electronics

MDPI

Article

PGMCR: Polarity-Guided Multimodal Consistency Regularization with Reliability Fusion for Multimodal Sentiment Analysis

Hua Yang [1], Junying Guo [1,*], Ming Zhao [1], Mingzhi Mu [1], Yuanyuan Li [1], Baozhou Chen [1,*] and Min Zhou [2]

[1] School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430040, China; huayang@whpu.edu.cn (H.Y.); zhaoming202307@163.com (M.Z.); xiaomumu1203@gmail.com (M.M.); likyuan888@163.com (Y.L.)

[2] School of Food Science and Engineering, Wuhan Polytechnic University, Wuhan 430040, China; mzhou268@163.com

* Correspondence: guojunying02@163.com (J.G.); bzchen@whpu.edu.cn (B.C.)

Abstract

Multimodal sentiment analysis must combine textual, visual, and acoustic evidence whose reliability varies across utterances. We propose PGMCR, which uses a predicted three-way polarity distribution to compose sample-level modality-reliability weights and applies a training-only consistency constraint to text-preserving visual and acoustic perturbations. Experiments on CMU-MOSI and CMU-MOSEI show that polarity-guided fusion is the principal architectural contribution, while the text-guided refinement and consistency branch have metric-dependent effects. The learned weights vary across polarity groups, but proxy-corruption diagnostics also reveal strong dependence on textual evidence and non-negligible polarity-estimation errors. The results support polarity as a useful conditioning signal for multimodal fusion while delimiting the scope of the nonverbal robustness claim.

Keywords: multimodal sentiment analysis; adaptive consistency regularization; modality reliability; multimodal fusion; polarity-guided fusion; CBAM-inspired vector refinement

1. Introduction

Multimodal sentiment analysis (MSA) estimates sentiment from heterogeneous textual, visual, and acoustic evidence. In video-based opinion data, these modalities offer complementary information with unequal predictive utility. Text often conveys explicit semantic content, whereas facial behavior and vocal prosody can provide additional affective cues when the verbal content is ambiguous, sarcastic, or only weakly emotional. This complementarity has motivated extensive work on multimodal fusion for benchmarks such as CMU-MOSI and CMU-MOSEI [1,2].

A central difficulty is that the contribution of each modality varies across utterances. Visual evidence can be informative when affect is expressed through facial behavior, but it may be weak or unreliable under occlusion or limited expression. Acoustic cues can reflect prosody and emotional intensity, yet may be degraded by low-energy speech or environmental noise. Text is often highly informative, but remains susceptible to speech-recognition errors, informal language, and implicit sentiment. Consequently, fixed modality weighting or generic attention may not adequately represent the conditions under which each modality should influence the final prediction.

Check for updates

Academic Editor: Ioannis Hatzilygeroudis

Received: 5 August 2026
Revised: 1 October 2026
Accepted: 7 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Existing MSA methods address this problem through interaction modeling, cross-modal attention, representation disentanglement, sentiment-aware learning, and dynamic fusion [3–11]. However, predicted polarity is less commonly used as an explicit conditioning variable for sample-level modality weighting.

This paper presents PGMCR, a multimodal sentiment model that combines text-guided nonverbal feature refinement, polarity-conditioned fusion, and a training-only consistency regularizer. The specific design investigated here uses a predicted three-way polarity distribution to mix learnable modality-weight prototypes, together with a text-preserving visual/acoustic consistency objective whose activation is controlled by a validation-only gate. TGCBAM adapts CBAM-inspired channel gating to pooled nonverbal vectors under textual guidance; MDC-KD(VA) is an auxiliary training constraint and does not introduce a missing-modality inference pathway. These components are evaluated as a combination, and their individual effects are interpreted through controlled removals.

In addition to polarity-guided fusion, PGMCR introduces Adaptive Dynamic MDC-KD(VA), a training-only consistency regularizer over text-preserving visual and acoustic perturbation views. The full-modality prediction is used as a detached teacher, while visual-dropped, audio-dropped, and visual-audio-dropped views are treated as student views. To avoid manually selecting a dataset-specific warmup epoch, the consistency branch is activated by a validation-only adaptive gate. This gate uses validation loss normalized by validation-label scale. The regularizer is therefore designed as an auxiliary optimization constraint.

We evaluate PGMCR on CMU-MOSI and CMU-MOSEI using removal-based ablations, comparisons with published reference results, polarity-conditioned interpretability analysis, and fixed-severity proxy-corruption diagnostics. The observed ablation values suggest that Polarity-Guided Fusion is the most consequential component in the evaluated configurations, whereas TGCBAM and Adaptive Dynamic MDC-KD(VA) have metric- and dataset-dependent effects. The interpretability analysis shows structured variation in the learned weights across polarity groups.

The main contributions are summarized as follows:

- We investigate a polarity-conditioned fusion parameterization in which predicted three-way polarity probabilities mix learnable prototypes to produce relative text, visual, and acoustic weights. The contribution is this explicit polarity-to-weight mapping, not dynamic weighting in general.
- We evaluate Adaptive Dynamic MDC-KD(VA) as a training-only consistency regularizer over text-preserving visual/acoustic perturbation views, with activation determined by a validation-only gate; it is not a test-time missing-modality method.
- We evaluate a text-guided, CBAM-inspired refinement of pooled visual and acoustic vectors. This is an adaptation of established gating ideas rather than a claim to a new attention family; removal comparisons quantify its observed, metric-dependent contribution.
- We evaluate the shared backbone, Polarity-Guided Fusion, and final PGMCR configuration on CMU-MOSI and CMU-MOSEI using removal-based ablations, polarity-conditioned weight analysis, qualitative case studies, and proxy-corruption diagnostics.

2. Related Work

2.1. Multimodal Sentiment Analysis

Existing MSA studies can be grouped into benchmark construction, multimodal feature extraction, fusion mechanism design, sentiment-aware modeling, and robustness-oriented learning. CMU-MOSI and CMU-MOSEI are widely used English benchmarks

with continuous sentiment annotations, where regression metrics and discretized or binarized classification metrics are commonly reported. Common nonverbal feature pipelines include facial behavior analysis and acoustic feature extraction tools such as OpenFace and openSMILE [12,13]. Lexical and commonsense sentiment resources such as SenticNet provide another line of sentiment-oriented representation support [14].

A recurring challenge in MSA is that different modalities have different noise patterns, sampling rates, semantic densities, and temporal structures. Text features often dominate clean benchmark performance because language carries explicit semantic content, while visual and acoustic features provide affective cues with varying reliability. This imbalance creates a central fusion problem: the model must exploit complementary nonverbal information without allowing noisy modalities to degrade the fused representation. PGMCR follows this line of work but focuses specifically on using predicted sentiment polarity as an explicit reliability cue.

Survey work has established the main MSA axes as multimodal representation, temporal or cross-modal interaction, fusion, and robustness. Early audio-visual-text systems demonstrated the value of joint multimodal learning, while later contextual and conversational models made the role of linguistic context more explicit. Together, these lines of work motivate a reliability-aware fusion mechanism, but do not by themselves specify how sentiment state should parameterize per-utterance modality weights. The relevant surveys and representative early studies are cited here [15–25].

2.2. Multimodal Fusion Methods

Early neural fusion approaches include Tensor Fusion Network (TFN), which models unimodal, bimodal, and trimodal interactions through tensor products [3], and Low-rank Multimodal Fusion (LMF), which reduces the computational cost of tensor fusion with low-rank factorization [4]. These approaches provide expressive multimodal interaction modeling but can be computationally expensive or sensitive to feature dimensionality.

Attention-based and Transformer-based methods further improve cross-modal interaction modeling. For example, MulT employs directional cross-modal attention to model interactions across unaligned multimodal sequences without requiring explicit alignment [5]. MAG-BERT incorporates multimodal adaptation gates into a BERT-based architecture, allowing visual and acoustic information to modulate language representations [26]. These methods demonstrate the benefit of text-aware or cross-modal attention, but the learned interaction weights are not explicitly conditioned on sentiment polarity as a reliability variable.

Representation-learning approaches tackle multimodal fusion from a different perspective. MISA projects each modality into modality-invariant and modality-specific subspaces to reduce modality gaps while preserving modality-specific characteristics [6]. Self-MM generates self-supervised unimodal labels to improve modality-specific representation learning [7]. These methods improve the quality of learned representations, but they do not directly formulate polarity-conditioned modality reliability prototypes. PGMCR differs by using predicted polarity probabilities to compose modality weights in an explicit and interpretable fusion mechanism.

Recent MSA research can be grouped by the source of its improvement. The first line models interaction and temporal structure: MFN and MARN use multimodal memory or recurrent attention, MCTN learns cross-modal agreement through cyclic translation, and TCSP and MTAG encode shared-private or modal-temporal dependencies [24,27–30]. Together, these studies show that when and how modalities interact matters, but their attention or latent interactions do not explicitly quantify sample-level modality reliability.

The second line reshapes the training signal. HMIM preserves task-relevant information hierarchically, CLMLF combines contrastive learning with multi-layer fusion, UniMSE transfers supervision between sentiment and emotion recognition, and ConFEDE separates useful contrastive factors from modality-specific variation [31–34]. Their common implication is that fusion quality depends on the information retained by the objective; modality reliability nevertheless remains implicit rather than an interpretable, polarity-conditioned quantity.

The third line makes modality contribution adaptive through language or affective knowledge. LAHR uses language-guided hyper-modality representations, MMLF coordinates modality-specific and fused predictions with multiple losses, and KuDA uses sentiment knowledge to emphasize a dominant modality [8,35,36]. These approaches support adaptive rather than fixed contribution, but differ from PGMCR's continuous mixing of learnable modality prototypes by a predicted three-way polarity distribution while retaining all modalities at inference.

The fourth line targets incomplete or degraded input: TFRNet reconstructs corrupted features, MPLM uses prompts for missing modalities, ConKI and sentiment-knowledge-enhanced self-supervision inject affective knowledge, and TF-Mamba develops a text-enhanced missing-modality architecture [37–41]. This robustness objective is broader than PGMCR's bounded setting, which preserves the clean full-modality inference path and uses visual/acoustic masked views only as a training-time consistency regularizer.

Overall, the literature progresses from interaction modeling, to objective-level representation learning, adaptive knowledge-guided fusion, and missing-modality robustness. PGMCR addresses the narrower question of whether predicted polarity can provide an explicit sample-level reliability condition in full-modality fusion.

2.3. Sentiment-Aware and Dynamic Modality Modeling

Recent MSA work has increasingly explored sentiment-aware learning strategies, dynamic modality weighting, and multi-objective optimization. KuDA explicitly uses sentiment knowledge to select the dominant modality, CLGSI incorporates sentiment-intensity guidance into contrastive learning, and MMLF investigates multi-loss training for multimodal sentiment learning [8,36,42]. These studies are important because they show that affective knowledge and modality contribution should not be treated as fixed or uniform.

