electronics

MDPI

Article

DMBT Decoupled Multi-Modal Binding Transformer for Multimodal Sentiment Analysis

Rui Guo [1], Gu Gong [1,*] and Fan Jiang [2]

[1] School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China; 3020234238@jsnu.edu.cn
[2] School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; jiangfan@xupt.edu.cn
* Correspondence: gchxtxy@jsnu.edu.cn

Abstract

The performance of Multimodal Sentiment Analysis (MSA) is commonly hindered by two major bottlenecks: the complexity and redundancy associated with supervised feature disentanglement and the coarse granularity of static fusion mechanisms. To systematically address these challenges, a novel framework, the Decoupled Multi-modal Binding Transformer (DMBT), is proposed. The framework first introduces an Unsupervised Semantic Disentanglement (USD) module, which resolves the issue of complex redundancy by cleanly separating features into modality-common and modality-specific components in a lightweight, parameter-free manner. Subsequently, to tackle the challenge of coarse-grained fusion, a Gated Interaction and Fusion Transformer (GIFT) is constructed as the core engine. The exceptional performance of GIFT is driven by two synergistic components. The first is a Multi-modal Binding Transposed Attention (MBTA) that employs a hybrid convolutional and attention model to concurrently perceive both global context and local fine-grained features, and then a Dynamic Fusion Gate (DFG) that performs final, adaptive decision-making by re-weighting all deeply enhanced representations. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmarks demonstrate that the proposed DMBT framework surpasses existing state-of-the-art models across all key evaluation metrics. The efficacy of each innovative component is further validated through comprehensive ablation studies.

Keywords: multimodal sentiment analysis; feature disentanglement; transformer; cross-modal fusion; attention mechanism

check for updates

Academic Editors: Arkaitz Zubiaga, Alberto Postiglione, Francesco Orciuoli, Giuseppe Fenza and Mario Monteleone

Received: 29 September 2025
Revised: 27 October 2025
Accepted: 29 October 2025
Published: 31 October 2025

Citation: Guo, R.; Gong, G.; Jiang, F. DMBT Decoupled Multi-Modal Binding Transformer for Multimodal Sentiment Analysis. *Electronics* **2025**, *14*, 4296. https://doi.org/10.3390/electronics14214296

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Multimodal Sentiment Analysis (MSA) emerges as a pivotal tool for interpreting cross-modal affective semantics, demonstrating critical value in applications such as mental health monitoring and immersive interactive systems [1]. As shown in Figure 1, unlike unimodal approaches, MSA enhances the robustness of sentiment understanding by fusing heterogeneous information from language, vision, and audio [2]. In recent years, the Transformer architecture, owing to its powerful context-modeling capabilities, has become the mainstream framework for this task, with seminal works like MulT [3] successfully addressing challenges such as multimodal asynchrony.

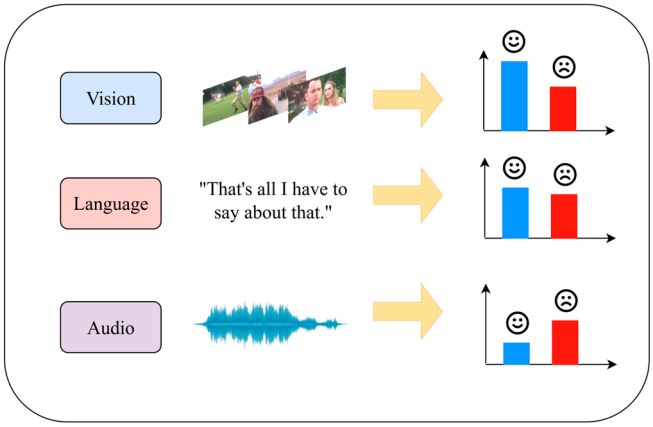


Figure 1. An illustration of multimodal sentiment analysis. Different modalities can yield varying sentiment interpretations for the same event.

