POINTWISE MULTIMODAL INFORMATION BUDGET REGULARIZATION FOR MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning systems integrate complementary information from different modalities and consequently demonstrate superior performance over unimodal methods. However, most existing multimodal fusion approaches assume that the modalities invariably provide synergistic benefits, overlooking the modality conflicts and joint failures that commonly arise in real-world deployment. This paper proposes a pointwise multimodal information budget framework that systematically distinguishes three regimes of multimodal interaction at the information-theoretic level: *the synergistic regime*, in which the modalities provide mutually consistent positive evidence; *the conflicting regime*, in which some modalities introduce contradictory information that interferes with the prediction; and *the joint-failure regime*, in which none of the modalities contains discriminative signals. The pointwise multimodal information budget adjusts the strength of the constraints according to the effective information contribution of each modality for the current sample. The multimodala dynamic routing mechanism uses the sign of pointwise mutual information as a supervisory signal to learn sample-wise modality reliability and enable branch selection. Extensive experiments on three multimodal benchmarks demonstrate the effectiveness and superiority of our proposed PMIB.

1 INTRODUCTION

Multimodal learning (Zadeh et al., 2017; Tsai et al., 2019; Zhang et al., 2023a; Wang et al., 2026) integrates heterogeneous signals from different perceptual channels, leveraging cross-modal complementarity to enhance a model's ability to capture complex semantics. For example, in multimodal sentiment analysis, text conveys explicit sentiment, speech provides prosodic information, and vision captures facial expressions. Existing methods generally assume that different modalities consistently provide mutually supportive, task-relevant information, implicitly expecting fusion to yield consistent gains. However, signal acquisition quality, including noise, occlusion, and temporal misalignment, and the limitations of feature extraction models can cause modality contributions to vary substantially for a given sample: i) **synergy**: modalities may provide complementary information that supports the same prediction, ii) **conflict**: produce contradictory judgments due to semantic ambiguity, or iii) **joint failure**: simultaneously lack discriminative information. From the information-theoretic perspective (Church & Hanks, 1990; Belghazi et al., 2018; Lee et al., 2026), although adding observations cannot decrease their overall mutual information with the target at the distribution level, this property does not guarantee that a finite-capacity model will benefit from fusion on every sample. Understanding the direction and strength of the information provided by each modality for a given sample is therefore central to effective fusion.

Existing studies on cross-modal information utilization has primarily followed two directions: multimodal fusion and multimodal representation learning. Multimodal fusion (Zadeh et al., 2017; Liu et al., 2018; Zhao et al., 2022b; Lv et al., 2025; Zhao et al., 2026c) captures dependencies across modalities through feature concatenation, tensor interactions, cross-modal attention, or decision aggregation to produce joint predictions. Multimodal representation learning (Hazarika et al., 2020; Yu et al., 2021; Zhao et al., 2022a; Fang et al., 2025; Zhao et al., 2026a), by contrast, organizes information before fusion through strategies such as shared–private representation decomposition, cross-modal alignment, contrastive learning, and information bottlenecks, preserving task-relevant information while reducing redundancy. The former focuses on how information interacts, whereas the latter focuses on how information is encoded; together, they have advanced multimodal learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

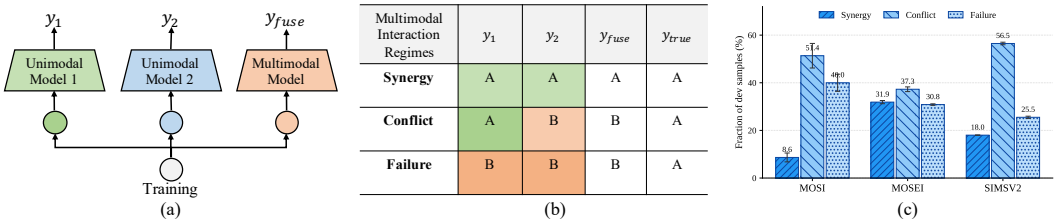


Figure 1: Multimodal Interaction Regimes: a) prediction results of unimodal and multimodal models; b) three multimodal interaction regimes: synergy, conflict, and joint failure; and c) the proportions of samples belonging to the three multimodal interaction regimes in the multimodal datasets.

However, these methods typically prioritize exploiting complementarity and enhancing consistency. Even when dynamic weighting or noise suppression is incorporated, they do not necessarily distinguish whether sample-specific information supports or opposes the target, or explicitly account for the fusion behaviors required by different interaction regimes.

When modalities conflict or jointly fail (Zhang et al., 2023a; Wang et al., 2026; Jiang et al., 2026), the uniform fusion strategy designed to enhance synergy faces significant challenges. As illustrated in Figure 1, in MSA task, **for synergistic samples**(AA), positive language, upbeat intonation, and cheerful facial expressions jointly support positive sentiment. **For conflicting samples**(AB), positive wording in a sarcastic utterance may bias the text branch toward a positive prediction, while a cold tone and a dissatisfied expression support the true negative sentiment. **For joint-failure samples**(BB), transcription errors, background noise, and facial occlusion may prevent all unimodal branches from extracting useful information. In these cases, agreement among representations does not necessarily imply correct predictions, nor do differences among representations necessarily indicate exploitable complementarity. Directly integrating such representations may allow misleading information to override useful information or further amplify biases shared across branches. Therefore, multimodal learning requires sample-wise assessment of the direction and strength of information, followed by regime-specific integration: reinforcing useful information under synergy, prioritizing reliable branches under conflict, and suppressing unsupported error amplification under joint failure.