PGMCR falls within the broader family of sentiment-aware and dynamically weighted fusion methods. KuDA emphasizes knowledge-guided dominant-modality selection, CLGSI uses sentiment-intensity guidance in contrastive learning, and MMLF studies multi-loss optimization. PGMCR instead uses a predicted three-way polarity distribution within the fusion layer to mix learnable modality-weight prototypes while retaining all modalities in the inference path. Thus, the distinction claimed here is a particular explicit parameterization and its evaluation. Since a polarity-conditioned prototype mixture can also be viewed as a form of dynamic weighting, we make no claim that this operation is categorically separate from dynamic fusion methods.

This distinction also bounds the claim made in this paper. PGMCR does not decompose the final task into separate intensity and polarity streams, nor does it compute modality weights from unimodal-versus-multimodal score differences. Instead, it predicts a three-way polarity distribution and uses this distribution to compose learnable reliability prototypes for text, visual, and acoustic modalities. This design keeps the reliability mechanism simple, directly tied to predicted polarity, and available for interpretation through sample-level modality weights.

Recent modality-consistency and missing-modality methods also use view construction, modality dropout, or auxiliary consistency objectives to improve robustness. Their primary objectives are generally representation completion, cross-view invariance, or recovery under missing inputs. PGMCR differs in two bounded respects: its consistency branch preserves text and perturbs only visual and acoustic views, and its activation is selected by a validation-only normalized-loss gate rather than by a fixed manually chosen warmup.

2.4. Robustness and Consistency Regularization

Robustness has become an important issue in multimodal learning because real-world multimodal inputs may contain corrupted text, occluded visual frames, weak acoustic signals, or missing modalities. Prior work has studied cyclic translation, feature reconstruction, prompt-based missing-modality handling, contrastive knowledge injection, sentiment knowledge-enhanced learning, and text-enhanced state-space fusion [28,37–41]. These methods are valuable, but they often target test-time missing-modality settings or reconstruction-oriented objectives.

PGMCR uses a different regularization setting from Adaptive Dynamic MDC-KD(VA), a regularizer designed for full-modality inference. It preserves the text stream and perturbs only visual and acoustic views, then penalizes inconsistency between the full-modality teacher prediction and degraded-view student predictions. The regularizer is applied only to confident non-neutral samples and only after a validation-driven activation gate is satisfied. Its goal is to improve or preserve generalization while keeping the inference path unchanged.

This distinction also limits the scope of the comparison: PGMCR is not presented as a replacement for reconstruction-based or missing-modality methods. Instead, it regularizes the full-modality model during training and leaves the clean full-modality inference path unchanged.

3. Methodology

Figure 1 provides a roadmap for the proposed PGMCR framework. Solid black arrows indicate the full-modality computation path shared by training and inference, including modality encoding, text-guided CBAM-inspired vector refinement, Polarity-Guided Fusion, and regression prediction. Dashed arrows mark the auxiliary Adaptive Dynamic MDC-KD(VA) consistency branch, which is attached only during training and removed at inference. The following subsections first define the prediction problem and encoders, and then describe CBAM-inspired vector refinement, polarity-guided fusion, MDC-KD regularization, and the final objective.

3.1. Problem Formulation

Given a multimodal utterance, let x_t, x_v, and x_a denote the text, visual, and acoustic inputs, respectively. The goal is to predict a continuous sentiment score $y \in \mathbb{R}$. The model prediction is written as

$$\hat{y} = F_\Theta(x_t, x_v, x_a), \tag{1}$$

where F_Θ denotes the complete PGMCR model with parameters Θ. Following common MSA evaluation practice, the continuous prediction is evaluated by regression metrics and by classification metrics obtained from discretized or binarized sentiment labels.

3.2. Multimodal Encoders

The text stream is encoded by a BERT-based language encoder [43]. Let H_t denote the token-level hidden states. The utterance-level text representation is obtained from the `[CLS]` token

$$H_t = \text{BERT}(x_t), \qquad h_t = H_t^{[\text{CLS}]}. \tag{2}$$

In the implementation, h_t is a 768-dimensional representation.

The visual and acoustic streams are encoded by lightweight Transformer encoders [44]. Here, Transformer is used as a generic sequence encoder; it operates on frame-level visual features and on acoustic feature sequences. For modality $m \in \{v, a\}$, the encoded utterance-level representation is

$$h_m = \text{Pool}(\text{Transformer}_m(x_m)), \qquad m \in \{v, a\}, \tag{3}$$

where $\text{Pool}(\cdot)$ denotes masked mean pooling and excludes padded positions. The visual and acoustic encoders use one Transformer layer, hidden size 64, four attention heads, and output projections to 16-dimensional features in the final experiments.

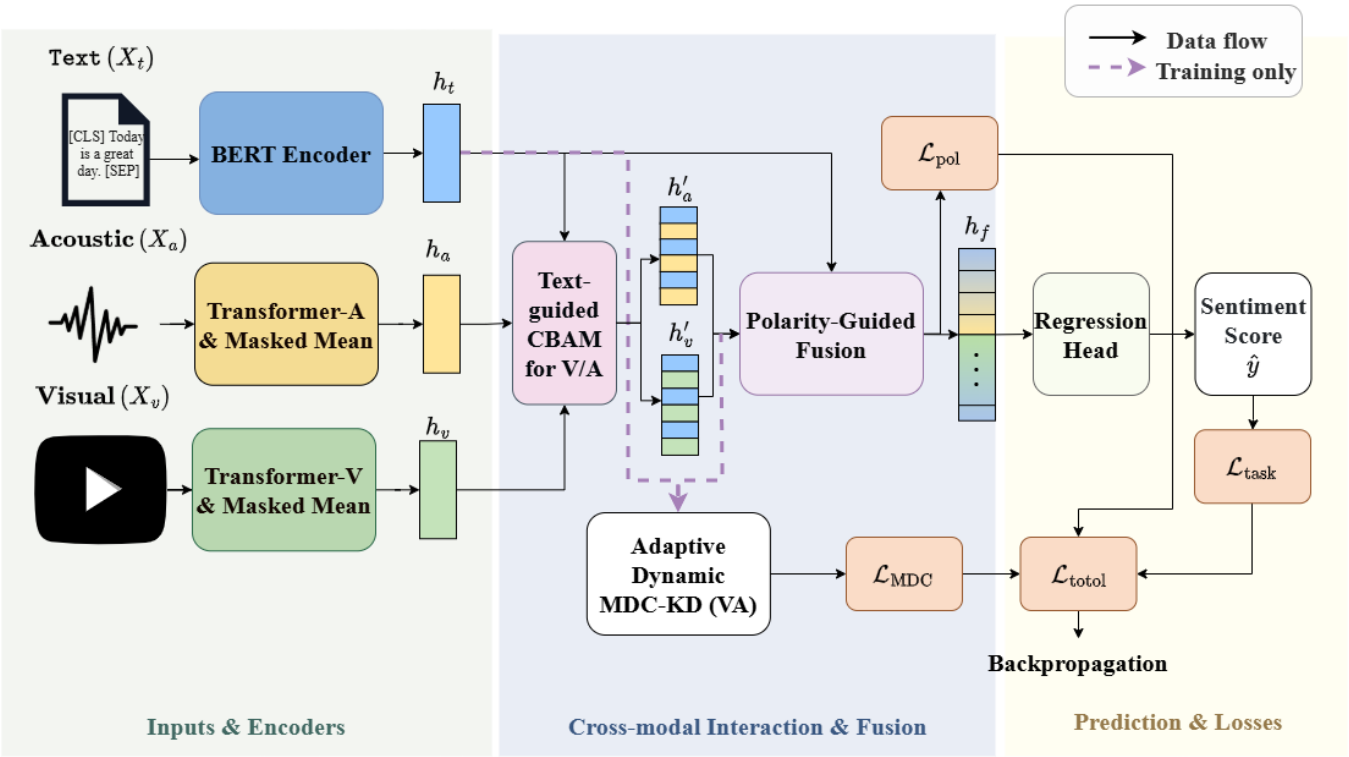


Figure 1. Overall architecture of PGMCR. Solid black arrows denote the full-modality path used in both training and inference; dashed arrows denote the training-only MDC-KD(VA) consistency branch, which is removed at inference.

3.3. Text-Guided CBAM-Inspired Vector Refinement

Before fusion, PGMCR applies TGCBAM, a text-guided CBAM-inspired vector refinement, to the visual and acoustic utterance-level features [45]. The global text representation $h_t \in \mathbb{R}^{d_t}$ ($d_t = 768$) guides each nonverbal representation $h_m \in \mathbb{R}^{d_m}$ ($d_m = 16$ in the final setting), where $m \in \{v, a\}$. The module adapts the sequential channel-gating idea of CBAM to pooled utterance-level vectors. It uses text-conditioned channel-wise gating followed by feature-wise gating. TGCBAM adapts the sequential gating principle of CBAM to utterance-level vectors. Original CBAM combines channel and spatial attention on convolutional feature maps. TGCBAM operates after masked temporal pooling, when each nonverbal stream is represented by one vector; the feature-wise gate therefore reweights vector dimensions, and the additional text signal conditions the channel gate. This description identifies

TGCBAM as a task-specific vector-level adaptation and clarifies the scope of its relationship to CBAM.

$$\begin{aligned} \bar{h}_m &= h_m, \qquad \hat{h}_m = h_m, \\ z_m &= [\bar{h}_m; \hat{h}_m; h_t], \\ g_c^{(m)} &= \sigma\Big(W_{c2}^{(m)}\,\mathrm{ReLU}\Big(W_{c1}^{(m)} z_m\Big)\Big), \\ h_m^{(2)} &= h_m \odot g_c^{(m)}. \end{aligned} \tag{4}$$

Because the input has already been pooled to one vector, the average-statistic branch and the maximum-statistic branch are both equal to h_m; this is a vector adaptation of the two-statistic CBAM input. The text vector is concatenated with these two vector branches, so the channel gate is text-guided. Here, $W_{c1}^{(m)}$ and $W_{c2}^{(m)}$ denote the two linear layers of the channel MLP, and $\sigma(\cdot)$ is the sigmoid function.

The second gate reweights the feature dimensions

$$\begin{aligned} g_f^{(m)} &= \sigma\Big(W_f^{(m)} h_m^{(2)}\Big), \\ h_m^{(3)} &= h_m^{(2)} \odot g_f^{(m)}, \qquad h'_m = \mathrm{LN}\Big(h_m^{(3)} + h_m\Big). \end{aligned} \tag{5}$$

Here, $W_f^{(m)}$ is the feature-wise linear projection, and $\odot$ denotes element-wise multiplication. Thus, both gates act on the 16-dimensional encoded utterance-level nonverbal vector; the second gate is feature-wise. The complete refinement is denoted as

$$h'_m = \mathrm{TGCBAM}_m(h_m, h_t), \qquad m \in \{v, a\}. \tag{6}$$

Figure 2 illustrates how the global text representation guides the CBAM-inspired vector refinement of visual and acoustic features before fusion.

