A Text-Routed Sparse Mixture-of-Experts Model with Explanation and Temporal Alignment for Multi-Modal Sentiment Analysis

Dongning Rao[1], Yunbiao Zeng[1, 2], Zhihua Jiang[3*], Jujian Lv[2*]

[1] School of Computer, Guangdong University of Technology, Guangzhou 510006, China
[2] School of Computer Science, Guangdong Polytechnic Normal University
[3] Department of Computer Science, Jinan University, Guangzhou 510632, China
raodn@gdut.edu.cn, 2112405257@mail2.gdut.edu.cn, tjiangzhh@jnu.edu.cn, jujianlv@gpnu.edu.cn

Abstract

Human-interaction-involved applications underscore the need for **M**ulti-modal **S**entiment **A**nalysis (MSA). Although many approaches have been proposed to address the subtle emotions in different modalities, the power of explanations and temporal alignments is still underexplored. Thus, this paper proposes the **T**ext-routed sparse mixture-of-**E**xperts model with e**X**planation and **T**emporal alignment for MSA (TEXT). TEXT first augments explanations for MSA via **M**ulti-modal **L**arge **L**anguage **M**odels (MLLM), and then novelly aligns the representations of audio and video through a temporality-oriented neural network block. TEXT aligns different modalities with explanations and facilitates a new text-routed sparse mixture-of-experts with gate fusion. Our temporal alignment block merges the benefits of Mamba and temporal cross-attention. As a result, TEXT achieves the best performance across four datasets among all tested models, including three recently proposed approaches and three MLLMs. TEXT wins on at least four metrics out of all six metrics. For example, TEXT decreases the mean absolute error to 0.353 on the CH-SIMS dataset, which signifies a 13.5% decrement compared with recently proposed approaches.

Code — https://github.com/fip-lab/TEXT

1 Introduction

Applications in healthcare and human-computer interaction rely heavily on multi-modal sentiment analysis. Thus, popular datasets for MSA like MOSI (Zadeh et al. 2016), are proposed. However, more comprehensive approaches are needed to understand the subtle emotional nuances conveyed in audio and video (Wu et al. 2025).

The left part of Fig. 1 is an example from MOSI, where MSA demands us to predict not only the polarity but also a score for short videos. For this example, only the text modality can correctly predict the polarity, and the estimated score using all modalities from previous studies (Zhang et al. 2023) is 1.080. Considering that the label of this sample is 1.400 (deviation: 0.320), there is still room for improvement.

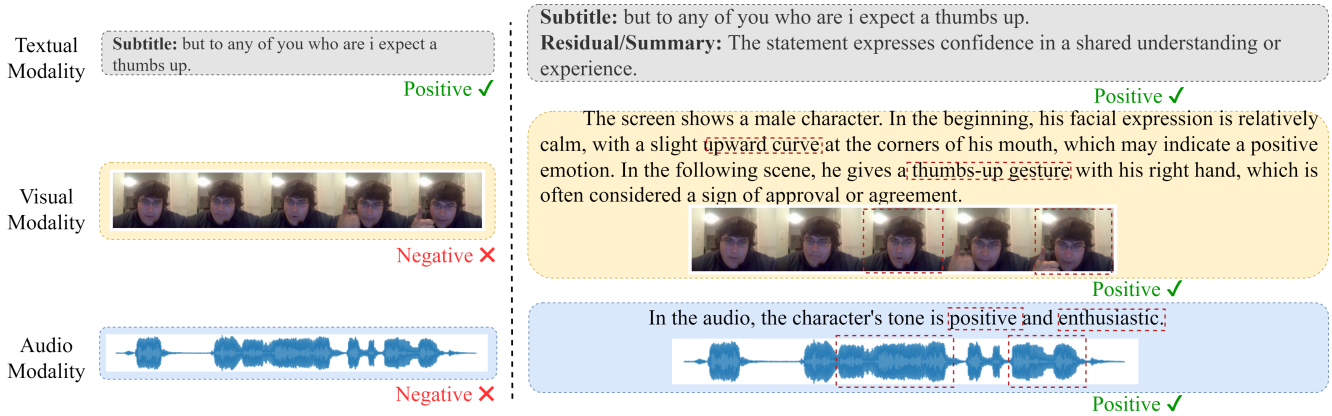
While fusion is the key to a comprehensive understanding (Zhang et al. 2023), recent studies notice that different modalities contribute disparately to MSA (Wu et al. 2025). For example, there is always a dominant modality (Feng et al. 2024) (e.g., the text in Fig. 1) and text-guide fusion is promising (Wu et al. 2025). However, as two-thirds of the modalities might cause misjudgment in Fig. 1, we posit that alignment is the crucial link between representation learning and multimodal fusion. Moreover, despite the aforementioned advances, the power of text in this large language model (LLM) (Bai et al. 2025) era has not been fully explored.

Therefore, we propose the **T**ext-routed sparse mixture-of-**E**xperts model with e**X**planation and **T**emporal alignment for multi-modal sentiment analysis (TEXT) in this paper.

