channel attention via KANs

Channel-wise attention focuses on modeling inter-modality and inter-feature dependencies. Each channel corresponds to a latent semantic dimension derived from the fused multimodal representation. Global pooling operations aggregate global contextual information, producing a compact descriptor that summarizes feature activation patterns across modalities.

Given an input multimodal feature tensor $F_{fusion} \in R^{C \times H \times W}$, we apply global average pooling (GAP) to extract a channel descriptor:

$$z = GAP(F_{fusion}) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(c, i, j), z \in R^C. \tag{19}$$

In the conventional CAM, this descriptor is passed through a two-layer MLP. In MEGAKANs, we replace the MLP with a two-layer Kolmogorov-Arnold network K_θ, which transforms z into a channel attention vector:

$$Attention_{chan} = \sigma(K_\theta(z)), Attention_{chan} \in R^C, \tag{20}$$

where $K_\theta(\cdot)$ represents a two-layer KAN with learnable spline parameters θ, and $\sigma(\cdot)$ is the sigmoid activation function, ensuring weights are bounded in [0,1]. Each KAN layers l implements the transformation:

$$h^{(l)} = \sum_{d=1}^{D} w_d^{(l)} \cdot \phi_d{}^{l}(h^{(l-1)}), \tag{21}$$

where ϕ_d^l is a univariate spline or piecewise polynomial basis function, $w_d^{(l)}$ is a learnable scalar, $h^{(0)} = z$, and $h^{(L)} = K_\theta(z)$. Channels receiving higher attention weights correspond to feature dimensions that contribute more strongly to sentiment prediction during training, whereas lower weights indicate less informative or redundant features.

This descriptor is then passed through a lightweight nonlinear transformation to generate channel-wise attention weights. These weights act as adaptive gates, amplifying channels that encode sentiment-relevant information (e.g., emotionally salient facial expressions or emphasized acoustic patterns) and suppressing channels that contribute marginally or inconsistently. Importantly, this process is entirely learnable and does not rely on any handcrafted heuristics.

The resulting attention vector is broadcasted and applied to the original feature map:

$$ChannelAttention(F_{fusion}) = \text{Attention}_{chan}(\text{z}) \cdot F_{fusion}. \tag{22}$$

3.5.3. Spatial attention via KANs

While channel attention emphasizes what features matter, spatial attention focuses on where sentiment-relevant information is located within the feature map. In the context of multimodal sentiment analysis, spatial dimensions may correspond to temporal segments, visual regions, or aligned feature positions.

In the original SAM, the spatial attention map is generated by applying a 7 × 7 convolution to a concatenation of average and max-pooled spatial descriptors. In MEGAKANs, we replace this convolution with a KANs-based mapping to model spatial dependencies more flexibly.

Let $F_{avg}, F_{max} \in R^{1\times H\times W}$ be the average and max-pooled maps along the channel axis:

$$F_{avg} = \frac{1}{C}\sum_{c=1}^{C} F'(c,:,:), F_{max} = \max_{c=1}^{C} F'(c,:,:). \tag{23}$$

We concatenate them and flatten the spatial dimension to form an input vector:

$$s = Flatten([F_{avg}; F_{max}]) \in R^{2HW}. \tag{24}$$

This vector is fed into a second KANs:

$$Attention_{Spat} = \sigma(K_\theta(s)), Attention_{Spat} \in R^{H\times W}. \tag{25}$$

The spatial attention map is reshaped and applied elementwise:

$$SpatialAttention(F_{fusion}) = \text{Attention}_{Spat}(i,j) \cdot F_{fusion}. \tag{26}$$

Spatial attention aggregates local and global cues to generate a spatial importance map, highlighting regions that are strongly correlated with sentiment expression. For instance, frames containing pronounced facial expressions or temporal segments with strong vocal emphasis may receive higher attention weights. Conversely, neutral or noisy regions are down-weighted during fusion.

3.5.4. Final fusion and output

The refined multimodal feature representation $F_{MEGAKANs}$ encapsulates both channel- and spatial-wise dependencies. This is subsequently fed into the final fusion KANs for nonlinear integration and sentiment prediction. This functional decomposition enables the model to explicitly represent nonlinear interactions between individual feature dimensions originating from different modalities. The mathematical expression for the GCSA mechanism is as follows:

$$F_{MEGAKANs} = GCSA(F_{fusion}) = \alpha \cdot \text{ChannelAttention}(F_{fusion}) + \beta \cdot SpatialAttention(F_{fusion}), \quad (27)$$

where $\text{ChannelAttention}(F_{fusion})$ and $\text{SpatialAttention}(F_{fusion})$ represent the attention operations on the channel and spatial dimensions, respectively, and α and β are learnable weight parameters. The GCSA mechanism effectively adjusts the weights of different modalities based on the input emotional features, enhancing attention to key emotional features. Through joint channel-spatial reweighting, GCSA enables the model to dynamically focus on emotionally relevant cues—such as expressive facial regions, emphasized acoustic patterns, or sentiment-bearing textual components—without relying on explicit supervision. This data-driven evaluation mechanism ensures that feature importance is context-sensitive and task-dependent, aligning with the inherent subjectivity and variability of emotional expression.

This structure enables end-to-end optimization of emotional saliency, inter-modal correlation, and nonlinear transformation under the unified KANs paradigm. The KAN-based fusion in MEGAKANs serves as a principled alternative to conventional MLP-based fusion, offering explicit, adaptive, and interpretable modeling of intermodal dependencies rather than implicit feature aggregation.

3.6. Training optimization and notation clarification

The proposed MEGAKANs framework is trained in a supervised manner to support both classification and regression-based sentiment prediction, following the standard evaluation protocols of the CMU-MOSI benchmark.

For classification tasks, including binary (Acc_2) and multi-class (Acc_5, Acc_7) sentiment prediction, the model is optimized using the categorical cross-entropy loss:

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

where C denotes the number of sentiment classes, y_c is the ground-truth label, and $\hat{y}_c$ is the predicted class probability.

For regression-based sentiment intensity prediction, the MSE loss is employed:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2 \ , \quad (29)$$

where s_i and $\hat{s}_i$ denote the ground-truth and predicted sentiment scores, respectively.

During training, the optimization objective is selected according to the task configuration. No auxiliary losses or task-specific heuristics are introduced, ensuring that performance gains arise from the proposed fusion mechanism rather than optimization tricks.

All model parameters, including modality-specific encoders, the GCSA module, and the

KAN-based fusion network, are jointly optimized using the Adam optimizer with a fixed learning rate. Early stopping based on validation performance is applied to prevent overfitting and ensure stable convergence.

To assist readers in navigating the mathematical formulations presented in this section, Table 2 summarizes the key symbols and notations used in the proposed method, including modality-specific features, attention weights, fusion operators, and loss-related variables. This table serves as a unified reference to facilitate comprehension and avoid redundancy across subsections.

Table 2. Summary of symbols and notations used in the proposed method.

Symbol	Description	Symbol	Description
f_v	Visual feature representation after ViT	F_v	Visual feature representation after ViT-MDFA
f_a	Audio feature representation after mel	F_a	Audio feature representation after mel-YAMNet
f_t	Text feature representation after BERT	F_t	Text feature representation after BERT-TextCNN/BiGRU
f_t^{local}	Text local feature representation after TextCNN	F_v^{concat}	Visual feature maps concatenated
f_t^{global}	Text global feature representation after BiGRU	W_i	Convolution kernel weight matrix for text feature
F_v^{KANs}	Visual feature representation after ViT-MDFA-KANs	$h_i(\cdot)$	Learnable nonlinear transformations
F_t^{KANs}	Text feature representation after BERT-TextCNN/BiGRU-KANs	$w(\cdot)$	Window function for audio feature
F_a^{KANs}	Audio feature representation after mel-YAMNet-KANs	$g_i(\cdot)$	Feature mixing function
$H_m(k)$	Frequency response of the *m*-th Mel filter	$F_m^{'}$	Normalized feature of modality
F_{fusion}	Concatenated multimodal feature vector	$Attention_{Chan}$	Channel attention
$K_\theta(\cdot)$	A two-layer KANs with learnable spline parameters θ	$Attention_{Spat}$	Spatial attention
$\sigma(\cdot)$	Sigmoid activation function	$F_{MEGAKANs}$	Multimodal final output

4. Experiments

In the following sections, we first describe the public benchmark datasets and introduce the experimental settings, ten baseline models, and twenty-two multimodal sentiment analysis models. To evaluate the effectiveness of the KANs in multimodal sentiment analysis, we conducted a series of experiments using three modalities: text, audio, and visual data. The models were evaluated under four conditions:

(1) Comparison of results between MEGAKANs and other baseline models.
(2) Multimodal fusion with KANs.
(3) Modality-specific performance.
(4) Impact of KANs on modal contributions.