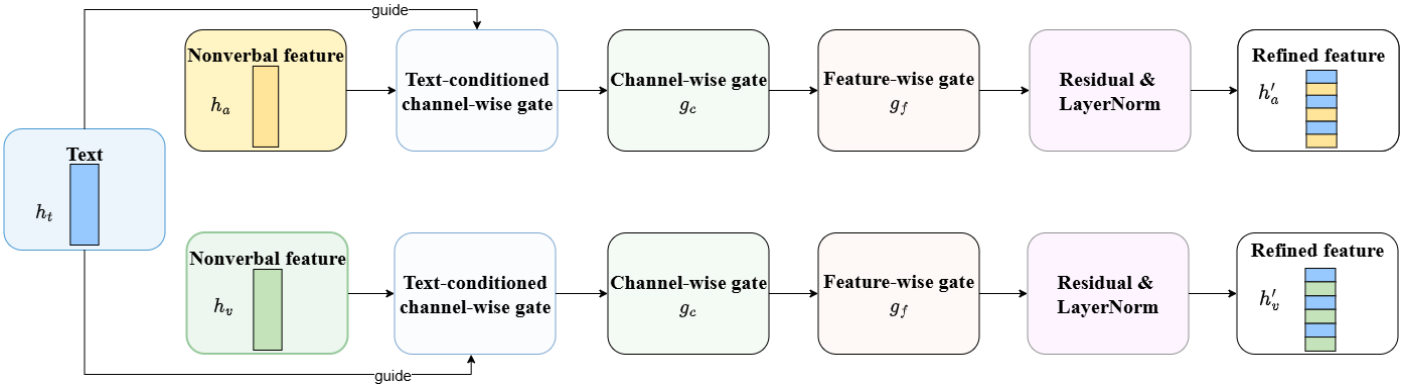


Figure 2. Text-guided CBAM-inspired vector refinement. The global text representation guides vector-level channel-wise and feature-wise gating of visual and acoustic utterance representations before polarity-guided fusion.

3.4. Polarity-Guided Fusion

The core component of PGMCR is Polarity-Guided Fusion. It uses predicted sentiment polarity as a compact conditioning signal for estimating sample-level modality reliability. The model first estimates a three-way polarity distribution from the text representation

$$p = g_p(h_t), \qquad p \in \mathbb{R}^3, \tag{7}$$

where $g_p(\cdot)$ is a multilayer polarity estimator with a final softmax layer, and $p = [p_{\mathrm{pos}}, p_{\mathrm{neg}}, p_{\mathrm{neu}}]$ represents the probabilities of positive, negative, and neutral polarity

states. The final implementation uses the text representation for polarity estimation, which reflects the strong semantic role of the language stream in sentiment polarity prediction.

Let $R \in \mathbb{R}^{3\times 3}$ denote the learnable polarity-specific modality reliability prototypes. Each row corresponds to one polarity state, and each column corresponds to one modality. The sample-level modality weights are computed as

$$w = \mathrm{softmax}(p^\top R), \qquad w = [w_t, w_v, w_a], \tag{8}$$

where w_t, w_v, and w_a are the text, visual, and acoustic reliability weights. The weighted modality features are concatenated and projected into the fusion space: The prototype parameterization has an identifiable degrees-of-freedom constraint that is useful when interpreting R. Adding an arbitrary constant to every entry of any row of R changes $p^\top R$ only by a sample-dependent scalar, which softmax removes. Thus, the nine raw entries contain six identifiable degrees of freedom under this row-shift invariance. Moreover, when p is close to one-hot for polarity state k, w approaches $\mathrm{softmax}(R_{k,:})$; intermediate polarity distributions interpolate among these three prototype weight vectors. The mechanism therefore provides three limiting reliability patterns and continuous interpolation mainly for uncertain polarity estimates.

$$z_f = \left[w_t h_t;\, w_v h'_v;\, w_a h'_a\right], \qquad h_f = \phi_f(z_f), \tag{9}$$

where $\phi_f(\cdot)$ denotes the projection network. The prediction head maps h_f to the final sentiment score

$$\hat{y} = f_\psi(h_f), \tag{10}$$

where f_ψ denotes the regression head.

The polarity supervision is derived from the continuous sentiment label

$$y^{\mathrm{pol}} = \begin{cases} \mathrm{pos}, & y > 0.5, \\ \mathrm{neg}, & y < -0.5, \\ \mathrm{neu}, & \text{otherwise}. \end{cases} \tag{11}$$

The auxiliary polarity loss is

$$\mathcal{L}_{\mathrm{pol}} = \mathrm{CE}(p, y^{\mathrm{pol}}), \tag{12}$$

where $\mathrm{CE}(\cdot)$ denotes cross-entropy loss.

To assess the dependence of the polarity branch on this discretization, we additionally evaluate symmetric thresholds of ± 0.3 and ± 0.7 against the default ± 0.5 setting. The resulting sensitivity analysis will be reported in Section 6.3.

Because the softmax in Equation (8) constrains $w_t + w_v + w_a = 1$, these values represent relative allocation of the available fusion contribution. Consequently, PGF cannot express the state in which all three modalities are simultaneously unreliable. This is a deliberate limitation of the simplex parameterization; the model relies on the downstream prediction loss and the training-time consistency branch to limit the effect of unreliable views.

In this implementation, Polarity-Guided Fusion has two coupled training outputs: the fused sentiment prediction $\hat{y}$ and the polarity distribution p. The task loss $\mathcal{L}_{\mathrm{task}}$ supervises $\hat{y}$, while $\mathcal{L}_{\mathrm{pol}}$ supervises the polarity branch that provides the reliability prior. Figure 3 summarizes the polarity-probability estimation, reliability-prototype composition, and weighted feature fusion steps.

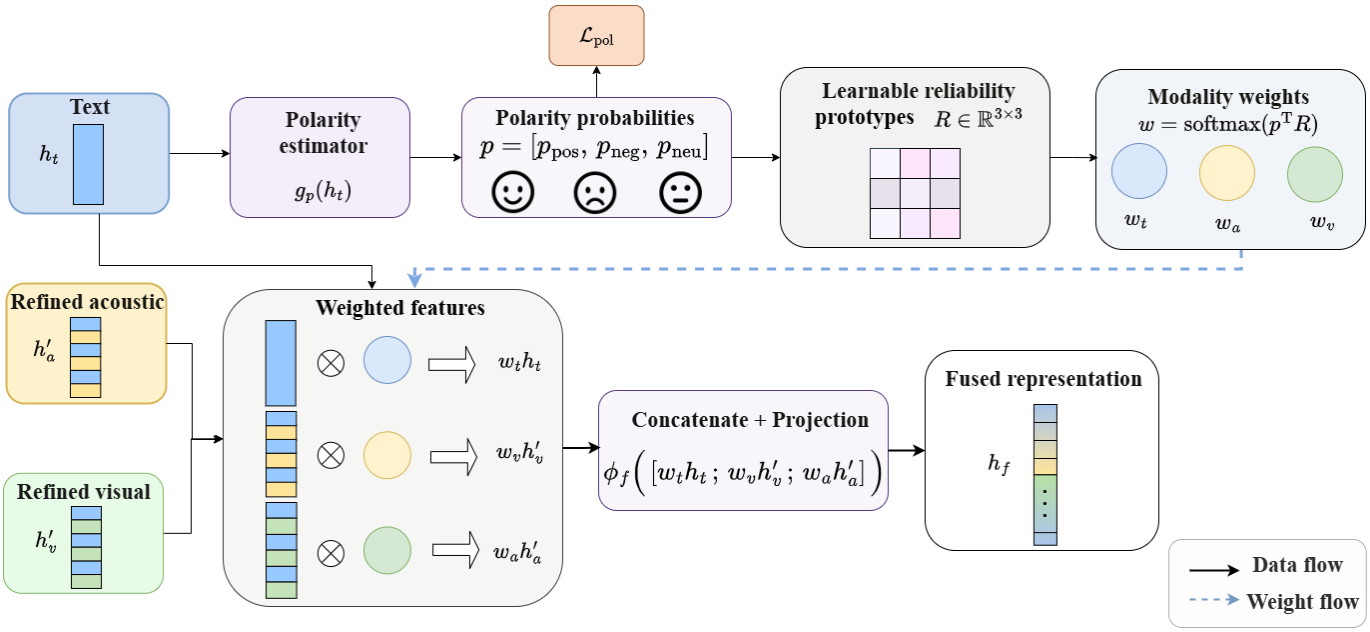


Figure 3. Polarity-Guided Fusion. The model predicts polarity probabilities and uses them to compose modality reliability weights from learnable polarity-specific prototypes.

3.5. Adaptive Dynamic MDC-KD(VA)

Adaptive Dynamic MDC-KD (VA) is a text-preserving consistency regularizer. It is applied after TGCBAM refinement and before Polarity-Guided Fusion: the encoded-and-refined visual and acoustic vectors are selectively set to zero, while the text representation remains intact. For $r \in \mathcal{R} = \{v, a, va\}$, the corresponding student inputs are

$$(\bar{h}_v^{(r)}, \bar{h}_a^{(r)}) = \begin{cases} (\mathbf{0}, h'_a), & r = v, \\ (h'_v, \mathbf{0}), & r = a, \\ (\mathbf{0}, \mathbf{0}), & r = va. \end{cases} \tag{13}$$

The full-modality prediction $\hat{y}_i$ is the detached teacher target, and $\hat{y}_i^{(r)}$ is the prediction for student view r. All views reuse the same polarity estimator, fusion module, and regression head; in particular, their polarity probabilities are computed from the shared, preserved text representation. Figure 4 illustrates these text-preserving visual, acoustic, and joint visual-acoustic student views. Gradients from $\mathcal{L}_{\text{MDC}}$ are stopped on the teacher prediction and propagated through the student paths only. The confident non-neutral mask is computed inside the MDC loss branch and gates which samples contribute to this auxiliary objective.

The consistency loss is applied only to confident non-neutral samples. The sample mask is

$$m_i = \mathbb{1}(|y_i| \geq \gamma_y) \cdot \mathbb{1}(|\hat{y}_i| \geq \gamma_p), \tag{14}$$

where $\gamma_y = 0.5$ is the label-strength threshold and $\gamma_p = 0.4$ is the prediction-confidence threshold. The implementation uses a Smooth L1 consistency loss averaged over the three text-preserving degradation modes

$$\mathcal{L}_{\text{MDC}} = \frac{1}{|\mathcal{R}|(\sum_i m_i + \epsilon)} \sum_{r\in\mathcal{R}} \sum_i m_i \,\text{SmoothL1}\left(\hat{y}_i^{(r)}, \text{sg}(\hat{y}_i)\right), \tag{15}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and ϵ prevents division by zero. The perturbation set $\mathcal{R} = \{v, a, va\}$ corresponds to visual dropout, acoustic dropout, and joint visual-acoustic dropout. This construction has a structural limitation. It does not reconstruct a missing

modality or provide a missing-modality inference path: the perturbations are zero masks applied to encoded visual and acoustic vectors, while the clean full-modality path remains the deployment path. The detached teacher can also transmit its own text-anchored errors to the student views, and the confident non-neutral mask together with the mini-batch probability q limits which samples receive the auxiliary constraint.

