Dual-Path Dynamic Fusion with Learnable Query for Multimodal Sentiment Analysis

Miao Zhou[1], Lina Yang[1]*, Thomas Wu[2]*, Dongnan Yang[1], Xinru Zhang[1]
[1]School of Computer and Electronic Information, Guangxi University
[2]School of Electrical Engineering, Guangxi University
{lnyang, xwu}@gxu.edu.cn
{2313301064, 2313301054, 2313301061}@st.gxu.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) is the task of understanding human emotions by analyzing a combination of different data sources, such as text, audio, and visual inputs. Although recent advances have improved emotion modeling across modalities, existing methods still struggle with two fundamental challenges: balancing global and fine-grained sentiment contributions, and over-reliance on the text modality. To address these issues, we propose DPDF-LQ (Dual-Path Dynamic Fusion with Learnable Query), an architecture that processes inputs through two complementary paths: global and local. The global path is responsible for establishing cross-modal dependencies, while the local path captures fine-grained representations. Additionally, we introduce the key module Dynamic Global Learnable Query Attention (DGLQA) in the global path, which dynamically allocates weights to each modality to capture their relevant features and learn global representations. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmarks demonstrate that DPDF-LQ achieves state-of-the-art performance, particularly in fine-grained sentiment prediction by effectively combining global and local features. Our code will be released at https://github.com/ZhouMiaoGX/DPDF-LQ.

1 Introduction

Multimodal sentiment analysis (MSA) has emerged as a critical research area aimed at comprehensively understanding human emotions by analyzing data from multiple modalities—primarily text, audio, and visual information (Poria et al., 2020). Unlike traditional text-based sentiment analysis, MSA captures the rich tapestry of human emotional expression through facial expressions, vocal intonations, and linguistic content, offering a more holistic view of sentiment (Yuan et al., 2021; Gandhi et al., 2023; Wu et al., 2025).

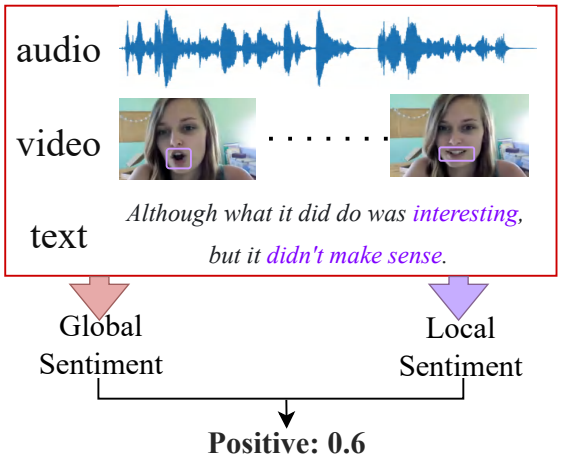


Figure 1: Illustration of Our Objective: Pink Boxes Denote Global Sentiment, Purple Boxes and Text Indicate Local Sentiment.

Recent advances in deep learning have driven MSA research along two primary directions: representation learning-centered methods (Zhang et al., 2023; Yang et al., 2023; Wang et al., 2025) and multimodal fusion-centered methods (Zadeh et al., 2017; Tsai et al., 2019; Wu et al., 2025). Representation learning-centered methods focus on extracting meaningful features to create unified global representations, utilizing modality-specific encoders to capture global semantics across modalities. In contrast, multimodal fusion-centered methods prioritize designing integration mechanisms that effectively combine local information from each modality, capturing fine-grained details that enhance task-specific performance.

Despite significant progress, current MSA methods still face issues such as the inability to properly balance global and fine-grained sentiment contributions, and over-reliance on the language modality. Methods like ALMT (Zhang et al., 2023) primarily capture global sentiment, whereas DEVA (Wu et al., 2025) emphasizes local representations. As shown

* Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11367–11377
November 4-9, 2025

in Figure 1, ULMD (Zhu et al., 2025) attempts to combine global and local features but struggles to effectively integrate them. Additionally, models like DLF (Wang et al., 2025) overemphasize text features, neglecting audio and visual information, limiting their use of non-verbal emotional cues.

To address these challenges, we propose DPDF-LQ (Dual-Path Dynamic Fusion with Learnable Query). Inspired by human emotion perception, which combines holistic impressions with detailed assessments, our framework introduces two complementary parallel processing paths: (1) the global path, which uses Dynamic Global Learnable Query Attention (DGLQA) to model cross-modal dependencies through a learnable query token. (2) the local path, which uses cross-modal transformers with concatenated audio-visual features as queries to capture fine-grained interactions. These representations are adaptively integrated through a dynamic gating fusion module that balances their contributions from the global and local paths. Furthermore, the DGLQA module, introduced in the global path, plays a pivotal role in learning global representations by dynamically allocating weights to each modality and effectively integrating them to capture global features.

Our contributions are summarized as follows:

- We propose a novel multimodal sentiment analysis method, Dual-Path Dynamic Fusion with Learnable Query (DPDF-LQ), which enhances cross-modal and feature fusion, showing strong performance in fine-grained sentiment prediction. Specifically, the global path captures cross-modal semantic dependencies via modality-joint attention, while the local path focuses on fine-grained information. Ultimately, a dynamic gating mechanism coordinates cross-path feature complementarity.