Finally, we present and discuss the results obtained.

4.1. Experimental setup

4.1.1. Dataset

The statistical characteristics of the CMU-MOSI are summarized in Table 3, and their details are as follows:

We used the CMU-MOSI dataset, which is a widely used benchmark in the field of multimodal sentiment analysis. It consists of 2199 short video clips extracted from 93 YouTube videos, where individuals express opinions on various movie topics. Each video clip is annotated with a sentiment score ranging from −3 (strongly negative) to +3 (strongly positive), following the Stanford sentiment treebank annotation method.

Table 3. Statistics for the dataset.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199

The CMU-MOSI dataset includes three modalities: linguistic (text), visual (facial expressions), and acoustic (voice tone), making it particularly valuable for studying the interplay between different communication modes in sentiment expression. When dealing with CMU-MOSI datasets, text data and label values can be read directly from the given table, while the audio data and image data are both in the video file of the original file, and the files need to be processed by themselves, which is an important part of this dataset in the data prediction. As a robust resource for sentiment analysis, CMU-MOSI is often employed in research to evaluate the performance of multimodal models in capturing and interpreting sentiment from diverse sources of information.

The CMU-MOSI dataset is publicly available at: https://github.com/pliang279/multibenchand. Accessed using a standardized preprocessing pipeline provided by the Software Development Kit (SDK).

4.1.2. Evaluation criteria

To comprehensively evaluate the effectiveness of multimodal sentiment analysis models, especially on complex datasets such as CMU-MOSI, we adopt a diverse set of evaluation metrics tailored to both classification and regression tasks. These metrics collectively assess a model's ability to correctly interpret sentiment polarity, intensity, and nuanced affective states.

(1) F1-score

The F1-score measures the harmonic meaning between precision and recall, making it particularly valuable for imbalanced datasets. It is given by:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{30}$$

where

$$Precison = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}. \tag{31}$$

Here, *TP* is true positive, *FP* is false positive, and *FN* is false negative. A higher F1-score indicates that the model achieves both high correctness and completeness in sentiment detection.

(2) Binary classification accuracy (Acc_2)

Acc_2 evaluates the model's performance in binary sentiment classification, where sentiment labels are divided into positive and negative categories. This metric is defined as:

$$Acc\_2 = \frac{1}{N}\sum_{i=1}^{N} \text{II}(\hat{y}_i = y_i), \tag{32}$$

where N is the total number of test samples, $\hat{y}_i \in \{0,1\}$ is the predicted label for the i^{th} sample, y_i is the corresponding ground truth label, and $\text{II}(\bullet)$ is the indicator function.

In CMU-MOSI, there are two common schemes for binarization: negative vs. non-negative (which includes neutral scores), and negative vs. positive (which excludes neutral scores (scores $\in$ $(-0.5, 0.5]$)). The latter setting is widely used in recent literature, as it yields more distinct sentiment separation and typically increases accuracy by 2–3%.

(3) Five-class classification accuracy (Acc_5)

Acc_5 divides the sentiment score into five discrete categories, usually mapped from the continuous range $[-3, +3]$ as follows: strongly negative: $[-3, -2]$, negative: $(-2, -1]$, neutral: $(-1, +1]$, positive: $(+1, +2]$, and strongly positive: $(+2, +3]$.

The accuracy is computed as the percentage of correctly classified samples within these five categories:

$$Acc\_5 = \frac{1}{N}\sum_{i=1}^{N} \text{II}(\hat{c}_i = c_i), \tag{33}$$

where $\hat{c}_i$ and c_i are the predicted and true sentiment categories.

(4) Seven-class classification accuracy (Acc_7)

Acc_7 offers finer granularity, dividing sentiment into seven classes corresponding to the discrete values:

$$\{-3, -2, -1, 0, +1, +2, +3\}.$$

This formulation enables precise sentiment intensity evaluation but also increases task difficulty. The classification accuracy is similarly defined:

$$Acc\_7 = \frac{1}{N}\sum_{i=1}^{N} \text{II}(\hat{c}_i = c_i),\ c_i \in \{-3,...,+3\}. \tag{34}$$

Due to higher class imbalance and the inclusion of neutral/ambiguous samples (class 0), this metric is considered challenging yet informative.

(5) Mean absolute error (MAE)

MAE is used in regression-based sentiment prediction, quantifying the average absolute difference between predicted sentiment scores $\hat{y}_i$ and ground truth values y_i:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i = y_i |. \tag{35}$$

Lower MAE indicates higher accuracy in predicting the sentiment intensity and is particularly useful when sentiment is represented on a continuous scale.

(6) Pearson correlation coefficient (Corr)

Corr measures the linear correlation between predicted and true sentimental values. It reflects how well the trend of predictions aligns with the true scores:

$$Corr = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}. \tag{36}$$

Here, $\bar{\hat{y}}$ and $\bar{y}$ denote the means of the predicted and ground truth sentiment values, respectively. Corr values range from −1 (perfect inverse correlation) to +1 (perfect direct correlation), with higher values indicating stronger predictive alignment.

4.1.3. Baseline models

To ensure a fair and comprehensive evaluation, MEGAKANs is compared against a diverse set of representative baseline models covering unimodal and multimodal sentiment analysis paradigms. All baselines are selected based on three criteria: (1) relevance to multimodal sentiment analysis, (2) frequent adoption in prior CMU-MOSI studies, and (3) publicly available or reproducible implementations.

The baseline models span multiple fusion strategies, including tensor-based fusion, attention-based fusion, graph-based modeling, and transformer-driven cross-modal alignment. This diversity allows for systematic evaluation of the proposed method under different architectural assumptions.

For clarity and brevity, detailed architectural descriptions of these models are provided in the Related Work section, while this section focuses on their role in experimental comparison. Table 4 summarizes all baseline models used in the experiments.

The baseline models are grouped into classical fusion, transformer-based, graph-based, memory-driven, and self-supervised categories, as summarized in Table 4. This categorization highlights the differences in fusion strategy, temporal modeling capability, and representation learning mechanisms, providing a structured comparison with the proposed MEGAKANs framework.

Our proposed MEGAKANs is positioned at the intersection of nonlinear modeling and dynamic attention recalibration, bridging interpretability with adaptability. Through this comparison, we aim to demonstrate that MEGAKANs not only integrates the strengths of prior work but also surpasses them in both interpretive capability and sentiment prediction performance.

These baselines provide a comprehensive benchmark for validating our model's contributions. In the following sections, we will present detailed experimental results comparing these models quantitatively and qualitatively.

Table 4. Categorization and comparison of baseline multimodal sentiment analysis models.

