Fast Retrieval and Slow Reasoning for Explainable Multimodal Sentiment Analysis

Aoqiang Zhu[1,2], **Min Hu**[1,2*], **Yan Xing**[3]
[1]Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine
[2]School of Computer Science and Information Engineering, Hefei University of Technology
[3]School of Mathematics, Hefei University of Technology
aoqzhu@gmail.com, {jsjxhumin, yanxing}@hfut.edu.cn

Abstract

Most existing Multimodal Sentiment Analysis (MSA) methods rely on holistic fusion, treating all modalities and temporal segments equally. Such strategies often introduce redundant information and obscure the decision process, limiting both robustness and interpretability. Inspired by dual-process theory, we propose **FRSR** (**F**ast **R**etrieval and **S**low **R**easoning), an interpretable framework that decomposes multimodal sentiment modeling into two cooperative pathways. The Fast Pathway acts as a lightweight evidence selector, using context-aware convolution and auxiliary supervision to retrieve a sparse set of Top-K sentiment-relevant cues from noisy multimodal inputs. Based on these cues, the Slow Pathway performs deeper cross-modal reasoning through learnable reasoning tokens, enabling hierarchical sentiment inference. By separating salient evidence retrieval from multimodal reasoning, FRSR improves interpretability while reducing computational cost. Experiments on three benchmark datasets show that FRSR achieves competitive performance, higher efficiency, stronger robustness to noise, and clearer decision transparency than existing holistic fusion methods.

1 Introduction

Multimodal Sentiment Analysis (MSA) is an important research topic in affective computing and human-centered AI (Singh et al., 2024). By jointly modeling language, visual, and acoustic information, MSA has been widely applied to conversational agents (Li et al., 2025), social media analytics (Zeng et al., 2023), and intelligent interaction systems (Dzedzickis et al., 2025). With the development of Transformer-based architectures, recent multimodal models have achieved remarkable progress on a variety of sentiment benchmarks, benefiting from stronger representation capacity and more flexible cross-modal interaction mechanisms (Zhang et al., 2022; Zhu et al., 2024; Shi and Huang, 2023; Yang et al., 2023). Nevertheless, despite these advances, most existing MSA methods still share a common limitation: they rely on unified fusion strategies that process all multimodal observations in an undifferentiated manner.

In most existing frameworks, informative emotional signals, such as facial expressions or vocal intonation, are fused together with irrelevant contextual content inside the same modeling pipeline (Zhu et al., 2025b; Tsai et al., 2019; Huang et al., 2023; Zhang et al., 2023a; Wu et al., 2025; Qian et al., 2023). As a result, the model must implicitly determine which information is useful while simultaneously learning sentiment representations from highly redundant inputs. This design leads to two major issues. **(1) Sensitivity to Redundant Information:** Models can easily become biased toward irrelevant modality-specific patterns or environmental artifacts. For example, repeated backgrounds, speaker-specific habits, or stable acoustic rhythms may be incorrectly treated as meaningful sentiment evidence, weakening robustness under diverse real-world conditions. **(2) Limited Interpretability:** Since all information is entangled during fusion, it is difficult to explicitly identify which multimodal cues contribute to the final prediction. Even when the output is correct, the reasoning basis behind the decision often remains unclear. These challenges indicate that simply building larger or more complicated fusion modules does not fundamentally solve the problems of robustness and interpretability in MSA.

Unlike existing computational models, human affective understanding typically follows a staged cognitive process. According to Dual-Process Theory (Kahneman, 2011, 2003; Stanovich and West, 2000), people often make judgments through the cooperation of two complementary systems. System 1 rapidly focuses on prominent perceptual pat-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 30381–30391

terns, such as sudden changes in tone or short-lived facial movements, while System 2 further interprets these observations through deeper semantic analysis and contextual reasoning (Stanovich and West, 2000). This cognitive mechanism suggests that an effective MSA model should avoid treating all multimodal content equally. Instead, it should first locate potentially useful affective evidence and then conduct more deliberate reasoning on the selected information.

Motivated by this idea, we introduce FRSR (Fast Retrieval and Slow Reasoning), an interpretable framework that decomposes multimodal sentiment modeling into evidence selection and structured reasoning. Fast Pathway (System 1). As a lightweight perception module, it employs a Conv1D-based local modeling network to estimate token-level importance and preserve only Top-K informative cues from redundant multimodal sequences. In addition, auxiliary unimodal supervision is adopted to improve preliminary sentiment awareness and stabilize the salience estimation process. Slow Pathway (System 2). Rather than processing the entire multimodal sequence, the Slow Pathway operates only on the retained evidence. By introducing learnable reasoning tokens, it progressively captures cross-modal dependencies and performs hierarchical sentiment inference over the selected cues. Since intensive multimodal interaction is restricted to a compact subset of salient information, the proposed framework can substantially reduce unnecessary computation while maintaining strong predictive ability.