- We design a Dynamic Global Learnable Query Attention (DGLQA) layer that achieves joint semantic fusion of video, audio, and text through dynamic weight allocation. It adaptively balances multimodal contributions in the global path, which is responsible for learning global representations.

- We validate our approach through comprehensive experiments on two benchmark datasets, showing that DPDF-LQ consistently outperforms strong baselines across both benchmark datasets.

2 Related Work

In this section, we review previous work on multimodal sentiment analysis from two perspectives: representation learning-centered methods and multimodal fusion-centered methods.

2.1 Representation Learning-Centered Methods

Representation learning-centered methods focus on extracting meaningful features to create unified global representations. By using modality-specific encoders, they capture global semantics across modalities, facilitating more comprehensive sentiment analysis. Yang et al., 2022 presents FDMER, which strategically decomposes multimodal information into modality-invariant and modality-specific representations through dedicated common and private encoders. Zhang et al., 2023 uses language modality to guide the representation learning of other modalities, laying the a hierarchical framework where linguistic features serve as anchors for cross-modal alignment. Yang et al., 2023 introduces a contrastive learning framework called ConFEDE that decomposes features into modality-specific and shared components, enabling more effective cross-modal integration while preserving unique modality characteristics. Moreover, Wang et al., 2025 employs disentanglement techniques to separate sentiment-relevant from irrelevant information in multimodal representations, addressing a key limitation in previous approaches by explicitly modeling and filtering modality-specific noise.

2.2 Multimodal Fusion-Centered Methods

Multimodal fusion-centered methods focus on combining local information from multiple modalities. These approaches capture complementary signals and inter-modal dynamics, improving sentiment analysis by emphasizing fine-grained details within each modality. Early methods such as Zadeh et al., 2017 employed outer product to capture complex inter-modal interactions, establishing fundamental approaches for combining heterogeneous data sources. More sophisticated techniques subsequently emerged, including Tsai et al., 2019, which introduced cross-modal attention mechanisms for unaligned sequences, enabling more dynamic integration of temporal information across modalities. Building upon these foundations, Wu et al., 2025 enhances fusion by translating visual-audio content into textual descriptions, creating a unified

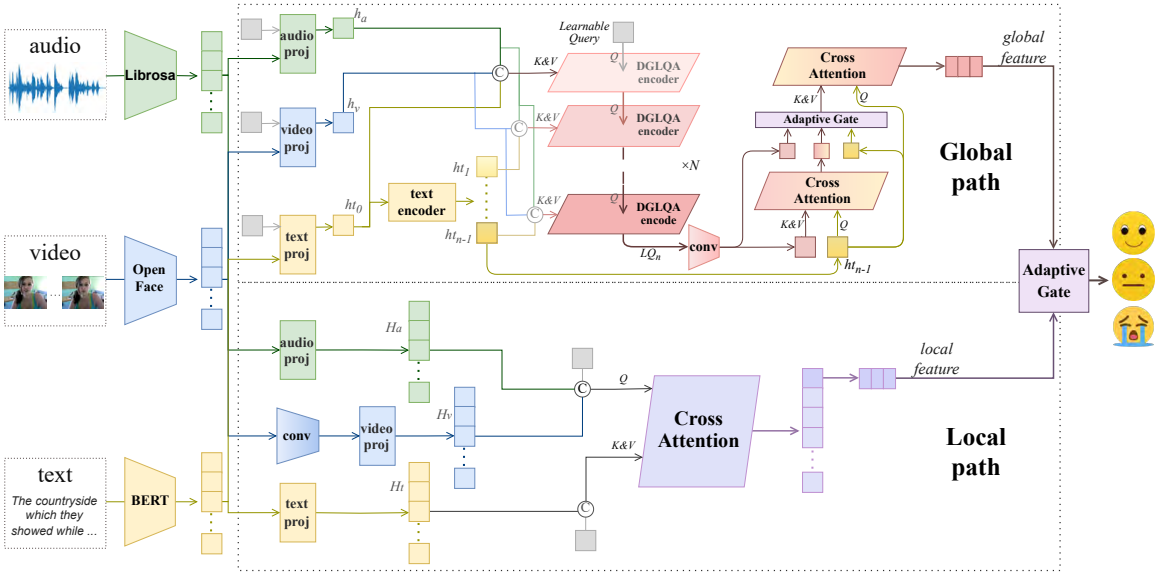


Figure 2: The architecture of our proposed DPDF-LQ for multimodal sentiment analysis. The framework processes inputs through two parallel paths: (1) a global path using Dynamic Global Learnable Query Attention (DGLQA) to capture cross-modal dependencies, and (2) a local path extracting fine-grained features using cross-attention. These complementary representations are adaptively integrated by a dynamic gating module for final sentiment prediction.

representation space that facilitates more coherent integration of multimodal cues. These advancements in fusion strategies complement representation learning approaches by providing frameworks to effectively combine the learned features for improved sentiment analysis performance.

In addition, some researchers have adopted hybrid approaches that combine elements of both representation learning and fusion, such as Zhu et al., 2025, which integrates label generation with decomposition techniques to leverage advantages from both methodological categories. However, existing hybrid approaches often lack a mechanism to dynamically balance global and local contributions. In contrast, DPDF-LQ introduces a dynamic gating mechanism that adaptively balances both global and local contributions, thereby enhancing the overall robustness and stability of the model.