- To explore the power of text, TEXT first facilitates multimodal LLMs (MLLM) to generate explanations that will be encoded by BERT (Devlin et al. 2019). The MLLM is VideoLLaMA 3 (Zhang et al. 2025), which is fine-tuned with the EMER-fine dataset (Lian et al. 2025).
- Then, considering how video or audio can mislead (such as in Fig. 1), this study aligns audio (Librosa (McFee et al. 2015)) and video (OpenFace (Baltrusaitis et al. 2018)) encoding with explanations for the first time. This module uses a **C**ross-**A**ttention(CA)-based alignment.
- To improve temporal fusion, we propose a novel temporal alignment between the aligned audio/video representations. This block combines, simplifies, and outperforms Mamba (Dao and Gu 2024) and temporal CA (Zhang et al. 2024).
- Then, for better decisions, we implement a text-routed sparse mixture-of-experts (SMoE) (Shazeer et al. 2017) because of the dominance of text. That is, experts are activated based on the text.
- At last, a multi-layer perceptron (MLP) with gate fusion (GF) (Qiu et al. 2025) is applied as a classifier.

The right part of Fig. 1 is the result of TEXT. With explanations, both visual and audio modalities can correctly predict the polarity. Furthermore, the final score given by TEXT is 1.390, which is close to 1.400 (deviation: 0.010). By contrast, the representative MLLM, QWen 2.5-vl (Bai et al. 2025) predicts a score of 2.500 (deviation: 1.100). The deviation is significantly reduced from 1.100 to 0.010, indicating a substantial improvement in prediction accuracy.

*Corresponding authors who contributed equally.



(a) Score Predicted by QWen 2.5-vl: 2.500 (b) Score Predicted by TEXT (with explanations) : 1.390

Figure 1: An example of MSA from MOSI. On the left side, from top to bottom, are the textual, visual, and audio modalities of the short video. The corresponding modalities with MLLM-generated explanations are on the right. On the left side, only text can correctly predict the polarity, with a predicted score of 2.500. As a comparison, the label of this example is 1.400, and the TEXT's prediction is 1.390. With explanations, both video and audio modalities can correctly predict the polarity.

To test the effectiveness of TEXT, we compare TEXT with three recently proposed models and three MLLMs on four datasets. The three compared models are ALMT (Zhang et al. 2023), KuDA (Feng et al. 2024),, and DEVA (Wu et al. 2025). The three MLLMs are Qwen2.5-vl (Bai et al. 2025), GPT-4o (Achiam et al. 2023) and VideoLlama3-7B (Zhang et al. 2025); the four datasets are MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018), CH-SIMS (Yu et al. 2020), and CH-SIMSv2 (Liu et al. 2022). Experiment results show that TEXT outperforms all compared models. For example, on CH-SIMS TEXT improve the mean absolute error (MAE) from 0.449 (ALMT), 0.408 (KuDA), and 0.424 (DEVA) to 0.353 (i.e., a decrease of 13.5%). Further, our ablation study suggests that temporal alignment is the most crucial component for MAE.

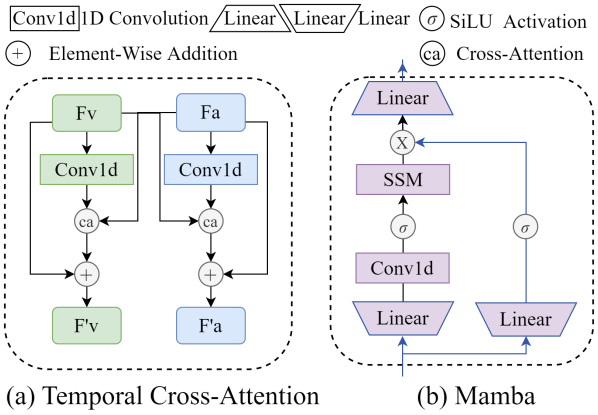
Our contribution can be summarized as follows:

1. TEXT aligns encoded audio and video through a novel temporality-oriented neural network block;
2. TEXT first augments data for MSA via MLLM with explanations that can be aligned with audio and video;
3. TEXT uses a new text-routed SMoE;
4. TEXT is the winner among all tested models on four datasets across six metrics.

2 Related Work

2.1 Multi-modal Sentiment Analysis

MSA has been investigated for its attractive applications like fraud detection (Park, Kim, and Choi 2021) trading systems (Chen and Huang 2021), human-machine interaction (Rozanska and Podpora 2019), and health care applications (Shah et al. 2020). Existing methods for MSA can be classified into two categories: representation learning-centered methods (e.g., ALMT and KuDA) and multi-modal fusion-centered methods (Zhang et al. 2023) (e.g., DEVA). Both KuDA and DEVA are claimed to be state-of-the-art (Feng et al. 2024; Wu et al. 2025).



(a) Temporal Cross-Attention (b) Mamba

Figure 2: Two temporal alignment block designs: (a) TCA, and (b) Mamba. Legends are also applicable for Fig. 5∼6. In (a), F_v, F_a, and F_t represent the input vectors to each module, while F'_v, F'_a, and F'_t represent the output vectors of the module. In (b), $\otimes$ represents nonlinearity.

Formally, inputs of MSA models include text (t), audio (a), and visual (v). Our goal is to fuse the data from different modalities and predict the sentimental polarity $\hat{y}$ along with a sentiment score between [-1, 1] or [-3, 3]. A score greater than, equal to, or less than zero represents positive, neutral, and negative, respectively. As a regression problem, the basic optimization objective function of MSA is the MSE loss.