Category	Model	Core idea	Key mechanism	Temporal modeling
Classical Fusion	TFN [35]	Exhaustive outer-product fusion	Full tensor interaction	✗
	LMF [36]	Low-rank tensor factorization	Rank-constrained fusion	✗
Transformer-based	MulT [37]	Modality-asynchronous cross-attention	Cross-modal transformers	✓
	MTFN [38]	Unified self-attention fusion	Transformer on joint space	✓
Graph & Attention-based	BIMHA [39]	Hierarchical bimodal attention	Intra-/inter-modal attention	✗
	GraphCAGE [40]	Context-aware sentiment graph	Gated GCN propagation	✓
	w/GCN [41]	Fixed-structure graph modeling	Vanilla GCN	✗
	w/HGCN [41]	Hierarchical graph abstraction	Multi-level GCN	✗
	MDH [41]	Dynamic hierarchical weighting	Context-aware fusion	✓
Hybrid & Feature Matching	FmIMSN [42]	Inter-modality feature alignment	Feature matching	✗
	CMHFM [43]	Hybrid semantic fusion	Cross-modal attention	✗
	LMF_MulT [44]	Low-rank + attention	Tensor + transformer	✓
	SWAFN [45]	Sliding-window fusion	Adaptive temporal fusion	✓
Memory & Temporal	MG [46]	Long-term memory gating	Modality-specific memory	✓
	VCAN_R [47]	Visual refinement via residuals	Visual context modeling	✗
	UniVL+MELTR [48]	Vision-language pretraining	Temporal reasoning	✓
Self-supervised	Self-M / Self-MM [49]	Reconstruction-based pretraining	Self-supervised learning	✗
	MMIM [50]	Mutual information maximization	Contrastive & invariant learning	✗
	TEASEL [51]	Transformer-based audio modeling	Audio self-attention	✓
Ours	MEGA	GCSA-based dynamic fusion	Channel-spatial attention	✓
	MEGAKANs	KAN-based nonlinear fusion + GCSA	Functional decomposition + attention	✓

4.1.4. Implementation details

The experiments were conducted on a high-performance computational cluster, with GPU acceleration for efficient training. The models were optimized using Adam with a learning rate of 0.00005. Early stopping was applied to prevent overfitting. The recorded results are the average of over 5 independent runs. Detailed results are presented in Table 5.

Table 5. Hyperparameter settings.

Dataset	Batch size	Epochs	Optimizer	Dropout	Learning rate
CMU-MOSI	32	40	Adma	0.3	0.00005

4.2. Results and analysis

4.2.1. Performance gains by KAN integration

In this section, we show the results of integrating the baseline model into KANs (Table 6). It can be seen from the table that the baseline models fall into four categories, including text, image, audio, and mixed models, and each model has reinforcement models incorporating KANs. The reported results presented are the average of over 5 independent runs. "/" is consistent in accuracy and F1-score, the left side of "/" is "negative / non-negative", and the right is "negative / positive". The effectiveness of the whole model, and how much to improve it, which modes contribute more, and which modes have more influence on the effectiveness of KANs will be discussed in the following subsections. The Acc_2, F1, MAE, Corr, Acc_5, and Acc_7 distributions are shown in Figure 3. In the model, the odd subscript is the model without KANs, and the even subscripts are all the models with KANs.

This result highlights the advantage of KANs in modeling high-order nonlinear feature interactions, which are essential for fine-grained sentiment discrimination. Compared to baseline fusion strategies, the regression improvements suggest that MEGAKANs not only improves classification boundaries but also enhances the overall sentiment representation quality across modalities.

Overall, MEGAKANs consistently outperforms existing methods across all evaluation metrics, with particularly strong gains observed in binary accuracy, fine-grained classification, and regression stability. These results confirm that replacing conventional MLP-based fusion with KANs leads to more expressive and robust multimodal sentiment modeling.

Table 6. Experimental results of the integrated KANs presented under a supervised scenario on the CMU-MOSI dataset.

Modality	Models	CMU-MOSI					
		F1-score ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	MAE ↓	Corr ↑
Text	W2V	48.08/45.92	55.90/54.17	29.92	29.53	1.4375	0.6725
	W2V_KANs	52.46/51.71	56.13/54.17	30.18	29.96	1.4370	0.6787
	Roberta	44.79/41.87	59.83/57.41	31.81	30.35	1.4348	0.6269
	Roberta_KANs	52.84/45.72	61.04/57.69	32.47	30.91	1.4353	0.6935
	BERT_TC_BG	81.55/83.14	81.66/83.33	46.29	36.68	1.0280	0.6508

Continued on next page

Modality	Models	CMU-MOSI					
		F1-score ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	MAE ↓	Corr ↑
	BERT_TC_BG_KANs	83.75/85.45	83.72/85.50	46.73	37.19	1.0132	0.6965
Vison	Cnn	47.30/44.52	58.08/55.56	21.03	21.03	1.4543	0.1410
	Cnn_KANs	66.18/65.92	67.25/67.13	23.58	23.14	1.4368	0.1488
	Transform	54.09/51.98	59.83/57.87	20.09	20.09	1.4156	0.1526
	Transform_KANs	59.31/58.22	63.32/62.50	20.96	20.96	1.4065	0.1599
	ViT_MDFA	61.67/60.73	65.94/65.28	21.72	21.21	1.4176	0.2109
	ViT_MDFA_KANs	67.97/67.91	68.76/68.89	23.61	23.19	1.4023	0.2720
Audio	Spec	53.01/52.44	53.64/52.90	15.45	15.45	1.4698	0.1299
	Spec_KANs	53.99/53.06	54.66/53.51	15.51	15.51	1.4299	0.1454
	Raw	44.79/41.87	59.83/57.41	20.41	20.41	1.4193	0.1596
	Raw_KANs	45.27/42.81	60.14/58.27	21.46	21.46	1.4172	0.1888
	Mel_YAMNet	50.86/48.56	60.26/58.33	22.74	22.51	1.4091	0.1698
	Mel_YAMNet_KANs	55.44/54.30	60.70/58.80	22.92	21.78	1.3984	0.1976
Multimodal	MEGA	83.08/85.20	82.97/85.19	54.15	44.18	0.9107	0.7889
	MEGAKANs(Our)	**84.35/87.02**	**84.28/87.04**	**57.28**	**47.12**	**0.7265**	**0.8024**

*The reported results presented are the average of over 5 independent runs. Note that the left side of "/" is "negative/non-negative" and the right is "negative/positive". (↑ means higher is better and ↓ means lower is better.)