Our main contributions are as follows:

- We propose FRSR, an interpretable framework for MSA that decouples multimodal learning into salient evidence retrieval and sentiment reasoning, thereby alleviating the redundancy and opacity issues of conventional holistic fusion methods.

- We develop a dual-path architecture consisting of a Fast Pathway and a Slow Pathway. The former quickly identifies sparse Top-K sentiment-related cues through context-aware convolution, while the latter conducts cross-modal interaction and hierarchical reasoning on the retrieved evidence.

- Extensive experiments on three public benchmarks demonstrate that FRSR achieves highly competitive predictive performance. Furthermore, compared to state-of-the-art methods, FRSR provides intuitive, transparent evidence for its predictions while significantly improving computational efficiency and robustness to noise.

2 Related Work

2.1 Multimodal Sentiment Analysis

Research on Multimodal Sentiment Analysis has long focused on how to effectively combine heterogeneous information from language, visual, and acoustic modalities. Earlier studies mainly explored dedicated fusion schemes, including tensor-based interaction modeling (Zadeh et al., 2017; Liu et al., 2018), quantum-inspired fusion strategies (Zhang et al., 2020; Phukan et al., 2024), and contrastive representation alignment (Yu et al., 2023; Mai et al., 2022). In recent years, Transformer-driven frameworks have become the dominant paradigm due to their ability to capture long-range dependencies and flexible cross-modal relationships through attention mechanisms (Hazarika et al., 2020; Yang et al., 2022; Zhu et al., 2026; Tao et al., 2025; Wen et al., 2025; Wang et al., 2025; Zhang et al., 2024; Zhu et al., 2025a). Typical examples include *DashFusion* (Wen et al., 2025), which combines temporal-semantic alignment with hierarchical bottleneck fusion, and *DEVA* (Wu et al., 2025), which improves multimodal representations by incorporating text-derived sentiment priors into visual and acoustic features.

Although these methods have achieved encouraging results, most of them still depend on end-to-end fusion pipelines that mix multimodal features across different sources and temporal positions without distinction. As a consequence, irrelevant contextual patterns are often retained together with sentiment-bearing information. Since the fusion process itself does not explicitly separate useful evidence from noisy content, model decisions are easily influenced by redundant cues and remain difficult to interpret.

2.2 Dual-Process Cognition

Dual-Process Theory characterizes human cognition as the interaction between a fast, intuitive mechanism (*System 1*) and a slower, analytical mechanism (*System 2*) (Kahneman, 2003; Stanovich and West, 2000). In *Thinking, Fast and*