To this end, we propose a multimodal information budget framework based on a pointwise mutual information, combining sample-level information assessment with information constraints during fusion. Specifically, we estimate pointwise mutual information using unimodal predictions and class priors to characterize whether a modality increases or decreases support for the true class relative to the prior. We then distinguish synergy, conflict, and joint failure according to the signs of these modality-specific information estimates. Based on the information provided by individual modalities, we construct a pointwise multimodal information budget and impose two-sided constraints that encourage the fusion model to preserve useful information while limiting unsupported information amplification. Furthermore, we introduce multimodal dynamic routing, using the signs of pointwise mutual information during training to supervise reliability learning. This enables the model to dynamically adjust modality contributions based solely on input representations at inference time. The information budget constrains fusion outputs, while modality routing regulates information inputs. Together, they translate the three interaction regimes into learnable fusion behaviors, providing a unified information-theoretic perspective on synergy enhancement, conflict resolution, and joint-failure suppression in multimodal learning. Experiments on multiple multimodal sentiment analysis datasets demonstrate that the proposed method maintains predictive performance while more effectively distinguishing synergistic, conflicting, and joint-failure samples

The contributions of this paper are summarized as follows:

- We propose a fusion framework based on a pointwise mutual information budget that enables information accumulation under synergy, reliable branch selection under conflict, and error suppression under joint failure.
- We use pointwise mutual information as a meta-supervisory signal to guide modality routing in learning reliability assessment, enabling dynamic modality weight adjustment without supervision.

- Extensive experiments demonstrate that the proposed method significantly outperforms existing baselines on most metrics, validating the effectiveness of our proposed PMIB

2 Relation Work (Extended Version in Appendix A)

Existing studies can be broadly divided into two categories: multimodal fusion methods and multimodal representation learning methods. Both approaches focus on multimodal collaboration to learn task-relevant multimodal fused representations. A detailed introduction to these methods is provided in Appendix A.

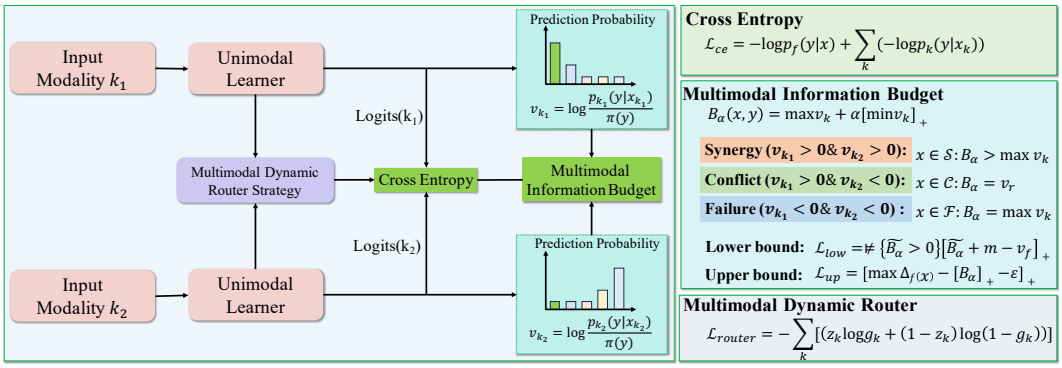


Figure 2: Schematic illustration of the proposed PMIB framework.

3 Method

In this section, as shown in Figure 2, we introduce PMIB, a pointwise multimodal information budget regularization framework. Specifically, PMIB comprises three core components: 1) multimodal pointwise mutual information estimation, which measures each modality's information contribution through log probability ratios relative to the class prior, identifying positive-information and negative-information modalities; 2) multimodal information budget regularization, which encourages synergistic information enhancement when modalities agree, guides the model to select reliable branches when modalities conflict, and suppresses unsupported information amplification during joint failure; 3) sample-level adaptive multimodal fusion, which estimates modality reliability using pointwise information signs and dynamically downweights negative-information modalities, thereby enabling synergy enhancement, conflict resolution, and joint-failure suppression.

3.1 Pointwise Multimodal Mutual Information

Let $x = (x_1, \ldots, x_M)$ be a multimodal input with label $y \in \mathcal{Y} = \{1, \ldots, C\}$. Modality encoders produce $E_k = g_k(x_k)$, and unimodal classifier h_k together with a multimodal fusion classifier h_f yield,

$$p_k(\cdot \mid x_k) = \mathrm{softmax}\big(h_k(E_k)\big), \qquad p_f(\cdot \mid x) = \mathrm{softmax}\big(h_f([E_1, \ldots, E_M])\big), \tag{1}$$

where p_k and p_f denote the probability distributions of unimodal and multimodal predictions, respectively. For each modality and for the fused predictor we define the pointwise mutual information of the true class (Church & Hanks, 1990),

$$v_k(x, y) = \log \frac{p_k(y \mid x_k)}{\pi(y)}, \qquad v_f(x, y) = \log \frac{p_f(y \mid x)}{\pi(y)}, \tag{2}$$

and, for an arbitrary class c, the information *claimed* by the multimodal fused classifier,

$$\Delta_f(c) = \log \frac{p_f(c \mid x)}{\pi(c)}, \qquad v_f = \Delta_f(y). \tag{3}$$

The Eq 2 measures how far the log-probability of the true class is lifted above the prior, and its *sign* is discriminative: $v_k > 0$ identifies modality k as an information source, whereas $v_k < 0$ identifies it as

a source of misinformation. Subtracting $\log \pi$ mirrors logit adjustment in long-tail learning (Menon et al., 2020) and removes the evidence-independent bias contributed by the prior, making readings comparable across classes.

