# PIDFUSION: SELECTIVE NON-TEXT SUPERVISION VIA PARTIAL INFORMATION DECOMPOSITION FOR MULTIMODAL SENTIMENT ANALYSIS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal sentiment analysis (MSA) integrates language, acoustic behavior, and visual expression, yet the dominance of language can cause models to underuse informative non-text cues. Applying additional non-text supervision uniformly is also suboptimal because audio and vision do not contribute unique sentiment evidence in every utterance. To address this issue, we propose PIDFusion, a partial-information-decomposition-guided framework that treats text and the joint audio–visual stream as two information sources for the sentiment target. A neural label-information probe combines source and joint label predictors with label-conditioned Sinkhorn transport to characterize redundant, unique, and synergistic information in continuous representations. The probe is optimized on detached features, separating information measurement from predictive representation learning. We then calibrate a sample-wise audio–visual uniqueness score and use it to selectively allocate an auxiliary text-blind regression objective, thereby strengthening non-text representations where they provide distinct sentiment evidence. In parallel, modality-specific and joint audio–visual regression heads provide direct auxiliary supervision to the corresponding representations. Extensive experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS demonstrate the effectiveness of PIDFusion across English and Chinese MSA benchmarks. Ablation studies further show that PID-guided sample selection is more effective than uniform or matched-random allocation and validate the contributions of the proposed training objectives.

## 1 INTRODUCTION

Multimodal sentiment analysis (MSA) predicts affect from language, acoustic behavior, and visual expression. These modalities provide qualitatively different evidence: words convey explicit semantic content, prosody expresses emphasis and uncertainty, and facial behavior can reinforce or contradict the spoken message (Das & Singh, 2023; Liang et al., 2024). This heterogeneity has motivated architectures ranging from tensor and low-rank interaction models to cross-modal attention, modality-specific representation learning, and adaptive expert routing (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Li et al., 2023; Wang et al., 2025; Fang et al., 2025).

Strong language encoders, however, make it possible to achieve a low training loss while only weakly exploiting audio and vision (Yu et al., 2021; Wang et al., 2025). The resulting challenge is not simply to increase non-text utilization, but to determine when non-text cues provide task-relevant evidence beyond the transcript. At the utterance level, audio–visual evidence may repeat what text already conveys, contribute information unavailable from text, become informative only through its interaction with text, or consist primarily of nuisance variation. Uniform non-text supervision treats these distinct cases alike. Attention weights and feature magnitudes describe model operations, while unimodal accuracy summarizes aggregate predictiveness; none directly identifies the type of target information contributed by a source. This motivates our central question: *can the target-information relationship between text and audio–visual cues determine which training examples should receive additional non-text supervision?*



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.

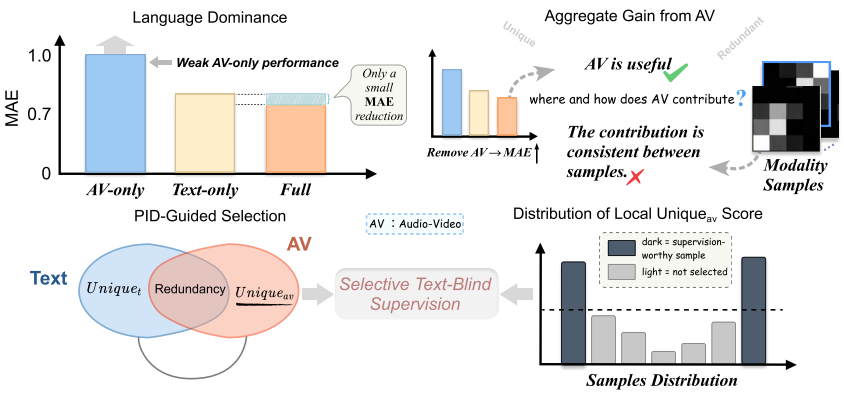



Figure 1: **Schematic motivation for information-aware supervision**. Top: strong text performance and a modest aggregate gain from AV do not identify where or how AV contributes. Bottom: PIDFusion uses an operational text–AV decomposition and allocates an AV-only auxiliary loss when the calibrated AV-unique score exceeds a fixed threshold.

Figure 1 summarizes this motivation schematically. Its top row contrasts language dominance with the aggregate gain from adding AV cues: a strong text pathway can support prediction while AV remains weak in isolation, yet a small full-modality gain still does not reveal which samples benefit from AV or what information it contributes. The bottom row motivates PID-guided selective supervision because samples containing the same modalities need not exhibit the same non-text information structure.

We address this question with PIDFusion, a partial-information-decomposition (PID)-guided framework for selective non-text supervision. Audio and vision are summarized as a joint non-text source $AV$, and a neural probe estimates operational redundant, text-unique, AV-unique, and synergistic quantities with respect to the sentiment target (Williams & Beer, 2010; Liang et al., 2023). The probe combines source and paired-source label predictors with a label-conditioned coupling whose marginals are enforced by Sinkhorn projection. Because it operates on detached features and is optimized separately, the probe observes the learned representations without providing an unrestricted gradient path into the task encoder. The resulting quantities are feature-level mini-batch estimates rather than claims of exact population PID.

PIDFusion turns this diagnostic into a concrete training decision. A one-time controlled injection sweep calibrates the estimator scale, after which a batch-centered local AV-unique proxy provides one score per example. Only examples whose detached, calibrated scores exceed a fixed threshold receive a text-blind regression loss. The auxiliary head predicts sentiment from the same AV representation used by the full model, strengthening that path without altering the full-modality inference interface. PID therefore affects task learning through the allocation of auxiliary supervision, rather than by directly defining the main fusion weight. A matched-random control preserves the expected PID-induced selection rate while changing sample identities, whereas a uniform control applies the auxiliary loss to every sample.

Our contributions are:

- We formulate text and joint audio–visual representations as two PID sources and develop a BATCH-inspired neural probe with label-conditioned Sinkhorn coupling. Detaching the probe inputs prevents its objective from back-propagating into the task encoder, while its scores can still govern auxiliary-loss allocation.
- We convert the AV-unique estimate into a calibrated, sample-wise selection rule that applies text-blind supervision only when its local proxy exceeds a fixed threshold, while retaining standard full-modality inference.
- We evaluate PIDFusion on MOSI, MOSEI, and CH-SIMS within MMSA, a unified open-source framework for MSA. The model achieves competitive results across all three benchmarks, while matched-random, uniform-selection, module, and objective controls on MOSI support the importance of both sample identity and selective non-text supervision.

## 2 Related Work

**Fusion, representation, and modality balance.** Early MSA models capture high-order interactions explicitly or through low-rank factors (Zadeh et al., 2017; Liu et al., 2018), whereas MuLT uses directional cross-modal attention for unaligned sequences (Tsai et al., 2019). Subsequent work separates modality-invariant and modality-specific factors (Hazarika et al., 2020), preserves task-relevant information with hierarchical mutual-information objectives (Han et al., 2021), derives pseudo-labels for unimodal supervision (Yu et al., 2021), or makes fusion explicitly language-guided (Zhang et al., 2023; Wang et al., 2025). Recent approaches further address dataset bias or modality imbalance through counterfactual inference, gradient modulation, and contribution-aware optimization (Yang et al., 2024; Guo et al., 2024; He et al., 2026). These lines of work improve representations, fusion, or optimization; in contrast, PIDFusion uses a task-conditioned information decomposition to decide, sample by sample, where an AV-only auxiliary loss is applied.

**Information decomposition for supervision.** PID decomposes the target information carried by two sources into redundant, source-unique, and synergistic components (Williams & Beer, 2010; Bertschinger et al., 2014). BATCH extends this framework to high-dimensional multimodal distributions with scalable estimators based on marginal-compatible couplings (Liang et al., 2023). PIDFusion builds on its batch-level formulation for text and joint AV representations, using Sinkhorn normalization (Cuturi, 2013). Its main distinction is operational: a calibrated local proxy for AV-unique information, rather than an attention weight or an aggregate dataset statistic, controls selective auxiliary supervision. An expanded review is provided in Appendix A.