3 Methodology

3.1 Overview

The overall workflow of our proposed Dual-Path Dynamic Fusion Network with Learnable Query (DPDF-LQ) for multimodal sentiment analysis is shown in Figure 2. DPDF-LQ projects inputs into a unified space through parallel paths. The global path processes the first token from each modality through a DGLQA encoder to capture cross-modal dependencies, then refines them via text-guided cross-attention. The local path extracts fine-grained features using cross-attention with audio-video features as queries. A dynamic gating module integrates these complementary representations for sentiment prediction.

3.2 Multimodal Input Representation

Our model processes three modalities: text (t), video (v), and audio (a). We use pre-extracted feature sequences for each modality. For any modality $m \in \{t, v, a\}$, the feature $X_m \in \mathbb{R}^{l_m \times d_m}$ comes from corresponding pre-processing tools: BERT (Devlin et al., 2019) for text, OpenFace (Tadas et al., 2018) for video, and Librosa (McFee et al., 2015) for audio. Here, l_m and d_m represent sequence length and feature dimension, respectively.

3.3 Global Path

The global path captures holistic multimodal sentiment representation through hierarchical attention-driven feature extraction, integrating all modalities within a structured attention framework, enabling modeling of cross-modal dependencies.

3.3.1 Modality-Specific Projection Layers

We first project heterogeneous features into a common semantic space:

$$X_m = \mathrm{FC}\left(X_m\right) \in \mathbb{R}^{l_m \times d_m}, m \in \{v, a, t\} \quad (1)$$

$$H_m = \mathrm{Trans}\left(X_m\right) \in \mathbb{R}^{l_m \times d_m} \quad (2)$$

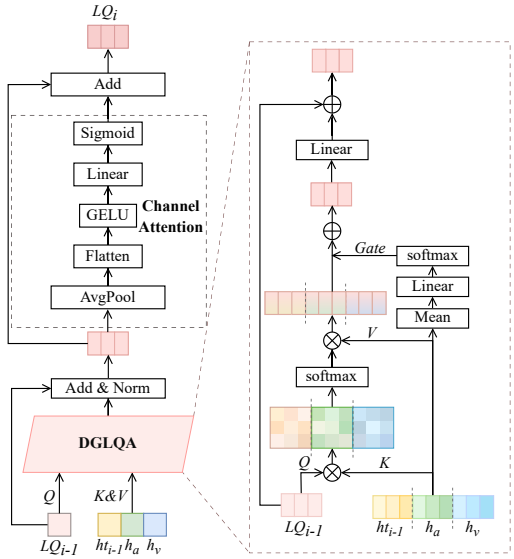


Figure 3: The workflow of the $i-th$ layer in our N-layer DGLQA encoder.

where Trans denotes the Transformer layer and FC represents a linear operation.

We use Transformer layers with global tokens, placing a learnable token at the beginning of each sequence to capture global information, similar to the approach used in the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2021). From the projected sequence, we extract these tokens:

$$h_m = H_m[0,:] \in \mathbb{R}^{l_t \times d} \tag{3}$$

where h_m encapsulates the holistic information of each modality, and l_t is the standardized sequence length (8 in our implementation).

3.3.2 Dynamic Global Learnable Query Attention

After obtaining modality-specific representations, we propose a Dynamic Global Learnable Query Attention (DGLQA) encoder consisting of: (1) a DGLQA mechanism that adaptively attends to features from different modalities, and (2) a Channel Attention Block that enhances important feature channels.

To enable multi-level attention, we pass text features h_t through a text encoder with $n-1$ layers:

$$h_t^i = \text{TextEncoder}_{i-1}\left(h_t^{i-1}\right), i \in (1, n-1) \tag{4}$$

where h_t^i is the i-th layer output. We mark the original feature h_t as h_t^0, and concatenate these into $h_{t-list} = \left(h_t^0, h_t^1, ..., h_t^{n-1}\right)$.

Figure 3 shows the DGLQA encoder workflow. It uses a learnable query $LQ \in \mathbb{R}^{l_t \times d}$ as an interaction hub, where LQ_0 is randomly initialized. In each layer, query LQ_{i-1} updates by interacting with all modalities and is enhanced through channel attention.

DGLQA concatenates features from all modalities:

$$Con = \text{Concat}\left(h_t^{i-1}, h_a, h_v\right) \in \mathbb{R}^{l_t \times 3d} \tag{5}$$

We compute queries, keys, and values:

$$Q_c = W_c^Q \cdot LQ_{i-1} \in \mathbb{R}^{n_h \times l_t \times d_h} \tag{6}$$

$$K_c = W_c^K \cdot Con \in \mathbb{R}^{3 \times n_h \times l_t \times d_h} \tag{7}$$

$$V_c = W_c^V \cdot Con \in \mathbb{R}^{3 \times n_h \times l_t \times d_h} \tag{8}$$

where d_h is the dimension of each attention head, and n_h is the number of attention heads.