3.2 Multimodal Interaction regimes

Based on the pointwise mutual information defined above, we jointly analyze the signs and magnitudes of the information contributions from each modality to define sample-level multimodal interaction regimes, including modality agreement, modality conflict, and joint failure. These regimes provide the foundation for constructing a regimes-adaptive multimodal information budget.

Definition 1 (Multimodal Interaction Regimes)**.** *Given (x, y), the sample belongs to*

$$\mathcal{S}:\ v_1 > 0 \wedge v_2 > 0, \qquad \mathcal{C}:\ v_1 v_2 < 0, \qquad \mathcal{F}:\ v_1 \le 0 \wedge v_2 \le 0, \tag{4}$$

termed the agreement*,* conflict *and* joint-failure *regimes. The three sets are mutually exclusive and exhaustive. On $\mathcal{C}$ we write $r = \arg\max_k v_k$ for the reliable branch ($v_r > 0$) and w for the misinformative branch ($v_w < 0$).*

The definition is determined entirely by the sign pattern of the PMI. When $C > 2$ or the prior is imbalanced, the branch may satisfy $\hat{y}_k = y$ while $v_k < 0$; equation 4 correctly assigns such nominally correct but uninformative branches to the misinformative side.

Multimodal Pointwise Co-Information. Building on this, we further introduce pointwise co-information to quantify synergy and redundancy among modalities. We define

$$ci(x, y) = v_1 + v_2 - v_f, \tag{5}$$

the pointwise form of interaction information Under the partial information decomposition $I(Y; X_1, X_2) = R + U_1 + U_2 + S$ into redundancy, unique and synergistic terms and neglecting the gap in Eq 23,

$$\mathbb{E}[ci] = I(Y; X_1) + I(Y; X_2) - I(Y; X_1, X_2) = R - S, \tag{6}$$

so $ci < 0$ indicates synergy dominance and $ci > 0$ redundancy dominance. Equation equation 6 is the population-level counterpart Liang et al. (2023) report that mainstream multimodal benchmarks are redundancy dominated, which is why the redundancy–synergy weight in Section 3.3 must be adaptive rather than fixed.

By the chain rule of mutual information,

$$v_f - v_k = \log \frac{p_f(y \mid x)}{p_k(y \mid x_k)} = \mathrm{pmi}(x_{\bar{k}}; y \mid x_k), \qquad \mathbb{E}[v_f - v_k] = I(Y; X_{\bar{k}} \mid X_k) = U_{\bar{k}} + S \ \ge 0. \tag{7}$$

Equation equation 7 is the pivot of our construction. Conditional mutual information is non-negative in expectation, yet its pointwise density can be negative, and negative values concentrate on failure cases. Using the expectation in equation 7 as a pointwise justification for "fusion should be more confident" therefore misapplies a population statement. PMIB instead models the pointwise violation itself and deliberately waives the requirement on $\mathcal{F}$.

3.3 The Multimodal Information Budget

Rather than attaching a separate constraint to each regime, we summarise the evidence available at a sample into a single scalar, the *information budget*,

$$B_\alpha(x, y) = \max_k v_k + \alpha \left[\min_k v_k \right]_+, \qquad \alpha \in [0, 1], \quad [\cdot]_+ = \max(0, \cdot). \tag{8}$$

Here α is a non-redundancy ratio: $\alpha = 0$ assumes full redundancy and sets the budget to the strongest modality, whereas $\alpha = 1$ assumes class-conditional independence so that evidence is approximately additive. The truncation applied to $\min_k v_k$ is the key design choice, since it makes the budget attain the intended semantics in all three regimes automatically:

$$x \in \mathcal{S}:\ B_\alpha > \max_k v_k; \qquad x \in \mathcal{C}:\ B_\alpha = v_r; \qquad x \in \mathcal{F}:\ B_\alpha = \max_k v_k \le 0. \tag{9}$$

Under agreement the budget exceeds the strongest modality, under conflict it equals exactly the reliable branch (the misinformative branch contributes nothing), and under joint failure it is non-positive. All three behaviours are realised by one scalar, with no explicit regime dispatch.

Multimodal information lower and upper bounds. The budget induces a pair of hinge constraints that act in opposite directions from a common reference point:

$$\mathcal{L}_{\text{low}} = \mathbb{1}\{\tilde{B}_\alpha > 0\} \left[\tilde{B}_\alpha + m - v_f\right]_+, \tag{10}$$

$$\mathcal{L}_{\text{up}} = \left[\max_c \Delta_f(c) - [B_\alpha]_+ - \varepsilon\right]_+, \tag{11}$$

$$\mathcal{L}_{bound} = \mathcal{L}_{\text{low}} + \mathcal{L}_{\text{up}}, \qquad \varepsilon > m > 0. \tag{12}$$

The lower bound $\mathcal{L}_{\text{low}}$ requires the fused predictor to realise the information the evidence supports: on $\mathcal{S}$ it demands a synergistic gain, and on $\mathcal{C}$ it degenerates, via equation 9, to information preservation with respect to the reliable branch. The indicator deactivates it on $\mathcal{F}$, where no positive budget exists and no gain should be demanded. The upper bound $\mathcal{L}_{\text{up}}$ ranges over *all* classes rather than the true class alone; were it restricted to the true class, the fused head could place its excess claim on a wrong class and evade the constraint entirely. On $\mathcal{F}$ we have $[B_\alpha]_+ = 0$, so $\mathcal{L}_{\text{up}}$ forces p_f back into the bounds $p_f(c|x)$ above by $e^\epsilon \pi(c)$.