## 3 Method

### 3.1 Problem formulation

Each sample consists of a token sequence $x^T$, an acoustic sequence $x^A$, a visual sequence $x^V$, and a sentiment regression target $y \in [-3, 3]$ for MOSI/MOSEI or $y \in [-1, 1]$ for CH-SIMS. We use an unaligned setting: the three sequences have independent lengths and are encoded separately. Let $z_T$, $z_A$, and $z_V$ denote the pooled modality representations. We construct

$$z_{AV} = g_{AV}([z_A; z_V]), \qquad z_J = g_J([z_T; z_A; z_V]), \tag{1}$$

where $z_{AV}$ is the non-text source used by the PID probe and $z_J$ is the joint representation used by the main fusion path.

Let $I(Z; Y)$ denote label information available in representation $Z$. For two sources, PID seeks four atoms satisfying

$$I(Z_T; Y) = R + U_T, \tag{2}$$

$$I(Z_{AV}; Y) = R + U_{AV}, \tag{3}$$

$$I(Z_T, Z_{AV}; Y) = R + U_T + U_{AV} + S. \tag{4}$$

where $R$ is redundant information, $U_T$ and $U_{AV}$ are source-unique information, and $S$ is synergistic information. Equations (2)–(4) define the population-level decomposition that motivates the probe below. Our BATCH-inspired neural approximation operates on mini-batches of learned features; it is not an exact population PID solver. We distinguish its operational estimates from the ideal atoms using hats and evaluate the AV-unique estimate, the only atom used for task training, through matched controls.

Figure 2 gives the complete architecture corresponding to the released three-file implementation. The model contains four stages: modality-specific representation learning, two-source PID probing and calibration, gated multimodal construction, and multi-head regression.

### 3.2 Modality-specific encoders

For text, a fine-tuned BERT encoder (Devlin et al., 2019) first produces contextual token features. All three feature sequences are then projected with modality-specific one-dimensional convolutions

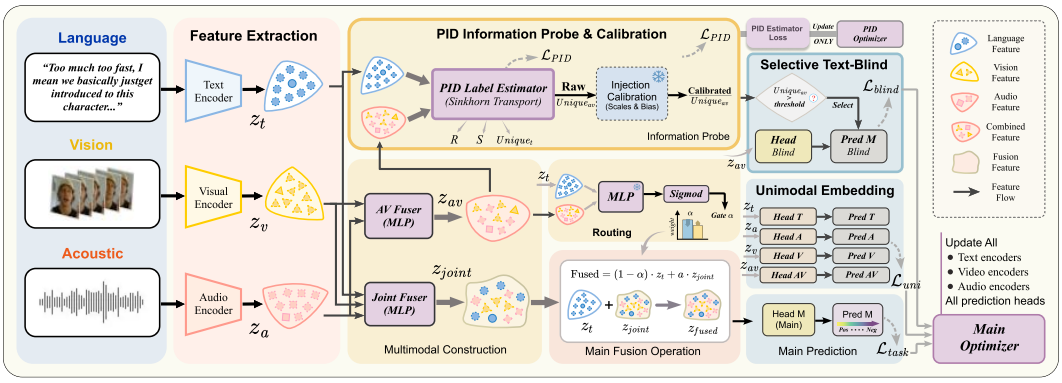



Figure 2: **Overview of PIDFusion**. Text, audio, and vision are encoded independently and pooled into $z_T$, $z_A$, and $z_V$. Audio and vision form $z_{AV}$, while all three modalities form $z_J$. A detached PID probe estimates the two-source information structure and a one-time injection sweep calibrates the unique and synergy scales. The active fusion route uses a frozen sigmoid MLP gate over $[z_T; z_{AV}]$; its output is detached before mixing $z_T$ and $z_J$. The calibrated local unique AV score selects the text-blind auxiliary objective.

and processed by independent Transformer encoders. For an encoded sequence $H_m$, we use a length-aware mean and maximum pool,

$$r_m = [\mathrm{mean}(H_m); \max(H_m)], \qquad z_m = g_m(r_m), \quad m \in \{T, A, V\}, \tag{5}$$

where pooling masks are derived from the text attention mask and the recorded acoustic and visual lengths, capped at the encoded sequence length. The modality mappings $g_m$ and the two fusers $g_{AV}, g_J$ in Equation 1 are LayerNorm–GELU MLPs.

### 3.3 Two-source PID label-information probe

The probe treats $(z_T, z_{AV})$ as two sources and the discretized sentiment label as the target. MOSI and MOSEI use seven bins corresponding to the rounded values $\{-3, \ldots, 3\}$; CH-SIMS uses three bins corresponding to $\{-1, 0, 1\}$. Three MLP predictors produce logits $\ell_T$, $\ell_{AV}$, and $\ell_P$ from $z_T$, $z_{AV}$, and $[z_T; z_{AV}]$, respectively. Their untempered softmax probabilities are $p_T$, $p_{AV}$, and $p_P$; temperature $\tau$ is used only in the probe classification loss. The subscript $P$ denotes the paired-source predictor, which is distinct from the main fusion representation $z_J$.

To obtain a marginal-compatible joint coupling, the probe builds label-conditioned transport embeddings $h_T(i, c)$ and $h_{AV}(j, c)$ and forms

$$a_{ijc} = \exp\left(\frac{h_T(i,c)^\top h_{AV}(j,c)}{\sqrt{d_h}}\right). \tag{6}$$

For a batch of size $B$, the shared class prior is

$$\pi(c) = \frac{1}{2B}\sum_{i=1}^{B}\left[p_T(c \mid z_{T,i}) + p_{AV}(c \mid z_{AV,i})\right]. \tag{7}$$

The class-specific row and column targets are

$$r_{ic} = \pi(c)\frac{p_T(c \mid z_{T,i})}{\sum_{k=1}^{B} p_T(c \mid z_{T,k})}, \qquad t_{jc} = \pi(c)\frac{p_{AV}(c \mid z_{AV,j})}{\sum_{k=1}^{B} p_{AV}(c \mid z_{AV,k})}. \tag{8}$$

Both targets have class mass $\pi(c)$. Alternating row and column projections for $K$ Sinkhorn iterations approximately matches these marginals, followed by global normalization to obtain $q$. The coupled joint-label information is

$$I_q = \sum_{i,j,c} q_{ijc} \log \frac{q_{ijc}}{q_{ij\cdot} q_{\cdot\cdot c}}. \tag{9}$$

Here $q_{ij\cdot} = \sum_c q_{ijc}$ and $q_{\cdot\cdot c} = \sum_{i,j} q_{ijc}$. For target bin $c_i$, the posterior-based label-information estimates are

$$\widehat{I}_m = \frac{1}{B}\sum_{i=1}^{B} \log \frac{p_m(c_i \mid v_{m,i})}{\pi(c_i)}, \qquad m \in \{T, AV, P\}, \tag{10}$$

where $v_{T,i} = z_{T,i}$, $v_{AV,i} = z_{AV,i}$, and $v_{P,i} = [z_{T,i}; z_{AV,i}]$. The operational decomposition is

$$\widehat{U}_T = I_q - \widehat{I}_{AV}, \tag{11}$$

$$\widehat{U}_{AV} = I_q - \widehat{I}_T, \tag{12}$$

$$\widehat{S} = \widehat{I}_P - I_q, \tag{13}$$

$$\widehat{R} = \widehat{I}_T + \widehat{I}_{AV} - I_q. \tag{14}$$

These relations follow the marginal-preserving PID construction (Bertschinger et al., 2014; Liang et al., 2023): $I_q$ is the coupled joint-label information, not redundancy itself. The estimated atoms satisfy the algebraic identities in Equations (2)–(4) when the corresponding posterior estimates are substituted, but finite-batch approximation does not guarantee exact population values or nonnegativity.