Despite this success, Transformer-based approaches in MSA are increasingly confronted with a critical dilemma balancing performance with efficiency. On one hand, many existing fusion strategies suffer from parameter redundancy and computational inefficiency. As highlighted by recent studies, the parameter count in many fusion models can grow exponentially with the number of modalities, severely impeding their training and practical deployment [4]. To facilitate more granular interaction, methods that separate features into modality-common and modality-specific components have become prevalent. However, this has introduced another layer of complexity, as mainstream paradigms rely heavily on cumbersome supervised learning. This approach not only results in convoluted model architectures and training instability but also fails to guarantee the purity of the separated features, thereby compromising the efficacy of subsequent fusion.

Concurrently, the rise in Multimodal Large Language Models (MLLMs) has introduced a new paradigm, demonstrating remarkable flexibility in generating personalized, high-quality content by integrating text and vision [5]. However, their widespread adoption is currently hindered by prohibitive computational costs and the need for massive-scale datasets, making them impractical for many real-world applications [6]. This creates a clear and pressing research gap: there is an urgent need for a framework that strikes a balance between performance, efficiency, and accessibility. Such a framework must achieve lightweight, unsupervised feature disentanglement while simultaneously performing intelligent, dynamic feature fusion, all without relying on the colossal resources required by MLLMs.

To fill this void, this paper proposes a novel Decoupled Multi-modal Binding Transformer (DMBT) framework. To address the dilemma of disentanglement efficiency and purity, an Unsupervised Semantic Disentanglement (USD) module is first introduced. This module discards complex supervised paradigms to achieve clean feature separation in a lightweight, parameter-free manner. Subsequently, to overcome the limitations of coarse and static fusion, a Gated Interaction and Fusion Transformer (GIFT) is constructed. Through its core components—the Multi-modal Binding Transformer Attention (MBTA) and the Dynamic Fusion Gate (DFG)—GIFT synergistically resolves the challenges of fine-grained feature extraction and dynamic decision-making.

The main contributions of this work are summarized as follows:

- A novel, end-to-end framework for MSA, termed DMBT, is proposed. It systematically addresses the limitations of existing methods through two core innovative modules for feature disentanglement and fusion.
- An efficient Unsupervised Semantic Disentanglement (USD) module is designed and validated. It integrates an unsupervised, parameter-free paradigm with a language-

focused framework, significantly enhancing model efficiency, stability, and disentanglement purity.

- A new Gated Interaction and Fusion Transformer (GIFT) is constructed, which serves as the core engine of the DMBT model. Its superior performance stems from the synergy between its two integral components: the Multi-modal Binding Transposed Attention (MBTA) and the Dynamic Fusion Gate (DFG).

2. Related Work

2.1. Multimodal Fusion Strategies

The core challenge in MSA is the effective fusion of heterogeneous information [7–9]. Foundational research established early (feature-level) and late (decision-level) fusion paradigms, utilizing methods from simple concatenation to more complex tensor fusion to model high-order interactions [9–12]. Although foundational, these approaches exhibit limitations in managing modal heterogeneity, noise, and dynamic temporal semantics.

To address these limitations, Transformer-based methods have become the mainstream due to their powerful sequence modeling capabilities. MulT [3] stands as a milestone, enabling end-to-end modeling of temporally unaligned sequences via a directional cross-modal attention mechanism. Subsequent research has refined this framework from various perspectives, such as cyclical modality translation [13], hierarchical modeling, and knowledge distillation [14,15]. More recently, a notable trend has been the exploration of language-centric fusion, where the textual modality serves as an anchor to guide the integration of information from other modalities [16–18].

However, the increasing sophistication of these models has highlighted a critical trade-off with efficiency, spurring a trend toward lightweight multimodal Transformers. This concern is not unique to MSA. For instance, in medical imaging, MACTFusion proposes a lightweight cross-Transformer that uses a novel cross-axis attention to reduce computational complexity [19]. By employing a cycle-attention mechanism, Cy-Atten achieves a linear O(N) complexity for adding new modalities, demonstrating the feasibility of creating highly efficient yet powerful fusion models [4].