Keys and values are reshaped to separate modalities $m \in \{t, a, v\}$:

$$K_m \in \mathbb{R}^{n_h \times l_t \times d_h}, V_m \in \mathbb{R}^{n_h \times l_t \times d_h} \tag{9}$$

We compute attention maps and outputs for each modality:

$$Attn_m = \text{Softmax}\left(\frac{Q_c \cdot K_m^T}{\sqrt{d_h}}\right) \tag{10}$$

$$Out_m = Attn_m \cdot V_m \in \mathbb{R}^{n_h \times l_t \times d_h} \tag{11}$$

A dynamic gating mechanism weights the contribution of each modality:

$$Mean = \text{Mean}\left(Con\right) \in \mathbb{R}^{3d} \tag{12}$$

$$Gate = \text{Softmax}\left(W_g \cdot Mean\right) \in \mathbb{R}^3 \tag{13}$$

The function Mean here is used to calculate the arithmetic mean value of the elements within the input.

The gate values are applied to the attention outputs:

$$Fused = \sum_{m \in \{t,a,v\}} Gate[m] \cdot Out_m \tag{14}$$

The updated query is obtained through:

$$LQ_i = LQ_{i-1} + \text{FC}\left(Fused\right) \tag{15}$$

Channel Attention enhances important feature dimensions:

$$z = \frac{1}{l_t} \sum_{j=1}^{l_t} LQ_i[j,:] \in \mathbb{R}^d \tag{16}$$

$$s = \sigma\left(W_2\left(\text{GELU}\left(W_1\left(\mathbf{z}\right)\right)\right)\right) \tag{17}$$

$$LQ_i' = LQ_i \odot s \tag{18}$$

where $W_1 \in \mathbb{R}^{d \times d/r}$, $W_2 \in \mathbb{R}^{d/r \times d}$, r is the reduction ratio (typically 4), and $\odot$ represents channel-wise multiplication.

The final updated query uses a residual connection:

$$LQ_i = LQ_i' + LQ_i \in \mathbb{R}^{l_t \times d} \tag{19}$$

3.3.3 Cross-Attention and Feature Refinement

After obtaining the enhanced query representations through the DGLQA encoder, we apply a series of refinement operations to further improve the feature representation before final prediction.

To capture fine-grained local patterns within the feature space, we first apply a depthwise convolution to the learnable query:

$$LQ' = \text{DepthwiseConv}\left(LQ_n\right) \tag{20}$$

where the depthwise convolution uses a kernel size of 3, followed by GELU activation and batch normalization. This operation is computationally efficient while enhancing the local receptive field of each feature channel.

Next, to integrate global semantic information from text, we use the refined learnable query as the source and the highest-level text features as the target in a cross-attention mechanism. This cross-attention mechanism is detailed in Zhang et al., 2023, which references the work of Tsai et al., 2019 :

$$f = \text{CrossAttn}\left(LQ', h_t^{n-1}\right)[0, :] \tag{21}$$

Where h_t^{n-1} is the output from the final layer of the text encoder, and we extract the first token (CLS token) as the global representation.

To dynamically balance contributions from the learnable query and text features, and suppress potential redundancy, we introduce an adaptive gating mechanism:

$$G_{\text{in}} = \text{Concat}\left(LQ', h_t^{n-1}, f\right) \tag{22}$$

$$G = \sigma\left(FC\left(G_{\text{in}}\right)\right) \tag{23}$$

$$LQ'' = G \odot LQ' + (1 - G) \odot h_t^{n-1} \tag{24}$$

This gating mechanism adaptively adjusts the balance between the learnable query and text features.

Finally, we use cross-attention once again to extract the global features.

$$f_{\text{global}} = \text{CrossAttn}(LQ'', h_t^{n-1})[0, :] \tag{25}$$

This completes the global path, which produces $f_{\text{global}} \in \mathbb{R}^d$, a comprehensive representation that captures complex cross-modal interactions. This feature will later be combined with local feature to produce the final sentiment prediction, allowing global semantics and fine-grained details to complement each other.

3.4 Local Path

While the global path captures cross-modal dependencies, the local path focuses on preserving emotional local details and extracting fine-grained features. This complementary design ensures that fine-grained audio-visual cues, which may be underrepresented in text, are effectively captured, enabling the model to balance global context and local precision.

3.4.1 Modality-Specific Projections

Similar to the global path, we project each modality's features into a common feature space. For visual features, we first apply spatial attention using convolutional operations to enhance spatial patterns:

$$X_v = \text{SpatialAttention}\left(X_v\right) \tag{26}$$

SpatialAttention uses depthwise separable convolutions to capture local visual dependencies.

We then project all modalities into a unified representation space:

$$X_m = \text{FC}\left(X_m\right) \in \mathbb{R}^{l_m \times d_m}, m \in \{v, a, t\} \tag{27}$$

$$H_m = \text{Trans}\left(X_m\right) \in \mathbb{R}^{l_m \times d_m} \tag{28}$$

Unlike the global path, where only the first token (CLS token) is processed for each modality, the local path processes the entire sequence of projected features, preserving the complete temporal and sequential information from each modality.

3.4.2 Cross-Modal Transformer Fusion

The key characteristic of the local path is its use of cross-attention with audio and visual features as queries:

$$Con_{a+v} = \text{Concat}\left(H_a, H_v\right) \tag{29}$$

$$f_{\text{local}} = \text{CrossAttn}\left(H_t, Con_{a+v}\right)[0, :] \tag{30}$$

Here, the text features serve as the source and the concatenated audio-visual features as targets, allowing the text to guide the integration of fine-grained local cues while preserving complementary information from audio and visual modalities.

