Tri-Subspaces Disentanglement for Multimodal Sentiment Analysis

Chunlei Meng[1] Jiabin Luo[3] Zhenglin Yan[1] Zhenyu Yu[2]
Rong Fu[4] Zhongxue Gan[1] Chun Ouyang[1*]

[1] The College of Intelligent Robotics and Advanced Manufacturing (CIRAM), Fudan University
[2] The College of Computer Science and Artificial Intelligence, Fudan University
[3] MOE Key Lab of HCST, School of Computer Science, Peking University. [4] University of Macau

clmeng23@m.fudan.edu.cn oy_c@fudan.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) integrates language, visual, and acoustic modalities to infer human sentiment. Most existing methods either focus on globally shared representations or modality-specific features, while overlooking signals that are shared only by certain modality pairs. This limits the expressiveness and discriminative power of multimodal representations. To address this limitation, we propose a Tri-Subspaces Disentanglement (TSD) framework that explicitly factorizes features into three complementary subspaces: a common subspace capturing global consistency, submodally-shared subspaces modeling pairwise cross-modal synergies, and private subspaces preserving modality-specific cues. To keep these subspaces pure and independent, we introduce a decoupling supervisor together with structured regularization losses. We further design a Subspace-Aware Cross-Attention (SACA) fusion module that adaptively models and integrates information from the three subspaces to obtain richer and more robust representations. Experiments on CMU-MOSI and CMU-MOSEI demonstrate that TSD achieves state-of-the-art performance across key metrics, reaching 0.691 MAE on CMU-MOSI and 54.6% ACC_7 on CMU-MOSEI under the unaligned setting, and also transfers well to multimodal intent recognition tasks. Ablation studies confirm that tri-subspaces disentanglement and SACA jointly enhance the modeling of multi-granular cross-modal sentiment cues.

1. Introduction

As human-centered AI applications such as empathetic chatbots, social robots, image understanding systems, and virtual assistants become increasingly pervasive, accurately modeling human emotions is becoming more critical [13–17, 27]. Consequently, Multimodal Sentiment Analysis (MSA) has attracted growing attention in both academia and industry for its ability to interpret emotions by jointly leveraging acoustic, visual, and linguistic modalities [35]. However, MSA remains challenging: different modalities follow heterogeneous feature distributions, emotional cues are often sparse or noisy, and informative signals are not always aligned across modalities. These issues call for more effective ways to represent and fuse complementary information across diverse modalities [11, 24].

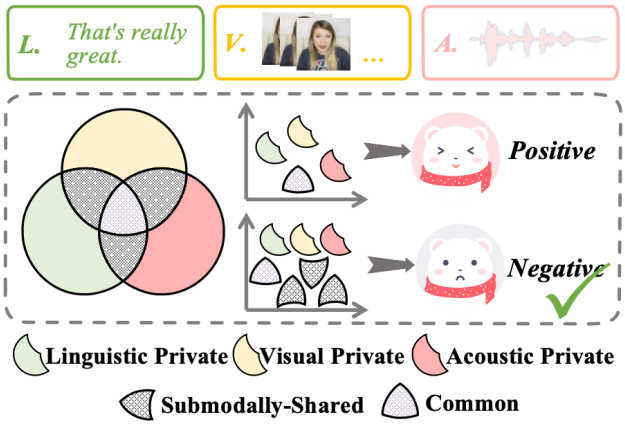

Figure 1. An example of a submodally shared cue: the utterance "that's really great" is delivered with a sarcastic tone and a disdainful facial expression. While the lexical content suggests positive sentiment, the conflicting acoustic and visual cues jointly convey negative affect (sarcasm).

A major line of work addresses these challenges by factorizing features into common (modality-invariant) and private (modality-specific) subspaces. For example, MISA [7] explicitly separates a shared latent space from private spaces capturing each modality's unique aspects. Follow-

*Corresponding Author

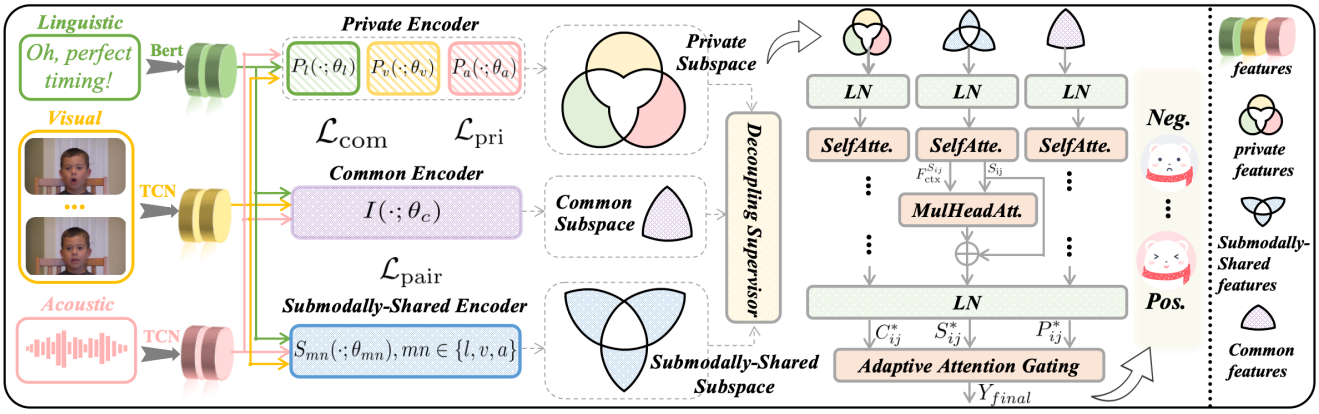

Figure 2. Overview of the proposed TSD framework. Given multimodal inputs, TSD disentangles features into three complementary subspaces and fuses them via a Subspace-Aware Cross-Attention (SACA) module.

up studies such as FDMER [27], FDRL [20], and DLF [24] refine this binary decomposition to obtain more fine-grained disentangled representations. Other approaches employ self-attention and dynamic fusion to model cross-modal interactions and adaptively combine modalities [1, 3, 10, 18], or incorporate contrastive learning and semi-supervised objectives to better exploit data and auxiliary signals [5, 11, 26, 28]. While these methods have advanced MSA performance, they are built on an implicit two-subspace assumption: features are either fully shared across all modalities or entirely private to a single modality, leaving partially shared patterns under-explored.