This landscape reveals two primary limitations that our work addresses. First, most specialized frameworks still fuse raw, coupled features, limiting performance. Second, there is a clear need for models that are both performant and computationally efficient. These are the precise challenges that the proposed DMBT framework aims to resolve.

2.2. Disentangled Representation Learning

To address modality heterogeneity at a deeper level, disentangled representation learning, pioneered by Tsai et al. [20], has become a key research area. The goal is to decompose features into modality-invariant (shared) and modality-specific subspaces before fusion.

One major line of work is supervised learning based on constraint optimization. These methods typically employ parallel encoders and a suite of meticulously designed loss functions (e.g., similarity, difference, reconstruction, triplet loss) to enforce the separation of shared and specific features [21–23]. However, they often suffer from architectural complexity and training instability.

More recently, contrastive learning has emerged as a powerful paradigm for achieving more robust and fine-grained disentanglement. This trend is evident across different domains. Wei et al. used a cross-temporal contrastive model (CCDM) to disentangle ancient and modern Chinese into shared semantics, shared syntax, and different syntax, effectively bridging the cross-temporal gap. This shows how disentanglement is a fundamental strategy for handling any form of data heterogeneity, be it across time or across

modalities [24]. In computer vision, CoDeGAN goes a step further by entirely replacing traditional mutual information objectives with a contrastive loss for unsupervised class disentanglement in GANs, arguing that this leads to more stable training and avoids issues like mode collapse [25].

These cutting-edge works demonstrate that contrastive learning is becoming a mainstream approach for effective disentanglement. However, a common thread among most existing paradigms—whether based on complex supervision like MISA and DLF [21,23] or advanced techniques like contrastive learning is their reliance on parameter-heavy parallel networks and intricate training procedures. This heavy dependence incurs significant computational costs and fails to guarantee complete feature separation, often resulting in residual semantic entanglement. This limitation motivates our exploration of a more lightweight and direct unsupervised disentanglement approach, which is the foundational principle of our USD module.

The design of the DMBT framework is informed by a systematic analysis of the aforementioned literature. It is predicated on the insight that an ideal MSA framework must combine the efficiency and purity of front-end disentanglement with the intelligence and dynamism of back-end fusion. Therefore, this work does not represent an incremental improvement in a single method but rather an organic synthesis of the strengths of different technical routes. For feature disentanglement, inspired by the advantages of alternative, non-learning paradigms, the proposed USD module is parameter-free and similarity-based, ensuring highly efficient and pure feature separation. These separated representations are then processed by the GIFT core, which synergistically resolves the challenges of fine-grained feature extraction and dynamic decision-making through its hybrid attention mechanism and gating controls. In essence, DMBT offers a novel, high-performance solution for MSA by uniquely combining efficient unsupervised disentanglement with a deeply enhanced, dynamically adaptive fusion framework.

3. Proposed Approach

To systematically address the prevalent challenges of supervised disentanglement and static fusion in existing MSA frameworks, a novel Disentangled Multimodal Binding Transformer (DMBT) is proposed. As illustrated in Figure 2, the DMBT architecture is organized as a four-stage, end-to-end pipeline: (1) Feature Extraction, (2) Unsupervised Disentanglement, Language-centric Feature Enhancement, and (4) Gated Fusion and Prediction.