By extracting the first token of the output, we obtain a local feature $f_{\text{local}} \in \mathbb{R}^d$ that preserves fine-grained features and local details while allowing for cross-modal interactions.

3.5 Dynamic Gate Fusion and Prediction

After obtaining the global representation $f_{\text{global}} \in \mathbb{R}^d$ from the global path and the local representation $f_{\text{local}} \in \mathbb{R}^d$ from the local path, we employ a dynamic gate fusion mechanism to adaptively integrate these complementary features. We first concatenate these representations:

$$Z = \text{Concat}\left(f_{\text{global}}, f_{\text{local}}\right) \in \mathbb{R}^{2d} \tag{31}$$

The dynamic fusion gate is then computed through a nonlinear transformation:

$$Z' = \text{GELU}\left(\text{FC}\left(Z\right)\right) \tag{32}$$

$$FG = \text{Softmax}\left(\text{FC}\left(Z'\right)\right) \in \mathbb{R}^2 \tag{33}$$

This adaptive gate mechanism dynamically adjusts the weight of each path based on the specific input sample.

The fused representation comes into being via a weighted sum calculation:

$$f_{\text{fused}} = FG[0] \cdot f_{\text{global}} + FG[1] \cdot f_{\text{local}} \tag{34}$$

The final sentiment prediction is generated by:

$$\hat{y} = \text{FC}(f_{\text{fused}}) \in \mathbb{R}^1 \tag{35}$$

In practice, we use 8-head attention in all transformer components to model the relationships between modalities. The dynamic gate adaptively balances global and local features, suppressing redundant signals while emphasizing complementary information, allowing the model to focus on the most informative aspects for each input.

3.6 Overall Learning Objectives

Our model is trained using a straightforward mean squared error loss function for sentiment regression:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} ||y_n - \hat{y}_n||_2^2 \tag{36}$$

where N represents the batch size, y_n is the ground truth sentiment score, and $\hat{y}_n$ is our model's prediction.

Our simple optimization objective makes DPDF-LQ easy to train compared to methods with multiple optimization goals, without requiring extensive hyperparameter tuning.

4 Experiments

4.1 Datasets

We evaluate our approach on two widely used benchmark datasets for multimodal sentiment analysis: MOSI and MOSEI.

MOSI. The CMU Multimodal Opinion Sentiment Intensity (MOSI) dataset (Zadeh et al., 2016) contains 2,199 short video segments extracted from 93 YouTube movie review videos involving 89 speakers. Each segment is annotated with sentiment scores ranging from -3 (strongly negative) to 3 (strongly positive). Following standard practice, we use the official split with 1,284 segments for training, 229 segments for validation, and 686 segments for testing. The dataset provides aligned multimodal features across language, visual, and acoustic modalities.

MOSEI. The CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset (Bagher Zadeh et al., 2018) is a larger-scale benchmark containing 23,453 video segments from 1,000 YouTube speakers (57% male, 43% female) discussing various topics. The dataset features greater diversity in terms of speakers, topics, and recording conditions (including variations in illumination, head poses, and occlusions). Each segment is labeled with sentiment scores from -3 (strongly negative) to 3 (strongly positive). We follow the standard data split of 16,326 segments for training, 1,871 for validation, and 4,659 for testing.

4.2 Evaluation Metrics

Following established practices in multimodal sentiment analysis (Yu et al., 2020, 2021; Zhang et al., 2023), we evaluate our approach using multiple metrics to provide a comprehensive assessment of performance. For classification tasks, we report binary classification accuracy (Acc-2), F1 score, five-class accuracy (Acc-5), and seven-class accuracy (Acc-7). For regression tasks, we report Mean Absolute Error (MAE) and Correlation (Corr). Furthermore, Acc-2 and F1 are reported under two settings: negative/non-negative (including zero) and negative/positive (excluding zero) (Hazarika et al., 2020). For all metrics except MAE, higher values indicate better performance.

4.3 Baselines

To rigorously evaluate the effectiveness of our DPDF-LQ framework, we conduct extensive experiments under the same settings on a range of

Model	MOSI						MOSEI					
	Acc-2	F1	Acc-5	Acc-7	MAE↓	Corr	Acc-2	F1	Acc-5	Acc-7	MAE↓	Corr
LMF†	77.9/79.18	77.8/79.15	38.13	33.82	0.95	0.651	80.54/83.48	80.94/83.36	52.99	51.79	0.576	0.717
MFN†	77.67/78.87	77.63/78.90	40.47	35.83	0.927	0.67	78.94/82.86	79.55/82.85	52.76	51.34	0.573	0.718
MuIT†	79.71/80.98	79.63/80.95	42.68	36.91	0.88	0.702	81.15/84.63	81.56/84.52	54.18	52.84	0.559	0.733
MISA†	81.84/83.54	81.82/83.58	47.08	41.37	0.777	0.778	80.67/84.67	81.12/84.66	53.63	52.05	0.558	0.752
Self-MM†	83.44/85.46	83.36/85.43	53.47	46.67	0.708	0.796	83.76/85.15	83.82/84.9	55.53	53.87	0.531	0.765
TETFN†	83.24/85.37	83.13/85.33	53.64	45.77	0.708	0.798	**84.12/86.21**	**84.35**/86.11	55.78	53.9	0.537	0.767
ConFEDE*	**84.4**/85.82	84.36/85.82	52.62	46.27	0.741	0.783	82.83/85.53	83.09/85.38	54.86	53.06	0.538	0.771
ALMT*	83.38/85.82	83.17/85.7	53.25	46.79	0.725	0.787	83.28/85.44	82.87/85.25	53.25	53.04	0.543	0.765
ULMD*	83.09/85.82	82.88/85.71	54.23	47.81	0.7	0.799	82.59/85.75	83/85.71	55.31	53.81	0.531	0.771
DEVA	**84.4**/86.29	**84.48**/86.3	51.78	46.32	0.730	0.787	83.26/86.13	82.93/**86.21**	55.32	52.26	0.541	0.769
DLF	–/85.06	–/85.04	52.33	47.08	0.731	0.781	–/85.42	–/85.27	55.7	53.9	0.536	0.764
DPDF-LQ	84.11/**86.59**	83.88/**86.45**	**54.81**	**48.54**	**0.682**	**0.803**	83/**86.21**	83.36/86.14	**55.93**	**54.07**	**0.529**	**0.774**