In practice, many sentiment cues are neither fully global nor strictly modality-specific, but are instead shared across only a subset of modalities. Ignoring these submodally shared signals can lead to incomplete or even misleading sentiment interpretation. As illustrated in Fig. 1, a speaker may say *"That's really great"* with a sarcastic tone and a disdainful facial expression. The linguistic content alone suggests positive sentiment, whereas the combination of vocal and facial cues clearly expresses a negative attitude. Here, the true emotion is revealed by the audio–visual pair: these two modalities share a negative sentiment cue (sarcasm) that is not reflected in the language. Prior works [2, 31] have shown that such intermediate, submodally shared features are semantically meaningful and should be preserved. However, existing common-vs-private frameworks tend to push any signal not present in all modalities into private channels, effectively discarding valuable pairwise or subgroup correlations.

To address this limitation, we propose a **Tri-Subspaces Disentanglement (TSD)** framework that explicitly factorizes the multimodal feature space into three complementary subspaces: (1) a **common subspace** capturing globally consistent information shared by all modalities, (2) **submodally shared subspaces** modeling synergies between specific modality pairs, and (3) **private subspaces** retaining modality-specific characteristics. We introduce a tri-subspace decoupling supervisor together with structured regularization losses to encourage each subspace to remain distinct, i.e., to focus on its designated type of information with minimal leakage. On top of these disentangled representations, we introduce a **Subspace-Aware Cross-Attention (SACA)** module that dynamically attends to and fuses information from the three subspaces, enabling the model to exploit high-level commonalities, intermediate cross-modal cues, and fine-grained modality-specific details within a unified prediction framework. Experiments on CMU-MOSI, CMU-MOSEI, and a multimodal intent recognition benchmark show that TSD achieves state-of-the-art performance and generalizes well across tasks, while ablation studies further verify the contributions of both tri-subspace disentanglement and SACA fusion. Overall, our contribution lies in an architectural and representation design that explicitly models submodally shared signals, rather than in proposing a new theoretical framework.

2. Related Work

Multimodal Sentiment Analysis. MSA aims to recognize emotions by leveraging language, visual, and acoustic signals. Early methods relied on simple feature-level fusion (e.g., concatenation or summation), which struggles to capture complex cross-modal dependencies. Subsequent works introduced tensor-based models such as TFN [31] and attention mechanisms to better model intra- and inter-modal interactions. However, MSA remains challenging due to modality-specific noise, temporal misalignment, and semantic inconsistencies across modalities, which can severely degrade fusion quality.

Disentangled Representation Learning. To mitigate modality heterogeneity, recent approaches adopt disentan-

gled representations that separate shared and modality-specific components [17, 18]. MISA [7], for example, employs a shared-private framework to align modality-invariant features while preserving modality-specific signals, and FDMER [27] further introduces contrastive objectives to strengthen this separation. These methods improve robustness and generalization, but typically impose a binary partition (common vs. private) and overlook partially shared signals. In practice, many sentiment cues (e.g., the intensity of anger in speech and facial expressions) are expressed only through specific modality pairs, which cannot be adequately modeled by purely global or purely private subspaces.

Fusion Strategies. Fusion design is another key component in MSA. Beyond early fusion by concatenation or summation, attention-based fusion, modality-specific transformations, and decision-level fusion with gating have been proposed to better capture cross-modal interactions. Representative methods include EMOE [3], which employs emotion-specific experts, DEVA [26], which uses auxiliary textual prompts for contextual reasoning, and CMAF [27], which performs cross-modal alignment fusion. Most of these methods, however, still assume that modality contributions are either fully shared or independent, implicitly collapsing partially shared signals into private spaces or enforcing global sharing. In contrast, our TSD framework explicitly introduces pairwise shared subspaces between modalities and couples them with the SACA fusion module, which performs subspace-level cross-attention and channel re-weighting over common, pairwise shared, and private representations, enabling a more fine-grained modeling of multimodal sentiment cues.

3. Proposed Method

3.1. Model Overview

Given an utterance U, our goal is to predict its sentiment polarity by jointly modeling three modalities: linguistic (l), visual (v), and acoustic (a). Different from previous methods [3, 24], and as illustrated in Fig. 2, our framework explicitly disentangles multimodal representations into three complementary subspaces: (i) a fully shared (*common*) subspace that captures global semantic cues shared across all modalities, (ii) *submodally shared* subspaces that model interactions present only in subsets of modalities, and (iii) modality-specific (*private*) subspaces that preserve unique and discriminative information for each modality.

Concretely, we first encode U_l, U_v, and U_a into modality-specific high-level representations and project them into a unified feature space. Dedicated decoupling modules then map these representations into the three subspaces. The common subspace aggregates modality-invariant information critical for overall affective inference. The submodally shared subspaces capture associations present in two out of three modalities, enabling the model to exploit pairwise correlations. The private subspaces retain complementary features unique to each modality, ensuring that unimodal cues are not lost. Given the representations from all three subspaces, we fuse their outputs and map the fused representation to either a categorical label ($y \in \mathbb{R}^C$) or a continuous sentiment score ($y \in \mathbb{R}$). The following sections formalize the feature extraction, tri-subspace construction, and their fusion in detail.

3.2. Feature Representation

To capture temporal patterns within each modality, we process the visual and acoustic inputs using independent temporal convolutional networks, while language features are extracted by a pre-trained BERT encoder. For each modality $m \in \{v, a, l\}$, the input sequence is denoted as $\mathbf{X}_m \in \mathbb{R}^{T_m \times d_m^{in}}$, where T_m is the sequence length and d_m is the feature dimension. Each modality-specific encoder produces a sequence of hidden representations:

$$\mathbf{H}_m = \text{Encoder}_m(\mathbf{X}_m; \theta_m^{\text{enc}}) \in \mathbb{R}^{T_m \times d_m}, \tag{1}$$

where θ_m^{enc} are the learnable parameters of the encoder for modality m.