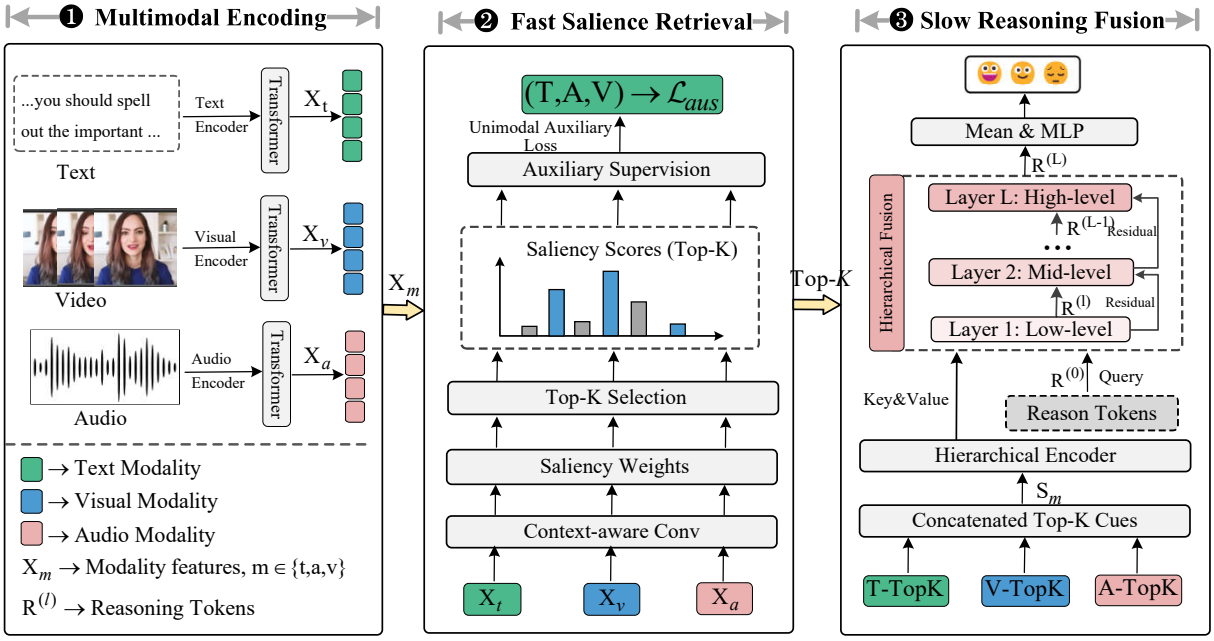

Figure 1: **Overview of the proposed FRSR framework.** Inspired by dual-process theory, FRSR consists of three main components: ❶ **Multimodal Encoding** for initial feature extraction and alignment; ❷ **Fast Salience Retrieval** for extracting sparse top-K critical cues; and ❸ **Slow Reasoning Fusion**, which employs learnable reasoning tokens to facilitate hierarchical cross-modal interaction and sentiment reasoning.

Slow (Kahneman, 2011), Kahneman describes System 1 as rapidly identifying salient perceptual cues, while System 2 performs more deliberate reasoning and contextual analysis. Inspired by this perspective, several MSA methods introduce selective attention or salience estimation into multimodal modeling (Zhang et al., 2023b; Tellamekala et al., 2023). However, most remain limited to attention reweighting or shallow feature filtering (Yang et al., 2023; Yuan et al., 2021), where attention scores are often treated as explanations despite not necessarily reflecting the evidence that truly supports the final prediction.

Different from these approaches, FRSR explicitly builds a Fast–Slow reasoning structure for interpretable sentiment analysis. Instead of performing direct holistic fusion over the entire multimodal sequence, the framework first retrieves sparse sentiment-related evidence and then carries out deeper reasoning only on the retained cues. This design allows the prediction process to be associated with more transparent multimodal evidence rather than hidden statistical patterns.

3 Methodology

3.1 Overall Framework

Fig. 1 illustrates the overall architecture of FRSR. Inspired by Dual-Process Theory, FRSR decomposes conventional holistic fusion into a two-stage retrieval-and-reasoning framework for multimodal sentiment analysis.

Fast Salience Retrieval (Sec. 3.3) corresponds to System 1 and acts as a lightweight evidence selection module. It uses a Conv1D-based network to estimate token-level importance and retain only a sparse set of Top-K sentiment-relevant cues. **Slow Reasoning Fusion** (Sec. 3.4) corresponds to System 2 and performs reasoning over the selected evidence rather than the full sequence. By introducing learnable Reasoning Tokens, it progressively models cross-modal dependencies and conducts hierarchical sentiment inference over the retained multimodal cues.

3.2 Multimodal Encoding

Experiments are conducted on three widely used multimodal sentiment benchmarks: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and SIMS (Yu et al., 2020). Following prior work (Wu et al., 2025; Zhang et al., 2024), textual, acoustic, and visual features are extracted using BERT (Devlin et al., 2019), librosa (Baltrusaitis et al., 2018), and OpenFace (McFee et al., 2015), respectively.

To align feature dimensions across modalities, each modality is projected into a shared hidden space through a linear embedding layer followed

by two Transformer layers. For modality $m \in \{t, a, v\}$, the encoded representation is denoted as $X^m \in \mathbb{R}^{L_m \times d_m}$, where L_m and d_m represent the sequence length and hidden feature dimension.

3.3 Fast Salience Retrieval

The Fast Pathway aims to rapidly identify informative affective patterns from long and noisy multimodal sequences. Since sentiment-related cues often appear in short temporal segments, such as brief facial changes or sudden vocal variations, the model should emphasize locally important regions while avoiding unnecessary processing of redundant content. To this end, we design a lightweight salience estimation module based on temporal convolution.