Table 1: Performance comparison on MOSI and MOSEI datasets. The best results are highlighted in bold; † denotes results obtained from Mao et al., 2022; * indicates our reproduced results; unmarked results are directly cited from original papers. – denotes that the metric was not reported in the original work.

state-of-the-art approaches, such as LMF (Liu et al., 2018), MFN (Zadeh et al., 2018), MuIT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), TETFN (Wang et al., 2023), ConFEDE (Yang et al., 2023), ALMT (Zhang et al., 2023), ULMD (Zhu et al., 2025), DEVA (Wu et al., 2025), and DLF (Wang et al., 2025).

4.4 Comparison of Results

Table 1 compares our DPDF-LQ with state-of-the-art methods on the MOSI and MOSEI datasets.

On MOSI, DPDF-LQ achieves SOTA performance in fine-grained metrics: Acc-7 (48.54%, representing a +1.46% improvement over DLF), MAE (0.682), and Corr (0.803), while maintaining competitive binary accuracy (86.59%). The improvements in Acc-5 (54.81%) and F1 score (86.45%) further validate our approach's effectiveness in capturing nuanced sentiment expressions.

On MOSEI, DPDF-LQ achieves superior performance across most key metrics: Acc-7 reaches 54.07% (+0.17% over the previous best), MAE is 0.529, and Corr is 0.774. It also achieves competitive F1 (86.14%) and strong Acc-5 (55.93%), highlighting its strength in modeling sentiment intensity at a fine-grained level.

The dual-path architecture consistently outperforms single-path baselines (e.g., +2.27% Acc-7 over ConFEDE) by capturing both local nuances and global context. While some baselines perform well on binary classification, DPDF-LQ excels in fine-grained sentiment analysis and is vital for applications that model sentiment intensity in detail.

4.5 Ablation Study

Table 2 presents our ablation studies on the MOSI and MOSEI benchmarks, evaluating the impact of different modalities, components, and attention mechanisms to validate the key innovations in our proposed DPDF-LQ framework for multimodal sentiment analysis.

4.5.1 Effect of Modalities

We first examine the contribution of each modality (T: text, A: audio, V: visual) by systematically removing them from the multimodal input. Results show that while removing visual (w/o V) or audio (w/o A) modalities results in only minor performance degradation (less than 1%), removing text (w/o T) leads to substantial performance drops (over 32% in accuracy), confirming text as the dominant modality for sentiment analysis.

4.5.2 Effect of Components

Next, we evaluate the impact of key components: Local Path (LPath), Global Path (GPath), Gate Mechanism, and Dynamic Global Learnable Query Attention (DGLQA). Removing any component causes performance drops, with GPath removal being most impactful on MOSI (5.10% drop) and LPath removal on MOSEI (3.78% drop). These results validate our dual-path architecture design and fusion mechanism.

4.5.3 Attention Mechanism Variants

We further investigate alternative attention mechanisms: (1) Cross-Attention (CA): our specific flow design outperforms alternatives; (2) LPath

Method	MOSI		MOSEI	
	Acc-5	MAE↓	Acc-7	MAE↓
DPDF-LQ	**54.81**	**0.682**	**54.07**	**0.529**
Effect of Modalities				
w/o V	54.63	0.684	53.34	0.534
w/o A	54.61	0.685	53.77	0.531
w/o T	22.35	1.406	26.72	0.945
w/o A&V	54.55	0.686	53.21	0.537
w/o T&V	23.18	1.406	26.36	0.944
w/o T&A	21.14	1.408	28.80	0.920
Effect of Components				
w/o LPath	52.92	0.718	50.29	0.560
w/o GPath	49.71	0.746	51.77	0.542
w/o Gate	53.21	0.702	53.60	0.532
w/o DGLQA	51.17	0.724	53.12	0.539
Attention Mechanism Variants				
w/o GPath CA	53.17	0.707	53.29	0.542
LPath CA↔GPath CA	48.10	0.752	51.77	0.547
DGLQA→Std.A	31.78	0.844	49.24	0.574
DGLQA→AHL	52.33	0.730	50.57	0.558

Table 2: Results of ablation studies on MOSI and MOSEI. Metrics (Acc-5, Acc-7, and MAE) are reported for different ablations.