2.2 Cross-Modal Temporal Fusion

Because audio and video are sequential, designing neural network blocks for temporal alignment is important. Cross-modal fusion, especially cross-modal attention mechanisms, is a popular temporal alignment approach. Cross-modal fusion captures both similar and dissimilar information between uni-modal representations ($U_m, m \in \{t, v, a\}$) and

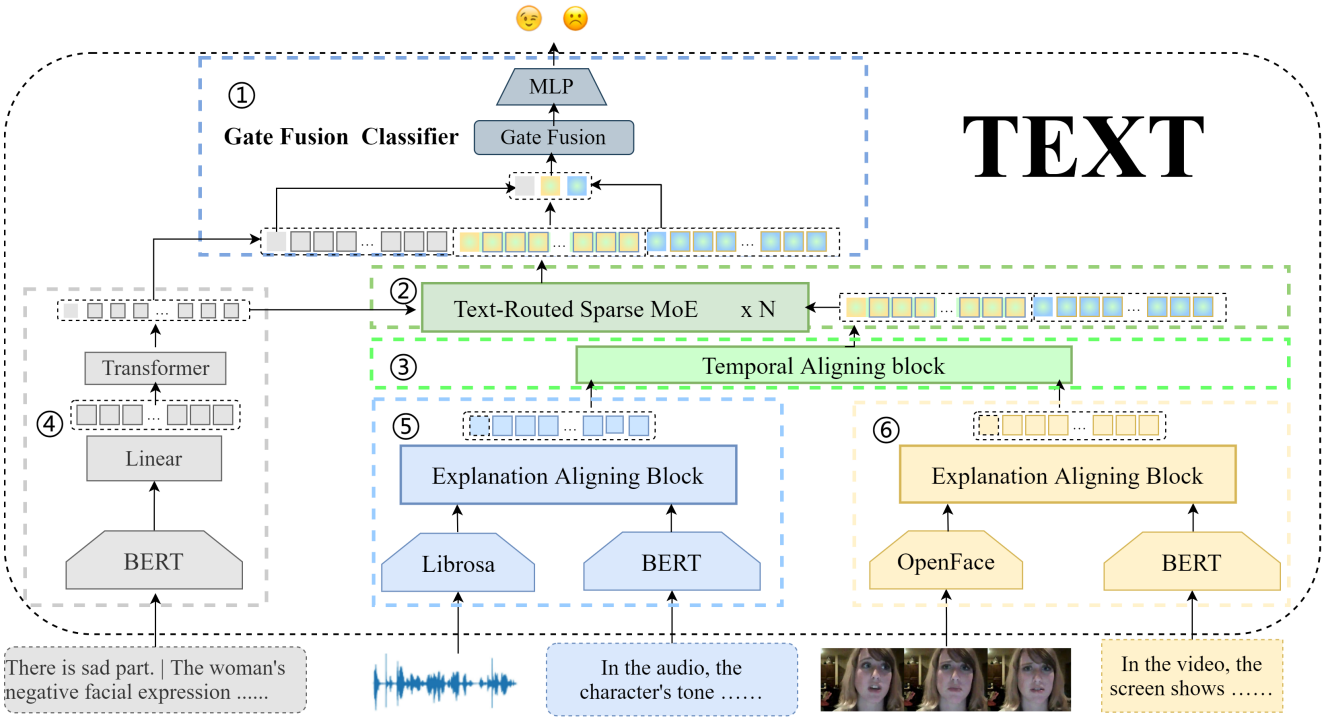


Figure 3: The architecture of TEXT. From top to bottom, TEXT comprises six functional modules (①∼⑥). Its sequential processing workflow centers on three core components: ① the Gate Fusion Classifier Module, which performs the final decision-making; ② the Text-Routed SMoE Module, designed to model cross-modal interactions; and ③ the Temporal Alignment Module, responsible for synchronizing audio and video streams. In addition, three uni-modal feature extraction modules (④∼⑥) operate in parallel. Notably, both the Audio Feature Extraction Module (⑤) and the Video Feature Extraction Module (⑥) incorporate an Explanation Alignment Block. All modalities are processed using pre-trained encoders: textual data (including explanation annotations) is encoded with BERT, audio signals are extracted using Librosa, and visual information—derived from video frames—is processed via OpenFace.

generates a multi-modal embedding.

For example, as shown in Fig. 2 (a), the temporal CA (TCA) is a specific emotion-oriented and CA-based block for short video (Zhang et al. 2024). By contrast, Fig. 2 (b) illustrates the Mamba, which is a linear-time sequence model with sophisticated **S**tructured **S**tate-space **M**odels (SSM) for long video (Gu and Dao 2023). However, samples of MSA are often short videos that might exhibit dynamic emotional transitions across frames (e.g., Fig. 1). That is, neither Mamba nor TCA is designed for MSA.

2.3 Sparse Mixture-of-Experts and Gate Fusion

To reduce computation overhead and combine multimodal features, two prominent techniques should be considered. First, SMoE (Touvron et al. 2023) is a technique that avoids unnecessary computation by selectively activating relevant experts. Considering the dominance of text (Wu et al. 2025), we can use text for SMoE routing. Second, GF is a commonly used structure for integrating features from different modalities (Ren et al. 2018). That is, we can employ GF to combine text, audio, and video features in MSA (Cheng et al. 2025). However, to our knowledge, neither SMoE nor GF has attracted adequate attention regarding MSA.