Batch-level atoms alone cannot distinguish examples for selective supervision. For target bin $c_i$, define the local AV label-information score

$$\iota_{AV,i} = \log \frac{p_{AV}(c_i \mid z_{AV,i})}{\pi(c_i)}. \tag{15}$$

The code constructs a nonnegative, batch-centered local proxy anchored at the global AV-unique estimate,

$$u^{\mathrm{loc}}_{AV,i} = \left[\widehat{U}_{AV} + \iota_{AV,i} - \frac{1}{B}\sum_{j=1}^{B} \iota_{AV,j}\right]_+ . \tag{16}$$

where $[x]_+ = \max(x, 0)$. Its purpose is to provide a sample-wise selection score, not to assign exact PID bits to individual samples. This distinction is important because local PID is not identifiable from the mini-batch coupling alone.

The probe is applied to detached $z_T$ and $z_{AV}$. Its combined prediction and transport objective is

$$\mathcal{L}_{\mathrm{PID}} = \frac{1}{3}\sum_{m\in\{T,AV,P\}} \mathrm{CE}(\ell_m/\tau, y^{\mathrm{bin}}) + I_q. \tag{17}$$

It is optimized by a separate AdamW optimizer. The cross-entropy terms fit the three label posteriors. For fixed source-label marginals, minimizing $I_q$ is equivalent to maximizing conditional label entropy, as in the BATCH construction. Here the posterior predictors and transport networks are optimized jointly, so this objective provides an approximate rather than an exact constrained PID solution. Consequently, the PID loss updates the probe parameters but does not back-propagate through the main representation encoder.

**Computational cost.** For a mini-batch of size $B$, $C$ sentiment bins, transport width $H$, and $K$ Sinkhorn iterations, the label-conditioned coupling has shape $B \times B \times C$. Its dominant time cost is $\mathcal{O}(B^2CH + KB^2C)$ and its transport memory cost is $\mathcal{O}(B^2C)$. In our experiments, $C \in \{3, 7\}$ and $K = 12$ are small fixed constants. The selective branch reuses the encoded AV representation, so it adds only a lightweight regression head and does not require a second BERT forward pass.

### 3.4 Injection calibration and selective text-blind learning

The absolute scale of a neural PID estimate depends on the representation scale. At the configured calibration epoch, we perform one controlled sweep on up to 512 real training representations. For strengths $\alpha \in \{0.25, 0.5, 1, 2\}$, a fixed label-indexed probe direction is added to $z_{AV}$ and the resulting unique and synergy estimates are measured. The sweep uses synthetic bin indices obtained by rounding evenly spaced values over the dataset's sentiment range, rather than the ground-truth

labels of these training examples. A clipped affine fit maps each raw quantity to the known injection strength:

$$\widehat{u} = \mathrm{clip}(a_u u + b_u, 0, \infty), \qquad \widehat{s} = \mathrm{clip}(a_s s + b_s, 0, \infty). \tag{18}$$

The prescribed strengths serve only as scale-calibration targets; they are not treated as ground-truth PID annotations. The calibration is executed once, immediately before the configured warm-up epoch begins, and the fitted scales are stored as buffers. If the four raw measurements are effectively constant (standard deviation below $10^{-4}$), the implementation retains the identity map instead of amplifying numerical noise. Otherwise, least squares fits the slope and intercept, which are clipped to $[0.1, 10]$ and $[-2, 2]$, respectively. The fitted unique-information map is also applied to the local proxy: $\widehat{u}_i = \mathrm{clip}(a_u u^{\mathrm{loc}}_{AV,i} + b_u, 0, \infty)$.

For a sample $i$, the detached local unique score $\widehat{u}_i$ selects the text-blind objective,

$$b_i = \mathbb{1}[\widehat{u}_i > \delta]. \tag{19}$$

The threshold is applied independently to every example. Here $\delta$ specifies the minimum calibrated AV-unique evidence required to activate auxiliary supervision and is treated as a dataset-specific hyperparameter. Dataset-specific settings are reported in Appendix C. The main forward pass always computes the AV representation. For selected samples, a blind head predicts from $z_{AV}$ and receives the original sentiment label, yielding

$$\mathcal{L}_{\mathrm{blind}} = \frac{\sum_i b_i |\widehat{y}^{\mathrm{blind}}_i - y_i|}{\max(\sum_i b_i, 1)}. \tag{20}$$

The selection is detached, but the blind prediction head and the AV encoder remain trainable. This makes the objective selective rather than a separate AV-only model. The selector and blind loss are used only during training; standard validation and test inference use the full-modality main prediction.

### 3.5 Detached gated fusion and prediction heads

The active fusion route uses a sigmoid MLP,

$$g = \sigma\left(\mathrm{MLP}([z_T; z_{AV}])\right), \qquad \widetilde{g} = \mathrm{stopgrad}(g), \tag{21}$$

and the fused representation is

$$z_F = (1 - \widetilde{g}) z_T + \widetilde{g} z_J. \tag{22}$$

The gate parameters remain at their initialization because its output is detached and no separate gate-training objective is used. Its values still depend on the evolving input representations. Task gradients update the encoders, fusers, and prediction heads through the weighted mixture, but not through the gate. The PID output controls the selective objective rather than this main fusion route. Five regression heads produce $\widehat{y}_T$, $\widehat{y}_A$, $\widehat{y}_V$, $\widehat{y}_{AV}$, and the main prediction $\widehat{y}_M$ from $z_T, z_A, z_V, z_{AV}, z_F$, respectively. Each head is a LayerNorm–GELU MLP followed by a scalar output and a bounded $3\tanh(\cdot/3)$ activation.

### 3.6 Training objective

The task loss combines an L1 term and a Smooth-L1 term with mixing coefficients $\eta_1$ and $\eta_2$ and transition parameter $\beta$:

$$\mathcal{L}_{\mathrm{task}} = \eta_1 \|\widehat{y}_M - y\|_1 + \eta_2 \,\mathrm{SmoothL1}_\beta(\widehat{y}_M, y). \tag{23}$$

The four auxiliary heads share the sentiment target:

$$\mathcal{L}_{\mathrm{uni}} = \frac{1}{4}\left(\|\widehat{y}_T - y\|_1 + \|\widehat{y}_A - y\|_1 + \|\widehat{y}_V - y\|_1 + \|\widehat{y}_{AV} - y\|_1\right). \tag{24}$$

The main optimizer minimizes

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{uni}}\mathcal{L}_{\mathrm{uni}} + \lambda_{\mathrm{blind}}\mathcal{L}_{\mathrm{blind}}, \tag{25}$$

where the task loss has unit coefficient. The PID probe is updated separately using $\lambda_{\mathrm{PID}}\mathcal{L}_{\mathrm{PID}}$. Loss weights are reported with the remaining optimization settings in Appendix C.

**Optimization sequence.** For each mini-batch, the model first computes modality representations and PID outputs. The PID optimizer updates the probe using $\lambda_{\mathrm{PID}}\mathcal{L}_{\mathrm{PID}}$. The main optimizer then updates the encoders, fusers, and regression heads using $\mathcal{L}_{\mathrm{main}}$; all PID scores and routing values entering this step are detached. This ordering preserves a clear gradient boundary while still allowing PID-derived threshold decisions to change which samples contribute to $\mathcal{L}_{\mathrm{blind}}$.