We then project the encoder outputs into a unified feature space via a linear layer:

$$\mathbf{Z}_m = \text{Proj}_m(\mathbf{H}_m) \in \mathbb{R}^{T_m \times d_z}, \tag{2}$$

where Proj_m is a modality-specific fully connected layer and d_z is the shared feature dimension. These unified representations $\mathbf{Z}_m$ serve as the inputs to the tri-subspace disentanglement module.

3.3. Tri-Subspaces Disentanglement

Existing multimodal representation learning methods can capture both global consistency and modality-specific information, but often overlook collaborative signals that are shared only by subsets of modalities. Without explicit modeling, these submodally shared cues may be misattributed to either common or private subspaces, limiting the expressiveness and discriminative power of the learned features. We therefore introduce a tri-subspace disentanglement framework that factorizes multimodal representations into common, submodally shared, and private subspaces, enabling multi-granular modeling of global consistency, pairwise collaboration, and individual modality characteristics. Let $\mathcal{M} = \{l, v, a\}$ denote the modality set, with $|\mathcal{M}| = 3$, and let $\mathcal{P}_2(\mathcal{M})$ be the set of all unordered modality pairs.

3.3.1. Tri-Subspaces Encoders

Common encoder $I(\cdot; \theta_c)$. This encoder extracts modality-invariant, globally consistent features and is shared across

all modalities. It consists of two fully connected layers with GELU activation and LayerNorm. The output is

$$\mathbf{C}_m = I(\mathbf{Z}_m; \theta_c) \in \mathbb{R}^{T_m \times d_c}. \tag{3}$$

Submodally shared encoder $S_{mn}(\cdot; \theta_{mn})$. This encoder captures specific correlations between a modality pair (m, n) and extracts the information shared only between the two modalities. For each unordered modality pair $(m, n) \in \mathcal{P}_2(\{l, v, a\})$, we define a shared-parameter encoder with one fully connected layer followed by a sigmoid activation. Its outputs are

$$\mathbf{S}_{mn}^{(m)} = S_{mn}(\mathbf{Z}_m; \theta_{mn}), \quad \mathbf{S}_{mn}^{(n)} = S_{mn}(\mathbf{Z}_n; \theta_{mn}), \tag{4}$$

where the superscript (m) indicates that the representation is derived from modality m but lives in the subspace shared with modality n. In practice, we construct $\mathbf{S}ij$ by concatenating $\mathbf{S}ij^{(i)}$ and $\mathbf{S}_{ij}^{(j)}$ along the temporal dimension.

Private encoder $P_m(\cdot; \theta_m)$. For each modality, we also define a private encoder composed of one fully connected layer and a sigmoid activation, which extracts modality-specific representations:

$$\mathbf{P}_m = P_m(\mathbf{Z}_m; \theta_m). \tag{5}$$

3.3.2. Decoupling Supervisor

To prevent representation mixing across subspaces, we propose a tri-subspace decoupling supervisor. The supervisor contains three discriminative branches corresponding to the common, submodally shared, and private subspaces, respectively. Each branch is a two-layer perceptron with GELU activation [8], taking the corresponding subspace embedding as input. During training, embeddings from the three subspaces are pooled over time and mixed within a mini-batch, and the supervisor predicts the true source subspace of each embedding.

Let $\rho(\cdot)$ be a temporal pooling operator (mean pooling) that maps a token sequence to an utterance-level vector. We define

$$\mathbf{c}_m = \rho(\mathbf{C}_m), \quad \mathbf{s}_{mn}^{(m)} = \rho(\mathbf{S}_{mn}^{(m)}), \quad \mathbf{p}_m = \rho(\mathbf{P}_m). \tag{6}$$

Let D_{com}, D_{sub}, and D_{pri} denote the probabilities assigned by the supervisor to the "common", "submodally shared", and "private" labels, respectively. The supervision loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{sup}} = -\frac{1}{M} \sum_{m=1}^{M} \Bigg[& \mathbb{E}_{x \sim \mathcal{D}} [\log D_{\text{com}}(\mathbf{c}_m)] \\ & + \sum_{n \neq m} \mathbb{E}_{x \sim \mathcal{D}} \left[\log D_{\text{sub}}(\mathbf{s}_{mn}^{(m)}) \right] \\ & + \mathbb{E}_{x \sim \mathcal{D}} [\log D_{\text{pri}}(\mathbf{p}_m)] \Bigg], \end{aligned} \tag{7}$$

where $\mathcal{D}$ denotes the data distribution (in practice, expectations are approximated by mini-batch averages). Backpropagation through this loss explicitly enforces clear boundaries among the three subspaces and suppresses information leakage. Compared with the modality discriminators used in FDMER [27] and related works, our supervisor jointly discriminates between three subspaces, enforcing stricter disentanglement beyond the shared private dichotomy.

3.3.3. Subspace Loss Functions

To further regularize each subspace and capture multi-granular relationships, we design a set of subspace-specific losses. All losses are computed on utterance-level representations obtained via $\rho(\cdot)$.

Common consistency loss. For the common subspace, we adopt a consistency loss $\mathcal{L}_{\text{com}}$ that encourages alignment of modality-invariant features and reduces distribution gaps across modalities:

$$\mathcal{L}_{\text{com}} = \frac{1}{|\mathcal{P}_2(\mathcal{M})|} \sum_{(m,n) \in \mathcal{P}_2(\mathcal{M})} \left\| \mathbf{c}_m - \mathbf{c}_n \right\|_2^2. \tag{8}$$

Pairwise collaboration loss. To model the pairwise collaboration between modalities, we introduce a pairwise consistency loss $\mathcal{L}_{\text{pair}}$ that encourages submodally shared representations from the two directions to encode only joint information of each modality pair while suppressing modality-specific noise:

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{P}_2(\mathcal{M})|} \sum_{(m,n) \in \mathcal{P}_2(\mathcal{M})} \left\| \mathbf{s}_{mn}^{(m)} - \mathbf{s}_{mn}^{(n)} \right\|_2^2. \tag{9}$$