3 Method

3.1 The Overall Architecture of TEXT

Fig. 3 is the overall architecture of TEXT. From the top down, TEXT has six modules: ① GF classifier; ② text-routed SMoE; ③ temporal aligning; ④ text encoding; ⑤⑥: explanation aligning blocks for audio and video (with uni-modal encoding). In this section, we will first introduce our explanation generation approach (§3.2) and then present the uni-modal encoding methods (§3.3). The explanation for aligning blocks in modules ⑤ and ⑥ is provided in §3.4, and details of the temporal aligning blocks can be found in §3.5. At last, the ① GF classifier will be stated in §3.7 after §3.6, which presents the ② text-routed SMoE.

3.2 Explanation Generation

To explore the power of MLLM, our explanation generation process is two-stage. See Fig. 4. We first generate raw explanations using VideoLLaMA 3, which is fine-tuned on the EMER-fine dataset (Lian et al. 2025). At this stage, the prompt separates explanations of audio, video, and comments. Then, with Qwen 3, we refine raw explanations to fine explanations with the checking prompt. The prompt for the second stage is called the reasoning prompt.

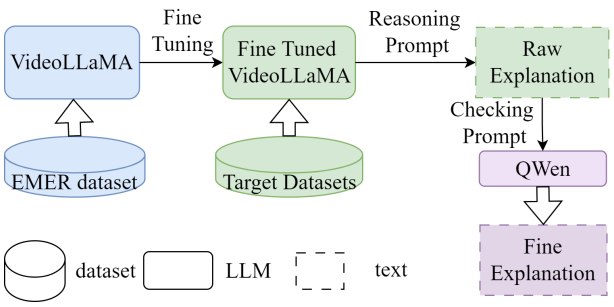


Figure 4: The procedure of explanation generation.

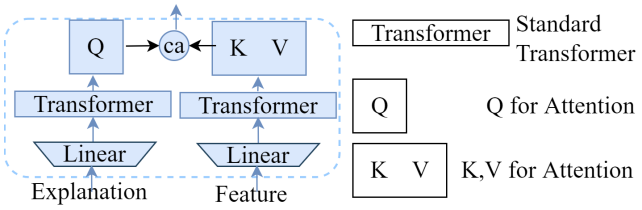


Figure 5: Explanation aligning block.

3.3 Uni-Modal Encoding

We encode subtitles and explanations using BERT, audio using Librosa, and video using OpenFace. See the lower part of Fig. 3. Viewing models as functions, we have $BERT(t), Librosa(a), OpenFace(v)$ $(B(t), Li(a), OF(v)$ for short). Because we separate the explanation for audio, video, and comments in §3.2, $B(t)$ can be further specified as the explanation for audio $B(e_a)$, the explanation for video $B(e_v)$ and the comments $B(c)$.

3.4 Explanation Aligning

In the explanation aligning block, audio $Li(a)$, video $OF(v)$ and subtitles $B(s)$ are aligned with corresponding explanations or comments $B(e_a)$/$B(e_v)$/$B(c)$ via CA ($\textcircled{ca}(\cdot)$). That is, for feature F and explanation E, $\textcircled{ca}(F, E)$ is as Eq. 1, where W_Q, W_K and W_V are weights and the transpose of a matrix is ^{T}. We illustrate this module in Fig. 5.

$$\textcircled{ca}(F, E) = softmax((W_Q E)(W_K F)^T)W_V F \quad (1)$$

As the lengths of $B(t), Li(a), OF(v)$ may differ, we use 50 tokens for all uni-modal encodes and a learnable token for feature aggregation. The final representation is a 51-dimensional embedding. The results are aligned representations of text, audio, and video. Namely, E_t, E_a, and E_v.

3.5 Temporal Aligning

TEXT uses a novel temporal alignment block, which includes one convolution ($Conv1d(\cdot)$) and two linear operations. It combines the advantages of Mamba and TCA. Fig. 6 shows this temporal aligning block. Comparing Fig. 6 with Fig. 2, we can see that our temporal aligning block is simpler than Mamba and TCA because it does not involve CA or SSM.

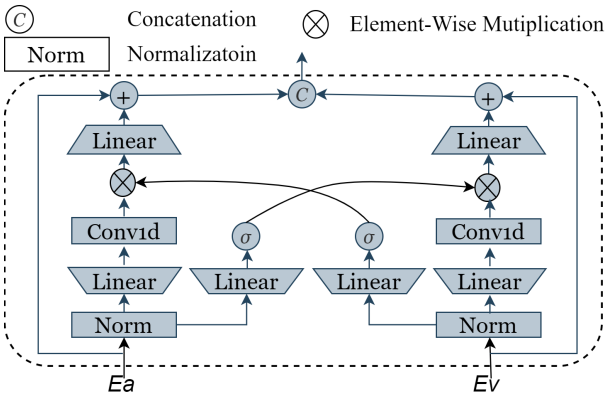


Figure 6: Temporal aligning block.