Since $v_f \le \max_c \Delta_f(c)$, the two constraints form a sandwich,

$$\tilde{B}_\alpha + m \;\le\; v_f \;\le\; \max_c \Delta_f(c) \;\le\; [B_\alpha]_+ + \varepsilon, \tag{13}$$

i.e. an information corridor of width $\varepsilon - m$ nats on positive-budget samples. Non-emptiness requires $\varepsilon \ge m$.

Multimodal Dynamic Routing and Adaptive α Estimation The sign of v_k requires labels and is unavailable at test time. To make reliable branch selection architectural rather than merely a loss preference, we introduce a sample-wise weight $g_k = \sigma(\phi_k(E_k)) \in (0, 1)$ supervised by the PMI sign,

$$\mathcal{L}_{router} = -\sum_k \big[z_k \log g_k + (1 - z_k) \log(1 - g_k)\big], \qquad z_k = \mathbb{1}\{v_k > 0\}, \tag{14}$$

$$E_f = [\, g_1 E_1, \ldots, g_M E_M \,], \qquad p_f = \mathrm{softmax}\big(h_f(E_f)\big). \tag{15}$$

The supervision z_k is the indicator of the sign ofequation 2 and requires no extra annotation. This is related to reliability-aware fusion based on Dirichlet evidence (Han et al., 2022) or energy scores (Zhang et al., 2023b); those methods use uncertainty proxies, whereas z_k is an exactly computable ground-truth signal at training time. Fixing $\alpha = 1$ is unattainable for strongly redundant modality pairs and would turn $\mathcal{L}_{\text{low}}$ into a persistent positive penalty conflicting with $\mathcal{L}_{\text{up}}$. Following equation 6, we estimate α online from batch statistics,

$$\hat{c} = \frac{1}{N} \sum_i c\big(x^{(i)}, y^{(i)}\big), \qquad \hat{\alpha} = \mathrm{clip}\left(1 - \frac{[\hat{c}]_+}{\frac{1}{N}\sum_i \left[\min_k v_k^{(i)}\right]_+ + \delta},\, 0,\, 1\right), \tag{16}$$

smoothed by an exponential moving average. Since interaction information determines only $R - S$ and does not identify R alone (Yu et al., 2024), equation 16 uses $[\hat{c}]_+$ as a lower-bound surrogate for redundancy. The overall objective is,

$$\mathcal{L} = -\log p_f(y \mid x) + \sum_k -\log p_k(y \mid x_k) + \mathcal{L}_{reg} + \lambda_p \mathcal{L}_{bound} + \lambda_r \mathcal{L}_{route}. \tag{17}$$

where $\mathcal{L}_{reg}$ is regression prediction in MSA tasks.

3.4 Theoretical Analysis

This section theoretically analyzes the information budget regularization mechanism of PMIB. Based on pointwise mutual information and pointwise co-information, we examine their respective roles in modality synergy, conflict resolution, and joint-failure suppression. We further explain the rationale behind the adaptive information budget and sample-level fusion strategy, providing a theoretical basis for the effectiveness of PMIB.

Proposition 1 (Synergy amplification). *Let $x \in \mathcal{S}$, $\alpha = 1$, $\tilde{B}_\alpha = B_\alpha$ and $\mathcal{L}_{\text{low}} = 0$. Then*

$$p_f(y \mid x) \;\geq\; e^m \, \frac{p_1(y \mid x_1) p_2(y \mid x_2)}{\pi(y)}, \qquad p_f(y \mid x) \;>\; e^m \max_k p_k(y \mid x_k). \tag{18}$$

Proof. $\mathcal{L}_{\text{low}} = 0$ yields $v_f \geq v_1 + v_2 + m$; exponentiating equation 2 gives the first inequality. Since $x \in \mathcal{S}$ implies $p_k(y \mid x_k)/\pi(y) > 1$, we obtain $p_f/p_1 \geq (p_2/\pi)e^m > e^m$, and symmetrically for $k = 2$. □

The right-hand side of equation 18 is the product-of-experts solution (Hinton, 2002). The synergy constraint therefore targets not "higher confidence" but the additive Bayesian form plus an m-nat gain, giving m an interpretable unit.

Proposition 2 (Conflict routing equals pointwise CMI non-negativity). *Let $x \in \mathcal{C}$. By equation 9, $B_\alpha = v_r$, hence*

$$\mathcal{L}_{\text{low}} = [m - (v_f - v_r)]_+ = [m - \text{pmi}(x_w; y \mid x_r)]_+ \,. \tag{19}$$

If $\mathcal{L}_{\text{low}} = 0$, then

$$p_f(y \mid x) \geq e^m p_r(y \mid x_r) > \pi(y), \qquad -\log p_f(y \mid x) \leq \min_k \big(-\log p_k(y \mid x_k)\big) - m, \tag{20}$$

$$p_f(c \mid x) \leq 1 - e^m p_r(y \mid x_r) \qquad \forall c \neq y. \tag{21}$$

Proof. equation 19 substitutes the definitions; equation 20 exponentiates and uses $v_r > 0$; equation 21 follows from $\sum_c p_f(c \mid x) = 1$. □

Equation equation 19 shows that on $\mathcal{C}$ the lower bound is exactly the pointwise counterpart of equation 7 with a margin, establishing a pointwise–population duality. Equation equation 21 shows that suppressing the misinformative branch requires no separate loss term: the lower bound together with simplex normalisation achieves it. By equation 20, information gain and logarithmic risk reduction coincide under a proper scoring rule (Gneiting & Raftery, 2007), so m simultaneously lower-bounds both.