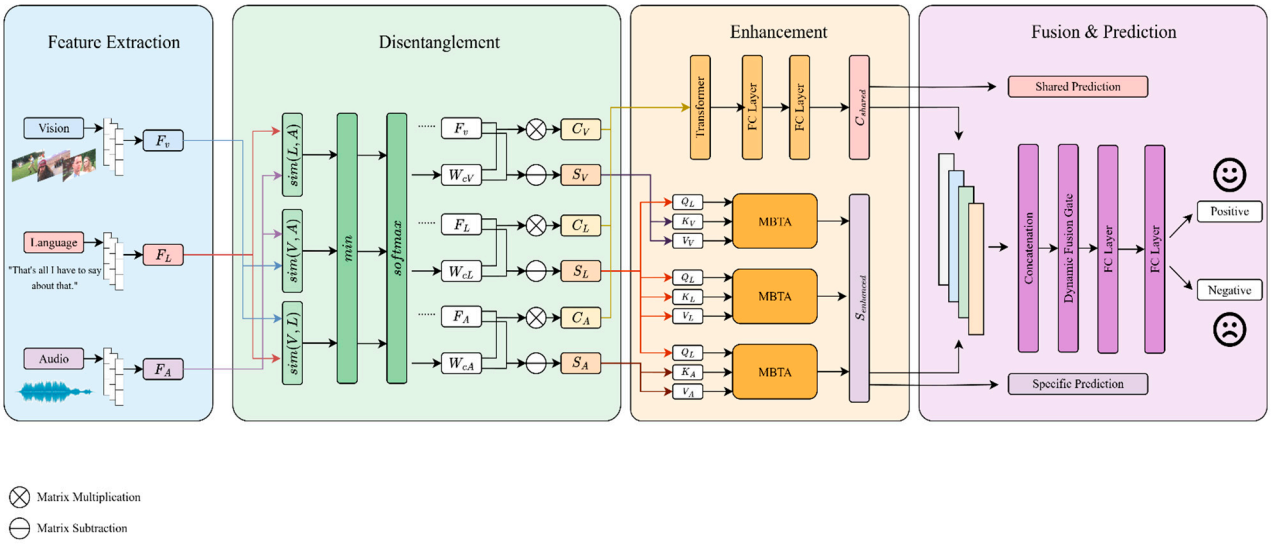


Figure 2. Overall architecture of the proposed DMBT framework.

The pipeline commences with standard encoders for initial Feature Extraction. Subsequently, to circumvent the complexities of supervised methods, the features are processed by an Unsupervised Semantic Disentanglement (USD) module. This parameter-free module decomposes the features of each modality into modality-invariant (common) and modality-specific components, establishing a clean foundation for subsequent processing. (3) The core of the framework is the Gated Interaction and Fusion Transformer (GIFT), which orchestrates the enhancement and fusion stages. Within GIFT, a language-centric paradigm is adopted where the textual modality serves as an anchor. The Multimodal Binding Transposed Attention (MBTA) mechanism performs targeted, unidirectional enhancement of the audio and visual features, addressing the issue of insufficient feature granularity. Finally, all common and enhanced specific features are fed into a Dynamic Fusion Gate (DFG). The DFG performs sample-wise dynamic weighting to render a final prediction, thereby overcoming the static fusion bottleneck.

3.1. Initial Feature Representation

Given a multimodal input X^m for modality $m \in \{t, a, v\}$, where t, a, v denote text, audio, and vision, respectively, the raw feature sequences are first projected into a unified dimension, d_{model}, via modality-specific 1D convolutional layers ($Proj^m$):

$$F^m = Proj^m(X^m) \in \mathbb{R}^{L_m \times d_{model}},\ m \in \{t, a, v\} \tag{1}$$

where L_m is the sequence length of modality m, and F^m is the resulting dimensionally aligned feature representation.

3.2. Unsupervised Semantic Disentanglement (USD)

To obviate the complexity and instability of supervised disentanglement, the proposed USD module operates in a parameter-free manner based on geometric similarity. It decomposes each projected feature representation F^m into a modality-common component C^m and a modality-specific component S^m.