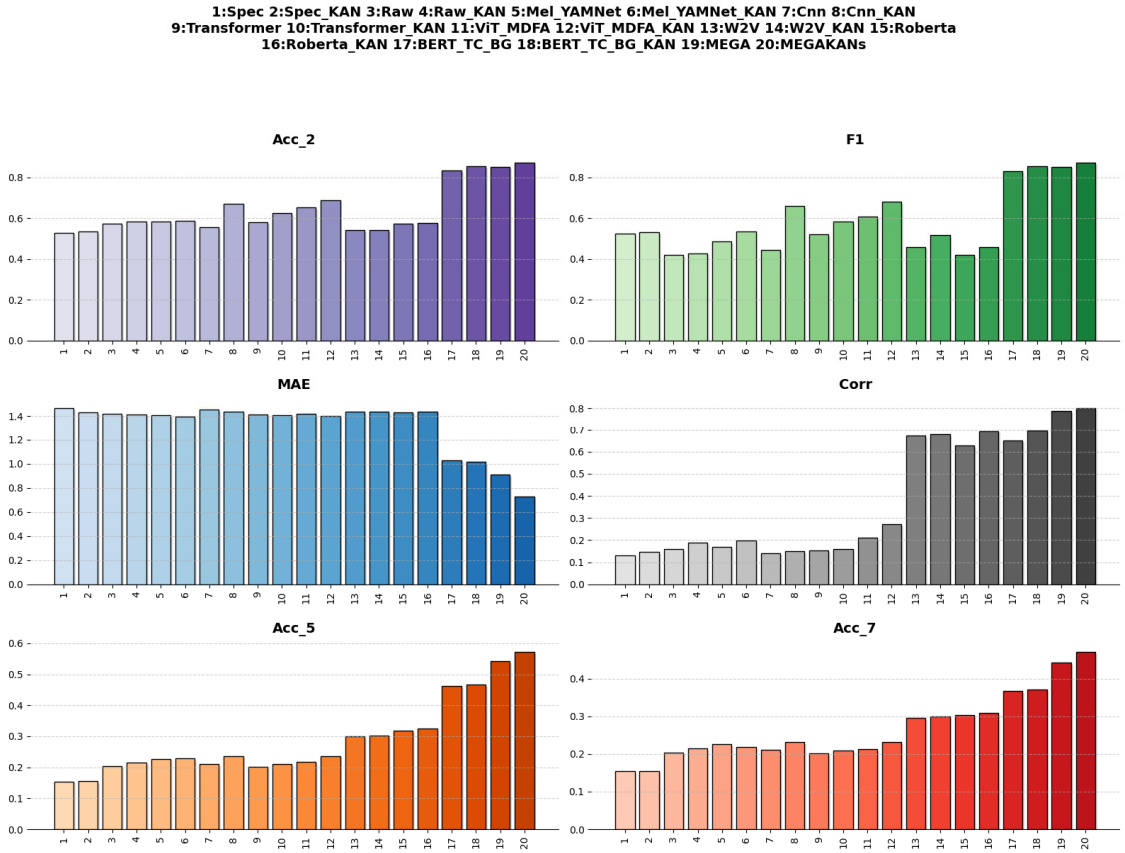


Figure 3. The distribution of accuracy and F1-score for each modality.

4.2.2. Comparison with state-of-the-art multimodal sentiment models

To further validate the superiority of MEGAKANs, we compare it against a collection of state-of-the-art multimodal sentiment analysis models, as presented in Table 7. These baselines include attention-based fusion strategies (e.g., MulT, LMF, MTFN), graph neural networks (e.g., w/GCN, GraphCAGE), memory-based architectures (e.g., MG), and self-supervised learning approaches (e.g., Self-M, MMIM).

Table 7. Experimental results of other multimodal sentiment analysis baseline models on the CMU-MOSI dataset.

Models	CMU-MOSI					
	F1-score ↑	Acc_2 ↑	Acc_5 ↑	Acc_7 ↑	MAE ↓	Corr ↑
VCAN_R	76	75.9	–	36.9	0.93	0.64
LMF	76.71 /77.83	76.76/77.80	38.51	34.17	0.963	0.663
BIMHA	77.53 /78.74	77.58/78.72	41.78	35.35	0.967	0.646
TFN	77.00 /79.37	78.18/79.45	39.85	35.02	0.935	0.658
LMF_MulT	77.9/–	77.9/–	–	32.4	1.016	0.647
FmIMSN	77.95/–	78.11/–	39.80	34.37	0.939	0.658
MTFN	78.49 /79.64	78.51 /79.60	39.87	34.90	0.955	0.659
MulT	78.57 /80.19	78.87 /80.27	40.21	36.27	0.918	0.683
SWAFN	–/80.1	–/80.2	–	40.1	0.88	0.697
GraphCAGE	–/80.27	–/80.18	36.88	37.46	0.9828	0.6688
MG	–/80.5	–/80.6	–	32.1	0.933	0.684
CMHFM	80.08 /81.52	80.18 /81.55	45.97	40.82	0.840	0.722
w/GCN	81.92/–	81.86/–	37.76	32.51	0.9104	0.688
w/HGCN	82.17/–	82.01/–	35.28	33.09	0.9503	0.6564
Self-M	82.68/–	82.54/–	–	45.79	0.712	0.795
Self-MM	82.52 /84.41	82.59 /84.42	–	–	0.719	0.719
MDH	–/82.69	–/82.47	38.63	34.99	0.8998	0.6994
MMIM	–/84	–/84.14	–	46.65	0.7843	0.741
UniVL+MELTR	–/85.4	–/85.3	–	–	0.759	0.789
TEASEL	–/85	–/86.5	–	47.02	0.894	0.736
MEGA	83.08/85.20	82.97/85.19	54.15	44.18	0.9107	0.7889
MEGAKANs	**84.35/87.02**	**84.28/87.04**	**57.28**	**47.12**	**0.7265**	**0.8024**

Overall, MEGAKANs achieves the highest scores across nearly all metrics, with an F1-score of 87.02, Acc_2 of 87.04, Acc_5 of 57.28, and Acc_7 of 47.12. Notably, MEGAKANs also records the lowest MAE (0.7265) and the highest Pearson correlation coefficient (0.8024), demonstrating both predictive accuracy and reliability across sentiment intensities (Table 7). The superiority of MEGAKANs can be attributed to its unique combination of adaptive attention recalibration (GCSA) and expressive nonlinear modeling (KANs), which collectively improve the model's ability to fuse and learn from multimodal signals.

In contrast, models like CMHFM and Self-MM show strong performance in certain dimensions—such as Acc_7 or Corr—but fall short in achieving consistent gains across all

evaluation criteria. For instance, while CMHFM reaches 40.82 in Acc_7, it lags in MAE (0.840) and Corr (0.722), suggesting weaker regression alignment. Similarly, MulT, MTFN, and TFN exhibit solid binary classification performance, yet their lower correlation values reflect a limited ability to model nuanced affective variations.

A particularly revealing contrast lies in the comparison with MMIM, which employs mutual information maximization to enhance modality alignment. While MMIM demonstrates effective correlation modeling (Corr: 0.741), it underperforms in classification tasks compared to MEGAKANs, particularly in Acc_5 and Acc_7. This suggests that although MMIM captures global representations, it lacks the fine-grained intermodal weighting and nonlinearity provided by the GCSA and KAN structure.

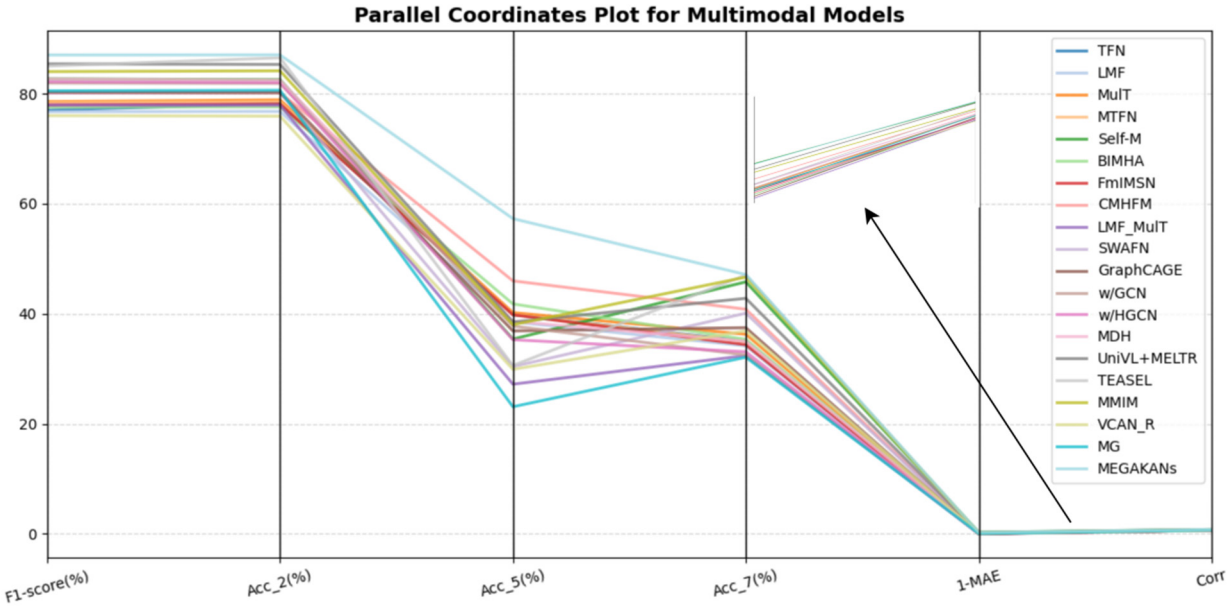


Figure 4. Parallel coordinates plot of multimodal models across six evaluation metrics.