Considering an individual modality stream $m \in \{T, V, A\}$, we denote its initial temporal features as $X_m \in \mathbb{R}^{B \times S \times D}$, where B, S, and D indicate the batch size, sequence length, and feature dimension, respectively. We initially apply a two-layer Conv1D architecture to evaluate the significance of each discrete token:

$$S_m = \text{Conv1D}_{k=1}\Big(\text{ReLU}\big(\text{Conv1D}_{k=3}(X_m)\big)\Big). \quad (1)$$

The first convolution layer uses a kernel size of 3 to model short-range temporal dependencies by jointly considering each token together with its neighboring positions. Following this, the second convolution layer performs a point-wise projection that compresses the intermediate hidden features into a single-channel salience score for each temporal position:

$$S_m \in \mathbb{R}^{B \times S \times 1}. \quad (2)$$

To transform the raw salience scores into normalized importance weights, we apply a Sigmoid function:

$$W_m = \text{Sigmoid}(S_m), \quad W_m \in [0, 1]. \quad (3)$$

Based on the derived weights, we extract only the Top-K instances exhibiting the highest salience values:

$$\mathcal{I}_m = \text{TopK}(W_m, K), \quad (4)$$

$$\hat{X}m = Xm, i \mid i \in \mathcal{I}_m \in \mathbb{R}^{B \times K \times D}. \quad (5)$$

The resulting subset $\hat{X}_m$ preserves only the most informative temporal segments for the corresponding modality and serves as the input to the subsequent Slow Pathway. In addition to selecting salient tokens, we further introduce an auxiliary unimodal prediction branch to encourage the Fast Pathway to learn sentiment-aware representations. Specifically, we first aggregate the selected Top-K features through weighted averaging:

$$h_m^{fast} = \sum_{i \in \mathcal{I}m} \tilde{w}iXm, i, \quad (6)$$

where the normalized weight $\tilde{w}_i$ is computed as:

$$\tilde{w}i = \frac{Wm, i}{\sum j \in \mathcal{I}mWm, j}. \quad (7)$$

Although the Top-K operation is discrete, the re-normalized weights provide a differentiable path through which the auxiliary loss can partially propagate gradient information back to the salience estimation module. The aggregated representation h_m^{fast} is then passed into a lightweight unimodal classifier:

$$\hat{y}m^{fast} = \text{MLP}aux(h_m^{fast}). \quad (8)$$

By imposing auxiliary supervision on each individual modality, the Fast Pathway is encouraged to capture preliminary sentiment information before multimodal fusion, which improves the stability and reliability of token selection.

3.4 Slow Reasoning Fusion

After salient multimodal cues have been retrieved, the Slow Pathway performs deeper reasoning on the selected evidence. Unlike conventional fusion models that process the entire sequence at once, our framework restricts reasoning to a much smaller set of informative tokens, making the interaction process both more efficient and more interpretable.

To construct a joint representation, the extracted Top-K segments from all modalities are grouped together along the temporal dimension, yielding:

$$\hat{X}_{all} = [\hat{X}_T; \hat{X}_V; \hat{X}_A]. \quad (9)$$

The composite sequence is then processed by a Transformer encoder spanning L layers. Rather than exclusively utilizing the computations from the terminal layer, our approach extracts the multiscale representations generated across all transformer blocks, $\mathcal{H} = H^{(l)}{}_{l=0}^{L}$, where

$$H^{(l)} = \text{TransformerLayer}_l(H^{(l-1)}). \quad (10)$$

Here, $H^{(0)}$ corresponds to the initial multimodal embedding, while deeper layers contain progressively richer contextual and semantic information.

To explicitly model hierarchical reasoning, we introduce a set of learnable reasoning tokens:

$$R^{(0)} \in \mathbb{R}^{K \times D}. \tag{11}$$

These learnable tokens serve as latent queries that repeatedly attend to hierarchical multimodal features extracted from different Transformer layers. At each layer l, the reasoning tokens attend to the corresponding multimodal representation through cross-attention:

$$R^{(l)} = \text{CrossAttn}\big(\text{LN}(R^{(l-1)}),\text{LN}(H^{(l)}), \text{LN}(H^{(l)})\big) + R^{(l-1)}. \tag{12}$$

Here, $\text{LN}(\cdot)$ denotes Layer Normalization. In the cross-attention operation, the reasoning tokens are used as queries, while the hidden multimodal representations serve as keys and values. Through repeated interactions across layers, the reasoning tokens gradually accumulate information from different modalities and temporal levels. Representations in lower layers mainly encode local contextual relationships, while deeper layers gradually capture more abstract cross-modal semantics and global sentiment structures.

Finally, the refined reasoning tokens from the last layer are aggregated through mean pooling and fed into a two-layer prediction head:

$$\hat{y} = \text{MLP}(\text{MeanPool}(R^{(L)})). \tag{13}$$

The final output $\hat{y}$ corresponds to the predicted sentiment score or category for the input sample.