Specifically, for a given source modality (e.g., text, t), its semantic affinity with a target modality $n \in \{a, v\}$ is quantified by a pairwise cosine similarity matrix:

$$Sim^{tn} = \frac{F^t (F^n)^T}{\|F^t\| \|F^n\|} \in \mathbb{R}^{L_t \times L_n} \tag{2}$$

The underlying principle is that a feature is considered "common" only if it exhibits high semantic relevance to all other modalities. Therefore, a shared weight matrix W_c^t is derived by retaining the minimum similarity score across modalities. This approach is more robust than an average, which could mask low alignment with one modality, or a maximum, which would be overly permissive. The resulting consensus scores are then normalized via a softmax function to produce the final weight matrix:

$$W_c^t = \text{Softmax}\big(\min\big(Sim^{ta}, Sim^{tv}\big)\big) \in \mathbb{R}^{L_t \times L_a} \text{ or } \mathbb{R}^{L_t \times L_v} \tag{3}$$

The weight matrix W_c^t is then used to project the original features F^t onto the modality-invariant subspace, yielding the common component C^t. The specific component S^t is subsequently derived via subtraction, which geometrically corresponds to an orthogonal decomposition. The modality-common (C^t) and modality-specific (S^t) features are then calculated as:

$$\begin{aligned} C^t &= W_c^t \cdot F^t \\ S^t &= F^t - C^t \end{aligned} \tag{4}$$

where $C^t, S^t \in \mathbb{R}^{L_t \times d_{\text{model}}}$. This process is applied symmetrically to obtain $\{C^a, S^a\}$ and $\{C^v, S^v\}$.

3.3. Gated Interaction and Fusion Transformer (GIFT)

Following disentanglement, the GIFT module performs deep feature interaction and fusion. It employs a language-centric strategy where the specific textual features S^t serve as an anchor to unidirectionally fuse complementary information from the audio (S^a) and visual (S^v) modalities. This asymmetric approach maximizes the semantic guidance of language while mitigating potential redundancy from multi-directional interactions.

3.3.1. Common Feature Stream Processing

The common features (C_t, C_a, C_v) are processed to form a robust consensus representation. They are first concatenated and then fed into a dedicated shared Transformer encoder to refine and integrate the shared information across modalities:

$$\begin{aligned} C_{\text{concat}} &= \text{Concat}(C_t, C_a, C_v) \\ C_{\text{encoded}} &= \text{SharedTransformer}(C_{\text{concat}}) \end{aligned} \tag{5}$$

The final state of the encoded sequence is then projected through a two-layer feed-forward network (Proj_c) to yield the final shared representation, C'_{shared}:

$$C'_{\text{shared}} = \text{Proj}_c(C_{\text{encoded}}) \tag{6}$$

This representation, C'_{shared}, encapsulates the unified, modality-invariant information and is subsequently used for both auxiliary prediction and final fusion.

3.3.2. Specific Feature Stream Processing

The core of GIFT is the Multimodal Binding Transposed Attention (MBTA) which performs deep modeling of the specific features (S^t, S^a, S^v). As shown in Figure 3, MBTA enhances standard multi-head self-attention by integrating a 1D convolutional layer into the Query (Q) and Key (K) paths. This allows for the simultaneous capture of both local temporal patterns and global contextual dependencies. The computation for a single attention head is:

$$\begin{aligned} Q' = \text{Conv1d}(Q),\ K' &= \text{Conv1d}(K) \\ \text{Attention}(Q, K, V) &= \text{Softmax}\left(\frac{Q'(K')^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{7}$$

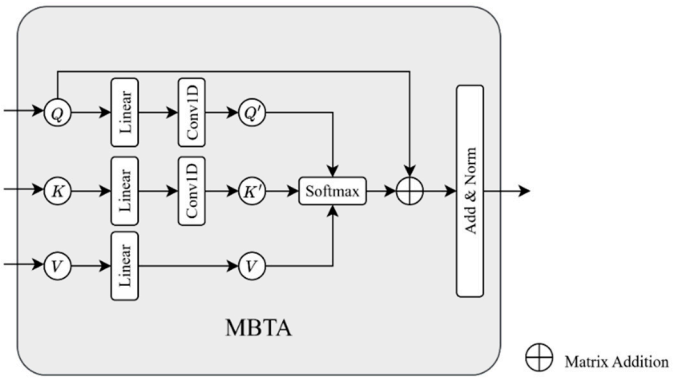


Figure 3. The architecture of the proposed Multimodal Binding Transposed Attention (MBTA) module.