Figure 4 illustrates a parallel coordinates visualization of selected multimodal sentiment analysis models, comparing their performance across six standardized evaluation metrics: F1-score, Acc_2, Acc_5, Acc_7, 1−MAE, and Corr. Each colored line traces an individual model's performance profile across all dimensions, allowing for simultaneous comparison of accuracy, regression precision, and consistency.

The trajectory of MEGAKANs, highlighted in cyan, remains consistently close to the upper bounds of all performance axes, demonstrating both superior accuracy and robust regression capability. Unlike many baseline models that exhibit sharp drops in multi-class accuracy or regression metrics, MEGAKANs maintains a well-balanced performance across all criteria, indicating its generalization ability across sentiment granularity levels and task types.

This visualization underscores the comprehensive advantage of MEGAKANs: it not only surpasses most models in absolute performance but also avoids the typical trade-offs seen in other approaches, thereby achieving multi-objective optimization across classification and regression tasks.

4.2.3. Multimodal fusion with KANs

The integration of KANs into the multimodal fusion framework significantly enhances the performance of sentiment analysis models across both classification and regression tasks. Leveraging the capacity of KANs to model nonlinear relationships and inter-modality dependencies, the KAN-enhanced model achieves superior accuracy, robustness, and generalizability.

Specifically, in binary classification, the proposed model attains an Acc_2 (negative/positive) of 87.02, a marked improvement over its non-KANs counterpart. As illustrated in Figure 5, models that incorporate KANs exhibit a substantial increase in Acc_2 and F1-scores. This suggests that KAN-enhanced models are better at distinguishing between positive and negative sentiments, likely due to their ability to capture more nuanced emotional expressions. Furthermore, F1-scores reflect a similar trend, with KAN-enhanced models consistently achieving higher precision and recall. This improvement in F1 indicates that KANs contribute to a more balanced classification, reducing the likelihood of misclassifications in edge cases, such as weakly positive or weakly negative sentiments. The observed increase in F1-score confirms that KANs improve not only the accuracy but also the reliability of sentiment classification.

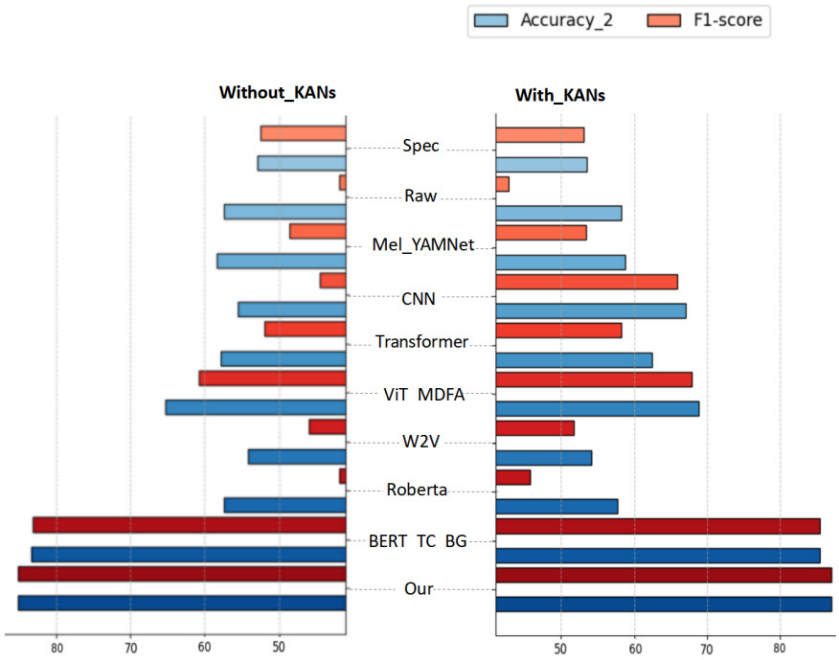


Figure 5. A comparison of Accuracy_2 and F1-score between proposed models and previous works for the CMU-MOSI.

Beyond binary classification, the model's performance on multi-class sentiment tasks further highlights the advantages of KAN-based fusion. For Acc_5, the model achieves a score of 57, while on the more fine-grained Acc_7, the performance rises from 15 to 47. These gains demonstrate KANs' ability to capture subtle sentiment gradients, allowing for more precise sentiment intensity differentiation—an aspect critical for real-world applications where emotional expression is often nuanced rather than binary. These improvements reflect KANs' ability to capture subtle distinctions

in sentiment, particularly in more fine-grained classification tasks.

However, the pronounced increase in Acc_7 also suggests a degree of task-specific optimization. While such improvement is desirable, it raises the possibility of overfitting the CMU-MOSI dataset. The robustness of KANs across datasets should therefore be further validated on larger and more diverse corpora (e.g., CMU-MOSEI) to ensure generalizability. Moreover, the observed enhancement in performance under both negative/non-negative and negative/positive binary settings implies that KANs may reduce the sentiment polarity ambiguity common in traditional classification boundaries, leading to more stable sentiment categorization.

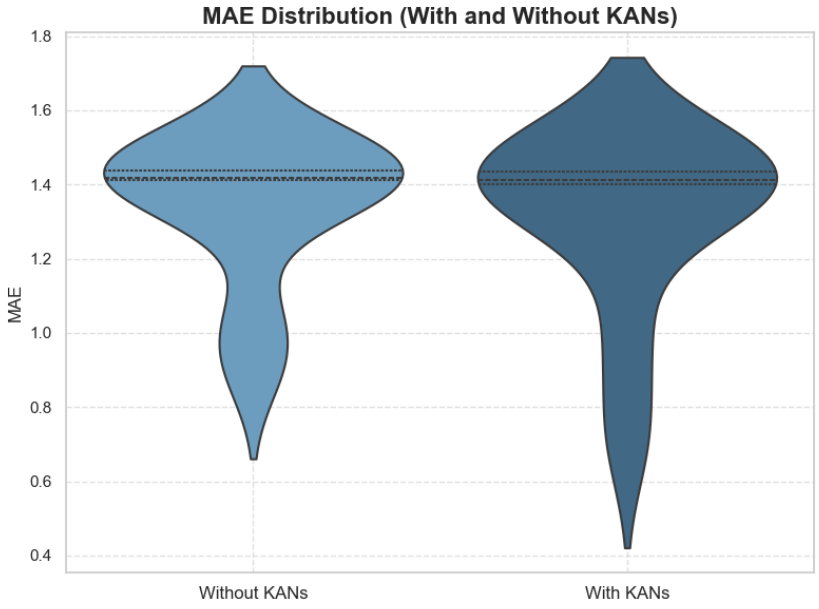


Figure 6. The distribution of MAE for the KAN and non-KAN models.

In regression tasks, the MAE exhibits substantial improvement. As illustrated in Figure 6, the MAE decreases from approximately 1.45 in non-KAN models to as low as 0.91 in the KAN-enhanced models, demonstrating finer sentiment intensity estimation. Simultaneously, the Corr increases, reaching up to 0.80, indicating stronger linear alignment between predicted and true sentimental values. Notably, the distribution of MAE scores across samples narrows under KANs fusion, reflecting reduced variance and enhanced prediction consistency.

These results collectively affirm that KANs not only boost the average performance but also contribute to more stable and reliable multimodal fusion, especially in tasks requiring high sensitivity to inter-modality alignment. Through its learnable spline functions and edgewise nonlinearity, KANs enable a more expressive fusion mechanism that adapts dynamically to modality-specific signal distributions.