3.5 Overall Learning Objectives

The optimization objective of FRSR consists of the primary sentiment prediction loss together with auxiliary supervision from the Fast Pathway:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}}(y, \hat{y}) + \lambda_{\text{aux}}(t)\mathcal{L}_{\text{aux}}. \tag{14}$$

Here, $\mathcal{L}_{\text{task}}$ denotes the loss for the final sentiment prediction. For regression settings, we adopt Mean Squared Error (MSE) between the ground-truth label y and the prediction $\hat{y}$ generated by the Slow Pathway:

$$\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{15}$$

To encourage the Fast Pathway to identify sentiment-related cues at the unimodal level, we further introduce an auxiliary loss over the unimodal predictions:

$$\mathcal{L}_{\text{aux}} = \sum_{m \in \{T,V,A\}} \mathcal{L}_m^{\text{aux}}(y, \hat{y}_m^{\text{fast}}), \tag{16}$$

where $\hat{y}_m^{\text{fast}}$ denotes the prediction obtained from the Fast Pathway for modality m.

Rather than keeping the auxiliary supervision strength fixed throughout training, we gradually reduce its contribution as training proceeds:

$$\lambda_{\text{aux}}(t) = \max\left(0.1, 1 - \frac{t}{T_{\max}}\right), \tag{17}$$

where t denotes the current training epoch and $T_{\max}$ is the maximum number of training epochs.

Datasets	Train	Valid	Test
MOSI	1284	229	686
MOSEI	16326	1871	4659
SIMS	1368	456	457

Table 1: Dataset statistics.

4 Experiments

4.1 Datasets

We evaluate FRSR on three widely used Multimodal Sentiment Analysis benchmarks: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and SIMS (Yu et al., 2020). These datasets cover both English and Chinese scenarios, allowing us to assess the generalization ability of the proposed framework across different languages and multimodal environments.

Detailed statistics of the training, validation, and testing splits are reported in Table 1.

4.2 Experimental Setup

Evaluation Metrics. For MOSI and MOSEI, we report binary accuracy (Acc-2), seven-class accuracy (Acc-7), F1 score, mean absolute error (MAE), and Pearson correlation (Corr). Following common practice, Acc-2 and F1 are reported under both the negative/non-negative and negative/positive settings. For SIMS, we report Acc-2, F1, MAE, and Corr. Except for MAE, higher values indicate better performance.

Implementation Details. All experiments are implemented in PyTorch and conducted on an RTX A40 GPU with 60GB memory. For each dataset,

Model	MOSI					MOSEI				
	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑
MISA	80.79/82.10	80.77/82.03	-	0.804	0.764	82.59/84.23	82.67/83.97	-	0.548	0.724
Self-MM	82.54/84.77	82.68/84.90	45.79	0.712	0.795	82.68/84.96	82.95/84.93	53.46	0.529	0.767
MMIM	84.14/86.06	84.00/85.98	46.65	0.700	0.800	82.24/85.97	82.66/85.94	54.24	0.526	0.772
HyCon	-/85.2	-/85.1	46.60	0.713	0.790	-/85.4	-/85.6	52.80	0.601	0.776
ConKI	84.37/86.13	84.33/86.13	48.43	0.681	0.816	82.73/86.25	83.08/86.15	54.25	0.529	0.782
EMT	83.30/85.00	83.20/85.00	47.40	0.705	0.798	83.40/86.00	83.70/86.00	54.50	0.527	0.774
GLoMo	84.10/86.70	83.90/86.60	**48.30**	0.718	0.782	83.70/86.50	84.00/86.40	**55.00**	0.539	0.771
DLF	- / 85.06	- / 85.04	47.08	0.731	0.781	- /85.42	- / 85.27	53.90	0.536	0.764
DashFusion	84.26/85.82	84.17/85.78	45.63	0.709	0.796	82.27/86.30	82.70/86.24	53.12	0.524	0.784
MFMB	82.70/85.70	83.20/86.00	45.80	0.709	0.798	84.70/85.10	**85.00**/85.10	54.20	0.532	0.758
DEVA	84.40/86.29	84.48/86.30	46.32	0.730	0.787	83.26/86.13	82.93/86.21	52.26	0.541	0.769
FRSR	**85.07/87.20**	**85.22/87.15**	47.63	**0.661**	**0.819**	**84.75/86.73**	84.35/**86.65**	54.85	**0.510**	**0.786**

Table 2: Results on MOSI and MOSEI. For Acc-2 and F1, results are reported as "negative/non-negative" and "negative/positive". Bold indicates the best performance.