# 4 Experiments

## 4.1 Datasets and Evaluation Metrics

We evaluate on three complete-modality MSA benchmarks. CMU-MOSI contains 2,199 opinion-video segments from 93 English-language videos (Zadeh et al., 2016). The original CMU-MOSEI corpus contains 23,453 annotated segments (Zadeh et al., 2018b); the MMSA split used in our experiments contains 22,856. CH-SIMS provides 2,281 Chinese video segments with both multimodal and modality-specific sentiment annotations (Yu et al., 2020). For all three datasets, we use the train/validation/test partitions and unaligned features distributed with MMSA. English BERT encodes MOSI/MOSEI text, and Chinese BERT encodes CH-SIMS text. Split sizes and feature dimensions are given in Appendix C.

We evaluate continuous sentiment predictions and their discretized classes. For MOSI and MOSEI, we report binary accuracy (Acc-2) and weighted F1 after excluding neutral targets, five- and seven-class accuracy (Acc-5/Acc-7), mean absolute error (MAE), and Pearson correlation (Corr). For CH-SIMS, we report Acc-2, Acc-3, Acc-5, weighted F1, MAE, and Corr using the MMSA evaluation protocol. Accuracy and F1 are percentages; MAE and Corr are reported to three decimal places.

## 4.2 Baselines and Comparison Protocol

We compare against tensor, low-rank, and temporal fusion methods (TFN, LMF, MFN), cross-modal attention (MuLT), modality-specific representation learning (MISA, Self-MM), and language-guided fusion (ALMT, DLF). DMD and EMOE provide additional comparisons from the related multimodal emotion-recognition literature; the CH-SIMS comparison also includes CENET and TETFN. We reproduce Self-MM, ALMT, and EMOE for MOSI/MOSEI, and CENET, TETFN, and ALMT for CH-SIMS, using public implementations in the MMSA setting. The remaining MOSI/MOSEI baseline values are taken from the results compiled by Wang et al. (2025). Table markers distinguish reproduced values from those reported by prior work; a dash denotes a metric unavailable from that source.

## 4.3 Implementation Details

We implement PIDFusion in PyTorch and train it on an NVIDIA RTX 4090D GPU with 24 GB of memory. The representation width is 256 on MOSI/MOSEI and 128 on CH-SIMS; all datasets use eight attention heads, with four Transformer layers on MOSI/MOSEI and three on CH-SIMS. Batch sizes are 16, 32, and 16 for MOSI, MOSEI, and CH-SIMS, respectively. We train for at most 100, 50, and 100 epochs, respectively, and select the checkpoint with the lowest validation loss. Training stops after 10 consecutive epochs without improvement on all three datasets. The main-model learning rates are $10^{-4}$ for MOSI/CH-SIMS and $5 \times 10^{-5}$ for MOSEI. Additional dataset-specific settings are reported in Appendix C.2.

# 5 Results and Analysis

Table 1 summarizes the principal MOSI and MOSEI metrics, and Table 2 presents all six CH-SIMS metrics. Complete MOSI and MOSEI results are provided in Appendix D. The best and second-best available values are shown in bold and underlined, respectively; unavailable entries are excluded when assigning these ranks.

**Results on CMU-MOSI and CMU-MOSEI.** On MOSI, PIDFusion attains 85.37 Acc-2, 85.36 F1, 47.96 Acc-7, and 0.705 MAE, ranking first on all four metrics among the listed methods. Relative to the reproduced EMOE baseline, these results correspond to gains of 0.92, 0.85, and 1.90 percentage

Table 1: Main regression results on CMU-MOSI and CMU-MOSEI. † represents reproduced results from public code released on THUIAR's GitHub page; ‡ represents the result from Wang et al. (2025); and – represents that the result reported in the original paper is not provided. Best results are highlighted in bold, and second-best results are underlined. Higher is better for Acc-2, F1 Score, and Acc-7; lower is better for MAE.

| | CMU-MOSI | | | | CMU-MOSEI | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Acc-2 (↑) | F1 Score (↑) | Acc-7 (↑) | MAE (↓) | Acc-2 (↑) | F1 Score (↑) | Acc-7 (↑) | MAE (↓) |
| TFN‡ (Zadeh et al., 2017) | 80.08 | 80.07 | 34.90 | 0.901 | 82.50 | 82.10 | 50.20 | 0.593 |
| EF-LSTM‡ (Williams et al., 2018) | 78.48 | 78.51 | 35.39 | 0.949 | 80.79 | 80.67 | 50.01 | 0.601 |
| LMF‡ (Liu et al., 2018) | 82.50 | 82.40 | 33.20 | 0.917 | 82.00 | 82.10 | 48.00 | 0.623 |
| MFN‡ (Zadeh et al., 2018a) | 78.87 | 78.90 | 35.83 | 0.927 | 82.85 | 82.85 | 51.34 | 0.575 |
| MuLT‡ (Tsai et al., 2019) | 83.00 | 82.00 | 40.00 | 0.871 | 82.50 | 82.30 | 51.80 | 0.580 |
| MISA‡ (Hazarika et al., 2020) | 83.54 | 83.58 | 41.37 | 0.777 | 84.67 | 84.66 | 52.05 | 0.558 |
| Self-MM† (Yu et al., 2021) | <u>85.06</u> | 85.03 | 46.36 | <u>0.710</u> | 82.22 | 82.41 | 52.76 | <u>0.536</u> |
| ALMT† (Zhang et al., 2023) | 83.23 | 83.22 | 44.31 | 0.743 | 83.93 | 84.10 | 53.29 | 0.542 |
| DMD‡ (Li et al., 2023) | 83.23 | 83.29 | 46.06 | 0.752 | 84.62 | 84.62 | 52.78 | 0.543 |
| DLF‡ (Wang et al., 2025) | <u>85.06</u> | <u>85.04</u> | <u>47.08</u> | 0.731 | **85.42** | <u>85.27</u> | <u>53.90</u> | <u>0.536</u> |
| EMOE† (Fang et al., 2025) | 84.45 | 84.51 | 46.06 | 0.718 | 84.73 | 84.69 | 52.93 | 0.539 |
| PIDFusion (ours) | **85.37** | **85.36** | **47.96** | **0.705** | <u>85.36</u> | **85.32** | **54.20** | **0.529** |

Table 2: Main regression results on CH-SIMS. † represents reproduced results from public code released on THUIAR's GitHub page. Best results are highlighted in bold, and second-best results are underlined. Higher is better for Acc-2, F1 Score, Acc-3, Acc-5, and Corr; lower is better for MAE.

| Model | Acc-2 (↑) | F1 Score (↑) | Acc-3 (↑) | Acc-5 (↑) | MAE (↓) | Corr (↑) |
|---|---|---|---|---|---|---|
| CENET† (Wang et al., 2023b) | 68.93 | 59.88 | 48.80 | 20.68 | 0.582 | 0.123 |
| TETFN† (Wang et al., 2023a) | <u>78.77</u> | <u>78.50</u> | **64.22** | <u>42.34</u> | <u>0.427</u> | 0.584 |
| ALMT† (Zhang et al., 2023) | 78.12 | 78.28 | <u>63.89</u> | 39.82 | <u>0.427</u> | <u>0.585</u> |
| PIDFusion (ours) | **79.43** | **79.21** | 63.68 | **43.11** | **0.412** | **0.589** |

points in Acc-2, F1, and Acc-7, respectively, together with a 0.013 reduction in MAE. On the substantially larger MOSEI benchmark, PIDFusion achieves the best F1 (85.32), Acc-7 (54.20), and MAE (0.529), while its Acc-2 of 85.36 ranks second. It outperforms the reproduced Self-MM, ALMT, and EMOE results on all four metrics. Compared with EMOE, the gains are 0.63 percentage points in both Acc-2 and F1 and 1.27 percentage points in Acc-7, with MAE reduced by 0.010. Relative to the reported DLF result, PIDFusion is 0.06 percentage points lower in Acc-2 but higher in F1 and Acc-7 and lower in MAE. Because the table combines reproduced and previously reported results, comparisons should be interpreted together with the corresponding source markers.