Proposition 3 (Provable rejection under joint failure). *Let $x \in \mathcal{F}$ and $\mathcal{L}_{\text{up}} = 0$. Then*

$$\max_c p_f(c \mid x) \;\leq\; e^{\varepsilon} \pi_{\max} < 1. \tag{22}$$

Consequently, for any threshold $\tau > e^{\varepsilon}\pi_{\max}$ (equivalently $\tau > e^{\varepsilon}/C$ for balanced classes), every joint-failure sample is rejected under maximum-softmax scoring, and its contribution to the false positive rate is identically zero.

Proof. $x \in \mathcal{F}$ implies $[B_\alpha]_+ = 0$, hence $\Delta_f(c) \leq \varepsilon$ for all c, so $p_f(c \mid x) \leq e^{\varepsilon}\pi(c) \leq e^{\varepsilon}\pi_{\max}$. □

4 Experiments

4.1 Performance Comparison

Multimodal Fusion Methods. Multimodal fusion methods have evolved from tensor-based fusion to attention-based fusion. These methods typically model intermodal interactions and dynamically adjust the contributions of different modalities, achieving strong performance. This demonstrates that modalities can contribute substantially differently to the task label, and that appropriately assigning modality weights can effectively improve fusion performance. Compared with existing multimodal fusion methods, PMIB achieves superior results on most evaluation metrics. This advantage is primarily attributed to its fine-grained modeling of multimodal interaction regimes. Specifically, PMIB distinguishes among modality synergy, modality conflict, and joint failure, and adaptively handles these regimes through a multimodal dynamic routing mechanism.

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
TFN [EMNLP17]	77.99	77.95	34.46	0.947	0.673	78.50	78.96	51.60	0.573	0.714
MulT [ACL19]	79.71	79.63	36.91	0.880	0.702	81.15	81.56	52.84	0.559	0.733
MAG-BERT [ACL20]	82.37	82.50	43.62	0.727	0.781	82.51	82.77	50.41	0.583	0.741
MISA [MM20]	81.84	81.82	41.37	0.777	0.778	83.60	83.80	52.20	0.555	0.756
Self-MM [AAAI21]	83.44	83.36	46.67	0.708	0.796	82.81	82.53	53.46	0.530	0.765
UniMSE [EMNLP22]	85.85	85.83	48.68	0.691	0.809	85.86	85.79	54.39	0.523	0.773
UniSA$_{\text{BART}}$ [MM23]	84.11	-	48.54	0.742	-	84.93	-	50.03	0.587	-
ITHP [ICLR24]	88.70	88.60	-	0.643	0.852	87.30	87.40	-	0.564	0.813
EMOE [CVPR25]	85.40	85.30	47.80	0.697	-	85.50	85.50	53.90	0.530	-
FUSE-Net [CVPR26]	85.83	85.96	49.27	0.688	0.798	85.88	86.14	54.32	0.527	0.772
TSD [CVPR26]	86.50	86.60	49.00	0.691	-	86.20	86.20	54.60	0.525	-
EBMC [CVPR26]	86.26	86.20	50.34	0.694	0.833	86.04	86.23	57.32	0.505	0.824
CACR [CVPR26]	87.04	86.90	48.69	0.710	-	85.37	85.36	54.41	0.532	-
DecAlign [ICLR26]	85.75	85.82	-	0.735	-	86.48	86.07	-	0.543	-
PMIB (Ours)	90.40	90.40	50.73	0.595	0.878	88.17	88.07	54.65	0.542	0.836

Table 1: Performance comparison on the MOSI and MOSEI datasets. The top three results are highlighted as best, second, and third, respectively.

Multimodal Representation Learning Methods. Multimodal representation learning methods primarily model multimodal representations from a geometric perspective. By mining fine-grained shared and modality-specific representations, they aim to improve the expressive capacity of multimodal information. Experimental results show that accurate and discriminative multimodal representations can effectively enhance downstream task performance. Compared with existing multimodal representation learning methods, PMIB also demonstrates strong competitiveness on most evaluation metrics. This is because PMIB characterizes the specific contribution of each modality to the task label from the perspective of sample-level pointwise mutual information. It constraints to guide each modality to preserve more task-relevant information, thereby producing more accurate multimodal representations.

Table 2: Performance comparison on the SIMSV2 dataset.

Model	SIMSV2			
	Acc-2	F1	MAE	Corr
TFN [EMNLP17]	76.51	76.31	0.323	0.667
MulT [ACL19]	79.50	79.59	0.317	0.703
MAG-BERT [ACL20]	79.79	79.78	0.334	0.691
MISA [ACMMM20]	80.53	80.63	0.314	0.725
Self-MM [AAAI21]	79.01	78.89	0.335	0.640
MMIM [EMNLP21]	80.95	80.97	0.316	0.707
AV-MC [ICMI22]	82.50	82.55	0.297	0.732
CMamba [NeurIPS24]	83.40	83.50	0.287	0.758
MSE (Q) [AAAI25]	80.44	80.24	0.311	0.678
MSE (L) [AAAI25]	75.53	75.44	0.382	0.553
MSE (C) [AAAI25]	83.77	83.76	0.296	0.720
FUSE-Net [CVPR26]	79.59	79.53	0.297	0.712
PMIB	86.91	86.90	0.255	0.799

4.2 Ablation Study

To evaluate the effectiveness of each component in PMIB, we remove the pointwise mutual information supervision signal, the lower information budget bound, and the upper information budget bound separately, constructing three variants: 1) **w/o Route**: removing the pointwise mutual information supervision signal, 2) **w/o Low**: excluding the lower information budget bound, and 3) **w/o Up** eliminating the upper information budget bound.

Table 3: Ablation study on three datasets.