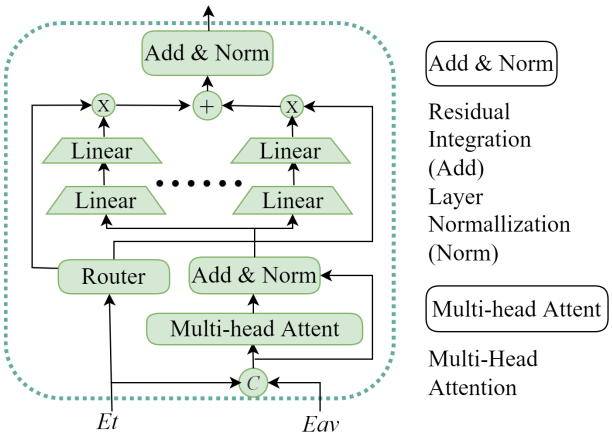


Figure 7: The SMoE block. The symbol $\textcircled{x}$ represents multiplication, and "Router" refers to the routing function.

Formally, $\otimes$ is used for element-wise multiplication, and $\oplus$ is for element-wise addition. We use $\textcircled{c}$ to represent the concatenation of features, let $L(\cdot)$ denote the linear layer, and let $N(\cdot)$ represent the normalization layer. Then, $LN(\cdot)$ is a linear layer after a normalization layer. The Sigmoid Linear Unit activation function (SiLU), $\textcircled{\sigma}$, is used for activation. Let Eq. 2 be the left part of Fig. 6, Eq. 3 be the right part of Fig. 6, and E_{av} be the temporal aligned representation. The temporal alignment can be defined as Eq. 4.

$$left = E_a \oplus L(Conv1d(LN(E_a)) \otimes \textcircled{\sigma}(LN(E_v))) \quad (2)$$

$$right = E_v \oplus L(Conv1d(LN(E_v)) \otimes \textcircled{\sigma}(LN(E_a))) \quad (3)$$

$$E_{av} = \textcircled{c}(left, right) \quad (4)$$

3.6 Text-Routed SMoE

This module is illustrated in Fig. 7.

An SMoE structure using text as the key for route decisions. Suppose the first parameter of the function $SMoE(\cdot)$

Model	MOSI						MOSEI					
	Acc-2	Acc-5	Acc-7	F1	MAE↓	Corr	Acc-2	Acc-5	Acc-7	F1	MAE↓	Corr
ALMT	83.10/85.23	50.41	45.01	83.20/85.37	0.716	0.773	82.39/85.87	53.96	52.16	82.18/85.95	0.542	0.767
KuDA[1]	84.40/86.43	N/A	**47.08**	84.48/86.46	0.705	0.795	83.26/86.46	N/A	**52.89**	82.97/86.59	0.529	0.776
DEVA	84.40/86.29	51.78	46.32	84.48/86.30	0.730	0.787	83.26/86.13	**55.32**	52.26	82.93/86.21	0.541	0.769
GPT-4o	85.71/86.74	52.59	44.61	85.68/86.68	0.682	0.823	84.77/86.08	50.53	48.38	84.82/86.08	0.637	0.744
Qwen	83.09/83.38	45.63	36.30	83.09/83.31	1.129	0.677	84.14/84.59	41.73	40.67	84.17/84.64	1.007	0.587
VL3	67.64/68.45	28.72	23.76	68.30/68.48	1.437	0.442	71.07/71.20	33.12	31.87	71.35/71.59	1.141	0.349
TEXT	**86.44/88.72**	**52.62**	45.92	**86.55/88.76**	**0.666**	**0.829**	**85.02/86.57**	54.05	52.29	**85.01/86.85**	**0.528**	**0.786**

[1] We use the results in previous studies for KuDA (Feng et al. 2024) and DEVA (Wu et al. 2025) in this paper. QWen: Qwen2.5-vl. VL3: VideoLLaMA3.

Table 1: Comparison on MOSI and MOSEI Datasets. Acc and F1 are shown in percentage scale. All results in our paper is statistical significant using T-test, i.e. $p < 0.05$. The best results are in bold, and the second best results are underlined. Acc-2 and F1-Score are computed in two settings: negative/non-negative (including zero) and negative/positive (excluding zero).

is the key of routing; this layer can be formalized as $SMoE(E_t, E_{av})$.

3.7 Gate Fusion Classifier

An MLP with a GF(ⓒ) comprises the classifier of TEXT, see Eq. 5. See the upper part of Fig. 3 for an intuitive understanding. Specifically, only tokens for feature aggregation (see §3.4) are used for the final decision.

$$L(ⓒ(SMoE(E_t, E_{av}))) \tag{5}$$

4 Experiment

4.1 Experiment Settings

4.2 Datasets

MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018), CH-SIMS (Yu et al. 2020) and CH-SIMSv2 (Liu et al. 2022) are four popular datasets of TEXT. The Multi-modal Opinion-Level Sentiment Intensity (MOSI) is a collection of YouTube monologues. It contains 2,199 subjective words and video clips, which are artificially labeled as consecutive opinion scores. The CMU Multi-modal Opinion Sentiment and Emotion Intensity (MOSEI) is an improvement on MOSI. It contains 22,856 YouTube monologues and video segments covering 250 distinct topics from 1,000 distinct speakers. The CHinese SIngle- and Multi-modal Sentiment analysis (CH-SIMS) is a Chinese TEXT dataset with fine-grained annotations of modality. It contains 2,281 human-labeled video clips collected from various sources, along with a sentiment score ranging from -1 (strongly negative) to 1 (strongly positive). At last, the CH-SIMSv2 is the updated version of CH-SIMS.