Within the language-centric paradigm, MBTA executes three key unidirectional enhancement operations:

$$\begin{aligned} S_{out}^{t} &= \mathrm{MBTA}(S^{t}, S^{t}, S^{t}) \\ S_{out}^{a\to t} &= \mathrm{MBTA}(S^{t}, S^{a}, S^{a}) \\ S_{out}^{v\to t} &= \mathrm{MBTA}(S^{t}, S^{v}, S^{v}) \end{aligned} \tag{8}$$

The final enhanced specific representation, $S_{enhanced}$, is the summation of these three outputs:

$$S_{enhanced} = S_{out}^{t} + S_{out}^{a\to t} + S_{out}^{v\to t} \tag{9}$$

3.3.3. Dynamic Fusion Gate (DFG)

The DFG serves as the final decision-making mechanism. It takes the shared representation C'_{shared} and the enhanced specific representation $S_{enhanced}$, concatenates them, and computes a dynamic gating weight $z \in [0, 1]$ via a Multi-Layer Perceptron (MLP) with a sigmoid activation:

$$F_{total} = \mathrm{Concat}(C'_{shared}, S_{enhanced}, \ldots) \tag{10}$$

$$z = \mathrm{Sigmoid}(\mathrm{MLP}(F_{total})) \tag{11}$$

This gate adaptively weights the fused features, enabling the model to dynamically balance the importance of different information sources on a per-sample basis.

3.4. Hierarchical Prediction and Training Objective

To foster robust intermediate representations, a deep supervision strategy with hierarchical prediction is employed. In addition to the final prediction from the DFG output (F_{final}), auxiliary prediction heads are attached to the intermediate common (C'_{shared}) and enhanced specific ($S_{enhanced}$) representations:

$$\begin{aligned} \mathrm{logits}_{c} &= \mathrm{Pred}_{c}(C'_{shared}) \\ \mathrm{logits}_{s} &= \mathrm{Pred}_{s}(S_{enhanced}) \\ \mathrm{logits}_{f} &= \mathrm{Pred}_{f}(F_{final}) \end{aligned} \tag{12}$$

where Pred_{c}, Pred_{s}, and Pred_{f} are independent predictors. The total training objective, L_{DMBT}, is a weighted sum of the L1 losses computed for each prediction against the ground truth label Y_{true}:

$$\begin{aligned} L_{task_c} &= \mathrm{L1Loss}(\mathrm{logits}_{c}, Y_{true}) \\ L_{task_s} &= \mathrm{L1Loss}(\mathrm{logits}_{s}, Y_{true}) \\ L_{task_f} &= \mathrm{L1Loss}(\mathrm{logits}_{f}, Y_{true}) \end{aligned} \tag{13}$$

$$L_{DMBT} = L_{task_f} + \alpha L_{task_c} + \beta L_{task_s} \tag{14}$$

where α and β are hyperparameters controlling the influence of the auxiliary losses. This hierarchical objective provides direct gradient signals to intermediate layers, mitigating potential gradient vanishing and ensuring that each stage of the model learns task-relevant features. This, in turn, supplies the DFG with high-quality inputs for its final adaptive fusion.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. Experiments are conducted on two standard benchmark datasets for MSA:

CMU-MOSI [26]: A dataset comprising 2199 monologue video clips. Each clip is annotated with a sentiment score from −3 (highly negative) to +3 (highly positive). The standard data partition consists of 1284 training, 229 validation, and 686 test samples.

CMU-MOSEI [27]: A significantly larger dataset containing 22,856 movie review clips, annotated identically to CMU-MOSI. The standard partition includes 16,326 training, 1871 validation, and 4659 test samples.