Private disparity loss. For the private subspaces, we employ the Hilbert–Schmidt Independence Criterion (HSIC) [4] to encourage independence between private representations of different modalities and prevent contamination from shared information. Given two private feature sets $\mathbf{p}_{m_1}$ and $\mathbf{p}_{m_2}$ from a mini-batch of size n, HSIC is defined as

$$\text{HSIC}(\mathbf{p}_{m_1}, \mathbf{p}_{m_2}) = (n-1)^{-2} \operatorname{Tr}(UK_{m_1}UK_{m_2}), \tag{10}$$

where K_{m_1} and K_{m_2} are the $n \times n$ Gram matrices of private features under a linear kernel, $U = I - \frac{1}{n}\mathbf{e}\mathbf{e}^\top$, I is the identity matrix, and $\mathbf{e}$ is an all-ones vector. The overall private disparity loss aggregates HSIC over all modality pairs:

$$\mathcal{L}_{\text{pri}} = \frac{1}{|\mathcal{P}_2(\mathcal{M})|} \sum_{(m_1,m_2) \in \mathcal{P}_2(\mathcal{M})} \text{HSIC}(\mathbf{p}_{m_1}, \mathbf{p}_{m_2}). \tag{11}$$

Orthogonality loss. To further disentangle the three subspaces within each modality, we introduce an orthogonality constraint $\mathcal{L}_{\text{ort}}$ that penalizes overlap among common, sub-

modally shared, and private representations:

$$\mathcal{L}_{\text{ort}} = \frac{1}{M} \sum_{m \in \mathcal{M}} \Big[\big\| \mathbf{C}_m^\top \mathbf{P}_m \big\|_F^2 + \sum_{n \neq m} \big(\big\| \mathbf{S}_{mn}^{(m)\top} \mathbf{P}_m \big\|_F^2 + \big\| \mathbf{S}_{mn}^{(m)\top} \mathbf{C}_m \big\|_F^2 \big) \Big], \tag{12}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

Tri-subspace training objective. The overall objective for tri-subspace disentanglement combines all the above losses:

$$\mathcal{L}_{\text{TS}} = \mathcal{L}_{\text{com}} + \lambda_1 \mathcal{L}_{\text{pair}} + \lambda_2 \mathcal{L}_{\text{pri}} + \lambda_3 \mathcal{L}_{\text{ort}} + \lambda_4 \mathcal{L}_{\text{sup}}, \tag{13}$$

where $\lambda_{1\text{–}4}$ are hyperparameters controlling the contribution of each regularization term.

3.4. Subspace-Aware Cross-Attention Fusion

The three types of feature subspaces (common, submodally shared, and private) capture complementary aspects of multimodal information, and their contributions to the final decision are generally unequal. Simple feature concatenation ignores these structural differences, leading to redundancy and loss of discriminative cues. To address this, we propose a Subspace-Aware Cross-Attention (SACA) module that performs context-aware interaction and adaptive fusion among all subspaces.

Let $\mathbf{C}_m$, $\mathbf{S}_{ij}^{(m)}$, and $\mathbf{P}_m$ be the tri-subspace representations from Sec. 3.3. We first apply self-attention for subspace-wise refinement:

$$\begin{cases} \tilde{\mathbf{C}}_m = \text{SelfAtt}\big(\text{LN}(\mathbf{C}_m)\big), & m \in \mathcal{M}, \\ \tilde{\mathbf{S}}_{ij} = \text{SelfAtt}\big(\text{LN}(\mathbf{S}_{ij})\big), & (i,j) \in \mathcal{P}_2(\mathcal{M}), \\ \tilde{\mathbf{P}}_m = \text{SelfAtt}\big(\text{LN}(\mathbf{P}_m)\big), & m \in \mathcal{M}, \end{cases} \tag{14}$$

where $\mathbf{S}_{ij}$ denotes the subspace shared between modalities i and j; in practice, we construct $\mathbf{S}_{ij}$ by concatenating $\mathbf{S}_{ij}^{(i)}$ and $\mathbf{S}_{ij}^{(j)}$ along the temporal dimension, and LN is layer normalization.

Context construction for cross-subspace attention. For each subspace, we dynamically construct a context set that collects the most relevant complementary signals:

$$\begin{aligned} F_{\text{ctx}}^{\tilde{\mathbf{S}}_{ij}} &= \big[\tilde{\mathbf{S}}_{ij};\ \tilde{\mathbf{C}}_i;\ \tilde{\mathbf{C}}_j;\ \tilde{\mathbf{P}}_i;\ \tilde{\mathbf{P}}_j\big], \\ F_{\text{ctx}}^{\tilde{\mathbf{C}}_m} &= \big[\tilde{\mathbf{C}}_m;\ \{\tilde{\mathbf{S}}_{mj} \mid j \neq m\};\ \tilde{\mathbf{P}}_m\big], \\ F_{\text{ctx}}^{\tilde{\mathbf{P}}_m} &= \big[\tilde{\mathbf{P}}_m;\ \{\tilde{\mathbf{S}}_{mj} \mid j \neq m\};\ \tilde{\mathbf{C}}_m\big], \end{aligned} \tag{15}$$

where $[\,\cdot\,]$ denotes concatenation along the temporal (token) dimension.

Cross-subspace attention enhancement. Taking the submodally shared subspace $\tilde{\mathbf{S}}_{ij}$ as an example, we define

$$Q = W_Q \tilde{\mathbf{S}}_{ij}, \quad K = W_K F_{\text{ctx}}^{\tilde{\mathbf{S}}_{ij}}, \quad V = W_V F_{\text{ctx}}^{\tilde{\mathbf{S}}_{ij}}, \tag{16}$$

where W_Q, W_K, and W_V are learnable projection matrices. The cross-attention output is computed as

$$\mathbf{O} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{17}$$

where d_k is the key dimension. For richer modeling, we use multi-head attention and obtain the enhanced subspace representation via a residual connection:

$$\mathbf{S}_{ij}^* = \text{LN}\big(\tilde{\mathbf{S}}_{ij} + \text{MultiHeadAttn}(\tilde{\mathbf{S}}_{ij}, F_{\text{ctx}}^{\tilde{\mathbf{S}}_{ij}}, F_{\text{ctx}}^{\tilde{\mathbf{S}}_{ij}})\big), \tag{18}$$

where MultiHeadAttn denotes the standard multi-head attention operator. The same procedure is applied to the common and private subspaces to obtain $\mathbf{C}_m^*$ and $\mathbf{P}_m^*$. The residual connections preserve the original information of each subspace while injecting cross-subspace context.