**Results on CH-SIMS.** PIDFusion obtains the best Acc-2, F1 Score, Acc-5, MAE, and Corr among the four methods in Table 2. These results cover binary polarity, fine-grained five-class prediction, regression error, and linear correlation. Its Acc-3 of 63.68 is slightly below TETFN and ALMT, indicating that the gains do not extend to every discretization granularity; nevertheless, PIDFusion is strongest on the other five reported metrics.

### 5.1 Ablation Studies

We evaluate architectural components and auxiliary training regularizers separately. All variants use MOSI with the same data split and evaluation protocol, and the full model is shown in bold. Table 3 reports the module and regularization results together.

The random-matched control separates the selected sample identities from the expected supervision budget. Within each batch, it samples identities at the selection rate induced by the PID rule and therefore matches the expected, rather than the exactly realized, number of selected examples. The full model outperforms this control on all six reported metrics, indicating that the PID selection criterion contributes beyond the expected amount of auxiliary supervision. Uniform selection instead applies the text-blind objective to every training example. Its lower performance, particularly the 3.06- and 1.90-percentage-point reductions in Acc-5 and Acc-7, supports selective rather than indiscriminate allocation of text-blind supervision. Removing the complete text-blind branch produces the largest simultaneous decreases in binary accuracy, binary F1, and Corr. Together, these comparisons support retaining the auxiliary AV branch and using PID-guided sample selection.

Table 3: Ablation study on CMU-MOSI. The module block compares PID-guided selection with matched-random and uniform controls and removes the complete text-blind branch. The regularization block removes one auxiliary training term at a time while retaining the other components. Best values within each block are bold; higher is better except for MAE.

| Module | | | | | | | Regularization | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Setting | Acc-2 (↑) | F1 (↑) | Acc-5 (↑) | Acc-7 (↑) | MAE (↓) | Corr (↑) | Setting | Acc-5 (↑) | Acc-7 (↑) | MAE (↓) |
| Full PIDFusion (ours) | **85.37** | **85.36** | **55.39** | **47.96** | **0.705** | **0.793** | Full PIDFusion (ours) | **55.39** | **47.96** | **0.705** |
| Random-matched selection | 84.45 | 84.50 | 54.37 | 47.52 | 0.723 | 0.787 | W/o $\mathcal{L}_{\text{uni}}$ | 54.37 | 46.65 | 0.713 |
| Uniform selection | 84.60 | 84.63 | 52.33 | 46.06 | 0.742 | 0.791 | W/o $\mathcal{L}_{\text{blind}}$ | 53.21 | 46.21 | 0.709 |
| W/o text-blind module | 83.38 | 83.46 | 52.62 | 46.36 | 0.731 | 0.779 | W/o $\mathcal{L}_{\text{PID}}$ | 54.23 | 47.52 | 0.710 |

In the regularization block, removing $\mathcal{L}_{\text{blind}}$ sets its coefficient to zero while preserving the model architecture, thereby isolating the contribution of its supervision. Removing $\mathcal{L}_{\text{uni}}$ similarly eliminates the direct auxiliary supervision of the modality-specific and joint AV representations. Each removal lowers both reported accuracies and increases MAE: excluding $\mathcal{L}_{\text{blind}}$ causes the largest Acc-5 and Acc-7 reductions, whereas excluding $\mathcal{L}_{\text{uni}}$ causes the largest MAE increase. Although $\mathcal{L}_{\text{PID}}$ is optimized separately rather than added to $\mathcal{L}_{\text{main}}$, it is included here because it trains the auxiliary probe that supplies the sample-selection scores. Removing it therefore alters supervision allocation indirectly and also degrades all three reported metrics.

### 5.2 Hyperparameter Sensitivity

We examine the AV-unique selection threshold $\delta$ on MOSI, varying it over $\{0.01, 0.02, 0.05, 0.10\}$ while fixing all other model and training settings. The threshold defines the sample-wise decision boundary for allocating text-blind supervision. Figure 3 reports both fine-grained classification and regression performance. At $\delta = 0.01$, Acc-7 is 46.36% and MAE is 0.714. The two intermediate settings perform similarly: $\delta = 0.02$ and $0.05$ yield Acc-7 values of 47.67% and 47.96%, respectively, with MAE of 0.704 and 0.705. Thus, the default $\delta = 0.05$ achieves the highest Acc-7 while retaining essentially the same MAE as $\delta = 0.02$. Increasing the threshold to $0.10$ lowers Acc-7 to 45.19% and raises MAE to 0.739. The sweep therefore indicates a stable intermediate range but sensitivity to an overly large threshold. We use $\delta = 0.05$ for the main benchmark results.

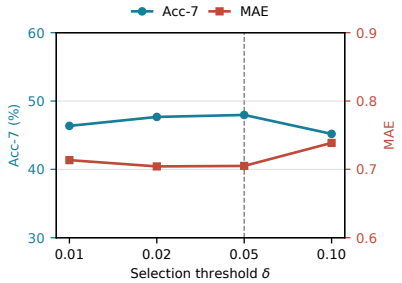



Figure 3: Sensitivity to $\delta$ on CMU-MOSI. Acc-7 (↑) and MAE (↓) use the left and right axes; the dashed line marks $\delta = 0.05$.

## 6 Discussion

The module and regularization controls support the intended role of PID-guided selection: allocating text-blind supervision according to the local AV-unique proxy outperforms matched-random and uniform allocation, while the main input-conditioned fusion gate remains separate from the PID probe. Thus, the task-facing role of PID is to determine where auxiliary non-text supervision is applied; redundancy and synergy remain diagnostic outputs. The learned atoms are treated as operational feature-level estimates rather than exact per-sample information quantities. Appendix B provides an extended discussion of the method, its limitations, reproducibility, and broader impact.

## 7 Conclusion

We presented PIDFusion, a PID-guided architecture for complete-modality sentiment analysis. The model estimates the label-relevant information structure between text and joint audio–visual representations, calibrates a local AV-unique score, and thresholds it to selectively allocate text-blind auxiliary supervision. Across MOSI, MOSEI, and CH-SIMS, PIDFusion achieves competitive regression and classification performance. Ablation studies on MOSI further support the central design choice: performance depends not only on whether non-text supervision is used, but also on which samples receive it. Thus, information decomposition serves as an operational training signal rather than solely as a post-hoc diagnostic.

### AI USE STATEMENT

In preparing this manuscript, we used an LLM-based assistant to improve the readability and wording of existing text in the abstract, introduction, and conclusion; to retrieve and identify related work on information decomposition and established multimodal sentiment-analysis baselines; to check bibliographic details such as publication years and related reference information; and to screen the proposed research direction against the literature for potential overlap. The assistant was not used to generate synthetic datasets, formulate or prove mathematical claims, develop or refine the research hypothesis, design or execute the methodology or experiments, implement the method, clean or reformat datasets, generate or alter experimental results, or draft new sections of the paper. All AI-assisted suggestions were reviewed by the authors, and citations and factual claims were checked against the original sources. The authors take full responsibility for the final manuscript, including all claims, analyses, and reported results.

### ETHICS STATEMENT

This work uses publicly available benchmark datasets, including CMU-MOSI, CMU-MOSEI, and CH-SIMS, under their respective usage terms. No new human data were collected, and we report only aggregate experimental results without attempting to identify individuals. As sentiment-analysis systems can reflect cultural, linguistic, and demographic biases in their data and pretrained representations, the method is intended for research use and should not be used for surveillance, profiling, or consequential decisions without appropriate bias evaluation, privacy review, and human oversight.

### REPRODUCIBILITY STATEMENT

We use the fixed MMSA data partitions and pre-extracted features described in the paper. The main text and Appendix C report the model structure, objective, feature dimensions, optimization settings, training limits, early-stopping rule, and evaluation protocol; the remaining benchmark results and analyses are provided in Appendices D–E and Appendix F. The implementation and dataset-specific configuration files needed to reproduce the reported experiments will be released publicly upon acceptance.