Evaluation Metrics. For a comprehensive and fair comparison with prior work, model performance is evaluated from both classification and regression perspectives. Classification metrics include 7-class accuracy (Acc-7), 5-class accuracy (Acc-5), binary accuracy (Acc-2), and the binary F1-Score. Regression metrics include the Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr). For all metrics except MAE, higher values indicate better performance.

4.2. Implementation Details

Feature Extraction. Following standard protocols, textual features are extracted using a bert-base-uncased model (768-D). Audio and visual features are extracted using the COVAREP and Facet toolkits, respectively.

Model and Training. The DMBT model is implemented in PyTorch 1.13.0 and trained on a single NVIDIA V100 GPU. Hyperparameters are tuned separately for each dataset. The best-performing configuration on MOSI, for instance, utilizes an internal feature dimension (d_model) of 60, 10 attention heads, and 2 layers (n_levels) within the GIFT module. The model is trained using the Adam optimizer with a batch size of 16 and an initial learning rate of 1×10^{-4}. An early stopping criterion is applied, terminating the training if the validation loss fails to improve for 10 consecutive epochs. To ensure the robustness and reproducibility of our findings, all reported experimental results are the average of five independent runs using different random seeds.

4.3. Main Results and Analysis

The proposed DMBT model is benchmarked against 12 representative baseline models, spanning from early tensor fusion methods to recent Transformer-based approaches, including TFN [10], LMF [11], EF-LSTM [28], LF-DNN [29], MFN [30], Graph-MFN [27], MulT [3], PMR [31], MISA [21], MAG-BERT [32], DMD [33], and DLF [23]. The comparison results on the CMU-MOSI and CMU-MOSEI datasets are presented in Tables 1 and 2, respectively.

Table 1. Performance comparison of DMBT against baseline models on the CMU-MOSI test set.

Method	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)	Corr (↑)	MAE (↓)
TFN *	34.9	39.39	80.08	80.07	0.698	0.901
LMF *	33.2	38.13	82.5	82.4	0.695	0.917
EF-LSTM †	35.39	40.15	78.48	78.51	0.669	0.949
LF-DNN †	34.52	38.05	78.63	78.63	0.658	0.955
MFN †	35.83	40.47	78.87	78.9	0.67	0.927
Graph-MFN †	34.64	38.63	78.35	78.35	0.649	0.956
MulT	40	42.68	83	82	0.698	0.871
PMR	40.6	-	83.6	83.6	-	-
MISA †	41.37	47.08	83.54	83.58	0.778	0.777
MAG-BERT	43.62	-	84.43	84.61	0.781	0.727
DMD **	44.3	49.34	84.33	84.22	0.769	0.741
DLF	**47.08**	52.33	85.06	85.04	0.781	0.731
DMBT(Ours)	46.48	**52.34**	**85.52**	**85.48**	**0.784**	**0.718**

Bold: The best results for each metric. † denotes results from THUIAR's GitHub page (THUIAR 2024), * denotes results from [21], - denotes results not provided in the original paper, and ** denotes replicated results from public code with hyperparameters provided in the original paper.

Table 2. Performance comparison of DMBT against baseline models on the CMU-MOSEI test set.

Method	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)	Corr (↑)	MAE (↓)
TFN *	50.2	53.1	82.5	82.1	0.7	0.593
LMF *	48	52.9	82	82.1	0.677	0.623
EF-LSTM †	50.01	51.16	80.79	80.67	0.683	0.601
LF-DNN †	50.83	51.97	82.74	82.52	0.709	0.58
MFN†	51.34	52.76	82.85	82.85	0.718	0.575
Graph-MFN †	51.37	52.69	83.48	83.43	0.713	0.575
MulT	51.8	54.18	82.5	82.3	0.703	0.58
PMR	52.5	-	83.6	83.4	-	-
MISA †	52.05	53.63	84.67	84.66	0.752	0.558
MAG-BERT	52.67	-	84.82	84.71	0.755	0.543
DMD **	53.68	54.52	85.25	85.11	0.759	0.54
DLF	**53.9**	55.7	85.42	85.27	0.764	0.536
DMBT(Ours)	53.73	**56.27**	**85.74**	**85.72**	**0.77**	**0.532**

Bold: The best results for each metric. † denotes results from THUIAR's GitHub page (THUIAR 2024), * denotes results from [21], - denotes results not provided in the original paper, and ** denotes replicated results from public code with hyperparameters provided in the original paper.