Variant	MOSI		MOSEI		SIMSV2	
	Acc-2	F1	Acc-2	F1	Acc-2	F1
PMIB	90.40	90.40	88.17	88.07	86.91	86.90
w/o Route	87.35	87.39	85.00	85.21	84.68	84.63
w/o Up	87.65	87.69	86.32	86.48	84.89	84.82
w/o Low	87.96	87.98	86.67	86.72	84.79	84.82

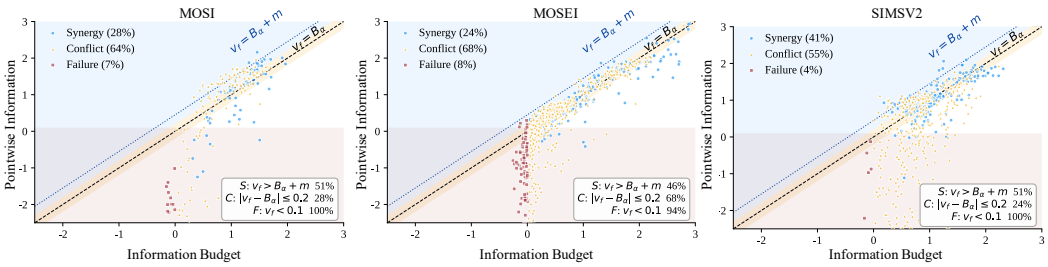


Figure 3: The visualization of the Pointwise Multimodal Information Budget.

As shown in Table 3, all three variants exhibit varying degrees of performance degradation on most evaluation metrics compared with the full model. The performance drop of **w/o Route** indicates that pointwise mutual information supervision effectively guides the router in identifying the sample-level information contributions of different modalities and improves the accuracy of dynamic fusion. The degradation observed for **w/o Low** suggests that the lower information budget bound helps preserve synergistic and complementary information across modalities. The performance drop of **w/o Up** demonstrates that the upper information budget bound suppresses the redundant accumulation of repeated information and the propagation of ineffective information in joint-failure scenarios. These results verify the complementary roles of dynamic routing, the lower budget bound, and the upper budget bound.

4.3 Analysis

RQ1. How does PMIB regulate fusion information under different interaction regimes? To evaluate whether PMIB's pointwise information budget can regulate fused information according to the regime of modality interaction, we plot information budget corridors on the development sets of MOSI, MOSEI, and SIMSv2, as shown in Figure 3. All three datasets exhibit clear regime separation. In the synergy regime, approximately 45.7%∼50.8% of the samples satisfy $v_f > B_\alpha + m$, indicating that fusion obtains additional synergistic information. In the joint-failure regime, 94.3%∼100% of the samples satisfy $v_f < 0.1$, demonstrating that fusion information is effectively suppressed. Samples in the conflict regime are primarily distributed around the budget corridor, with 67.8% of the MOSEI samples falling within the prescribed budget band. These results indicate that PMIB can adaptively regulate information capacity according to different modality interaction regimes, preserving synergistic gains while suppressing ineffective information in joint-failure scenarios.

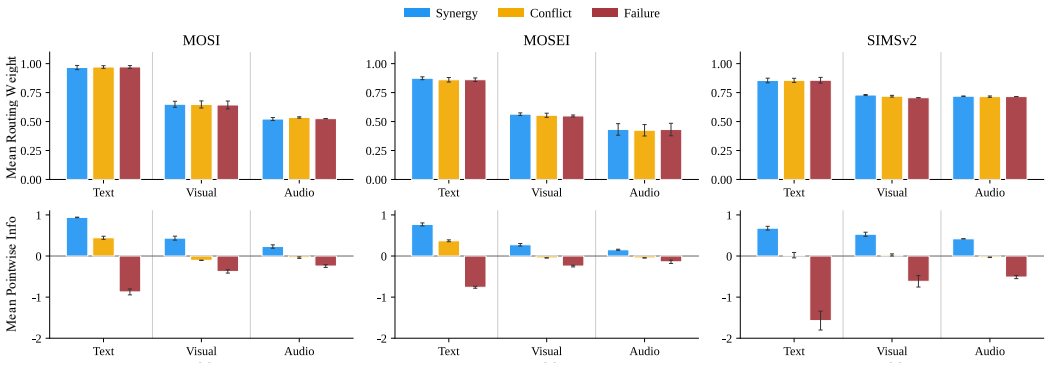


Figure 4: The correlation between the point-wise mutual information and dynamic routing weight.

RQ2. How does PMIB route modalities according to their pointwise information? To evaluate whether PMIB's dynamic routing reflects differences in the information provided by individual modalities, we calculate the average gating weights and pointwise information of each modality under different interaction regimes, as shown in Figure 4. The conflict regime accounts for the

largest proportion of samples across all three datasets (55.5% $\sim$ 67.1%), indicating that partial modality failure is the predominant interaction pattern in multimodal samples. Meanwhile, the overall hierarchy of gating weights is consistent with that of the modality information values, and the Spearman correlations between gating weights and pointwise information are positive on all three datasets ($\rho = 0.226 \sim 0.400$). These results indicate that the routing strategy learned by PMIB can identify the relative information contribution of each modality at the sample level and adjust its participation accordingly, thereby supporting adaptive fusion under different interaction regimes.