## REFERENCES

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38, 2023. doi: 10.1145/3586075.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14314–14324, 2025. doi: 10.1109/CVPR52734.2025.01335.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133328–133344, 2024. doi: 10.52202/079017-4238.

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. doi: 10.18653/v1/2021.emnlp-main.723.

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. doi: 10.1145/3394171.3413678.

Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. Pase: Prototype-aligned calibration and shapley-based equilibrium for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30960–30968, 2026. doi: 10.1609/aaai.v40i37.40355.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. Msamba: Exploring multimodal sentiment analysis with state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1309–1317, 2025. doi: 10.1609/aaai.v39i2.32120.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10074–10082, 2024. doi: 10.1609/aaai.v38i9.28871.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640, 2023. doi: 10.1109/CVPR52729.2023.00641.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Russ Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1192.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024. doi: 10.1145/3656580.

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. doi: 10.18653/v1/P18-1209.

Alexandra M. Proca, Fernando E. Rosas, Andrea I. Luppi, Daniel Bor, Matthew Crosby, and Pedro A. M. Mediano. Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *PLOS Computational Biology*, 20(6):e1012178, 2024. doi: 10.1371/journal.pcbi.1012178.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2503.11892`.

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 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, Lihuo He, and Xuemei Luo. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259, 2023a. doi: 10.1016/j.patcog.2022.109259.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023b. doi: 10.1109/TMM.2022.3183830.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21180–21188, 2025. doi: 10.1609/aaai.v39i20.35416.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pp. 11–19, 2018. doi: 10.18653/v1/W18-3302.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pp. 464–481, 2024. doi: 10.1007/978-3-031-73636-0_27.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020. doi: 10.18653/v1/2020.acl-main.343.

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. doi: 10.1609/aaai.v35i12.17289.

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. doi: 10.1109/MIS.2016.94.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 5634–5641, 2018a. doi: 10.1609/aaai.v32i1.12021.

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, 2018b. doi: 10.18653/v1/P18-1208.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.49.

# A Extended Related Work

**Fusion and structured representation learning.** Early MSA methods model high-order interactions with tensor fusion (Zadeh et al., 2017), reduce its cost through low-rank factorization (Liu et al., 2018), or maintain cross-view temporal memories (Zadeh et al., 2018a). MuLT instead learns directional cross-modal attention between unaligned sequences (Tsai et al., 2019). Later methods impose more structure on the learned representations: MISA separates invariant and modality-specific factors (Hazarika et al., 2020); MMIM maximizes mutual information hierarchically to preserve task-relevant content through fusion (Han et al., 2021); and Self-MM constructs unimodal pseudo-labels for multi-task supervision (Yu et al., 2021). In the broader multimodal emotion recognition setting, DMD decouples modality-irrelevant and modality-exclusive spaces for graph-based distillation (Li et al., 2023). ALMT suppresses irrelevant or conflicting AV information under language guidance (Zhang et al., 2023), while DLF combines shared–specific disentanglement with a language-focused attractor (Wang et al., 2025). MSAmba represents a more recent architectural direction, combining state-space sequence modeling with cross-modal interaction (He et al., 2025).

**Bias, dominance, and adaptive modality balance.** Recent work increasingly treats modality contribution as an optimization or sample-dependent problem. The MCIS framework uses factual and counterfactual outcomes to mitigate utterance-level label bias and word-level context bias (Yang et al., 2024). CGGM balances multimodal training by modulating both the magnitude and direction of gradients (Guo et al., 2024), whereas PaSE couples prototype-guided calibration with Shapley-based gradient modulation (He et al., 2026). EMOE, proposed for multimodal emotion recognition, learns sample-dependent modality weights and uses unimodal distillation to retain single-modality predictive ability (Fang et al., 2025). DecAlign aligns modality-common and modality-unique representations hierarchically while preserving their distinction (Qian et al., 2026). These methods motivate explicit control over modality contributions, but their weights, gradients, or alignment objectives do not themselves identify PID components. PIDFusion instead uses a label-conditioned text–AV decomposition and routes AV-only supervision according to a calibrated local unique-information proxy.

**Information-theoretic views of multimodal interaction.** Classical PID formalizes how two sources carry target information redundantly, uniquely, or synergistically (Williams & Beer, 2010; Bertschinger et al., 2014). This differs from pairwise or hierarchical mutual-information maximization, which preserves statistical dependence without separately identifying these components (Han et al., 2021). BATCH makes PID estimation practical for high-dimensional multimodal distributions by learning marginal-matching joint distributions at the mini-batch level and applying Sinkhorn normalization (Liang et al., 2023; Cuturi, 2013). It uses the resulting statistics to analyze datasets and models and to support model selection. Related work also uses information decomposition to study how synergy supports modality integration in multi-task neural networks (Proca et al., 2024). Building on the BATCH formulation, PIDFusion has a different operational goal: it derives a local AV-unique proxy from the learned coupling and uses that proxy to control a training intervention rather than treating PID only as an aggregate characterization.

**Auxiliary supervision and task setting.** Self-MM jointly trains multimodal and pseudo-supervised unimodal prediction heads (Yu et al., 2021), while CGGM shows more generally that optimization dynamics can determine whether all modalities are learned effectively (Guo et al., 2024). A separate literature studies robustness when one or more modalities are absent, for example through unified self-distillation under uncertain missingness (Li et al., 2024). That setting is distinct from ours: PIDFusion assumes all three modalities at inference, and its text-blind branch is only a training auxiliary applied to a PID-selected subset. The proposed selection mechanism therefore targets modality use under complete inputs, not reconstruction of or robustness to missing modalities.

# B Extended Discussion, Limitations, and Impact

The module controls support the intended role of PID-guided selection: under the matched-random control, selection based on the local AV-unique proxy performs better across all six reported metrics than random sample identities, while uniform allocation is also inferior to the full model. The regularization results further support the contributions of the text-blind objective and PID supervision.

In the implemented prediction route, only the local AV-unique proxy controls selective supervision; redundancy and synergy remain diagnostic outputs, and the main input-conditioned fusion gate is kept separate from the PID probe. This separation makes the role of each component clear and consistent with the released implementation.

### B.1 Limitations

The two-source formulation deliberately treats audio and vision jointly as a single AV source. It therefore does not distinguish audio-only from vision-only PID atoms, and extending the estimator to three sources may introduce additional identifiability and computational challenges. Moreover, the reported atoms are neural estimates over learned features rather than ground-truth information quantities. The local score in Equation 16 is a nonnegative, centered proxy anchored to a batch-level estimate and should not be interpreted as an exact per-sample number of bits.

The present study focuses on complete-modality sentiment analysis using three widely adopted short-video benchmarks. Its effectiveness in settings involving long-form content, multi-turn dialogue, additional modalities, or substantially different domains remains to be explored. Future work may extend the proposed selection strategy to these settings and investigate its applicability to a broader range of multimodal affective tasks.

### B.2 Reproducibility and Broader Impact

The implementation consists of a model, trainer, and dataset-indexed configuration in MMSA. The appendix reports the data partitions, feature dimensions, gradient boundaries, loss weights, and optimization settings needed to reconstruct the experiments. As with other sentiment-analysis systems, predictions may reflect cultural, linguistic, or demographic biases present in annotations and pretrained representations. The PID estimates describe learned model representations and should not be interpreted as psychological measurements of individuals. Deployment in consequential settings would therefore require appropriate bias evaluation, privacy review, and human oversight.

## C Reproducibility Details

### C.1 Dataset statistics and features

Table 4 summarizes the fixed MMSA data partitions and input feature shapes used in our experiments. Text dimensionality denotes the contextual output width of BERT, whereas the audio and visual dimensions describe the pre-extracted features provided by MMSA and consumed by the corresponding encoders. The listed sequence lengths are the maximum padded lengths. Before length-aware mean/max pooling, the model derives valid text lengths from the BERT attention mask and uses the recorded audio and visual lengths, capped at the corresponding encoded sequence length.