4.2.4. Modality-specific performance analysis

To comprehensively assess the contribution of each modality to multimodal sentiment analysis, we conducted a comparative evaluation of models trained independently on text, visual, and audio

features. The results, summarized in Figure 7, reveal distinct differences in predictive effectiveness across modalities, reflecting their varying capacity to capture sentiment-relevant signals.

Among the three modalities, text features consistently achieved the highest performance, with an Acc_2 (negative/positive) of 85.5 and a corresponding F1-score of 85.45. This dominance of textual information can be attributed to its semantic richness and explicit emotional content. Language, being a primary medium of human expression, encodes sentiment directly through word choice, syntactic structures, and context. As a result, textual data provides the most discriminative cues for classifying sentiment polarity. These findings align with prior studies underscoring the robustness of text in sentimental tasks, particularly when pretrained language models are leveraged.

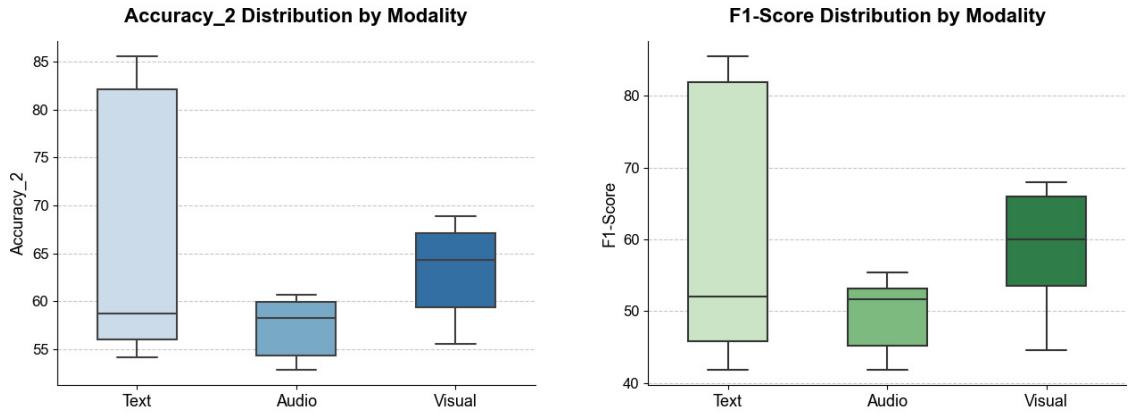


Figure 7. The distribution of Accuracy_2 and F1-scores for each modality.

In contrast, audio features exhibited the lowest performance among the three, with an Acc_2 of 60.7 and F1-score of 55.44. The comparatively limited discriminative power of the audio modality may be due to its susceptibility to noise, speaker variability, and the ambiguity inherent in prosodic cues. Emotional expression through voice—while valuable—is often subtle and context-dependent, making it more difficult for models to extract consistent sentiment indicators. Furthermore, spectral and temporal variations in speech, such as pitch, energy, and speaking rate, do not always correlate linearly with emotional states, posing challenges for feature representation.

Visual features, on the other hand, occupy a middle ground in terms of performance, surpassing audio in both Acc_2 and F1-score by approximately 8 and 14 points, respectively. This suggests that visual cues, particularly facial expressions and micro-expressions, contribute meaningfully to sentiment recognition. Facial expressions often serve as intuitive indicators of affective states and provide non-verbal context that reinforces or clarifies spoken or written sentiments. The improvement in F1-score also indicates enhanced model precision and recall, particularly in cases of subtle or ambiguous emotion expressions.

However, when analyzing the distribution of performance across different methods, text features exhibit significant variability, as shown by the wide range in both Acc_2 and F1-scores. This fluctuation may stem from the diversity of natural language processing techniques and their varying sensitivity to contextual nuances, which can result in substantial performance changes depending on

the model employed. Another possible explanation could be the noise or ambiguities in the textual data itself, where sentiments can be expressed ambiguously, leading to inconsistent model performance.

On the other hand, both audio and visual features demonstrated more stable performance, with relatively smaller fluctuations in Acc_2 and F1-scores. This stability may result from the inherent nature of these modalities, where the data is less likely to vary as dramatically as text due to the more direct association of speech patterns or facial expressions with sentiment. The robustness of these features, particularly in visual data, could be attributed to the more constrained and standardized representation of emotions through facial cues, which are less subject to interpretation compared to text.

This analysis reveals the complementary nature of combining these modalities, where textual information provides depth and granularity, while visual and audio data offer stability and additional context. Thus, a multimodal approach that effectively integrates all three modalities—text, audio, and visual—can lead to more comprehensive and accurate sentiment predictions.

4.2.5. Impact of KANs on modal contributions

In multimodal sentiment analysis, not all features contribute equally to emotional understanding. Emotional salience varies across modalities, spatial regions, and feature channels depending on context. A robust fusion strategy must therefore distinguish informative emotional cues from redundant or noisy signals.

To further understand the role of KANs in enhancing multimodal sentiment analysis, investigate their impact on the individual contributions of each modality—text, audio, and visual—within the fusion architecture. As shown in Figure 8, a comparative analysis is conducted to measure the performance of unimodal feature branches both before and after the introduction of KANs, with a specific focus on their respective Acc_2 metrics (negative vs. positive classification).

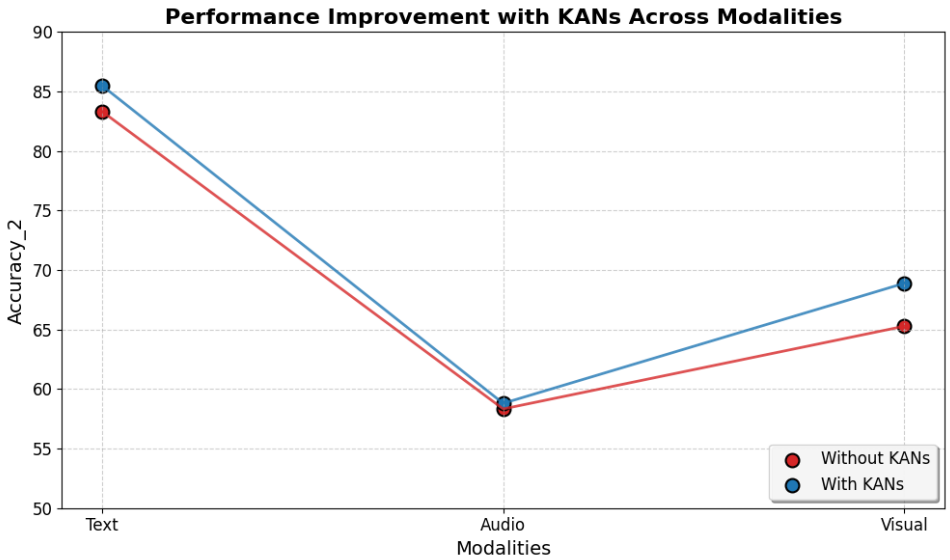


Figure 8. A comparison of modality-specific performance with and without KANs.

Text consistently serves as the dominant modality in sentiment analysis due to its high semantic density and clarity in emotional articulation. Even without KANs, the text modality achieves an

Acc_2 of 83.33, outperforming both audio and visual modalities. After integrating KANs, the accuracy further improves to 85.5, demonstrating a meaningful enhancement. This performance gain can be attributed to KANs' ability to learn nonlinear feature transformations and contextual interactions, thereby enabling more expressive modeling of latent linguistic sentiment cues. In particular, KANs strengthen the model's capacity to discern subtle distinctions such as sarcasm, negation, or mixed sentiment in text—phenomena that are notoriously difficult to capture using standard linear transformations or fully connected networks.

The audio modality exhibits relatively limited improvement following the incorporation of KANs. The unimodal Acc_2 increases modestly from 58.33 to 58.80, suggesting that while KANs enhance representational learning, the intrinsic constraints of the audio signal—such as lower semantic granularity, speaker variability, and the presence of background noise—may cap its standalone effectiveness. Audio features like prosody, rhythm, and intonation do convey affective information; however, their contribution is often context-dependent and less direct than text or facial expressions. Therefore, while KANs refine feature usage, the audio stream alone still lacks the richness needed for high-precision classification in nuanced affective settings.