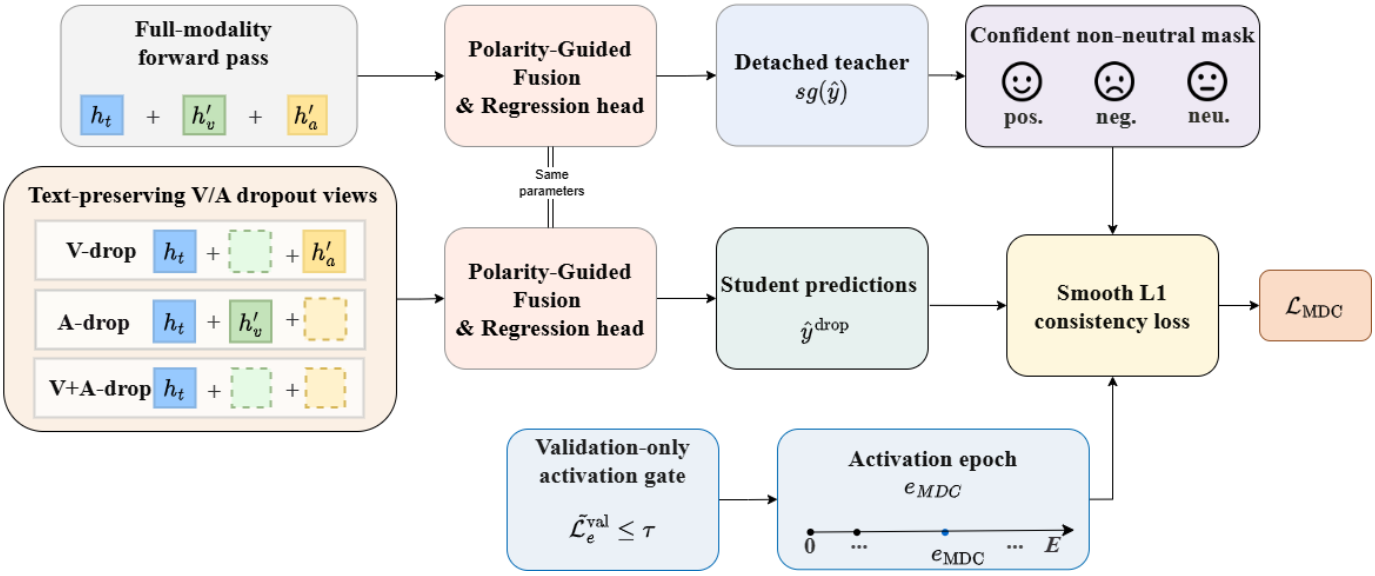


Figure 4. Adaptive Dynamic MDC-KD(VA). During training, the detached full-modality prediction supervises visual-masked, acoustic-masked, and jointly masked student views, where the dashed-outline boxes denote the modality streams that are set to zero in the corresponding view. The ellipsis on the epoch axis denotes the omitted intermediate epochs. The consistency loss is applied only to confident non-neutral samples after validation-driven activation; this auxiliary branch is removed at inference.

3.6. Adaptive Activation

A fixed consistency warmup strategy renders the regularizer vulnerable to dataset size and training dynamics. PGMCR thus adopts an adaptive activation schedule based on validation performance. Let $\mathcal{D}_{\text{val}}$ denote the validation set. The validation-label scale is

$$s_d = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} |y_i|. \tag{16}$$

At epoch e, the validation loss is scaled by the validation-label magnitude

$$\tilde{L}_e^{\text{val}} = L_e^{\text{val}}(1 + s_d), \qquad \tilde{L}_e^{\text{val}} \leq \tau \iff L_e^{\text{val}} \leq \frac{\tau}{1 + s_d}. \tag{17}$$

The activation epoch is selected as

$$e_{\text{MDC}} = \min\left(e_{\max}, \inf\left\{e + 1 \mid e \geq e_{\min},\ \tilde{L}_e^{\text{val}} \leq \tau\right\}\right), \tag{18}$$

where the infimum of an empty set is defined as $+\infty$. Thus, $e_{\max}$ provides a forced latest activation epoch when the validation-loss condition is not met earlier. The quantity $\tilde{L}_e^{\text{val}}$ is a label-scale-adjusted validation loss. The final settings are $e_{\min} = 2$, $e_{\max} = 8$, and $\tau = 1.0$. Here, τ is the target value for the scaled validation loss. In the reported runs, the gate condition is met on CMU-MOSEI at epoch 3, whereas on CMU-MOSI it is not met before the predefined latest-activation safeguard activates the branch at epoch 8. We further examine

the scheduler by varying $e_{\min}$, $e_{\max}$, and τ while keeping the remaining training settings fixed; the results are reported in Section 6.3.

3.7. Training Objective and Inference

Once MDC-KD(VA) is scheduled to be active, the consistency branch is evaluated independently for each mini-batch with probability q. Let $\xi \sim \text{Bernoulli}(q)$ denote this mini-batch indicator. Conditional on $\xi = 1$, the loss in Equation (15) averages the three visual, acoustic, and joint visual-acoustic student views; when $\xi = 0$, no MDC loss is added for that mini-batch. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{pol}}\mathcal{L}_{\text{pol}} + \lambda_{\text{MDC}}\mathbb{1}(e \geq e_{\text{MDC}})\xi\mathcal{L}_{\text{MDC}}, \tag{19}$$

where $\mathcal{L}_{\text{task}}$ is the main sentiment regression loss, $\lambda_{\text{pol}} = 0.1$, $\lambda_{\text{MDC}} = 0.01$, and $q = 0.25$ in the final setting. The three loss objectives play complementary roles. The task loss $\mathcal{L}_{\text{task}}$ optimizes the final sentiment prediction. The auxiliary polarity loss $\mathcal{L}_{\text{pol}}$ provides fine-grained supervision for polarity estimation and modality-reliability learning. The consistency loss $\mathcal{L}_{\text{MDC}}$ acts as a training-only cross-view regularizer that aligns degraded student predictions with the detached full-modality teacher prediction. In particular, the teacher prediction is detached from the gradient graph, while the polarity-guided fusion module and the regression head remain shared across the teacher and student paths.

4. Experiments

4.1. Datasets and Metrics

We evaluate PGMCR on CMU-MOSI and CMU-MOSEI. CMU-MOSI contains 1284 training, 229 validation, and 686 test utterances. CMU-MOSEI contains 16,326 training, 1871 validation, and 4659 test utterances. We report MAE, Pearson correlation (Corr), seven-class accuracy (Acc-7), binary accuracy and F1 on nonzero samples (Acc-2 Non0 and F1 Non0), and binary accuracy and F1 on all samples (Acc-2 All and F1 All). Lower MAE is better, and higher values are better for the other metrics.

4.2. Implementation Details

The text stream uses a pretrained BERT encoder, loaded locally through the AutoModel interface, and takes the `[CLS]` representation as the utterance-level text feature. This choice follows the broader shift from static word embeddings to pretrained contextual language representations in NLP and multimodal language analysis [43,46]. The visual and acoustic streams use lightweight one-layer Transformer encoders with hidden size 64, four attention heads, masked mean pooling, and 16-dimensional output projections. A capacity sensitivity comparison with a two-layer, hidden-size-256 alternative is included in Section 6.3. The experiments consume the pre-extracted non-aligned visual and acoustic features supplied in the MOSI/MOSEI pickle files. Models are trained for up to 40 epochs with patience 10, batch size 32, main learning rate 8×10^{-4}, BERT learning rate 3×10^{-5}, and main weight decay 2×10^{-4}. Visual and acoustic dropout are both 0.1. The polarity auxiliary weight is $\lambda_{\text{pol}} = 0.1$, the MDC consistency weight is $\lambda_{\text{MDC}} = 0.01$, and the mini-batch MDC application probability is $q = 0.25$. Under the validation-only adaptive gate, MDC-KD(VA) is activated at epoch 8 on CMU-MOSI and epoch 3 on CMU-MOSEI.

4.3. Comparison Protocol

We include two types of comparison. First, we conduct controlled comparisons among the local backbone, Polarity-Guided Fusion, and final PGMCR under the same training and reporting protocol. Second, we provide a reported-reference benchmark table for methods whose results are taken from their original papers or source summaries. The latter

is a contextual comparison under commonly used MOSI/MOSEI benchmark conventions and is not a controlled re-evaluation under one identical implementation, preprocessing pipeline, or checkpoint-selection procedure.

Table 1 summarizes the dataset splits, evaluation metrics, training configuration, and final PGMCR settings used in the controlled experiments.

Table 1. Experimental settings and hyperparameters on CMU-MOSI and CMU-MOSEI.

Setting	CMU-MOSI	CMU-MOSEI
Train/validation/test	1284/229/686	16,326/1871/4659
Hidden size in visual/acoustic encoders	64	64
Number of Transformer layers	1	1
Number of attention heads	4	4
Visual/acoustic output projection	16	16
Maximum training epochs	40	40
Early stopping patience	10	10
Batch size	32	32
Main learning rate	8×10^{-4}	8×10^{-4}
BERT learning rate	3×10^{-5}	3×10^{-5}
Main weight decay	2×10^{-4}	2×10^{-4}
Visual/acoustic dropout	0.1/0.1	0.1/0.1
Polarity loss weight λ_{pol}	0.1	0.1
MDC loss weight λ_{MDC}	0.01	0.01
MDC application probability q	0.25	0.25
Label-strength threshold γ_y	0.5	0.5
Prediction-confidence threshold γ_p	0.4	0.4
Adaptive activation range	$e_{\min} = 2, e_{\max} = 8, \tau = 1.0$	$e_{\min} = 2, e_{\max} = 8, \tau = 1.0$
Selected MDC activation epoch	8	3

All locally trained or reproduced models, including the local backbone, Polarity-Guided Fusion, PGMCR, and the reproduced Self-MM baseline, use the same benchmark-provided official train/validation/test partitions: 1284/229/686 utterances for CMU-MOSI and 16,326/1871/4659 utterances for CMU-MOSEI. These partitions are held fixed across the local comparisons. The remaining Table 2 entries are reported-reference values taken from prior studies or benchmark summaries; they were not reimplemented in this study, and their exact split, preprocessing, optimization, and checkpoint-selection protocols were not independently verified here. Accordingly, the contextual evidence presented in Table 2 and the numerical differences from the reported reference rows serve only as a reference.

Table 2. Contextual comparison with reported CMU-MOSI and CMU-MOSEI results. Results are taken from their cited sources and were not re-evaluated under a common implementation; Acc-2 and F1 are reported as Non0/All, and dashes indicate unavailable metrics.