CA↔GPath CA: swapping attention flows between paths significantly decreases performance; (3) Standard Attention (Std.A): replacing DGLQA with standard attention leads to a substantial performance drops (23.03% on MOSI); (4) Adaptive Hyper-modality Learning (AHL) (Zhang et al., 2023): although better than standard attention, it still lags behind our DGLQA.

These findings demonstrate the effectiveness and necessity of our DGLQA design.

4.5.4 Complexity Analysis

Our model contains 224.88M parameters and requires 8.70G FLOPs for a single forward pass, with an average inference time of 1.53ms per sample on an NVIDIA 3090 GPU. Although our model has a comparatively large parameter count, each of the two paths can operate independently.

4.6 Further Analysis

4.6.1 Dual-Path Attention

In Figure 4, we present the average cross-attention matrices from our dual-path model on the MOSEI dataset. The global path tends to focus on positions conveying overarching semantic information, while the local path prefers positions with rich fine-grained audio-visual details.

These complementary patterns are further supported by our ablation findings in Table 2: removing the local path (w/o LPath) reduces Acc7 on MOSEI from 54.07% to 50.29%, whereas removing the global path (w/o GPath) decreases it to 51.77%. Overall, the visualization and ablation results indicate that the dual-path design effectively captures complementary global and fine-grained local information.

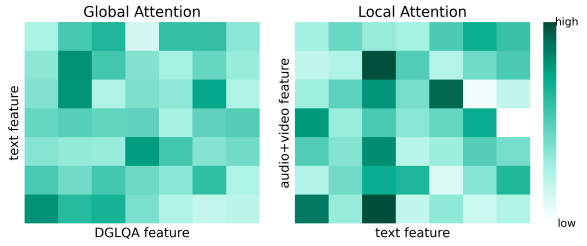


Figure 4: Visualization of average cross-attention weights on the MOSEI dataset. The left shows the global attention matrix, and the right shows the local attention matrix. Color intensity denotes attention weight (darker colors indicate higher values).

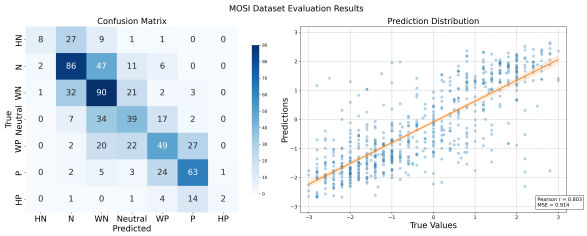


Figure 5: Fine-grained evaluation on MOSI. HN: Highly Negative; N: Negative; WN:Weakly Negative; WP: Weak Positive; P: Positive; HP: Highly Positive.

4.6.2 Fine-Grained Prediction

As shown in Figure 5, our model shows significant advances in fine-grained sentiment analysis, achieving strong performance on moderate sentiment classes ("WN": 90%, "N": 86%), while struggling with extreme sentiments ("HN": 8%, "HP": 14%). The high correlation (r=0.803) between predicted and ground-truth distributions confirms its ability to capture sentiment nuances, though further work is needed to better handle intensity extremes.

4.6.3 Multi-run Reliability

To ensure the robustness and statistical reliability of our proposed DPDF-LQ framework, we conducted five independent runs with different random seeds, repeating the training and evaluation process. Figure 6 shows the comparison of DPDF-LQ and ALMT across key metrics on the CMU-MOSI and CMU-MOSEI datasets. Each bar represents the average over five runs, with error bars indicating

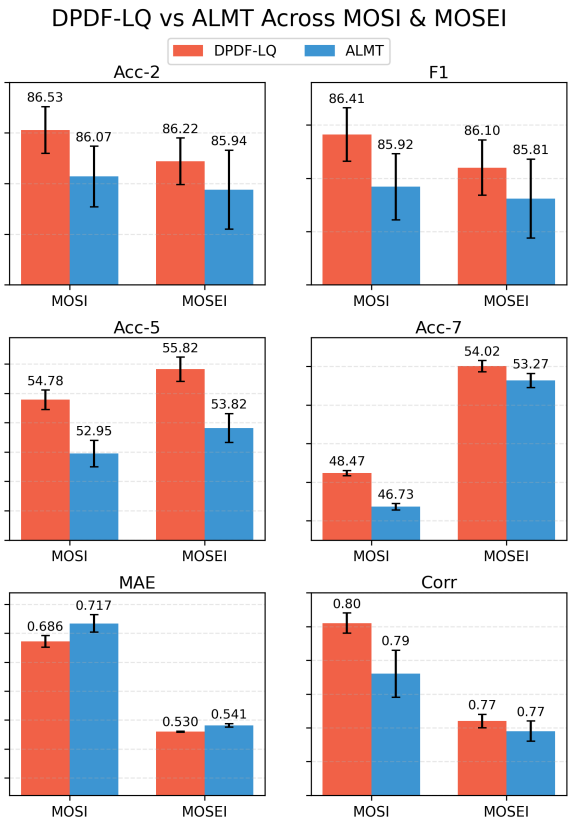


Figure 6: Multi-run performance on MOSI and MOSEI. Bars show the mean over five runs; error bars indicate standard deviation.

the standard deviation.