Hierarchical gated fusion. All context-enhanced subspaces are then aggregated:

$$F_{\mathcal{S}} = \big\{\mathbf{C}_l^*, \mathbf{C}_a^*, \mathbf{C}_v^*,\ \mathbf{S}_{la}^*, \mathbf{S}_{lv}^*, \mathbf{S}_{av}^*,\ \mathbf{P}_l^*, \mathbf{P}_a^*, \mathbf{P}_v^*\big\}. \tag{19}$$

A gating network computes the adaptive weight ψ_k for each subspace:

$$\psi_k = \frac{\exp\big(g_k(F_{\mathcal{S}})\big)}{\sum_{k' \in \mathcal{K}} \exp\big(g_{k'}(F_{\mathcal{S}})\big)}, \tag{20}$$

where $g_k(\cdot)$ is a learnable scoring function and $\mathcal{K}$ indexes all subspaces (here $|\mathcal{K}| = 9$). The final multimodal representation is obtained as a weighted sum:

$$\mathbf{Y}_{\text{final}} = \sum_{k \in \mathcal{K}} \psi_k \cdot F_{\mathcal{S}}^{(k)}, \tag{21}$$

where $F_{\mathcal{S}}^{(k)}$ denotes the enhanced output of the k-th subspace. Finally, $\mathbf{Y}_{\text{final}}$ is fed into fully connected layers for sentiment prediction.

3.5. Objective Optimization

For classification tasks, we use the cross-entropy loss, while for regression tasks the mean squared error serves as the task loss $\mathcal{L}_{\text{task}}$. The final optimization objective combines the task loss with the tri-subspace regularization terms:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{TS}}. \tag{22}$$

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate TSD on three widely used multimodal benchmarks. CMU-MOSI [30] contains 2,199 opinion video segments with aligned linguistic, acoustic, and visual features, split into 1,284 training, 229 validation,

Table 1. Performance comparison on CMU-MOSI and CMU-MOSEI. Each cell reports results under aligned/unaligned settings as a/b (a: aligned, b: unaligned). *: values from EMOE [3]. The last row lists the $\pm$ standard deviation of TSD over $n = 5$ random seeds.

Methods	CMU-MOSI				CMU-MOSEI			
	MAE(↓)	ACC_7(%)	ACC_2(%)	F1(%)	MAE(↓)	ACC_7(%)	ACC_2(%)	F1(%)
EF-LSTM* [25]	1.386 / 1.420	33.7 / 31.0	75.3 / 73.6	75.2 / 74.5	0.620 / 0.594	47.4 / 46.3	78.2 / 76.1	77.9 / 75.9
TFN* [31]	0.953 / 0.995	31.9 / 35.3	78.8 / 76.5	78.9 / 76.6	0.574 / 0.573	50.9 / 50.2	80.4 / 84.2	80.7 / 84.0
LMF* [12]	0.931 / 0.963	36.9 / 31.1	78.7 / 79.1	78.7 / 79.1	0.564 / 0.565	52.3 / 51.9	84.7 / 83.8	84.5 / 83.9
MFN* [32]	0.964 / 0.971	35.6 / 34.7	78.4 / 80.0	78.4 / 80.1	0.574 / 0.567	50.8 / 51.3	84.0 / 83.2	84.0 / 83.3
MuLT [22]	0.936 / 0.933	35.1 / 33.2	80.0 / 80.3	80.1 / 80.3	0.572 / 0.556	52.3 / 53.2	82.7 / 84.0	82.8 / 84.0
MISA [7]	0.754 / 0.742	41.8 / 43.6	84.2 / 83.8	84.2 / 83.9	0.543 / 0.557	52.3 / 51.0	85.3 / 84.8	85.1 / 84.8
FDMER [27]	- / 0.725	- / 44.2	- / 84.6	- / 84.7	- / 0.536	- / 53.8	- / 84.1	- / 84.0
ConFEDE [28]	- / 0.742	- / 46.3	- / 84.2	- / 84.2	- / 0.523	- / 54.9	- / 81.8	- / 82.3
Self-MM* [29]	0.738 / 0.724	45.3 / 45.7	84.9 / 83.4	84.9 / 83.6	0.540 / 0.535	53.2 / 52.9	84.5 / 85.3	84.3 / 84.8
DMD* [10]	0.721 / 0.721	46.2 / 46.7	83.2 / 84.0	83.2 / 84.0	0.546 / 0.536	52.4 / 53.1	84.8 / 84.7	84.7 / 84.7
DEVA [26]	- / 0.730	- / 46.3	- / 84.4	- / 84.5	- / 0.541	- / 52.3	- / 83.3	- / 82.9
DLF [24]	- / 0.731	- / 47.1	- / 85.1	- / 85.1	- / 0.536	- / 53.9	- / 84.4	- / 85.3
EMOE* [3]	0.710 / 0.697	47.7 / 47.8	85.4 / 85.4	85.4 / 85.3	0.536 / 0.530	54.1 / 53.9	85.3 / 85.5	85.3 / 85.5
TSD (Ours)	**0.701 / 0.691**	**48.8 / 49.0**	**86.3 / 86.5**	**86.3 / 86.6**	**0.529 / 0.525**	**54.9 / 54.6**	**85.8 / 86.2**	**85.9 / 86.2**
±Std. Dev.	±0.005 / ±0.005	±0.3/±0.3	±0.1/±0.1	±0.3/±0.3	±0.005 / ±0.006	±0.2/±0.2	±0.1/±0.1	±0.4/±0.4

and 686 test samples. CMU-MOSEI [33] includes 22,856 segments, with 16,326 for training, 1,871 for validation, and 4,659 for testing. Both datasets provide sentiment scores in $[-3, 3]$, where -3 and 3 denote strongly negative and strongly positive, respectively. MIntRec [34] is a Multimodal Intent Recognition (MIR) benchmark with 2,224 samples from 20 intent categories and standard train/validation/test splits.