Dataset	Method	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0/All ↑	F1 Non0/All ↑
MOSI	TFN [3]	0.901	0.698	34.90	80.80/–	80.70/–
	LMF [4]	0.917	0.695	33.20	82.50/–	82.40/–
	MulT [5]	0.861	0.711	–	84.10/81.50	83.90/80.60
	MISA [6]	0.804	0.764	–	82.10/80.80	82.00/80.80
	MAG-BERT [26]	0.712	0.796	–	86.1/84.2	86.0/84.1
	Self-MM [7]	0.712	0.795	45.80	84.80/82.50	84.90/82.70
	AMML [47]	0.723	0.792	46.30	84.90/–	84.80/–
	TETFN [48]	0.717	0.800	–	86.10/84.05	86.07/83.83
	TSPMG [49]	0.712	0.799	46.33	85.98/–	85.96/–
	Self-MM *	0.6968	0.7939	47.67	86.13/84.26	86.10/84.18
	PGMCR, ours	**0.6862**	**0.8164**	**48.10**	**86.43/85.42**	**86.42/85.38**

Table 2. *Cont.*

Dataset	Method	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0/All ↑	F1 Non0/All ↑
MOSEI	TFN [3]	0.593	0.700	50.20	82.50/–	82.10/–
	LMF [4]	0.623	0.677	48.00	82.00/–	82.10/–
	MulT [5]	0.580	0.703	–	82.50/–	82.30/–
	MISA [6]	0.568	0.724	–	84.20/82.60	84.00/82.70
	MAG-BERT [26]	–	–	–	–/84.7	–/84.5
	Self-MM [7]	**0.529**	0.767	53.50	85.00/82.70	84.90/83.00
	AMML [47]	0.614	**0.776**	52.40	85.30/–	85.20/–
	TETFN [48]	0.551	0.748	–	85.18/84.25	85.27/84.18
	TSPMG [49]	0.539	0.764	53.13	85.50/–	85.45/–
	Self-MM *	0.5465	0.7534	52.87	84.67/82.16	84.54/82.46
	PGMCR, ours	0.5292	0.7702	**53.72**	**85.72/85.15**	**85.49/85.15**

↓: lower values are better; ↑: higher values are better. Bold: the best value in each column (the lowest MAE; the highest value for the other metrics). *: obtained by our own local reproduction under the controlled experimental setting.

5. Results and Discussion

5.1. Reported-Reference Benchmark Comparison

Table 2 provides a contextual comparison with published CMU-MOSI and CMU-MOSEI results. Baseline values are collected from public benchmark tables under their corresponding dataset protocols, and the method column cites the relevant original papers or source reports. We include representative classic fusion, Transformer-based, pretrained-language, self-supervised, and recent adaptive-fusion baselines. Rows marked with * are obtained by our local reproduction under the controlled experimental setting. Acc-2 and F1 are reported as Non0/All: Non0 excludes samples with zero sentiment labels, whereas All includes the complete test set. Values for TFN, LMF, MulT, MISA, and the original Self-MM rows are taken from the complete-modality benchmark table of EMT-DLFR [50]; values for MAG-BERT, AMML, TETFN, and TSPMG are taken from their original papers [26,47–49]. Since the studies do not share a verified common experimental pipeline, boldface only identifies the best reported value in each column (the lowest MAE, and the highest value for the other metrics) and should not be read as a significance test or a matched-condition ranking.

5.2. Controlled Robustness Diagnostics

In addition to clean-test evaluation, we conduct controlled robustness diagnostics using proxy non-random modality corruptions. This is an internally defined, inference-only diagnostic protocol: the trained checkpoint is not updated, and fixed-severity corruption settings are applied to the test inputs before re-evaluation.

The protocol includes seven non-clean modes with fixed severity, covering ASR-like text corruption, visual occlusion, low-confidence visual dropout, audio noise, low-energy acoustic dropout, and combined multimodal degradation. For text corruption, valid BERT tokens are tokens with attention mask equal to one, excluding `[PAD]`, `[CLS]`, and `[SEP]`. *text-asr-drop* replaces 15% of valid tokens with the BERT `[MASK]` token, approximating missing-word or deletion-type ASR errors without changing the sequence length. *text-asr-substitute* replaces 10% of valid tokens with vocabulary ids sampled from the non-special BERT vocabulary range. For nonverbal corruption, *visual-occlusion* masks a centered contiguous 30% temporal span of the visual sequence. *visual-low-conf-drop* operationalizes low-confidence visual evidence by zeroing the lowest 30% visual frames ranked by feature norm. *audio-snr-noise* injects Gaussian noise at a target SNR of 10 dB, and *audio-low-energy-drop* zeros the lowest 30% acoustic frames ranked by feature norm. Finally, *multimodal-nonrandom* combines token masking, low-norm visual frame dropping, and 10 dB acoustic noise.

Table 3 therefore reports the clean metrics from the main-result checkpoint and the corresponding non-clean evaluations. To complement the output-level metrics, we measured the paired pre-CBAM acoustic encoder representation h_a under the same 10 dB perturbation and batch-wise noise protocol. The mean absolute change, averaged over test utterances and encoder dimensions, was 0.005983 on CMU-MOSI (686 utterances), with a mean per-sample relative L_2 change of 0.914%; the corresponding values on CMU-MOSEI were 0.000404 and 0.100% (4659 utterances). These small average changes are consistent with the negligible output-level changes. Under this protocol, PGMCR is comparatively stable under visual-only and audio-only degradation, whereas ASR-like text corruption and combined non-random degradation remain the dominant failure modes. Therefore, the results provide diagnostic evidence about the model's relative sensitivity to different corruption types. The asymmetry indicates that the evaluated model relies strongly on textual evidence under this protocol: the largest MOSI degradation occurs when text is corrupted, whereas the visual- and audio-only perturbations produce negligible changes in the reported aggregate metrics. The nonverbal streams are not thereby shown to be irrelevant; polarity-conditioned weights indicate that their relative contribution can vary across samples, but they do not generally compensate for severe text corruption in these tests. We therefore frame PGMCR as a text-anchored multimodal model with conditional nonverbal support. Importantly, these fixed-severity feature perturbations are controlled proxies, not measurements of naturally occurring ASR errors, acoustic environments, camera failures, or deployment conditions. Their ecological validity is unestablished, so the results support only relative sensitivity comparisons under the stated protocol.

Table 3. Inference-time proxy-corruption results. Metric drops are relative to the clean condition; all perturbations use fixed severity.

Dataset	Mode	MAE ↓	Corr ↑	F1 All ↑	ΔMAE ↓	ΔCorr ↓	ΔF1 ↓
MOSI	clean	0.6862	0.8164	85.38	–	–	–
	text-asr-drop	0.8011	0.7491	80.92	0.1150	0.0673	4.46
	text-asr-substitute	0.7617	0.7610	81.29	0.0756	0.0554	4.09
	visual-occlusion	0.6874	0.8159	84.94	0.0012	0.0005	0.44
	visual-low-conf-drop	0.6862	0.8162	85.23	0.0000	0.0002	0.15
	audio-snr-noise	0.6862	0.8164	85.38	0.0000	0.0000	0.00
	audio-low-energy-drop	0.6859	0.8163	85.38	−0.0002	0.0000	0.00
	multimodal-nonrandom	0.8007	0.7488	80.63	0.1145	0.0676	4.75
MOSEI	clean	0.5292	0.7702	85.15	–	–	–
	text-asr-drop	0.5650	0.7313	83.76	0.0357	0.0389	1.39
	text-asr-substitute	0.5733	0.7234	83.03	0.0440	0.0468	2.12
	visual-occlusion	0.5302	0.7694	85.08	0.0010	0.0008	0.07
	visual-low-conf-drop	0.5305	0.7697	85.07	0.0012	0.0005	0.08
	audio-snr-noise	0.5292	0.7702	85.15	0.0000	0.0000	0.00
	audio-low-energy-drop	0.5293	0.7702	85.15	0.0000	0.0000	0.00
	multimodal-nonrandom	0.5648	0.7313	83.65	0.0356	0.0389	1.49

↓: lower values are better; ↑: higher values are better.

5.3. Interpretability Analysis

Table 4 reports mean modality weights and predicted polarity probabilities, grouped by ground-truth polarity. On CMU-MOSI, the largest mean weight is assigned to audio in the negative group (0.3751), visual information in the positive group (0.3744), and text in the neutral group (0.3541). From negative to positive samples, the visual mean increases from 0.2941 to 0.3744, while the audio mean decreases from 0.3751 to 0.2959. CMU-MOSEI shows the same polarity-dependent ordering with smaller changes: visual weight rises by 0.0571 and audio weight falls by 0.0490. However, text has the largest mean in the negative

and neutral groups, and the visual and text means in the positive group are nearly equal (0.3484 and 0.3470). Thus, the observed pattern is more distinct on MOSI (Figure 5) and more modest on MOSEI (Figure 6).

Table 4. Polarity-conditioned modality weights for final PGMCR. Bold values indicate the largest mean weight among text, visual, and acoustic modalities within each ground-truth polarity group.

Dataset/Polarity	n	w_t	w_v	w_a	p_{pos}	p_{neg}	p_{neu}
MOSI neg	347	0.3309	0.2941	**0.3751**	0.0625	0.8475	0.0900
MOSI neu	106	**0.3541**	0.3253	0.3206	0.3049	0.2667	0.4284
MOSI pos	233	0.3297	**0.3744**	0.2959	0.8420	0.0933	0.0647
MOSEI neg	1023	**0.3551**	0.2913	0.3536	0.0727	0.6354	0.2919
MOSEI neu	1927	**0.3643**	0.3166	0.3191	0.2932	0.1296	0.5772
MOSEI pos	1709	0.3470	**0.3484**	0.3046	0.6153	0.0537	0.3310

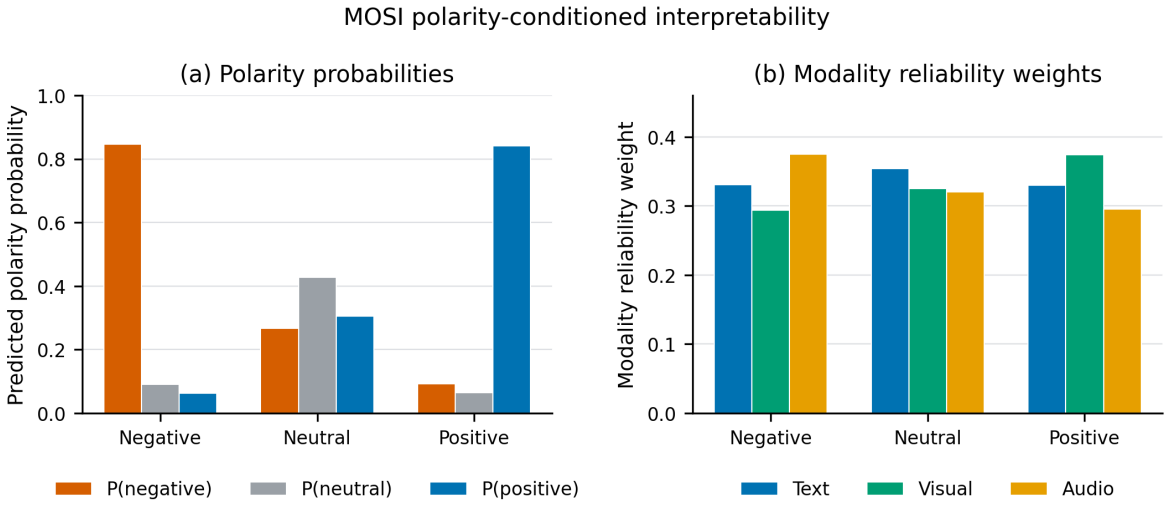


Figure 5. Combined interpretability analysis on CMU-MOSI. (**a**) Mean predicted polarity probabilities and (**b**) mean modality reliability weights, grouped by ground-truth polarity (negative, neutral, and positive) on CMU-MOSI. The panels provide descriptive summaries of modality-weight means and predicted polarity-probability distributions across ground-truth negative, neutral, and positive test groups.