Notably, the visual modality demonstrates a more substantial benefit from KANs integration. The Acc_2 improves from 65.28 to 68.89, reflecting a significant gain in model expressivity and discriminative power. Visual information—particularly facial expressions, eye movements, and subtle gestures—is inherently non-verbal and often correlates strongly with emotional states. However, without effective fusion mechanisms, such cues may be underutilized. The nonlinear mapping capabilities of KANs allow the model to better align visual and semantic cues, capture co-occurrence patterns (e.g., smile + positive words), and resolve inconsistencies in the emotional display. This facilitates more robust cross-modal integration and amplifies the contribution of the visual stream in the joint representation space. The comparatively moderate standalone performance of the visual modality further suggests that the proposed model does not overfit to visual cues alone. Instead, visual information primarily serves as complementary affective evidence, whose contribution is adaptively regulated during fusion.

The results collectively underscore that KANs are most impactful in modalities with rich yet structurally complex features, such as text and visual data. By contrast, modalities with more abstract or volatile information—such as audio—exhibit relatively lower responsiveness to KANs. Nevertheless, even modest improvements in these weaker modalities can contribute synergistically when integrated into a comprehensive multimodal system. Meanwhile, this suggests that KANs are effective in enhancing the synergy between different modalities, especially in capturing and fusing complex emotional signals. This finding aligns with previous research and further validates KANs' importance in multimodal sentiment analysis.

5. Discussion

The results clearly show that KANs enhance performance across all modalities. This improvement is primarily due to KANs' ability to dynamically adjust the contribution of each modality, leading to better sentiment predictions. Among the three modalities, text consistently demonstrated the highest contribution to sentiment prediction. However, with KANs, audio and visual modalities became more influential, especially in cases of sentimental ambiguity. This shift suggests that KANs facilitate a more holistic understanding of the input data. The performance gains

observed in MEGAKANs should not be interpreted merely as numerical improvements. Rather, they reflect a qualitative shift in how multimodal interactions are modeled—from static parametric fusion to adaptive functional composition—suggesting broader implications for multimodal learning beyond sentiment analysis.

The fusion of modalities using KANs resulted in more accurate and robust sentiment predictions. The analysis shows that without KANs, traditional fusion methods often underutilize the potential of non-text modalities. KANs help mitigate this by enabling better integration and use of all modalities.

Across all metrics, KAN-enhanced models demonstrate lower variance and better overall performance. This suggests that KANs not only improve accuracy but also stabilize model predictions across different tasks and metrics. This stability may result from KANs' ability to effectively capture nonlinear relationships in the data.

The interplay between the Pearson correlation and MAE highlights a potential area of interest. While KAN models perform better in both metrics, a detailed residual analysis might uncover whether these models are biased toward specific sentiment ranges (e.g., misclassifying neutral or extreme sentiments). The increasing accuracy in Acc_5 and Acc_7 reflects the models' growing capacity to handle complex, multi-class sentiment tasks. However, this improvement may come at the cost of overfitting, particularly as the models become more specialized for the dataset. Cross-validation on additional datasets would be a valuable next step to ensure generalizability.

These findings suggest that the advantage of MEGAKANs is not solely dataset-dependent but stems from its ability to represent intermodal relationships in a functionally decomposed and data-adaptive manner, which is difficult to achieve with conventional MLP-based fusion strategies.

In summary, the integration of KANs into multimodal fusion models offers significant performance improvements across classification and regression tasks, while also ensuring greater model stability. Future work should focus on extending the evaluation to additional datasets and refining multimodal fusion strategies to further enhance the robustness of KANs in real-world applications.

6. Conclusions

This paper presents MEGAKANs, a novel multimodal sentiment analysis framework that incorporates KANs into the mid-fusion stage to enhance intermodal representation learning. By leveraging KANs' ability to model high-order, nonlinear dependencies across modalities, this approach improves both classification and regression performance over strong baselines.

Empirical evaluations on the CMU-MOSI dataset demonstrate that MEGAKANs achieves consistent gains across multiple metrics, including Acc_2, Acc_5, Acc_7, F1-score, MAE, and the Pearson correlation. These improvements are primarily attributed to the KAN-based fusion mechanism, which facilitates more effective feature interactions, particularly between weakly correlated modalities. Notably, while textual features remain the dominant modality in sentiment prediction, the integration of KANs yields the largest relative performance gains in the visual modality. This suggests that KANs can amplify subtle affective cues—such as facial expressions and gestures—that are typically difficult to model using traditional MLP-based fusion architectures.

Despite these improvements, several limitations remain. The model exhibits signs of overfitting in fine-grained multi-class classification tasks, as reflected in the performance fluctuations under the Acc_7 metric. This indicates the need for broader empirical validation on larger and more diverse datasets such as CMU-MOSEI. Additionally, while textual and visual modalities benefited significantly from KAN-based modeling, improvements in the audio channel were relatively modest, highlighting

the need for more advanced audio processing strategies or modality-specific adaptation mechanisms.

From an application perspective, MEGAKANs holds promise for deployment in real-world sentiment-aware systems, such as conversational agents, human-computer interaction platforms, and emotion-driven media analysis. Its flexible architecture and learnable fusion mechanism allow for adaptive modulation of modality contributions, making it particularly suitable for settings with heterogeneous or missing inputs.

In future work, we plan to: extend MEGAKANs to temporal and streaming environments for real-time affective state tracking, explore its robustness under partial modality settings and noisy inputs, and investigate the theoretical expressiveness and approximation bounds of KANs in multimodal fusion tasks. Extending the proposed framework to multilingual sentiment analysis constitutes an important direction for future work, particularly to examine how language-specific semantic structures interact with multimodal fusion dynamics.

Overall, this study highlights the potential of functional-decomposition-based fusion models for advancing multimodal sentiment analysis. As research continues, we believe KANs will serve as a powerful tool not only for sentimental understanding but also for a broader range of multimodal reasoning tasks.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is partially supported by the Soft Science Research Project of Henan Province (Grant No. 262400410238).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A, (2023) Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf Fusion* 91: 424–444. https://doi.org/10.1016/j.inffus.2022.09.025
2. Zhao F, Zhang C, Geng B, (2024) Deep multimodal data fusion. *ACM Comput Surv* 56: 1–36. https://doi.org/10.1145/3649447
3. Jiao T, Guo C, Feng X, Chen Y, Song J, (2024) A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Comput Mater Continua* 80: 1. https://doi.org/10.32604/cmc.2024.053204
4. Pawłowski M, Wróblewska A, Sysko-Romańczuk S, (2023) Effective techniques for multimodal data fusion: A comparative analysis. *Sensors* 23: 2381. https://doi.org/10.3390/s23052381