Table 4: Dataset partitions and input feature shapes used by PIDFusion. Dimensions and maximum padded sequence lengths are reported as dim./len.

| Dataset | Train | Valid | Test | Text dim./len. | Audio dim./len. | Visual dim./len. |
|---|---|---|---|---|---|---|
| CMU-MOSI | 1,284 | 229 | 686 | 768 / 50 | 5 / 500 | 20 / 375 |
| CMU-MOSEI | 16,326 | 1,871 | 4,659 | 768 / 50 | 74 / 500 | 35 / 375 |
| CH-SIMS | 1,368 | 456 | 457 | 768 / 39 | 33 / 400 | 709 / 55 |

### C.2 Dataset-specific hyperparameters

Table 5 summarizes the dataset-specific hyperparameters. All three datasets use unaligned inputs and fine-tuned BERT encoders. The main model and PID probe are optimized with separate AdamW optimizers, and the BERT encoder forms a separate parameter group within the main optimizer. The PID optimizer uses a learning rate of $3 \times 10^{-4}$ and is updated every mini-batch; the transport estimator uses 12 Sinkhorn iterations. One-time injection calibration is performed immediately before epoch 3 using at most 512 training representations and strengths $\{0.25, 0.5, 1, 2\}$. Before each main-optimizer step, gradients are clipped to a maximum norm of 0.6.

Table 5: Dataset-specific hyperparameters used by PIDFusion. The final column reports the maximum number of training epochs and the early-stopping patience, respectively.

| Dataset | BERT | Width | Layers | Batch | Main LR | BERT LR | WD | Output Drop. | Grad. Clip | $\lambda_{\text{uni}}$ | $\lambda_{\text{PID}}$ | $\lambda_{\text{blind}}$ | $\delta$ | Max / Early stop |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOSI | `bert-base-uncased` | 256 | 4 | 16 | $10^{-4}$ | $2\times10^{-5}$ | $10^{-3}$ | 0.30 | 0.60 | 0.05 | 0.10 | 0.10 | 0.05 | 100 / 10 |
| MOSEI | `bert-base-uncased` | 256 | 4 | 32 | $5\times10^{-5}$ | $10^{-5}$ | $3\times10^{-3}$ | 0.40 | 0.60 | 0.02 | 0.05 | 0.03 | 0.10 | 50 / 10 |
| SIMS | `bert-base-chinese` | 128 | 3 | 16 | $10^{-4}$ | $2\times10^{-5}$ | $10^{-3}$ | 0.30 | 0.60 | 0.05 | 0.10 | 0.10 | 0.02 | 100 / 10 |

The one-dimensional projection kernels have size five on MOSI/MOSEI and three on CH-SIMS, and all Transformer encoders use eight attention heads. The PID probe uses seven sentiment bins on MOSI/MOSEI and three on CH-SIMS, with a 256-dimensional hidden layer for all datasets. The table reports output dropout. Text-input dropout is 0.30 on MOSI/MOSEI and 0.20 on CH-SIMS; attention dropout is 0.20 for text and 0.10 for audio and vision; activation, embedding, and residual dropout are all 0.10. The weight-decay values in the table apply to the non-BERT main parameters and the PID probe, while BERT uses a weight decay of $10^{-4}$ on all datasets. A ReduceLROnPlateau scheduler with a reduction factor of 0.5 and a patience of five validation epochs is applied to the main optimizer.

### C.3 Gradient audit

The trainer assigns the PID probe and the main model to separate optimizers. Before entering the probe, $z_T$ and $z_{AV}$ are detached; consequently, back-propagating $\mathcal{L}_{\text{PID}}$ updates only the probe's label predictors and transport networks, not the modality encoders or fusers. The main objective does not include $\mathcal{L}_{\text{PID}}$; a detached copy is retained only for logging.

In the opposite direction, the binary selector is computed from detached local AV-unique scores, so the main objective cannot update the PID probe through the selection decision. For selected samples, $\mathcal{L}_{\text{blind}}$ updates the audio and visual encoders, the AV fuser, and the blind regression head. The fusion-gate output is also detached before mixing. Its values remain input-dependent, but neither the task objective nor the PID objective updates the gate parameters. Finally, injection calibration runs without gradient tracking and stores the fitted scale and bias as model buffers rather than trainable parameters.

## D Complete Benchmark Results

Tables 6 and 7 provide the complete evaluation metrics for CMU-MOSI and CMU-MOSEI, respectively. The complete CH-SIMS metrics are reported directly in Table 2.

Table 6: Complete regression results on CMU-MOSI. † represents reproduced results from public code released on THUIAR's GitHub page; ‡ represents the result from Wang et al. (2025); and – represents that the result reported in the original paper is not provided. Best results are highlighted in bold, and second-best results are underlined. Higher is better for Acc-2, F1 Score, Acc-5, Acc-7, and Corr; lower is better for MAE.

| Model | Acc-2 (↑) | F1 Score (↑) | Acc-5 (↑) | Acc-7 (↑) | MAE (↓) | Corr (↑) |
|---|---|---|---|---|---|---|
| TFN‡ (Zadeh et al., 2017) | 80.08 | 80.07 | 39.39 | 34.90 | 0.901 | 0.698 |
| EF-LSTM‡ (Williams et al., 2018) | 78.48 | 78.51 | 40.15 | 35.39 | 0.949 | 0.669 |
| LMF‡ (Liu et al., 2018) | 82.50 | 82.40 | 38.13 | 33.20 | 0.917 | 0.695 |
| MFN‡ (Zadeh et al., 2018a) | 78.87 | 78.90 | 40.47 | 35.83 | 0.927 | 0.670 |
| MuLT‡ (Tsai et al., 2019) | 83.00 | 82.00 | 42.68 | 40.00 | 0.871 | 0.698 |
| MISA‡ (Hazarika et al., 2020) | 83.54 | 83.58 | 47.08 | 41.37 | 0.777 | 0.778 |
| Self-MM† (Yu et al., 2021) | <u>85.06</u> | 85.03 | <u>53.21</u> | 46.36 | <u>0.710</u> | 0.792 |
| ALMT† (Zhang et al., 2023) | 83.23 | 83.22 | 50.58 | 44.31 | 0.743 | 0.784 |
| DMD‡ (Li et al., 2023) | 83.23 | 83.29 | – | 46.06 | 0.752 | – |
| DLF‡ (Wang et al., 2025) | <u>85.06</u> | <u>85.04</u> | 52.33 | <u>47.08</u> | 0.731 | 0.781 |
| EMOE† (Fang et al., 2025) | 84.45 | 84.51 | 52.92 | 46.06 | 0.718 | **0.798** |
| PIDFusion (ours) | **85.37** | **85.36** | **55.39** | **47.96** | **0.705** | <u>0.793</u> |

Table 7: Complete regression results on CMU-MOSEI. † represents reproduced results from public code released on THUIAR's GitHub page; ‡ represents the result from Wang et al. (2025); and – represents that the result reported in the original paper is not provided. Best results are highlighted in bold, and second-best results are underlined. Higher is better for Acc-2, F1 Score, Acc-5, Acc-7, and Corr; lower is better for MAE.