The weights are generated from predicted polarity probabilities, whereas Table 4 groups samples by ground-truth polarity. The group means therefore show a descriptive association between the learned weights and true sentiment groups. The polarity-estimator accuracies are 79.01% on CMU-MOSI and 68.56% on CMU-MOSEI (Table 5), so prediction errors may also affect the resulting weights. The results support the presence of polarity-associated weight patterns in these test sets. Polarity-Guided Fusion is a form of polarity-conditioned dynamic weighting.

The confusion counts quantify the polarity-mismatch pattern used in the qualitative analysis. The polarity estimator is correct for 79.01% of CMU-MOSI test utterances and 68.56% of CMU-MOSEI test utterances; the corresponding mismatch rates are 20.99% and 31.45%. The specific positive-to-negative mismatch shown in the failure case occurs for 20 MOSI utterances (2.92%) and 46 MOSEI utterances (0.99%) under the reported evaluation. As a complementary calibration diagnostic, the expected calibration error (ECE) of the maximum polarity probability is 0.1583 on MOSI and 0.0503 on MOSEI; ECE is computed using 10 equal-width confidence bins, with confidence defined as the maximum predicted

polarity probability and accuracy defined by exact polarity-label agreement. These statistics show that polarity errors are non-negligible and can propagate to the reliability weights.

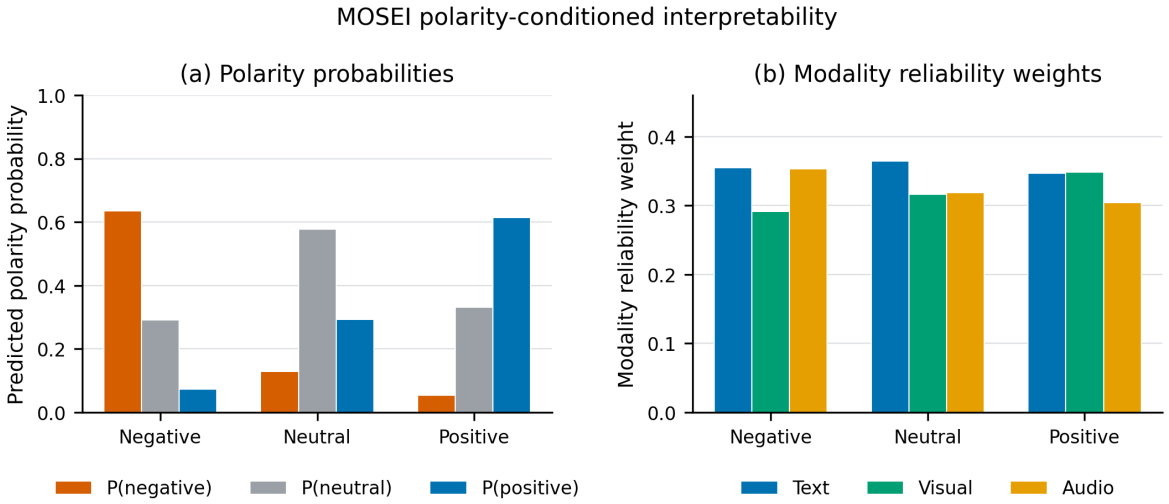


Figure 6. Combined interpretability analysis on CMU-MOSEI. (**a**) Mean predicted polarity probabilities and (**b**) mean modality reliability weights, grouped by ground-truth polarity (negative, neutral, and positive) on CMU-MOSEI. The panels provide descriptive summaries of modality-weight means and predicted polarity-probability distributions across ground-truth negative, neutral, and positive test groups.

Table 5. Polarity-estimator confusion counts on the test sets. Rows denote the true polarity and columns denote the predicted polarity.

Dataset/Polarity	Pred. Neg	Pred. Neu	Pred. Pos	Row Total
MOSI neg	299	25	23	347
MOSI neu	29	46	31	106
MOSI pos	20	16	197	233
MOSEI neg	681	308	34	1023
MOSEI neu	139	1368	420	1927
MOSEI pos	46	518	1145	1709

5.4. Case Study

To qualitatively illustrate the learned polarity-conditioned reliability behavior, we inspect a small set of representative test-set examples drawn from the interpretability exports. Table 6 includes correctly predicted positive, negative, and neutral cases together with one polarity-mismatch failure case, so that both typical behavior and an important limitation are visible. Figure 7 summarizes the sample-level reliability patterns, and Figure 8 visualizes the failure case in more detail. In the positive and negative examples, the dominant modality is consistent with the expected polarity-specific cue pattern, whereas in the neutral example text receives the largest weight. The failure example shows that the reliability prior remains informative but is not sufficient when polarity estimation is wrong or when the utterance is especially text-sensitive.

Paired Qualitative Comparison of TGCBAM

Figure 9 compares full PGMCR with standard CBAM without textual guidance and w/o TGCBAM on two post hoc CMU-MOSI test utterances. The source frames provide nonverbal context. For the negative and neutral cases, full PGMCR lowers AE by 0.2980 and 0.1981 relative to standard CBAM, and by 0.0333 and 0.0905 relative to w/o TGCBAM.

The two bar charts on the right report the final modality-reliability weights and predicted polarity probabilities of full PGMCR, respectively. These examples illustrate favorable case-level behavior of the text-guided refinement.

Because the cases were selected post hoc, they do not estimate the average module effect. The controlled ablations in Table 7 provide the quantitative model-level evidence for TGCBAM.

Table 6. Representative qualitative case-study examples from the test-set interpretability exports. The examples are provided for mechanism interpretation and include correctly predicted positive, negative, and neutral cases together with one polarity-mismatch failure case.

Case	Sample ID	Label/Pred.	True/Pred. Polarity	Observation
MOSI positive	`cW1FSBF59ik_18`	2.6000/2.6553	pos/pos	p_{pos} is dominant; visual weight is highest, and the prediction is correct.
MOSI negative	`d6hH302o4v8_36`	−2.2000/−2.2546	neg/neg	p_{neg} is dominant; audio weight is highest, and the prediction is correct.
MOSEI neutral	`u9I5WD3Nglk`	0.3333/0.0624	neu/neu	p_{neu} is dominant; text weight is highest, and the prediction is neutral.
MOSEI failure	`91844`	2.6667/−1.7662	pos/neg	Polarity mismatch; the utterance is positive, but the model predicts negative.

Table 7. Main module ablation on CMU-MOSI and CMU-MOSEI.

Dataset	Model	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0 ↑	Acc-2 All ↑	F1 Non0 ↑	F1 All ↑
MOSI	PGMCR	**0.6862**	**0.8164**	**48.10**	86.43	**85.42**	86.42	**85.38**
	w/o MDC-KD(VA)	0.6926	0.8042	47.81	**87.20**	85.13	**87.15**	85.03
	w/o Polarity-Guided Fusion	0.7113	0.7880	46.21	84.60	82.22	84.45	81.96
	w/o TGCBAM	0.7007	0.7998	45.77	86.28	84.26	86.15	84.06
	w/o text guidance in CBAM	0.6987	0.8001	46.06	84.76	83.38	84.78	83.37
MOSEI	PGMCR	0.5292	0.7702	53.72	**85.72**	**85.15**	85.49	**85.15**
	w/o MDC-KD(VA)	0.5354	0.7646	52.89	85.69	83.99	**85.58**	84.20
	w/o Polarity-Guided Fusion	0.5428	0.7600	52.52	84.89	81.88	84.94	82.41
	w/o TGCBAM	0.5346	0.7670	**54.02**	85.42	81.93	85.33	82.33
	w/o text guidance in CBAM	**0.5266**	**0.7708**	53.42	84.89	80.04	84.95	80.72

↓: lower values are better; ↑: higher values are better. Bold: the best value in each column (the lowest MAE; the highest value for the other metrics).

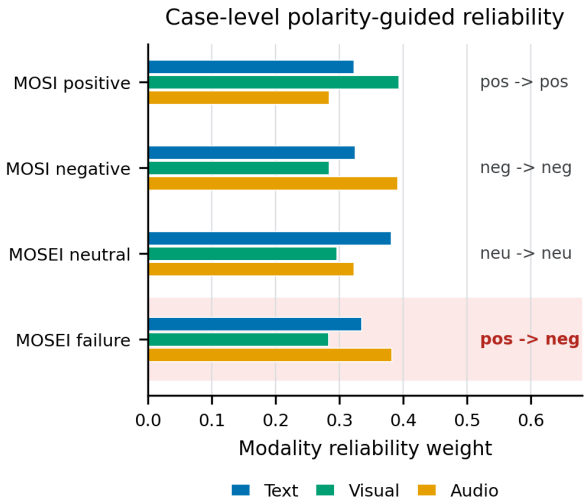


Figure 7. Illustrative case-level reliability summary. The representative examples show that polarity-guided reliability weights vary across samples and often follow the expected cue pattern, while the failure row highlights an important polarity-mismatch limitation at the sample level.

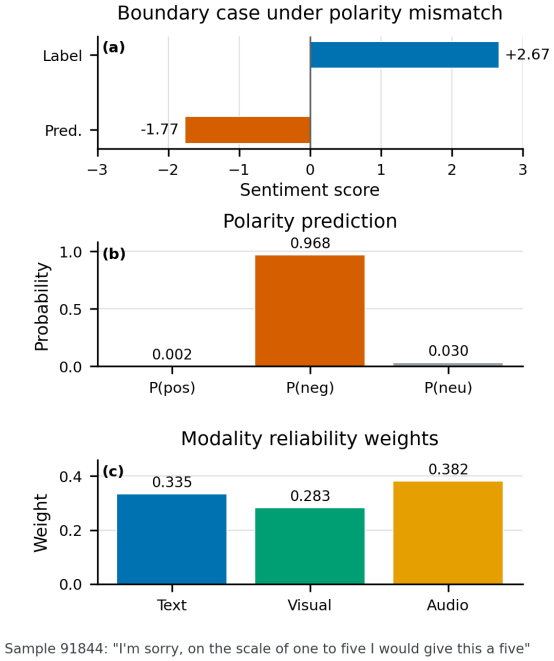


Figure 8. Failure-case visualization for qualitative diagnosis. (**a**) Ground-truth and predicted sentiment scores for the failure sample (91844); (**b**) predicted polarity probabilities, in which the negative class dominates (0.968, versus 0.030 neutral and 0.002 positive); and (**c**) modality reliability weights, where audio receives the largest weight (0.382). The model assigns a high negative polarity probability to a positive utterance, leading to a mismatched sentiment prediction and illustrating a limitation of the polarity prior when polarity estimation fails.