4.3 Compared Models

TEXT is compared with three models and three MLLMs.

The three recently proposed representative models are ALMT (Zhang et al. 2023), KuDA (Feng et al. 2024) and DEVA (Wu et al. 2025). First, ALMT learns an irrelevance/conflict-suppressing representation from visual and audio features, and each modality is first transformed into a unified form by using a Transformer (Vaswani et al. 2017) with initialized tokens. Second, KuDA argues that there is always a dominant modality, which is enhanced by sentiment knowledge. Third, DEVA incorporates the text-guided progressive fusion along with an emotional description generator. Many previous studies in this research line that have been compared with ALMT, KuDA, and DEVA is not listed in this paper.

The three MLLMs are Qwen2.5-vl (Bai et al. 2025), GPT-4o (Achiam et al. 2023) and VideoLlama3-7B (Zhang et al. 2025).

4.4 Metrics

Following the previous works (Zhang et al. 2023; Feng et al. 2024; Wu et al. 2025), we facilitate metrics from two classes. The first class is for classification, which includes the Weighted F1 score (F1-Score), binary classification accuracy (Acc-2), three-class classification accuracy (Acc-3), five-class classification accuracy (Acc-5), and seven-class classification accuracy (Acc-7). The second class focuses on regression, which includes the Mean Absolute Error (MAE) and Pearson correlation (r). For all metrics, except MAE, higher values indicate better performance.

For MOSI and MOSEI, we further compute Acc-2 and F1-Score in two settings, as in previous works (Zhang et al. 2023; Feng et al. 2024; Wu et al. 2025). That is, negative/non-negative (including zero) and negative/positive (excluding zero). Further, we calculate Acc-3 and Acc-5 on CH-SIMS and CH-SIMSv2.

4.5 Model Comparison

Table. 1 compared our model with compared methods on MOSI and MOSEI. The same comparisons on CH-SIMS and CH-SIMSv2 are illustrated in Fig. 8.

TEXT's advantages are clear in Table. 1 and Fig. 8, as it excels in almost every test. Specifically, five key results highlight the significance of TEXT:

1. For Acc-5, TEXT is better than KuDA (the 2^{nd} best model) on CH-SIMS, with 50.15% accuracy versus KuDA's 43.54%.
2. For the F1 score on CH-SIMS, TEXT reaches 86.75%, maintaining a clear lead over KuDA's 80.71%.

Settings	Method	Acc-2	Acc-5	Acc-7	F1	MAE	Corr
w explanations	**TEXT**	**85.02/86.57**	54.05	52.29	**85.01/86.85**	**0.528**	**0.786**
	A & V	75.51/77.71	43.76	43.57	75.02/77.94	0.694	0.582
	T & A	83.38/85.06	54.07	52.46	83.54/85.57	0.542	0.773
	T & V	83.01/85.75	50.55	48.62	83.89/85.94	0.566	0.776
	T	83.49/86.43	**54.56**	**52.84**	83.24/86.57	0.535	0.771
	A	75.90/74.00	39.88	39.30	77.15/75.89	0.776	0.492
	V	72.55/68.49	39.47	39.47	76.05/73.11	0.807	0.363
w/o explanations	TEXT	83.60/86.02	52.39	50.35	83.42/86.20	0.569	0.776
	A & V	70.59/62.71	41.36	41.36	82.21/76.73	0.834	0.146
	T & A	83.92/85.50	53.44	51.73	83.82/85.71	0.546	0.766
	T & V	83.32/85.83	50.96	48.98	83.05/85.94	0.573	0.765
	T	83.58/85.58	52.39	50.80	83.42/85.77	0.549	0.763
	A	71.02/62.85	41.27	41.27	73.06/70.19	0.838	0.153
	V	70.94/62.91	41.36	41.36	73.81/70.12	0.828	0.172
Component Ablation[1]	EA← Linear	84.25/86.57	49.58	48.21	84.11/86.77	0.577	0.762
	TA ← Ⓒ	83.77/85.42	49.52	48.40	83.83/85.80	0.580	0.749
	TA ← Mamba	84.80/86.41	52.65	50.65	84.75/86.63	0.562	0.780
	TA ← TCA	83.41/86.43	53.98	51.38	83.12/86.55	0.565	0.781
	SMoE ← Trans	83.73/85.33	52.22	50.29	83.70/85.62	0.573	0.769
	TEXT w/o Gating	84.40/86.35	51.41	49.07	84.35/86.62	0.571	0.780

[1] EA ← Linear : EA is replaced by a linear layer. TA ← ⓒ: TA is replaced by concatenation. TA ← Mamba: TA is replaced by Mamba. TA ← TCA: TA is replaced by TCA. SMoE ← Trans : SMoE is replaced by the Transformer.

Table 2: Ablation study results on MOSEI. Acc and F1 are shown in percentage scale, and best results are in bold. Acc-2 and F1-Score are computed in two settings: negative/non-negative (including zero) and negative/positive (excluding zero). A: audio; V: video; T: text. TA: Temporal Alignment; EA: Explanation Alignment. SMoE: Text-Routed Sparse Mixture-of-Experts.