Model	Acc-2↑	F1↑	MAE↓	Corr↑
MulT	75.62	75.84	0.485	0.504
MISA	75.49	75.85	0.472	0.542
Self-MM	77.37	77.54	0.458	0.535
ConKI	77.94	78.17	0.454	0.542
EMT	80.10	80.10	**0.396**	0.623
DashFusion	79.21	79.39	0.416	0.601
DEVA	79.64	80.32	0.424	0.583
FRSR	**81.71**	**81.54**	0.408	**0.626**

Table 3: Comparison with baselines on SIMS.

we repeat training three times using random seeds 1111, 2222, and 3333, and report the average performance. The Adam optimizer is used for training. The learning rate is set to 1×10^{-5} for the BERT encoder and 1×10^{-4} for the remaining parameters. The batch size is 32 and the total number of training epochs is 100.

4.3 Baselines

We compare FRSR with representative multimodal sentiment analysis methods, including MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), TFR-Net (Yuan et al., 2021), HyCon (Mai et al., 2022), ConKI (Yu et al., 2023), CENET (Wang et al., 2022), TETFN (Wang et al., 2023), GLoMo (Zhuang et al., 2024), EMT (Sun et al., 2023), LNLN (Zhang et al., 2024), DLF (Wang et al., 2025), DashFusion (Wen et al., 2025), MFMB (Tao et al., 2025), and DEVA (Wu et al., 2025).

4.4 Quantitative Results and Analysis

We compare FRSR with representative baseline methods on three benchmark datasets for multimodal sentiment analysis. As shown in Tables 2 and 3, FRSR achieves consistently strong performance across different evaluation settings.

Table 2 summarizes the results on MOSI and MOSEI. Compared with recent competitive methods, FRSR achieves consistent improvements on both datasets. Relative to DEVA (Wu et al., 2025), FRSR improves the average binary classification metrics (Acc-2 and F1) by approximately 0.8% on MOSI and 1.0% on MOSEI. FRSR also performs better on fine-grained sentiment metrics. Specifically, MAE is reduced by 9.4% on MOSI and 5.7% on MOSEI, while Corr is consistently improved. These results indicate that FRSR is effective not only for sentiment polarity prediction, but also for modeling subtle emotional intensity variations.

The improvements mainly come from the collaboration between the Fast and Slow Pathways. The Fast Pathway identifies sentiment-relevant cues while filtering out redundant or noisy information. The Slow Pathway then integrates these sparse but informative cues through Reasoning Tokens. By restricting cross-modal interaction to a compact set of salient tokens, FRSR reduces the influence of irrelevant context and achieves better alignment with ground-truth sentiment labels.

Table 3 presents the results on SIMS. Compared with MOSI and MOSEI, SIMS contains more realistic Chinese videos and exhibits stronger multimodal inconsistency, making it a more challenging benchmark.

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
FRSR	**85.07/87.20**	**85.22/87.15**	**47.63**	**0.661**	**0.819**	**84.75/86.73**	**84.35/86.65**	**54.85**	**0.510**	**0.786**
w/o FSR	84.15/86.15	84.26/86.31	45.20	0.695	0.811	83.50/86.15	83.45/86.20	53.10	0.532	0.772
w/o SRF	84.45/86.40	84.50/86.35	46.10	0.680	0.813	84.13/86.30	83.75/86.35	54.35	0.525	0.780

Table 4: Ablation experiments on MOSI and MOSEI datasets.

Testing Condition	MOSI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
{T}	83.53 / 85.12	83.64 / 84.97	50.14	45.42	0.693	0.793
{A}	57.48 / 59.23	73.08 / 72.46	22.53	21.78	1.254	0.158
{V}	57.29 / 58.82	72.31 / 71.88	21.81	21.19	1.282	0.141
{T, A}	84.26 / 86.14	84.37 / 86.09	51.08	46.47	0.677	0.806
{T, V}	84.08 / 85.91	84.23 / 85.82	50.82	46.29	0.684	0.798
{A, V}	58.89 / 60.46	74.02 / 74.88	23.44	22.71	1.218	0.187
{T, A, V}	**85.07 / 87.20**	**85.22 / 87.15**	**52.15**	**47.63**	**0.661**	**0.819**
Avg.	72.94 / 74.70	79.55 / 80.46	38.85	35.93	0.924	0.529

Table 5: Robustness evaluation results of FRSR under inter-modality missingness on the MOSI dataset. $\{T\}$, $\{A\}$, and $\{V\}$ denote language, audio, and visual modalities, respectively.

Method	SIMS			
	Acc-2	F1	MAE	Corr
FRSR	**81.71**	**81.54**	**0.408**	**0.626**
w/o FSR	80.20	80.15	0.422	0.592
w/o SRF	80.55	80.65	0.416	0.605

Table 6: Ablation experiments on the SIMS dataset.