Evaluation Metrics. Following prior work [3], we evaluate MSA on CMU-MOSI and CMU-MOSEI with Acc-2, Acc-7, F1-score, and Mean Absolute Error (MAE). For MIR on MIntRec, we report Accuracy, F1-score, Precision, and Recall.

Implementation Details. All models are implemented in PyTorch and trained on an NVIDIA A100 GPU. We use a batch size of 16, learning rate of 1e-4, a weight decay of 1×10^{-5}, and the Adam optimizer, and train up to 50 epochs with early stopping on validation performance. To reduce randomness, each method is run with $n = 5$ different random seeds, and we report averaged results.

4.2. Comparison with State-of-the-Art Methods

We compare TSD with a broad range of representative baselines, including early fusion methods, attention-based fusion models, and recent disentanglement-based approaches.

Results on MSA benchmarks. We follow the standard practice and evaluate all methods under aligned and unaligned settings. The aligned setting assumes perfect temporal synchronization across modalities, while the unaligned setting simulates more realistic asynchrony. In all tables, we report results as a/b, where a is the aligned score and b is the unaligned score. Table 1 summarizes the results on CMU-MOSI and CMU-MOSEI. TSD consistently achieves the best performance across key metrics and both settings. On CMU-MOSI (aligned), TSD reaches an MAE of 0.701 and an Acc-2 of 86.3%, outperforming the strongest baseline EMOE [3] by 0.009 MAE and 0.9% Acc-2. Under the unaligned setting, TSD further improves MAE to 0.691 and Acc-2 to 86.5%, showing strong robustness to temporal perturbations. We also observe consistent gains in Acc-7 and F1-score, indicating better fine-grained sentiment modeling. On CMU-MOSEI, TSD exhibits similar advantages. In the unaligned setting, TSD achieves an MAE of 0.525 and an Acc-2 of 86.2%, surpassing EMOE [3] by 0.005 MAE and 0.7% Acc-2 while also improving Acc-7 and F1. These improvements demonstrate that simply separating shared and private spaces is not sufficient: explicitly modeling common, submodally shared, and private subspaces, together with SACA fusion, yields more expressive and robust multimodal sentiment representations.

MIR benchmark results. Table 2 reports results on the MIntRec MIR task. TSD outperforms strong baselines on all metrics. In particular, it achieves 73.67% Accuracy and 72.57% F1, clearly outperforming EMOE [3] and other recent methods such as CAGC [21] and GsiT [9]. This confirms that the proposed tri-subspace architecture generalizes beyond sentiment regression to broader multimodal classification tasks.

4.3. Ablation Studies

We conduct ablations along four axes: modality contribution, representation spaces, fusion mechanisms, and regu-

Table 2. Performance comparison on the MIntRec dataset (%).

Method	Accuracy	F1	Precision	Recall
MAG-BERT [19]	70.34	68.19	68.31	69.36
MMIM [6]	71.21	68.70	69.20	68.90
MuLT [22]	72.58	69.36	70.73	69.47
MISA [7]	72.36	70.57	71.24	70.41
CAGC [21]	73.03	70.62	70.86	70.55
EMOE [3]	72.58	70.73	72.08	70.86
GsiT [9]	72.60	69.40	69.40	70.10
TSD (Ours)	**73.67**	**72.57**	**72.86**	**71.91**

Table 3. Ablation studies of TSD on the three benchmarks. For MOSI and MOSEI we report MAE (↓) and ACC_7 (%); for MIntRec we report ACC (%). † denotes the aligned setting for MOSI/MOSEI only.

Model	MOSI		MOSEI		MIntRec
	MAE	ACC_7	MAE	ACC_7	ACC
TSD† (Ours)	**0.701**	**48.8**	**0.529**	**54.9**	–
TSD (Ours)	**0.691**	**49.0**	**0.525**	**54.6**	**73.67**
Importance of Modality					
w/o Linguistic	1.010	35.5	0.830	39.8	54.3
w/o Acoustic	0.810	45.0	0.620	51.2	69.3
w/o Visual	0.850	44.1	0.650	50.0	67.8
Importance of Representations					
w/o Common	0.713	46.7	0.545	53.2	72.0
w/o Private	0.719	46.5	0.547	53.0	71.7
w/o Sub-Shared	0.705	46.9	0.543	53.4	72.3
Non-Disen.	0.722	46.1	0.550	52.6	71.1
Different Fusion Mechanisms					
Sum†	0.724	47.6	0.542	53.8	73.1
Sum	0.713	47.7	0.537	54.0	73.3
Concat†	0.727	47.2	0.544	53.7	72.7
Concat	0.715	47.4	0.538	53.9	73.0
CMAF [27]	0.710	47.5	0.533	53.8	73.0
Importance of Regularization					
w/o $\mathcal{L}_{\text{com}}$	0.719	46.7	0.546	53.0	71.9
w/o $\mathcal{L}_{\text{pair}}$	0.710	47.0	0.539	53.5	72.4
w/o $\mathcal{L}_{\text{pri}}$	0.712	47.2	0.541	53.6	72.6
w/o $\mathcal{L}_{\text{ort}}$	0.708	47.4	0.537	53.7	72.9
w/o $\mathcal{L}_{\text{sup}}$	0.721	46.9	0.548	52.8	71.9
Only Task Loss	0.735	45.8	0.555	52.1	70.6

larization design (Table 3).

Importance of modality. Removing any modality causes a clear performance drop. Excluding the linguistic modality leads to the largest degradation (e.g., MAE 1.010 and ACC_7 35.5 on MOSI), confirming its central role in sentiment and intent understanding. Dropping acoustic or visual inputs leads to milder but still substantial declines on all three benchmarks, showing that prosody and facial expressions provide complementary cues and that three-way fusion is beneficial for robust performance.