The results show that DPDF-LQ consistently outperforms ALMT across most metrics with low variance, indicating stable performance likely attributable to the dual-path design and Dynamic Global Learnable Query Attention.

5 Conclusion

We propose DPDF-LQ, a dual-path framework for multimodal sentiment analysis that integrates global understanding and fine-grained local information through Dynamic Global Learnable Query Attention (DGLQA) and adaptive fusion. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmarks show that DPDF-LQ achieves state-of-the-art performance, particularly in fine-grained sentiment prediction. Ablation studies validate the contribution of each component, and our method addresses key challenges in multimodal fusion. Overall, this work advances sentiment analysis by combining comprehensive global understanding with precise local feature extraction.

Limitations

Our model struggles with extreme sentiment predictions and requires careful hyperparameter tuning. The dual-path design is relatively complex, limiting deployment in resource-constrained settings. It has not been extensively tested on real-world data with diverse linguistic and cultural expressions.

Ethical Considerations

This work enhances multimodal sentiment analysis and may benefit fields like education, but raises privacy and ethical risks in behavioral monitoring. Users should follow regulations, obtain consent, and ensure oversight in high-stakes scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62371144 and 62461004).

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 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)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213, Dublin, Ireland. Association for Computational Linguistics.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE transactions on affective computing*, 14(1):108–132.

Baltrusaitis Tadas, Zadeh Amir, Lim Yao Chong, and Morency Louis-Philippe. 2018. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. 2023. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. 2025. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1601–1609.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1642–1651.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 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*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4400–4407.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore. Association for Computational Linguistics.

Linan Zhu, Hongyan Zhao, Zhechao Zhu, Chenwei Zhang, and Xiangjie Kong. 2025. Multimodal sentiment analysis with unimodal label generation and modality decomposition. *Information Fusion*, 116:102787.

A Appendix

A.1 Hyperparameter Settings

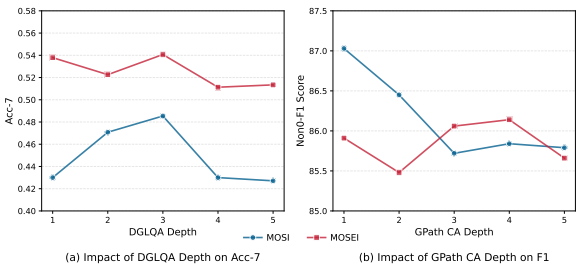
Table 3 presents the optimal hyperparameter configurations for our DPDF-LQ model on both MOSI and MOSEI datasets. Most hyperparameters remain identical across both datasets.

Parameter	MOSI	MOSEI
Learnable query length	8	8
Learnable query dimension	128	128
DGLQA depth	3	3
GPath CA depth	2	4
LPath CA depth	2	2
Hidden dimension	256	256
Learning rate	$1e-4$	$1e-4$
Weight decay	$1e-4$	$1e-4$
Batch size	64	64

Table 3: Optimal hyperparameters for DPDF-LQ

A.2 Impact of Hyperparameters

Figure 7 illustrates the effect of component depth on model's performance. For DGLQA, a depth of 3 provides the best trade-off, yielding the highest correlation (0.803) while maintaining a robust Acc-5 of 54.81%. For GPath CA, a depth of 2 yields the optimal results across both datasets, with deeper settings showing diminishing returns. GPath depth 2 is critical for MOSI, whereas MOSEI is less sensitive to this parameter.



(a) Impact of DGLQA Depth on Acc-7 (b) Impact of GPath CA Depth on F1

Figure 7: Impact of Model Depth on Performance

A.3 About Gate Weight

Our dynamic gate mechanism shows different weighting patterns between datasets. MOSI balances path contributions (Global: 0.435±0.161, Local: 0.565±0.161), with high variance for sample-specific adaptation. MOSEI favors local features more (Global: 0.316±0.062, Local: 0.684±0.062), with lower variance.

A.4 Impact of Larger Language Models

We further evaluate the influence of stronger language models by replacing BERT-base with BERT-large under identical experimental settings. This allows us to isolate the effect of increased language model capacity on the DPDF-LQ framework. Table 4 presents the results in vertical format for both the MOSI and MOSEI datasets.

Metric	DPDF-LQ (B)	DPDF-LQ (L)
	MOSI	
Acc-2	84.11 / 86.59	**86.15 / 88.26**
F1	83.88 / 86.45	**86.12 / 88.25**
Acc-5	54.81	**54.96**
Acc-7	48.54	**48.63**
MAE	0.682	**0.638**
Corr	0.803	**0.838**
	MOSEI	
Acc-2	83 / 86.21	**83.37 / 86.65**
F1	83.36 / 86.14	**83.77 / 86.63**
Acc-5	55.93	**56.66**
Acc-7	54.07	**54.88**
MAE	0.529	**0.517**
Corr	0.774	**0.789**

Table 4: Performance of DPDF-LQ with BERT-base (B) and BERT-large (L) on MOSI and MOSEI. Bold numbers indicate better metric.

Using BERT-large reduces MAE and slightly improves Acc-2 and Acc-7 on MOSI and MOSEI, reflecting more accurate regression performance and finer-grained sentiment modeling. In the experiments, some hyperparameters differed between the two setups, but the results show that our model can effectively leverage the larger language model. To maintain fair comparison with baselines and computational efficiency, we still report experimental results using BERT-base.