4.5 Ablation Study

To evaluate the contribution of each component in the proposed Fast–Slow framework, we conduct ablation studies on MOSI, MOSEI (Table 4), and SIMS (Table 6).

Impact of Fast Saliency Retrieval (FSR): Removing the Fast Pathway (w/o FSR) forces the model to process the entire multimodal sequence without salience filtering, resulting in clear performance degradation across all datasets. The effect is especially noticeable on SIMS, where F1 drops from 81.54 to 80.15. These results show that the Fast Pathway is important for suppressing redundant information and identifying informative emotional cues before multimodal reasoning.

Impact of Slow Reasoning Fusion (SRF): Replacing the Slow Pathway with simple feature concatenation over the retrieved tokens (w/o SRF) also degrades performance. While binary metrics such as F1 and Acc-2 decrease moderately, fine-grained metrics including Acc-7, MAE, and Corr are affected more significantly, especially on MOSI and MOSEI. This suggests that simple fusion is insufficient for modeling subtle sentiment variations, while the Slow Pathway is important for capturing deeper cross-modal semantic relationships.

4.6 Robustness under Modality Missingness

To evaluate modality-specific contributions and model robustness, we test FRSR under different modality missing settings on the MOSI dataset, as shown in Table 5.

First, the language modality ($\{T\}$) serves as the main semantic source, and any setting containing text performs much better than those without it. For example, using only $\{T\}$ achieves an Acc-7 of 45.42, while using $\{A, V\}$ without text only reaches 22.71. Second, although performance drops significantly when text is unavailable, FRSR still maintains relatively strong robustness under single-modality settings such as $\{A\}$ or $\{V\}$. This is mainly because the Fast Pathway filters out noisy or irrelevant information before reasoning. Finally, performance gradually improves from $\{T\}$ to $\{T, A\}$/$\{T, V\}$ and further to $\{T, A, V\}$, showing that the Slow Pathway can effectively leverage

Top-K	Acc-2	F1	Acc-7	MAE	Corr
1	84.50/86.58	84.55/86.60	45.85	0.685	0.802
2	84.95/87.05	85.05/87.02	**48.05**	0.665	0.815
4	85.07/**87.20**	**85.22/87.15**	47.63	**0.661**	**0.819**
8	**85.12**/87.18	85.10/87.05	47.15	0.672	0.815
16	84.75/86.80	84.85/86.85	46.55	0.680	0.808
32	84.35/86.45	84.45/86.40	45.80	0.688	0.801
S	84.05/86.15	84.11/86.12	45.20	0.695	0.795

Table 7: Sensitivity analysis of Top-K on MOSI.

complementary cross-modal cues to refine sentiment prediction.

4.7 Sensitivity Analysis of Top-K Evidence

We study the effect of the number of retrieved salient tokens on model performance using the MOSI dataset. Specifically, we evaluate $K \in \{1, 2, 4, 8, 16, 32\}$ and compare these settings with a full-sequence baseline S ($S = 50$), where all tokens are retained.

As shown in Table 7, performance first improves and then declines as K increases. When K is too small, the retained evidence is insufficient for complete sentiment understanding. The best results are achieved at $K = 4$, indicating that a compact set of salient cues is sufficient for effective multimodal reasoning.

Compared with the full-sequence baseline, the Top-4 setting consistently performs better, suggesting that the Fast Pathway effectively filters out redundant or noisy information before multimodal interaction. However, excessively large K values gradually reduce performance, likely because additional low-salience tokens reintroduce irrelevant contextual information.

Based on these results, we use a fixed setting of $K = 4$ for all experiments. Although dataset-specific tuning may yield further gains, a unified sparse setting better demonstrates the robustness of the proposed framework.

4.8 Complexity and Efficiency Analysis

To evaluate the computational cost, we compared FRSR with the recent strong baseline LNLN under identical hardware conditions.

As shown in Table 8, FRSR trains approximately 2.18× faster (7.856s vs. 17.143s per epoch) with fewer parameters (111.302M vs. 115.965M). This striking efficiency fundamentally stems from our decoupled dual-pathway design. Instead of performing computationally heavy, opaque holistic fusion over the entire sequence length, System 1 acts as a lightweight heuristic filter to discard redundant background noise. This allows the computationally intensive System 2 (the deep cross-modal Transformer) to operate exclusively on highly sparse Top-K tokens.

Method	Params (M)	Time/Epoch (s)
FRSR	111.302	7.856
LNLN	115.965	17.143

Table 8: Computational overhead comparison.