Importance of representations. We then ablate the common (global), private (modality-specific), and sub-shared (submodally shared) subspaces by removing their outputs from the fusion and prediction heads while keeping the encoders unchanged. Removing the common subspace harms performance across all benchmarks, indicating the importance of modality-invariant global semantics. Removing private subspaces also degrades performance, reflecting the need to preserve unimodal nuances. Removing the sub-shared space also causes a noticeable drop compared with the full model, particularly on MIntRec, indicating that pairwise shared cues contribute beyond global and private information. The Non-Disen. variant bypasses disentanglement and directly fuses features; although it remains close to some single-subspace variants, it consistently underperforms the full tri-subspace design, supporting the effectiveness of tri-subspace factorization.

Different fusion mechanisms. We compare SACA with several common fusion strategies: simple summation (Sum), concatenation (Concat), and the cross-modal alignment fusion CMAF [27], evaluating both aligned and unaligned variants for Sum and Concat. SACA consistently outperforms these baselines. For example, on MOSI (aligned), replacing SACA with Sum leads to higher MAE and lower ACC_7, and similar degradations are observed with Concat. Compared with CMAF, SACA also achieves lower MAE and higher ACC_7 on both MOSI and MOSEI, with consistent gains on MIntRec. These results indicate that SACA provides a more adaptive and discriminative integration of tri-subspace representations.

Importance of regularization. We study the effect of each regularization term. Removing any structural loss results in noticeable degradation. Dropping the common consistency loss $\mathcal{L}_{\text{com}}$ or the decoupling supervisor loss $\mathcal{L}_{\text{sup}}$ leads to the largest declines, showing their importance for global alignment and subspace purity. Removing pairwise consistency $\mathcal{L}_{\text{pair}}$, private disparity $\mathcal{L}_{\text{pri}}$, or orthogonality $\mathcal{L}_{\text{ort}}$ also harms performance, reflecting their roles in modeling pairwise interactions, preventing information leakage, and reducing redundancy. When all structural regularizers are removed and only the task loss is used (Only Task Loss), TSD degenerates into a generic fusion model and achieves the lowest performance on all three benchmarks, confirming the necessity of our regularization design.

4.4. Further Analysis

Qualitative analysis. To qualitatively examine the effect of submodally shared subspaces, we visualize prediction trajectories on CMU-MOSI. For utterances such as *"Oh,*

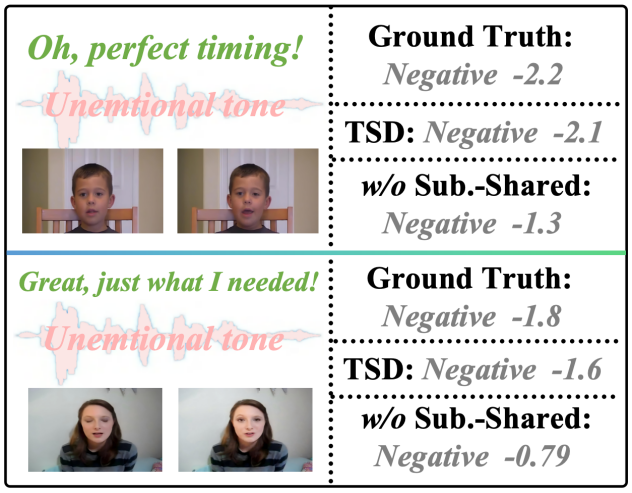

Figure 3. Qualitative examples illustrating that incorporating the submodally shared subspace enables TSD to better capture cross-modal cues (e.g., sarcasm), producing sentiment predictions closer to the ground truth.

perfect timing!", where the lexical content alone appears positive or nearly neutral but sarcasm is conveyed by tone and facial expression, variants that only use common and private subspaces tend to underestimate the negativity. In the w/o Sub-Shared variant, audio–visual cues are either over-smoothed in the common subspace or confined to private subspaces with limited influence on the final prediction, leading to clear deviations from the ground-truth score. With explicit sub-shared subspaces, TSD can preserve these cross-modal interactions and allow SACA to emphasize them when they are informative, yielding trajectories that follow the reference more closely. Similar behavior is observed in examples such as *"Great, just what I needed."*, where non-verbal signals dominate.

Visualization of feature distributions. We also visualize learned representations on CMU-MOSI using t-SNE [23], shown in Fig. 4. The variant without SACA and Sub-Shared produces scattered, irregular distributions with weak sentiment structure. In contrast, the full TSD model yields a more compact and continuous gradient from negative to positive sentiment. Since MSA is a regression task, such gradient-like patterns are more desirable than discrete clusters, suggesting that TSD learns more coherent and semantically ordered representations.

Visualization of subspace weights and importance. Finally, we analyze how TSD allocates attention across subspaces by visualizing learned fusion weights and estimated subspace contributions on MOSI and MOSEI (Fig. 5). On MOSI, the common and private subspaces receive higher average weights, whereas on MOSEI the weights are more balanced, reflecting the richer cross-modal patterns and the adaptive nature of SACA. The contribution analysis further

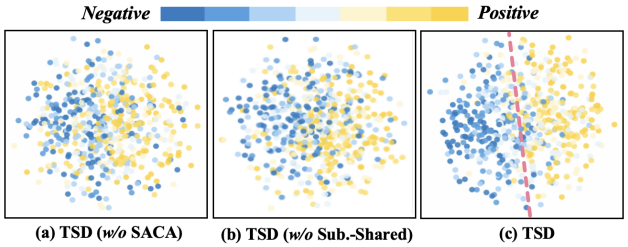

Figure 4. t-SNE visualization of feature distributions on CMU-MOSI. Colors indicate sentiment polarity from negative (dark) to positive (yellow). The full TSD model exhibits a clearer sentiment gradient than the variant without SACA and sub-shared subspaces.