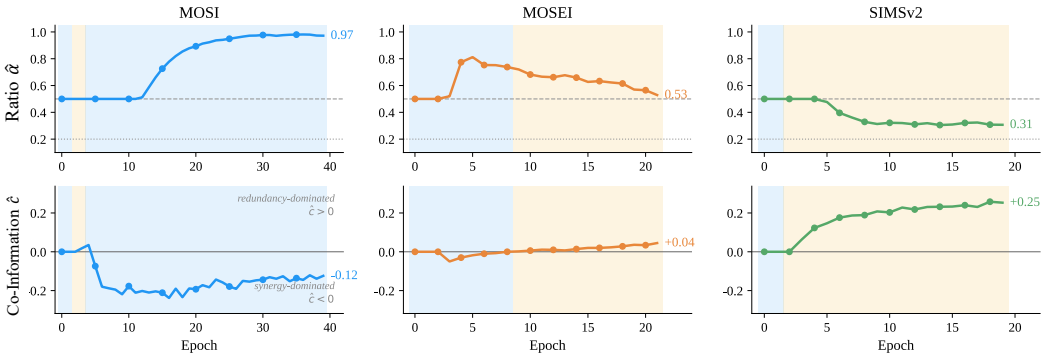


Figure 5: The evolution of the adaptive information budget coefficient α during training.

RQ3. How does PMIB adapt its information budget to dataset-specific redundancy? To evaluate whether the adaptive $\hat{\alpha}$ in PMIB can regulate the information budget according to the synergistic and redundant structures of different datasets, we track the trajectories of $\hat{\alpha}$ and the estimated co-information $\hat{c}$ during training, as shown in Figure 5. Different datasets converge to distinct stable regimes. For MOSI, $\hat{c}$ is approximately -0.12, while α converges to approximately 0.97. For MOSEI, $\hat{c}$ is close to zero, and $\hat{\alpha}$ stabilizes at approximately 0.53. For SIMSv2, $\hat{c}$ is approximately 0.25, and $\hat{\alpha}$ decreases to approximately 0.31. Overall, when synergistic structure is more pronounced, the model tends to relax the estimator to preserve complementary information; when redundancy is higher, it tightens the budget to reduce redundant counting. These results verify that adaptive $\hat{\alpha}$ can dynamically calibrate the budget constraint according to the information structure of each dataset, thereby improving PMIB's adaptability to different multimodal scenarios.

5 Conclusion

This paper proposes PMIB (Pointwise Information Budget Regularization), an information budget regularization method for multimodal learning. Taking pointwise mutual information as its core primitive, PMIB categorizes multimodal interactions into three unified regimes: synergy, conflict, and joint failure. It imposes bilateral constraints through lower and upper information bounds to encourage cross-modal synergistic gains, guide routing toward reliable branches in conflict scenarios, and suppress unsupported information amplification under joint failure. On this basis, PMIB introduces a sample-level multimodal dynamic routing mechanism that uses the signs of pointwise mutual information as training supervision, enabling the model to adaptively reduce the influence of negatively informative modalities during inference. Experimental results on three benchmark datasets, show that PMIB outperforms existing multimodal fusion and representation learning methods, demonstrating the effectiveness of information budget regularization.

6 AI USE STATEMENT

We use generative artificial intelligence to assist in writing, checking for grammatical errors and coding.

REFERENCES

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016.

David Barber and Felix Agakov. Information maximization in noisy channels: A variational approach. *Advances in Neural Information Processing Systems*, 16, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 960–964. IEEE, 2014.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14314–14324, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 14755–14766, 2024.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3394–3402, 2022.

Wei Han, Hui Chen, and Soujanya Poria. 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*, pp. 9180–9192, 2021.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

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

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30183–30193, 2026.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7037–7041. IEEE, 2022a.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022b.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021.

Zhongquan Jian, Wenhan Lv, Yanhao Chen, Guanran Luo, Wentao Qiu, Shaopan Wang, Bingbing Hu, and Qingqiang Wu. Mdf: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31292–31300, 2026.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. Cica: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37842–37851, 2026.

Nakyung Lee, Sangwoo Hong, and Jungwoo Lee. Efficient process reward modeling via contrastive mutual information. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37603–37617, 2026.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832, 2024.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 6132–6142, 2023.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

Ronghao Lin and Haifeng Hu. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 511–523, 2022.

Yadong Liu and Shangfei Wang. Fine: Factorized multimodal sentiment analysis via mutual information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1955–1963, 2026.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022a.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022b.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 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)*, pp. 2247–2256, 2018.

Xiongjian Lv, Yimin Wen, and Hang Yu. Diffufuse: Diffusion-driven dual-stream fusion framework for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8458–8467, 2025.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8791–8800, 2026.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. *arXiv preprint arXiv:2503.11892*, 2025.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2359–2369, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, pp. 6558, 2019.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15880–15889, 2026.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pp. 5240–5252, 2023.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *International Conference on Learning Representations*, volume 2024, pp. 30072–30100, 2024.

Yang Yang, Xunde Dong, and Yupeng Qiang. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25642–25650, 2025.

Zhilu Yang and Mingcheng Li. Factorize, reconstruct, enhance: A unified framework for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15860–15869, June 2026.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10006–10030, 2024.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 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, pp. 10790–10797, 2021.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pp. 2236–2246, 2018.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 756–767, 2023a.

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

Xianbing Zhao, Yinxin Chen, Sicen Liu, and Buzhou Tang. Shared-private memory networks for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(4):2889–2900, 2022a.

Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang. Mag+: An extended multimodal adaptation gate for multimodal sentiment analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4753–4757. IEEE, 2022b.

Xianbing Zhao, Lan Luo, Hengyang Lu, and Buzhou Tang. Prototype-as-prompt: Multimodal sentiment prototypes endowing large language models the capability to perform multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23010–23020, 2026a.

Xianbing Zhao, Shengzun Yang, and Buzhou Tang. Ccaf: Coarse-to-fine cross-modal alignment and fusion for multimodal sentiment analysis. In *Proceedings of the ACM Web Conference 2026*, pp. 7421–7430, 2026b.