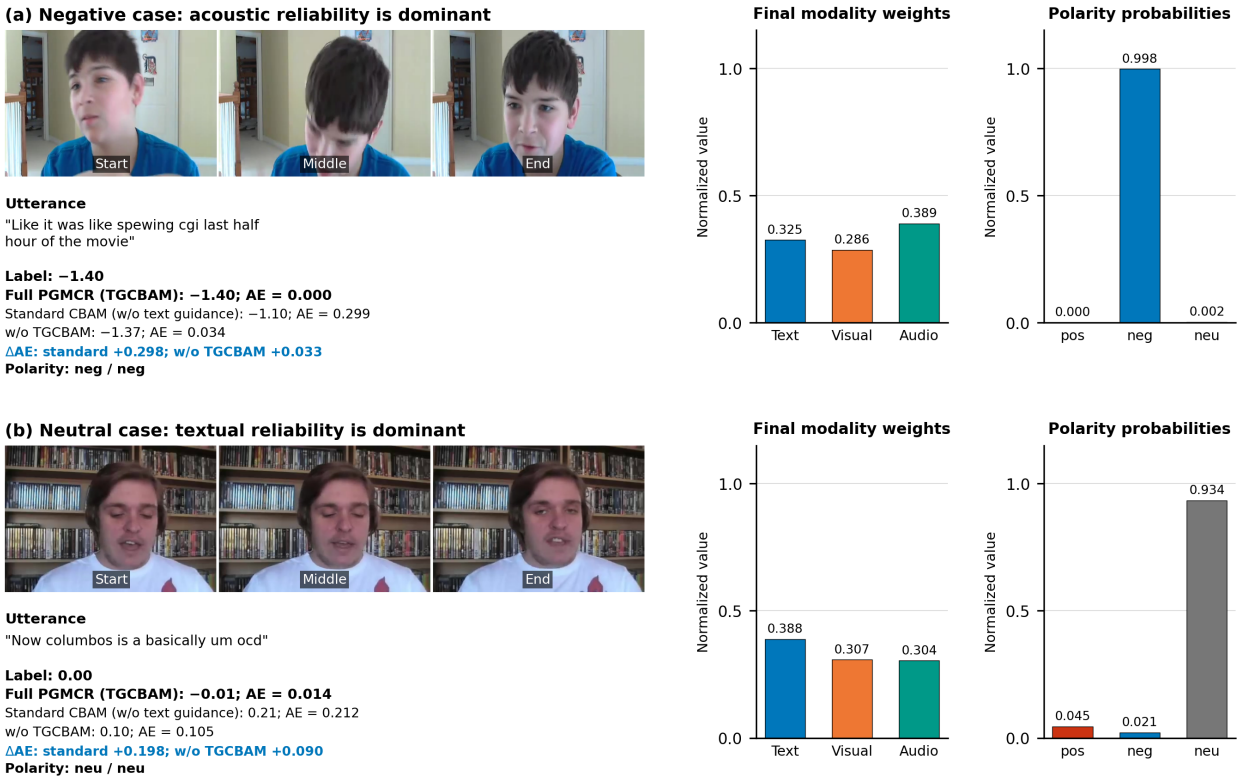


Figure 9. Paired qualitative comparison of TGCBAM on CMU-MOSI. Each row shows source frames and predictions from full PGMCR, standard CBAM without textual guidance, and w/o TGCBAM. AE denotes absolute error; positive ΔAE favors full PGMCR.

6. Ablation Study

6.1. Main Module Ablation

Table 7 reports removal-based comparisons among the evaluated configurations. On CMU-MOSI, the PGMCR point estimates are best on MAE, Corr, Acc-7, Acc-2 All, and F1 All, while removing MDC-KD(VA) gives the best Non0 binary metrics. Removing Polarity-Guided Fusion is associated with the broadest decrease across the displayed metrics. TGCBAM and its text guidance also show metric-dependent effects. On CMU-MOSEI, PGMCR has the largest displayed values for Acc-2 Non0, Acc-2 All, and F1 All, while removal variants retain isolated advantages on other metrics: w/o TGCBAM obtains the highest Acc-7, w/o text guidance obtains the best MAE and Corr, and w/o MDC-KD(VA) obtains the best F1 Non0. The variation across metrics supports the claim that these results reflect a trade-off across the evaluated configurations.

The MOSEI comparison also reveals a regression–classification trade-off for text guidance. Removing text guidance improves MAE from 0.5292 to 0.5266 and Corr from 0.7702 to 0.7708, but reduces Acc-2 All from 85.15 to 80.04 and F1 All from 85.15 to 80.72. A plausible interpretation is that text guidance changes the nonverbal representation in a way that is not uniformly beneficial for continuous-score calibration on the larger and more heterogeneous dataset, while still supporting the decision boundary used by all-sample binary classification.

6.2. MDC-KD Mechanism Ablation

Table 8 examines the MDC-KD mechanism variants. On CMU-MOSI, the complete MDC-KD(VA) setting yields the strongest overall profile, whereas removing MDC-KD(VA) retains an advantage on the Non0 binary metrics but weakens Acc-7 and the all-sample metrics. Removing the label gate or the joint VA perturbation causes a clear drop in Acc-7 and all-sample binary performance, while removing the confidence gate yields a milder degradation than the full design. The audio-only MDC-KD variant retains competitive all-sample binary scores but is weaker on regression and Acc-7.

On CMU-MOSEI, PGMCR gives the best MAE, Corr, Acc-2 All, and F1 All, whereas individual MDC-KD removals retain isolated advantages on Acc-7 or Non0 metrics. This pattern supports the complete gated visual/audio/VA consistency design as the most balanced MDC-KD configuration under full-sample evaluation. Figure 10 provides a compact MOSI comparison of the Acc-7 and F1 All trends.

Table 8. MDC-KD mechanism ablation on CMU-MOSI and CMU-MOSEI.

Dataset	Variant	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0/All ↑	F1 Non0/All ↑
MOSI	PGMCR	**0.6862**	**0.8164**	**48.10**	86.43/**85.42**	86.42/**85.38**
	w/o MDC-KD(VA)	0.6926	0.8042	47.81	**87.20**/85.13	**87.15**/85.03
	w/o label gate	0.6946	0.8070	46.06	85.67/83.67	85.66/83.60
	w/o confidence gate	0.7042	0.8058	46.36	85.98/84.55	85.99/84.53
	w/o VA pair	0.6951	0.8039	46.06	85.98/83.97	85.95/83.88
	Audio-only MDC-KD	0.7088	0.7988	46.36	86.89/85.13	86.92/85.12
MOSEI	PGMCR	**0.5292**	**0.7702**	53.72	85.72/**85.15**	85.49/**85.15**
	w/o MDC-KD(VA)	0.5354	0.7646	52.89	85.69/83.99	85.58/84.20
	w/o label gate	0.5347	0.7606	**54.15**	84.67/79.80	84.69/80.45
	w/o confidence gate	0.5316	0.7656	53.57	**85.83**/83.07	**85.71**/83.35
	w/o VA pair	0.5367	0.7680	53.75	85.03/84.59	84.81/84.61
	Audio-only MDC-KD	0.5366	0.7640	53.92	85.53/80.88	85.49/81.42

↓: lower values are better; ↑: higher values are better. Bold: the best value in each column (the lowest MAE; the highest value for the other metrics).

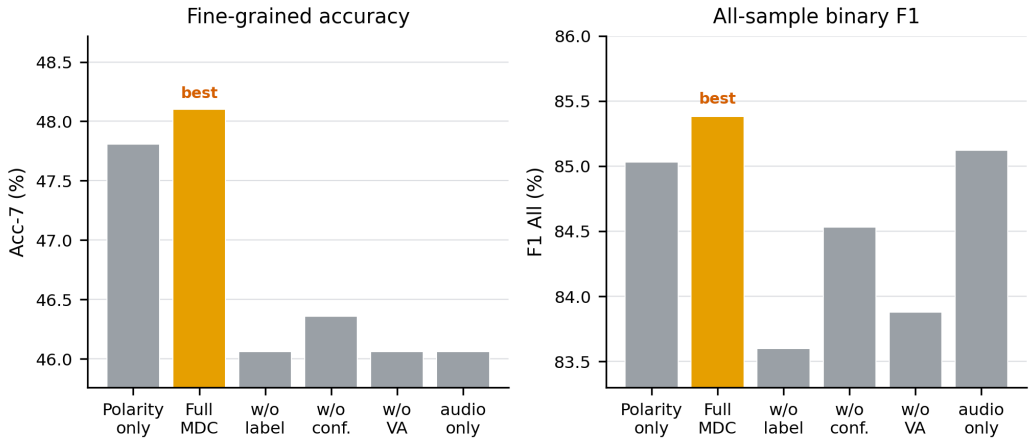


Figure 10. MDC-KD ablation visualization on CMU-MOSI. The figure summarizes the Acc-7 and F1 All results under PGMCR and the removal-based MDC-KD variants. The full PGMCR setting gives the best fine-grained accuracy and the best all-sample binary F1, while the remaining variants show different trade-offs after removing individual gates or perturbation components.

6.3. Sensitivity Analyses for Architecture and Adaptive Training Settings

We analyze sensitivity to encoder capacity, polarity thresholds, and adaptive-training settings. Each comparison retains the same data splits, optimization settings, polarity-guided fusion module, and MDC-KD(VA) objective; only the factor named in each table is changed. The lightweight Transformer configuration uses one layer with hidden size 64, whereas the larger-capacity comparison uses two layers with hidden size 256.

Table 9 reports the encoder-capacity comparison. The larger-capacity encoder does not improve the overall metric profile. On CMU-MOSI, the lightweight configuration is better in every reported metric, including MAE (0.6862 versus 0.7095), Corr (0.8164 versus 0.7970), and Acc-7 (48.10 versus 46.79). On CMU-MOSEI, the two configurations are closer, but the lightweight configuration remains better in MAE and Corr, while the larger-capacity configuration provides only a marginal Acc-7 difference. These results support the selected lightweight encoder on the reported metrics.

Table 9. Transformer capacity sensitivity on CMU-MOSI and CMU-MOSEI. The lightweight and larger-capacity encoders differ in depth and hidden size; the remaining settings are unchanged.