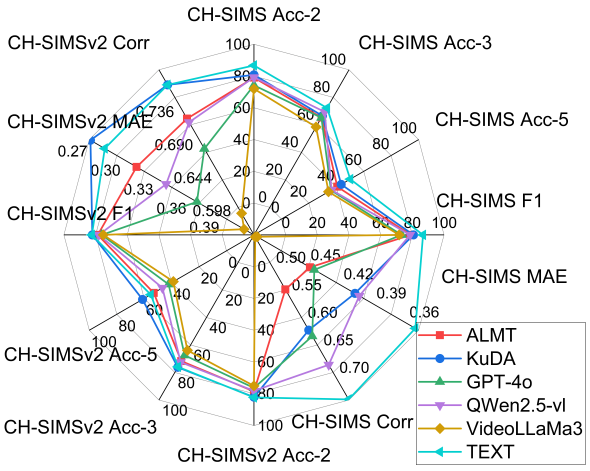


Figure 8: Comparison on CH-SIMS and CH-SIMSv2. Axes are metrics, and lines are compared models.

3. Regarding MAE on CH-SIMS, TEXT attains a notably lower error rate of 0.353—indicating superior performance—compared to QWen's 0.404.
4. In the Acc-2 evaluation on the MOSI dataset, TEXT achieves 88.72%, surpassing GPT-4o's 86.74%.
5. For the F1 score on MOSI, TEXT further solidifies its advantage with a score of 88.76%, outperforming GPT-4o's 86.68%.

However, there are two exceptions. First, KuDA is the best model for Acc-7 on MOSI and MOSEI. KuDA's success has also been proven for Acc-5 on CH-SIMSv2. Second, DEVA shows its advantages on Acc-5 on MOSEI. We conjecture that fine-grained scoring requires additional knowledge of sentiment (even domain-specific) to be effective.

4.6 Ablation Study

Table. 2 lists the results of the ablation study on MOSEI and Fig. 9 shows the corresponding results on CH-SIMS and CH-SIMSv2.

Analysis of Table. 2 yields five key insights into multimodal performance and the effects of ablation studies:

1. Text as the dominant uni-modal input: it outperforms audio and video across metrics like Acc-5 and Acc-7. For example, regarding Acc-7 on MOSEI using text with explanations yields 52.84%, while TEXT achieves only 52.29%. The dominance indicates that the text likely contains more accurate information than other modalities (see Fig. 1).
2. Audio and video contribute equally to performance: when integrated with text, both audio and video provide comparable performance boosts. That is, the results of "`T & V`" and "`T & A`" in Table. 2 are similar. These modalities complement textual information with distinct yet equally valuable contextual cues, such as prosody in audio and visual signals in video.
3. Explanations enhance prediction: the removal of explanations results in an approximate 2% performance decline across most metrics. This drop corroborates earlier findings, suggesting that explanations play a crucial role in integrating text with audio and video.

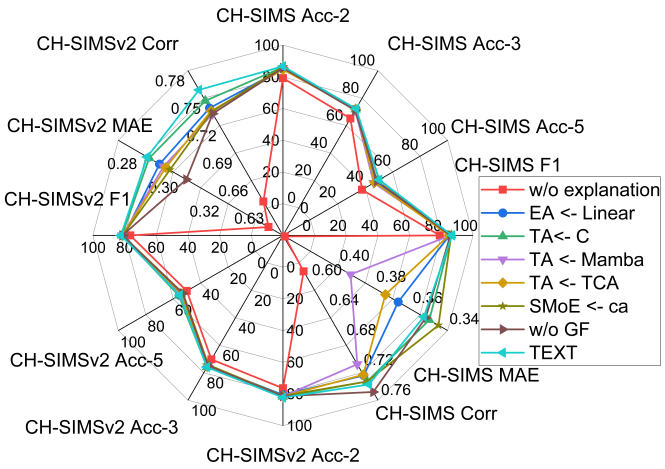


Figure 9: Ablation study on CH-SIMS and CH-SIMSv2. Axes are metrics, and lines are compared models.

4. Temporal alignment is crucial for multimodal integration: replacing our temporal alignment block with Mamba or TCA also causes a roughly 2% performance drop. For example, when we replaced temporal alignment with concatenation, the MAE was 0.580 in Table. 2. This suggests temporal alignment is crucial for MAE. That is, matching audio and text timing might be key for MSA.
5. SMoE demonstrates comparable value to explanations. We believe the effectiveness of SMoE stems from its keyword-sensitive expert activation mechanism. That is, some experts are trained for a specific topic, and text including corresponding keywords will activate these experts. Furthermore, it may foster cross-modal consistency for improved interpretability.

4.7 Qualitative Examples

Table. 3 shows the predicted scores evaluated by different settings of TEXT for the case in Fig. 1. The third column represents the predicted score, and the last column shows the deviation between the label and the evaluated score.

Table. 3 provides evidence for three key observations. First, in this specific case, TEXT shows the strongest alignment with human judgment, with a negligible discrepancy of only 0.01. Second, the textual modality acts as the dominant information channel. For example, relying solely on textual data—even without explanatory context—still results in a relatively small discrepancy of 0.390. Third, explanations effectively compensate for the limitations of audio and visual modalities, especially for audio. When explanations are incorporated, audio-based predictions achieve a discrepancy of 0.380; by contrast, omitting explanations leads to a significant performance decline, with the discrepancy rising to 1.440. Notably, although these statistical patterns are consistent, GF and SMoE have shown minimal impact in this particular scenario. For example, without GF or SMoE, the derivation increases by 0.050.