Xianbing Zhao, Shengzun Yang, Buzhou Tang, and Ronghuan Jiang. Towards multimodal sentiment analysis via contrastive cross-modal retrieval augmentation and hierachical prompts. *IEEE Transactions on Affective Computing*, 2026c.

A Related Work

Multimodal Fusion. Multimodal fusion integrates heterogeneous modalities through cross-modal interactions to exploit complementary semantics and enhance sentiment understanding. Existing methods mainly explore attention-based interaction fusion (Tsai et al., 2019; Rahman et al., 2020), structure-aware fusion (Hu et al., 2021; Guo et al., 2022) and dynamic modality fusion (Hu et al., 2022a; Feng et al., 2024; Fang et al., 2025) to improve multimodal information integration.

Multimodal Representation Learning. Multimodal representation learning methods can be broadly categorized into three directions: shared-specific representation learning, which disentangles modality-invariant and modality-specific information (Hazarika et al., 2020; Yu et al., 2021; Liu & Wang, 2026; Meng et al., 2026; Yang & Li, 2026); cross-modal alignment learning, which maps heterogeneous modalities into a shared semantic space to improve consistency (Han et al., 2021; Lin & Hu, 2022; Qian et al., 2025; Zhao et al., 2026b); and modality heterogeneity learning, which explicitly models modality differences, noise, and semantic conflicts to reduce heterogeneous interference (Zhang et al., 2023a; Wei et al., 2023; Wang et al., 2026; He et al., 2026; Jiang et al., 2026; Jian et al., 2026).

B THEORETICAL ANALYSIS

B.1 Pointwise Mutual Information

The Eq 2 is the pointwise form of a variational lower bound on mutual information (Barber & Agakov, 2003):

$$\mathbb{E}_{p(x_k,y)}[v_k] = I(Y;X_k) - \mathbb{E}_{x_k}\big[\mathrm{KL}\big(p(\cdot \mid x_k) \,\|\, p_k(\cdot \mid x_k)\big)\big] \;\le\; I(Y;X_k), \tag{23}$$

so $\mathbb{E}[v_k]$ is a conservative estimate of the true mutual information, with a gap equal to the calibration error of the unimodal head (Guo et al., 2017). The divergence in equation 23 characterises the estimation gap only and never enters the training objective: both equation 2 and equation 3 are differences of the logarithmic score, a strictly proper scoring rule (Gneiting & Raftery, 2007). Consequently every *pointwise* statement below is an algebraic fact about model outputs and requires no distributional assumption; only population-level statements rely on equation 23.

B.2 Extension to M Modalities

The proposed method can be naturally extended to M modalities. Let v_i denote the statistic corresponding to the i-th modality, and sort these statistics in nondecreasing order as

$$v_{(1)} \leq v_{(2)} \leq \cdots \leq v_{(M)}.$$

Here, $v_{(M)}$ represents the modality with the largest information contribution for the current sample, while $v_{(j)}$ denotes the relative information contribution of the remaining modalities. Replacing the two statistics in Eq. (10) with the above order statistics yields the information budget for the M-modality setting:

$$B_\alpha = v_{(M)} + \frac{\alpha}{M-1} \sum_{j<M} \left[v_{(j)}\right]_+, \tag{21}$$

C Experiments

C.1 Dataset and Metric Description

We evaluate PMIB on three widely used multimodal benchmarks. MOSI (Zadeh et al., 2016) and MOSEI Zadeh et al. (2018) are sentiment analysis datasets containing 2,199 and 22,856 video segments with sentiment scores ranging from -3 to +3, respectively. SIMSV2 Liu et al. (2022b) contains 4,403 video clips annotated with sentiment scores in [-1,+1]. Following previous works, we adopt dataset-specific evaluation protocols: for MOSI and MOSEI, we report seven-class accuracy (Acc-7), binary accuracy (Acc-2), mean absolute error (MAE), Pearson correlation coefficient (Corr), and F1 score (F1); for SIMSV2, we report binary accuracy (Acc-2), mean absolute error (MAE), Pearson correlation coefficient (Corr), and F1 score (F1).

C.2 Implementation Details

All models are implemented in PyTorch and trained on a single 80 GB GPU. We use AdamW as the optimizer. The textual, acoustic, and visual features are extracted using DeBERTa-V3 (He et al., 2021), COVAREP (Degottex et al., 2014), and OpenFace (Baltrušaitis et al., 2016), respectively. For MOSI, the learning rate, batch size, dropout rate, and number of training epochs are set to 5×10^{-5}, 64, 0.15, and 40, respectively. For MOSEI, these values are set to 2×10^{-5}, 64, 0.2, and 35, respectively. For SIMSv2, they are set to 2×10^{-5}, 32, 0.3, and 30, respectively. In addition, the information budget margin is set to $m = 0.05$, the upper-bound tolerance is set to $\epsilon = 0.5$, and the adaptive budget coefficient is constrained $\hat{\alpha} \in [0, 1]$. The λ_p and λ_r are set to 1. To ensure fair comparisons, we use [1111, 2222, 3333, 4444, 5555] as random seeds for multiple runs and report the averaged results.

C.3 Baselines

To more comprehensively evaluate the performance of PMIB, we compare it with representative state-of-the-art methods, including TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), UniMSE (Hu et al., 2022b), UniSA$_{\text{BART}}$ (Li et al., 2023), ITHP (Xiao et al., 2024), EMOE (Fang et al., 2025), FUSE-Net (Yang & Li, 2026), TSD (Meng et al., 2026), EBMC (He et al., 2026), CACR (Wang et al., 2026), DecAlign (Qian et al., 2025), MMIM (Han et al., 2021), AV-MC (Liu et al., 2022a), CMamba (Li et al., 2024), MSE Yang et al. (2025).