Dataset	Encoder	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0 ↑	Acc-2 All ↑	F1 Non0 ↑	F1 All ↑
MOSI	1 layer, $d = 64$	0.6862	0.8164	48.10	86.43	85.42	86.42	85.38
MOSI	2 layers, $d = 256$	0.7095	0.7970	46.79	84.91	82.65	84.84	82.51
MOSEI	1 layer, $d = 64$	0.5292	0.7702	53.72	85.72	85.15	85.49	85.15
MOSEI	2 layers, $d = 256$	0.5319	0.7663	53.83	85.14	81.37	85.13	81.88

↓: lower values are better; ↑: higher values are better.

Table 10 reports the threshold sensitivity analysis, which shows small and non-monotonic changes. On CMU-MOSI, ± 0.3 gives a lower MAE but lower Corr and Acc-7 than ± 0.5, whereas ± 0.7 improves Acc-7 but slightly reduces Corr. On CMU-MOSEI, ± 0.3 is favorable for MAE and Corr, while ± 0.7 is less favorable for both regression metrics. The default ± 0.5 threshold is therefore retained as a balanced operating point across regression and classification metrics.

Table 10. Polarity-threshold sensitivity on CMU-MOSI and CMU-MOSEI. The thresholds are symmetric around zero and are applied only to the auxiliary polarity targets.

Dataset	Threshold	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 Non0 ↑	Acc-2 All ↑	F1 Non0 ↑	F1 All ↑
MOSI	±0.3	0.6797	0.8076	46.06	87.04	84.69	86.94	84.52
MOSI	±0.5	0.6862	0.8164	48.10	86.43	85.42	86.42	85.38
MOSI	±0.7	0.6865	0.8106	49.42	86.28	84.84	86.27	84.79
MOSEI	±0.3	0.5284	0.7733	54.11	85.75	82.12	85.72	82.58
MOSEI	±0.5	0.5292	0.7702	53.72	85.72	85.15	85.49	85.15
MOSEI	±0.7	0.5333	0.7627	54.05	85.22	83.28	85.15	83.57

↓: lower values are better; ↑: higher values are better.

Table 11 shows that changing the scheduler parameters shifts the activation timing as intended. On CMU-MOSI, $e_{\min} = 5$ and $\tau = 1.5$ reproduce the baseline profile, whereas delaying the latest activation to epoch 12 produces a modest reduction in Corr and the all-sample classification metrics. On CMU-MOSEI, $e_{\max} = 12$ produces the same profile as the baseline because activation occurs before the upper bound is reached; $e_{\min} = 5$ delays activation and yields the largest observed decrease, while $\tau = 1.5$ remains close to the baseline. These results indicate that the adaptive scheduler exhibits stable behavior across the tested parameter settings.

Table 11. Adaptive MDC-KD(VA) scheduler sensitivity. The activation epoch is the observed epoch at which the consistency branch becomes active. The mini-batch application probability $q = 0.25$ is fixed across all settings.

Dataset	Setting	$(e_{\min}, e_{\max}, \tau)$	Act. Epoch	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 All ↑	F1 All ↑
MOSI	Baseline	$(2, 8, 1.0)$	8	0.6862	0.8164	48.10	85.42	85.38
MOSI	$e_{\min} = 5$	$(5, 8, 1.0)$	8	0.6862	0.8164	48.10	85.42	85.38
MOSI	$e_{\max} = 12$	$(2, 12, 1.0)$	12	0.6852	0.8083	47.38	84.40	84.24
MOSI	$\tau = 1.5$	$(2, 8, 1.5)$	8	0.6862	0.8164	48.10	85.42	85.38
MOSEI	Baseline	$(2, 8, 1.0)$	3	0.5292	0.7702	53.72	85.15	85.15
MOSEI	$e_{\min} = 5$	$(5, 8, 1.0)$	6	0.5354	0.7646	52.89	83.99	84.20
MOSEI	$e_{\max} = 12$	$(2, 12, 1.0)$	3	0.5292	0.7702	53.72	85.15	85.15
MOSEI	$\tau = 1.5$	$(2, 8, 1.5)$	3	0.5293	0.7698	53.53	83.67	83.93

↓: lower values are better; ↑: higher values are better.

7. Conclusions

The sensitivity analyses delimit the operating range of PGMCR: the lightweight nonverbal encoders remain competitive with a larger-capacity alternative, polarity thresholds produce non-monotonic changes, and scheduler settings alter activation timing with dataset-dependent metric effects.

The evidence supports a bounded conclusion. Polarity-Guided Fusion provides an explicit mapping from predicted polarity to relative modality allocation, while TGCBAM and Adaptive Dynamic MDC-KD(VA) are established design ideas adapted and evaluated here as auxiliary components whose effects vary by dataset and metric. The reported point estimates do not establish statistical significance, and the proxy-corruption diagnostics do not establish robustness to naturally occurring sensor failure. They show that, under the specified test protocol, text corruption causes larger degradation than the tested nonverbal perturbations on MOSI; polarity-estimation errors may also affect the reliability prior. Thus, the results support conditional nonverbal assistance in the evaluated setting, but not balanced modality reliance, absolute joint reliability, or real-world robustness. Future work should evaluate repeated training variability, real ASR outputs, ecologically grounded acoustic and visual degradations, and additional datasets and languages.

Author Contributions: Conceptualization, H.Y. and J.G.; methodology, H.Y. and J.G.; software, H.Y.; validation, H.Y., J.G., M.Z. (Ming Zhao), M.M. and Y.L.; formal analysis, H.Y. and J.G.; investigation, H.Y. and B.C.; resources, J.G.; data curation, H.Y.; writing—original draft preparation, J.G.; writing—review and editing, H.Y., J.G., M.Z. (Ming Zhao), M.M., M.Z. (Min Zhou) and Y.L.; visualization, J.G.; supervision, J.G. and B.C.; project administration, J.G.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hubei Provincial Natural Science Foundation (2025AFC122); the Ministry of Education Industry-University Cooperation Education Project (231106627155856); 2026 Wuhan Polytechnic University ESI Project (Engineering); the School-Enterprise Cooperation Project of Wuhan Polytechnic University (whpu-2025-kj-5347); University Graduate Teaching and Research Project, Research on Interdisciplinary Collaborative Innovation and Joint Training Mechanism in Artificial Intelligence (2025JY05); Hubei Provincial Technology Innovation Plan Project (2025BBB004); Hubei Provincial 2024 New Engineering Construction Project (XGK02078).

Data Availability Statement: The CMU-MOSI and CMU-MOSEI datasets used in this study are publicly available through the official CMU MultiComp release. The dataset files and code used in this study are publicly available through the authors' GitHub repository: https://github.com/Guojy-julia/PGMCR (accessed on 2 October 2026).

Acknowledgments: The authors thank the maintainers of the CMU-MOSI and CMU-MOSEI datasets and the developers of the open-source tools used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.-P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
2. Bagher Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. 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), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246. [CrossRef]
3. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114.
4. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Bagher Zadeh, A.; Morency, L.-P. 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), Melbourne, Australia, 15–20 July 2018; pp. 2247–2256. [CrossRef]
5. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569. [CrossRef] [PubMed]
6. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
7. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; Volume 35, pp. 10790–10797.
8. Feng, X.; Lin, Y.; He, L.; Li, Y.; Chang, L.; Zhou, Y. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 14755–14766.
9. Chen, M.; Liu, J.; Li, X.; Zhang, Y.; Liu, H.; Jiao, J.; He, H. DNMCN: Dual-stage normalization based modality-collaborative fusion network for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2025**, *16*, 2770–2787. [CrossRef]
10. Chen, S.; Zhou, D.; Wei, C.; Li, X.; Yao, H.; Luo, H. Dual-stream collaborative network (DSCN): Multimodal sentiment analysis via modality-invariant and modality-specific representation learning. *Neurocomputing* **2025**, *652*, 131059. [CrossRef]
11. Hou, Z.; Zhang, Q.; Lei, Z.; Zeng, Z.; Jia, R. KD-MSA: A multimodal implicit sentiment analysis approach based on KAN and asymmetric contribution-aware dynamic fusion. *Symmetry* **2025**, *17*, 1401. [CrossRef]

12. Baltrušaitis, T.; Robinson, P.; Morency, L.-P. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
13. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838.
14. Cambria, E.; Poria, S.; Hazarika, D.; Kwok, K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2018; Volume 32.
15. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
16. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
17. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
18. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intell. Syst.* **2018**, *33*, 17–25. [CrossRef]
19. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
20. Lai, S.; Hu, X.; Xu, H.; Ren, Z.; Liu, Z. Multimodal sentiment analysis: A survey. *Displays* **2023**, *80*, 102563. [CrossRef]
21. Pérez-Rosas, V.; Mihalcea, R.; Morency, L.-P. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 973–982.
22. Morency, L.-P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the Web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
23. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.
24. Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2018; Volume 32.
25. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* **2019**, *7*, 100943–100953. [CrossRef]
26. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2359–2369. [CrossRef] [PubMed]
27. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.-P. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2018; Volume 32.
28. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.-P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2019; Volume 33, pp. 6892–6899.
29. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.-N. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 4730–4738.
30. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.-P. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 1009–1021.
31. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-P.; Poria, S. 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, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 9180–9192.
32. Li, J.; Wang, X.; Lv, G.; Zeng, Z. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 July 2022; pp. 2282–2294.
33. Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7837–7851.

34. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. 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), Toronto, ON, Canada, 9–14 July 2023; pp. 7617–7630.
35. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 756–767.
36. Wu, Z.; Gong, Z.; Koo, J.; Hirschberg, J. Multimodal multi-loss fusion network for sentiment analysis. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 3588–3602.
37. Yuan, Z.; Li, W.; Xu, H.; Yu, W. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 4400–4407.
38. Guo, Z.; Jin, T.; Zhao, Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 1726–1736.
39. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. ConKI: Contrastive knowledge injection for multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 13610–13624.
40. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 12966–12978.
41. Li, X.; Cheng, X.; Miao, D.; Zhang, X.; Li, Z. TF-Mamba: Text-enhanced fusion Mamba with missing modalities for robust multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 4–9 November 2025; pp. 11252–11267. [CrossRef]
42. Yang, Y.; Dong, X.; Qiang, Y. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 16–21 June 2024; pp. 2099–2110.
43. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
44. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
45. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
47. Sun, Y.; Mai, S.; Hu, H. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Trans. Affect. Comput.* **2023**, *14*, 2209–2223. [CrossRef]
48. Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.* **2023**, *136*, 109259. [CrossRef]
49. Ma, G.; Ren, X.; Jiang, Y.; Guan, H.; Xu, B. From feature alignment to multimodal fusion: A two-stage primary modality-guided approach for MSA. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*; Association for Computing Machinery: New York, NY, USA, 2025; Article 95, pp. 1–8. [CrossRef]
50. Sun, L.; Lian, Z.; Liu, B.; Tao, J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2024**, *15*, 309–325. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.