| Model | Acc-2 (↑) | F1 Score (↑) | Acc-5 (↑) | Acc-7 (↑) | MAE (↓) | Corr (↑) |
|---|---|---|---|---|---|---|
| TFN‡ (Zadeh et al., 2017) | 82.50 | 82.10 | 53.10 | 50.20 | 0.593 | 0.700 |
| EF-LSTM‡ (Williams et al., 2018) | 80.79 | 80.67 | 51.16 | 50.01 | 0.601 | 0.683 |
| LMF‡ (Liu et al., 2018) | 82.00 | 82.10 | 52.90 | 48.00 | 0.623 | 0.677 |
| MFN‡ (Zadeh et al., 2018a) | 82.85 | 82.85 | 52.76 | 51.34 | 0.575 | 0.718 |
| MuLT‡ (Tsai et al., 2019) | 82.50 | 82.30 | 54.18 | 51.80 | 0.580 | 0.703 |
| MISA‡ (Hazarika et al., 2020) | 84.67 | 84.66 | 53.63 | 52.05 | 0.558 | 0.752 |
| Self-MM† (Yu et al., 2021) | 82.22 | 82.41 | 54.71 | 52.76 | <u>0.536</u> | 0.760 |
| ALMT† (Zhang et al., 2023) | 83.93 | 84.10 | 54.78 | 53.29 | 0.542 | <u>0.768</u> |
| DMD‡ (Li et al., 2023) | 84.62 | 84.62 | – | 52.78 | 0.543 | – |
| DLF‡ (Wang et al., 2025) | **85.42** | <u>85.27</u> | <u>55.70</u> | <u>53.90</u> | <u>0.536</u> | 0.764 |
| EMOE† (Fang et al., 2025) | 84.73 | 84.69 | 54.75 | 52.93 | 0.539 | 0.764 |
| PIDFusion (ours) | <u>85.36</u> | **85.32** | **56.02** | **54.20** | **0.529** | **0.771** |

# E Extended Analysis Protocols

This section provides the threshold-sweep protocol and additional computational-efficiency statistics that complement the main results.

## E.1 Hyperparameter sensitivity

The analysis in Section 5.2 varies only the AV-unique threshold $\delta$. All other architectural choices, optimizer settings, loss weights, and data splits remain fixed, and the sweep follows the same training and evaluation protocol as the main experiments. Acc-7 and MAE are plotted together in Figure 3 to cover both discretized sentiment intensity and continuous prediction error.

## E.2 Efficiency

The label-conditioned coupling has quadratic complexity in the mini-batch size and linear complexity in the number of target bins and Sinkhorn iterations. To quantify the overall model-level cost, we compare PIDFusion with Self-MM and EMOE on MOSI in Table 8. The profile covers the complete inference forward pass used by each implementation. For PIDFusion, this includes the PID probe, which remains active to expose diagnostic outputs, but excludes the training-only blind head.

Trainable parameters are counted from each instantiated model. FLOPs are measured at batch size one using maximum padded text, audio, and visual lengths of 50, 500, and 375, respectively, with one multiply–add counted as two operations. We use `fvcore` for supported operators and add the packed-LSTM matrix operations analytically. Embedding lookups, indexing, and element-wise nonlinearities are excluded consistently for all models. Epoch time is the median wall-clock interval between consecutive epoch-level training log entries, excluding the interval containing the one-time PID calibration. Because these timings were not obtained in a dedicated throughput benchmark, they provide approximate wall-clock context.

Table 8: Efficiency comparison on CMU-MOSI. FLOPs use batch size one, epoch times are approximate, and Acc-2 follows non-zero binary evaluation. Bold indicates the best predictive result.

| Model | Parameters (M) | FLOPs (G) | Epoch time | Acc-2 (↑) | Acc-7 (↑) |
|---|---|---|---|---|---|
| Self-MM (Yu et al., 2021) | 109.64 | 8.60 | 9.48 s | 85.06 | 46.36 |
| EMOE (Fang et al., 2025) | 317.30 | 9.99 | 13.49 s | 84.45 | 46.06 |
| PIDFusion (ours) | 122.83 | 16.05 | 15.19 s | **85.37** | **47.96** |

PIDFusion uses approximately 12% more parameters than Self-MM and 61% fewer than EMOE. Its complete forward pass nevertheless has the highest FLOP count and the longest approximate epoch

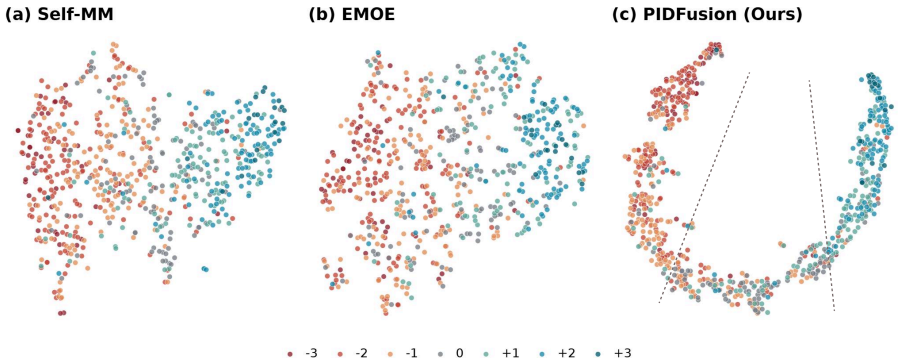



Figure 4: t-SNE visualization of fused MOSI test representations from Self-MM, EMOE, and PIDFusion. All three panels use the same 686 examples, color mapping by gold label, PCA dimensionality, and t-SNE settings; the dimensionality reductions are fitted independently. The PIDFusion projection shows a more continuous progression from negative through neutral to positive sentiment; the visualization is qualitative.

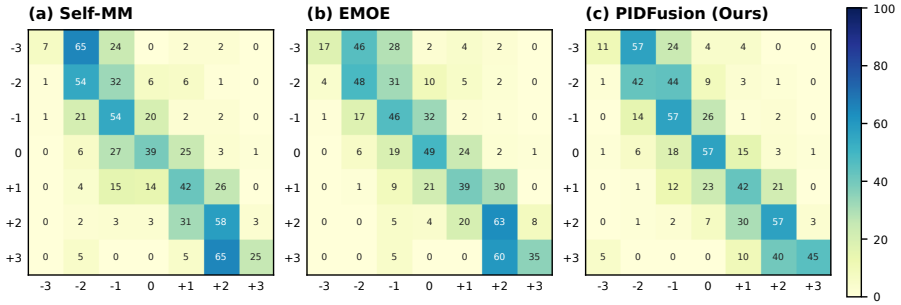



Figure 5: Row-normalized seven-class confusion matrices on the same 686 MOSI test examples for Self-MM, EMOE, and PIDFusion. Rows follow gold labels and columns follow predicted labels, both ordered from $-3$ to $+3$; cell values are percentages. PIDFusion exhibits stronger diagonal concentration across several central and positive intensity levels.

time among the three models. Thus, the gains in Acc-2 and Acc-7 are accompanied by additional computation, while the parameter count remains close to Self-MM and substantially below EMOE.

## F Representation Analysis

We visualize the fused representations learned by Self-MM, EMOE, and PIDFusion on the same MOSI test samples. For each model, the fused feature supplied to its final prediction stage is standardized, reduced to 50 dimensions by PCA, and projected independently using the same t-SNE settings. Gold labels are used only to color the points after projection. As shown in Figure 4, the PIDFusion projection exhibits a visually more continuous ordering from negative to positive sentiment, with neutral examples concentrated near the transition. Because the projections are fitted independently and t-SNE does not preserve global geometry, axes and distances should not be compared directly across panels. We therefore treat the figure as a qualitative diagnostic that complements, rather than substitutes for, the quantitative results in Table 1.

We further examine class-wise behavior through the row-normalized seven-class confusion matrices in Figure 5. PIDFusion achieves an exact seven-class accuracy of 47.96%, compared with 46.36% for Self-MM and 46.06% for EMOE. Relative to both baselines, it has higher diagonal recall for the $-1$, neutral, $+1$, and $+3$ classes; the clearest difference occurs for neutral sentiment, with 57% correctly classified compared with 39% and 49%. For all three models, most misclassifications fall into neighboring intensity levels, consistent with the ordinal, fine-grained nature of MOSI labels rather than systematic polarity reversal.