This demonstrates that the proposed retrieve-then-reason paradigm not only improves structural interpretability, but also achieves competitive performance with substantially better scalability and efficiency.

5 Conclusion

In this work, we introduce FRSR, an interpretable dual-stream architecture engineered for Multimodal Sentiment Analysis. By bifurcating the computational pipeline into heuristic evidence extraction and focused semantic analysis, FRSR effectively circumvents the chronic shortcomings of monolithic fusion models—namely, their vulnerability to environmental clutter and their structural opacity. Specifically, the rapid extraction module isolates a sparse subset of highly discriminative affective triggers from dense input streams, allowing the analytical module to execute profound inter-modal deductions exclusively on this distilled evidence. Comprehensive evaluations across the MOSI, MOSEI, and SIMS benchmarks confirm that FRSR delivers highly competitive predictive accuracy in both categorical and continuous sentiment tasks. Furthermore, by confining resource-intensive cross-modal interactions to a condensed token subset, the proposed methodology not only slashes computational overhead but also demonstrates remarkable resilience against irrelevant modality noise.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3604704, and in part by the National Natural Science Foundation of China under Grant 62176084 and Grant 62176083. The computation is completed on the HPC Platform of Hefei University of Technology.

Limitations

Although FRSR improves explainable multimodal reasoning, several limitations remain.

First, FRSR uses a fixed Top-K retrieval strategy. While efficient, a static sparsity level may not adapt well to inputs with different information densities. Moreover, if the Fast Pathway misses subtle but important cues, the Slow Pathway may struggle to recover the missing information from the retrieved subset alone. Although auxiliary supervision improves salience estimation, cascading errors cannot be fully avoided. Future work may explore adaptive retrieval or feedback mechanisms for dynamic evidence selection.

Second, although Reasoning Tokens improve interpretability by exposing cross-modal interaction patterns, their semantic meaning remains implicit. These tokens are still latent representations that are difficult to align with human-interpretable concepts such as sarcasm or negation. Future work could incorporate symbolic constraints or rule-based reasoning to provide more transparent explanations.

References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

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.

Andrius Dzedzickis, Gediminas Vaičiūnas, Karolina Lapkauskaitė, Darius Viržonis, and Vytautas Bučinskas. 2025. Recent advances in human–robot interaction: robophobia or synergy. *Journal of Intelligent Manufacturing*, 36(4):2281–2307.

Wei Han, Hui Chen, and Soujanya Poria. 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*, pages 9180–9192.

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.

Jian Huang, Yanli Ji, Zhen Qin, Yang Yang, and Heng Tao Shen. 2023. Dominant single-modal supplementary fusion (simsuf) for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:8383–8394.

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2025. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.

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.

Arpan Phukan, Santanu Pal, and Asif Ekbal. 2024. Hybrid quantum-classical neural network for multimodal multitask sarcasm, emotion, and sentiment analysis. *IEEE Transactions on Computational Social Systems*, 11(5):5740–5750.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the association for computational linguistics: ACL 2023*, pages 12966–12978.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Upendra Singh, Kumar Abhishek, and Hiteshwar Kumar Azad. 2024. A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9):1–38.

Keith E. Stanovich and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23:645–726.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325.

Chuanqi Tao, Jiaming Li, Tianzi Zang, and Peng Gao. 2025. A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1547–1555.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. 2023. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822.

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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

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.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.

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.

Yuhua Wen, Qifei Li, Yingying Zhou, Yingming Gao, Zhengqi Wen, Jianhua Tao, and Ya Li. 2025. Dashfusion: Dual-stream alignment with hierarchical bottleneck fusion for multimodal sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*.

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, Haopeng Kuang, Shuai Huang, and Lihua Zhang. 2022. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1708–1717. Association for Computing Machinery.

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.

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.

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.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. Conki: Contrastive knowledge injection for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624.

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.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

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.

Jiandian Zeng, Jiantao Zhou, and Caishi Huang. 2023. Exploring semantic relations for social media sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2382–2394.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023a. 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.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023b. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR.

Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. 2022. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437.

Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. 2020. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 62:14–31.

Aoqiang Zhu, Min Hu, Xiaohua Wang, and Fuji Ren. 2026. Beneficial noise learning for robust multimodal fusion. *IEEE Transactions on Multimedia*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Yan Xing, Yiming Tang, Jiaoyun Yang, Ning An, and Fuji Ren. 2025a. EaNet: Enhanced multimodal awareness alignment network for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025b. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. 2024. KEBR: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5732–5741.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 2024. Glomo: Global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1800–1809.