Settings	Model	Score	$\sigma \downarrow$
Human	N/A	1.400	N/A
Compared Model MLLM	ALMT	1.080	0.320
	GPT-4o	0.800	0.600
	Qwen2.5-vl	2.500	1.100
	VideoLLaMA3	2.000	0.600
Our Model	**TEXT**	**1.390**	**0.010**
Uni-modal	T	1.060	0.340
	A	1.780	0.380
	V	0.870	0.530
Two Modals	T & A	1.350	0.050
	T & V	1.330	0.070
	A & V	1.510	0.110
Uni-modal w/o explanation	T	1.010	0.390
	A	-0.040	1.440
	V	-0.380	1.780
Two Modals w/o explanation	T & A	1.050	0.350
	T & V	1.310	0.090
	A & V	0.070	1.330
Component Ablation[1]	w/o explanations	0.550	0.850
	EA← Linear	1.270	0.130
	TA ← ⓒ	1.420	0.200
	TA ← Mamba	1.220	0.180
	TA ← TCA	1.080	0.320
	SMoE ← Trans	1.320	0.080
	w/o GF	1.350	0.050

Table 3: Predicted scores for the case in 1. The best results are highlighted in bold. σ: deviation. See Table. 2 for more abbreviations.

4.8 Discussion

Our experiments on hyperparameters show that three layers are optimal for SMoE on CH-SIMS. On the other hand, while MLLMs can be very good at MSA on some datasets, these datasets might be memorized by MLLMs (Wang et al. 2024). As evidence, the power of GPT-4o diminishes on Chinese datasets. Although this paper only considers MSA for Chinese/English, we will expand TEXT to use more MLLMs in multiple languages.

5 Conclusion

In this paper, we propose a text-routed mixture-of-experts model with explanation and temporal alignment for multimodal sentiment analysis. With a novel temporality-oriented neural network block and cross-attention, our model performs explanation and temporal alignment. While the explanation comes from exploring the power of MLLMs, our temporal alignment block is a task-oriented neural network block design. Then, the aligned embedding is further processed by a new text-routed sparse mixture-of-experts with a gate fusion. As a result, we achieve the best performance across four datasets among all tested models, which include three state-of-the-art models and three leading MLLMs. However, as we rely on multiple MLLMs, eliminating cumulative error (e.g., from VideoLLaMA) is our future work.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baltrusaitis, T.; Zadeh, A.; Lim, Y.; and Morency, L. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International conference on automatic face & gesture recognition (FG 2018)*, 59–66.

Chen, Y.-F.; and Huang, S.-H. 2021. Sentiment-influenced trading system based on multimodal deep reinforcement learning. *Applied Soft Computing*, 112: 107788.

Cheng, C.; Xu, T.; Feng, Z.; Wu, X.; Tang, Z.; Li, H.; Zhang, Z.; Atito, S.; Awais, M.; and Kittler, J. 2025. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28102–28112.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and 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.

Feng, X.; Lin, Y.; He, L.; Li, Y.; Chang, L.; and Zhou, Y. 2024. Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14755–14766.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Lian, Z.; Chen, H.; Chen, L.; Sun, H.; Sun, L.; Ren, Y.; Cheng, Z.; Liu, B.; Liu, R.; Peng, X.; et al. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. *ICML (Spotlight)*.

Liu, Y.; Yuan, Z.; Mao, H.; Liang, Z.; Yang, W.; Qiu, Y.; Cheng, T.; Li, X.; Xu, H.; and Gao, K. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, 247–258.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, 18—24. SciPy.

Park, J.; Kim, M.-H.; and Choi, D.-G. 2021. Correspondence learning for deep multi-modal recognition and fraud detection. *Electronics*, 10(7): 800.

Qiu, Z.; Wang, Z.; Zheng, B.; Huang, Z.; Wen, K.; Yang, S.; Men, R.; Yu, L.; Huang, F.; Huang, S.; et al. 2025. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. *arXiv preprint arXiv:2505.06708*.

Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; and Yang, M.-H. 2018. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3253–3261.

Rozanska, A.; and Podpora, M. 2019. Multimodal sentiment analysis applied to interaction between patients and a humanoid robot Pepper. *IFAC-PapersOnLine*, 52(27): 411–414.

Shah, A. M.; Yan, X.; Shah, S. A. A.; and Mamirkulova, G. 2020. Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. *Journal of Ambient Intelligence and Humanized Computing*, 11(7): 2925–2942.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Bao, R.; Wu, Y.; Taylor, J.; Xiao, C.; Zheng, F.; Jiang, W.; Gao, S.; and Zhang, Y. 2024. Unlocking Memorization in Large Language Models with Dynamic Soft Prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9782–9796.

Wu, S.; He, D.; Wang, X.; Wang, L.; and Dang, J. 2025. Enriching Multimodal Sentiment Analysis through Textual Emotional Descriptions of Visual-Audio Content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1601–1609.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 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.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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)*, 2236–2246.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025.

VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 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*, 756–767.

Zhang, Y.; Chen, T.; Zhang, Y.; and Fu, Z. 2024. Enhanced Multimodal Hate Video Detection via Channel-wise and Modality-wise Fusion. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, 183–190. IEEE.