	CMU-MOSI				CMU-MOSEI			
Com.	0.28	0.30	0.33	0.39	0.32	0.30	0.31	0.23
Pri.	0.35	0.39	0.52	0.30	0.56	0.34	0.40	0.39
Sub.	0.37	0.31	0.15	0.31	0.12	0.36	0.29	0.38

(a) *Subspace Weights*

Com.	0.17	0.23	0.19	0.63	0.20	0.05	0.26	0.14
Pri.	0.25	0.56	0.71	0.15	0.63	0.20	0.51	0.46
Sub.	0.58	0.21	0.10	0.22	0.17	0.75	0.23	0.40

(b) *Subspace Importance*

Figure 5. (a) Average fusion weights of each subspace (Common, Private, Sub-Shared) in TSD on CMU-MOSI and CMU-MOSEI. (b) Estimated contribution of each subspace during fusion. Higher values and darker colors denote greater subspace significance.

shows that TSD tends to rely on private subspaces when one modality is dominant, and on common or Sub-Shared subspaces when multimodal cues are complementary. The strong agreement between learned weights and actual contributions supports both the effectiveness and interpretability of the tri-subspace design.

5. Conclusion

This study presented the Tri-Subspace Disentanglement (TSD) framework to address submodally shared signals in multimodal sentiment analysis. TSD factorizes multimodal features into common, submodally shared, and private subspaces, and uses a decoupling supervisor with structured regularization to keep them distinct, enabling unified modeling of global, intermediate, and modality-specific cues. Built on these representations, a subspace-aware cross-attention module adaptively fuses information across subspaces. Experiments on MSA and MIR benchmarks show that TSD achieves state-of-the-art performance and transfers well to related tasks. In the future, we will expand TSD to a wider range of human-computer interaction scenarios.

References

[1] Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. D2R: Dual-Branch Dynamic Routing Network for Multimodal Sentiment Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3536–3547, 2024. 2

[2] Weichen Dai, Xingyu Li, Zeyu Wang, Pengbo Hu, Ji Qi, Jianlin Peng, and Yi Zhou. MInD: Improving Multimodal Sentiment Analysis via Multimodal Information Disentanglement. *arXiv preprint arXiv:2401.11818*, 2024. 2

[3] 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 (CVPR)*, pages 14314–14324, 2025. 2, 3, 6, 7

[4] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, pages 63–77, 2005. 4

[5] Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 133328–133344, 2024. 2

[6] 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 (EMNLP)*, pages 9180–9192, 2021. 7

[7] 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 (ACM MM)*, pages 1122–1131, 2020. 1, 3, 6, 7

[8] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[9] Yijie Jin, Junjie Peng, Xuanchao Lin, Haochen Yuan, Lan Wang, and Cangzhi Zheng. Multimodal Transformers are Hierarchical Modal-wise Heterogeneous Graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pages 2188–2209, 2025. 6, 7

[10] 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 (CVPR)*, pages 6631–6640, 2023. 2, 6

[11] Jinhao Lin, Yifei Wang, Yanwu Xu, and Qi Liu. Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1411–1419, 2025. 1, 2

[12] 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) (ACL)*, pages 2247–2256, 2018. 6

[13] Chunlei Meng, Jiacheng Yang, Wei Lin, Bowen Liu, Hongda Zhang, Zhongxue Gan, and Chun Ouyang. CTA-Net: A CNN-Transformer Aggregation Network for Improving Multi-Scale Feature Extraction. *arXiv preprint arXiv:2410.11428*, 2024. 1

[14] Chunlei Meng, Wei Lin, Bowen Liu, Hongda Zhang, Zhongxue Gan, and Chun Ouyang. RTS-ViT: Real-Time Share Vision Transformer for Image Classification. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3576–3586, 2025.

[15] Chunlei Meng, Wei Lin, Jiacheng Yang, Yi Liu, Hongda Zhang, Yuning Chen, Bowen Liu, Ziqin Zhou, Chun Ouyang, Zhongxue Gan, Dunzhao Wu, and Zhihua Nie. CF-ViT: Cross-Feature Vision Transformer for Improving Feature Learning on Tiny Datasets. In *Proceedings of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 6919–6926, 2025.

[16] Chunlei Meng, Jiacheng Yang, Wei Lin, Linqiang Hu, Bowen Liu, Zhuo Zou, LiDa Xu, Zhongxue Gan, and Chun Ouyang. Multi-grained teacher–student joint representation learning for surface defect classification. *Journal of Industrial Information Integration*, 48:100958, 2025.

[17] Chunlei Meng, Guanhong Huang, Rong Fu, Runmin. Jian, Zhongxue Gan, and Chun Ouyang. CLCR: Cross-Level Semantic Collaborative Representation for Multimodal Learning. *arXiv preprint arXiv:2602.19605*, 2026. 1, 3

[18] Chunlei Meng, Ziyang Zhou, Lucas He, Xiaojing Du, Chun Ouyang, and Zhongxue Gan. Temporal-Spatial Decouple before Act: Disentangled Representation Learning for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2601.13659*, 2026. 2, 3

[19] 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 (ACL)*, pages 2359–2369, 2020. 7

[20] Haoqin Sun, Shiwan Zhao, Xuechen Wang, Wenjia Zeng, Yong Chen, and Yong Qin. Fine-Grained Disentangled Representation Learning For Multimodal Emotion Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11051–11055, 2024. 2

[21] Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26963–26973, 2024. 6, 7

[22] 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 (ACL), pages 6558–6569, 2019. 6, 7

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, pages 2579–2605, 2008. 8

[24] 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 (AAAI)*, pages 21180–21188, 2025. 1, 2, 3, 6

[25] 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)*, pages 11–19, 2018. 6

[26] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1601–1609, 2025. 2, 3, 6

[27] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1642–1651, 2022. 1, 2, 3, 4, 6, 7

[28] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7617–7630, 2023. 2, 6

[29] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10790–10797, 2021. 6

[30] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, pages 82–88, 2016. 5

[31] 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 (EMNLP)*, pages 1103–1114, 2017. 2, 6

[32] 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 (AAAI)*, 2018. 6

[33] 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 (ACL)*, pages 2236–2246, 2018. 6

[34] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1688–1697, 2022. 6

[35] 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 (EMNLP)*, pages 756–767, 2023. 1