5. Zheng Y, Xu Z, Wang X, (2021) The fusion of deep learning and fuzzy systems: A state-of-the-art survey. *IEEE Trans Fuzzy Syst* 30: 2783–2799. https://doi.org/10.1109/TFUZZ.2021.3062899
6. Zhu L, Zhu Z, Zhang C, Xu Y, Kong X, (2023) Multimodal sentiment analysis based on fusion methods: A survey. *Inf Fusion* 95: 306–325. https://doi.org/10.1016/j.inffus.2023.02.028
7. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S, (2018) Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl Based Syst* 161: 124–133. https://doi.org/10.1016/j.knosys.2018.07.041
8. Cheng H, Yang Z, Zhang X, Yang Y, (2023) Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. *IEEE Trans Affective Comput* 14: 3149–3163. https://doi.org/10.1109/TAFFC.2023.3265653
9. Wang H, Du Q, Xiang Y, (2025) Image-text sentiment analysis based on hierarchical interaction fusion and contrast learning enhanced. *Eng Appl Artif Intell* 146: 110262. https://doi.org/10.1016/j.engappai.2025.110262
10. Liu Z, Zhou B, Chu D, Sun Y, Meng L, (2024) Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf Fusion* 101: 101973. https://doi.org/10.1016/j.inffus.2023.101973
11. Hou Y, Ji T, Zhang D, Stefanidis A, (2024) Kolmogorov-Arnold networks: A critical assessment of claims, performance, and practical viability. preprint, arXiv:2407.11075. https://doi.org/10.48550/arXiv.2407.11075
12. Yu R, Yu W, Wang X, (2024) KAN or MLP: A fairer comparison. preprint, arXiv:2407.16674. https://doi.org/10.48550/arXiv.2407.16674
13. Zadeh A, Zellers R, Pincus E, Morency LP, (2016) Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. preprint, arXiv:1606.06259. https://doi.org/10.48550/arXiv.1606.06259
14. Poria S, Peng H, Hussain A, Howard N, Cambria E, (2017) Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* 261: 217–230. https://doi.org/10.1016/j.neucom.2016.09.117
15. Ezzameli K, Mahersia H, (2023) Emotion recognition from unimodal to multimodal analysis: A review. *Inf Fusion* 99: 101847. https://doi.org/10.1016/j.inffus.2023.101847
16. Li K, Huang Y, Zhong G, Nurmemet Y, Wushouer S, (2025) MHAN: Bottleneck fusion model based on hybrid attention network for multimodal emotion recognition. *J Shanghai Jiaotong Univ Sci* 2025: 1–8. https://doi.org/10.1007/s12204-025-2820-x
17. Gadzicki K, Khamsehashari R, Zetzsche C, (2020) Early vs late fusion in multimodal convolutional neural networks, In: *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 1–6. https://doi.org/10.23919/FUSION45008.2020.9190246
18. Fu Y, Zhang Z, Yang R, Yao C, (2024) Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing* 571: 127201. https://doi.org/10.1016/j.neucom.2023.127201
19. Krizhevsky A, Sutskever I, Hinton GE, (2012) Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012: 25.
20. Hochreiter S, Schmidhuber J, (1997) Long short-term memory. *Neural Comput* 9: 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., (2017) Attention is all you need. *Adv Neural Inf Process Systems* 30.
22. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. preprint, arXiv:2010.11929.
23. Devlin J, Chang MW, Lee K, Toutanova K, (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
24. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al., (2019) Roberta: A robustly optimized Bert pretraining approach. preprint, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
25. Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, et al. (2017) Audio set: An ontology and human-labeled dataset for audio events. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261
26. Hu Y, Wang K, Liu M, Tang H, Nie L, (2023). Semantic collaborative learning for cross-modal moment localization. *ACM Trans Inf Syst* 42: 1–26. https://doi.org/10.1145/3620669
27. Liu WX, Chen DX, Tan MQ, Chen KY, Yin Y, Shang WL, et al. (2024) Model parameter prediction method for accelerating distributed DNN training. *Comput Networks* 255: 110883. https://doi.org/10.1016/j.comnet.2024.110883
28. Schmidt-Hieber J, (2021) The Kolmogorov-Arnold representation theorem revisited. *Neural Networks* 137: 119–126. https://doi.org/10.1016/j.neunet.2021.01.020
29. Cheon M (2024) Demonstrating the efficacy of Kolmogorov-Arnold networks in vision tasks. preprint arXiv:2406.14916. https://doi.org/10.48550/arXiv.2406.14916
30. Liu J, (2024) Exploring the power of KANs: Overcoming MLP limitations in complex data analysis. *Appl Comput Eng* 83: 1–7. https://doi.org/10.54254/2755-2721/83/2024GLG0057
31. Tamotia A, Karmokar DS, Komal R, Nawas KK, Shahina A, Khan AN, (2025) Fusion of multimodal audio data for enhanced speaker identification using Kolmogorov-Arnold networks. *IEEE Access* 2025. https://doi.org/10.1109/ACCESS.2025.3569606
32. Lawan A, Pu J, Yunusa H, Lawan M, Umar A, Yahya AS, et al. (2026). DualKanbaFormer: An efficient selective sparse framework for multimodal aspect-based sentiment analysis. *IEEE Trans Emerging Top Comput Intell* 2026. https://doi.org/10.1109/TETCI.2026.3671067
33. Lawan A, Pu J, Yunusa H, Umar A, Lawan M, (2025) Enhancing long-range dependency with state space model and Kolmogorov-Arnold networks for aspect-based sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2176–2186.
34. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. (2024) KAN: Kolmogorov-Arnold networks. preprint, arXiv:2404.19756. https://doi.org/10.48550/arXiv.2404.19756
35. Zadeh A, Chen M, Poria S, Cambria E, Morency LP, (2017) Tensor fusion network for multimodal sentiment analysis. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. https://doi.org/10.18653/v1/D17-1115

36. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh AB, Morency LP, (2018) Efficient low-rank multimodal fusion with modality-specific factors. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256. https://doi.org/10.18653/v1/P18-1209
37. Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R, (2019) Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. https://doi.org/10.18653/v1/P19-1656
38. Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, et al., (2020) CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727. https://doi.org/10.18653/v1/2020.acl-main.343
39. Wu T, Peng J, Zhang W, Zhang H, Tan S, Yi F, et al., (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl Based Syst* 235: 107676. https://doi.org/10.1016/j.knosys.2021.107676
40. Wu J, Mai S, Hu H, (2021) Graph capsule aggregation for unaligned multimodal sequences. In: *Proceedings of the 2021 International Conference on Multimodal Interaction*, 521–529. https://doi.org/10.1145/3462244.3479931
41. Huang J, Pu Y, Zhou D, Cao J, Gu J, Zhao Z, et al., (2024) Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing* 565: 126992. https://doi.org/10.1016/j.neucom.2023.126992
42. Peng J, Wu T, Zhang W, Cheng F, Tan S, Yi F, et al. (2023) A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Exp Syst Appl* 221: 119721. https://doi.org/10.1016/j.eswa.2023.119721
43. Wang L, Peng J, Zheng C, Zhao T, Zhu LA, (2024) A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Inf Process Manage* 61: 103675. https://doi.org/10.1016/j.ipm.2024.103675
44. Sahay S, Okur E, Kumar SH, Nachman L, (2020) Low rank fusion based transformers for multimodal sequences. In: *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 29–34. https://doi.org/10.18653/v1/2020.challengehml-1.4
45. Chen M, Li X, (2020) Swafn: Sentimental words aware fusion network for multimodal sentiment analysis. In: *Proceedings of the 28th International Conference on Computational Linguistics*, 1067–1077. https://doi.org/10.18653/v1/2020.coling-main.93
46. Mai S, Xing S, He J, Zeng Y, Hu H, (2023) Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling. *ACM Trans Multimedia Comput Commun Appl* 19: 1–24. https://doi.org/10.1145/3542927
47. Chen R, Zhou W, Li Y, Zhou H, (2022) Video-based cross-modal auxiliary network for multimodal sentiment analysis. *IEEE Trans Circuits Syst Video Technol* 32: 8703–8716. https://doi.org/10.1109/TCSVT.2022.3197420
48. Ko D, Choi J, Choi HK, On KW, Roh B, Kim HJ, (2023) Meltr: Meta loss transformer for learning to fine-tune video foundation models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20105–20115. https://doi.org/10.1109/CVPR52729.2023.01925

49. Yu W, Xu H, Yuan Z, Wu J, (2021) Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 10790–10797. https://doi.org/10.1609/aaai.v35i12.17289
50. Han W, Chen H, Poria S, (2021) Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723
51. Arjmand M, Dousti MJ, Moradi H, (2021) Teasel: A transformer-based speech-prefixed language model. preprint, arXiv:2109.05522. https://doi.org/10.48550/arXiv.2109.05522

AIMS Press