The results indicate that DMBT consistently achieves state-of-the-art or highly competitive performance across all key metrics on both datasets. When compared with recent disentanglement-based models that rely on complex supervision (e.g., MISA, DMD, DLF), DMBT surpasses MISA and DMD on all reported metrics. While DLF shows a slight advantage on the Acc-7 metric for MOSI, which may stem from its complex supervision providing a stronger inductive bias for fine-grained classification, DMBT outperforms it on all other metrics. This suggests that the proposed lightweight, parameter-free USD module achieves a more effective and robust feature disentanglement.

Furthermore, DMBT demonstrates a consistent advantage over other strong Transformer-based baselines (e.g., MulT, MAG-BERT). Notably, on metrics that assess a deeper understanding of sentiment dynamics, such as Correlation (Corr) and Mean Absolute Error (MAE), DMBT sets a new state-of-the-art. This advantage is primarily attributed to the proposed GIFT framework. Its core engine, MBTA, captures richer, fine-grained local features in addition to global context, while its decision-making mechanism, DFG, enables more refined, sample-wise adaptive fusion. The strong performance on the larger and more complex MOSEI dataset further validates the scalability and robustness of the DMBT architecture.

4.4. Ablation Study

To verify the contribution of each key component, a series of ablation studies was conducted on the MOSI dataset by systematically removing the USD, MBTA, and DFG modules from the full DMBT model. The results are presented in Table 3.

Various Modalities Combinations. To assess modality contributions and fusion efficacy, we evaluated DMBT under various input configurations (Table 3). Unimodal analysis confirms the expected dominance of language, which significantly outperforms audio-only or visual-only inputs, establishing it as the primary source of sentiment information. Subsequent bi-modal experiments (L & A, L & V) demonstrate marked performance improvements over the language-only baseline, empirically validating that our language-centric framework effectively extracts and integrates valuable complementary information from the auxiliary visual and audio modalities via the GIFT module. Finally, the superior performance achieved by the full tri-modal model underscores that each modality offers

unique contributions, which our proposed architecture successfully synergizes to yield the most accurate sentiment predictions.

Table 3. Ablation study results on the CMU-MOSI dataset. The best results are in bold.

Method	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
DMBT(Ours)	**46.48**	**85.52**	**85.48**	**0.718**
Different Modalities				
only A	15.45	42.23	25.07	1.468
only V	15.12	42.68	28.14	1.470
only L	44.15	84.45	84.37	0.764
L & A	44.37	84.70	84.61	0.752
L & V	44.31	84.21	84.03	0.758
Different Components				
w/o USD	44.02	83.08	83.12	0.765
w/o MBTA	45.61	84.54	83.75	0.781
w/o DFG	44.31	83.62	83.56	0.760

Effectiveness of USD: Removing the USD module and reverting to a standard feature fusion approach (as in MulT) leads to significant performance degradation across all metrics. This confirms that the proposed lightweight unsupervised disentanglement provides purer, higher-quality features for downstream fusion.

Effectiveness of MBTA: Replacing the MBTA module with a standard Transformer encoder also results in a consistent performance drop. This indicates that the hybrid convolutional-attentional design of MBTA, which simultaneously models local and global dependencies, is more effective for complex multimodal interactions than standard self-attention alone.

Effectiveness of DFG: Eliminating the DFG and using simple feature concatenation for the final prediction causes a notable decline in performance, particularly on the MAE and Acc-7 metrics. This highlights the importance of the adaptive, sample-wise gating mechanism for intelligent decision-making.

To validate our design choice of using minimum similarity aggregation in the USD module, we conducted a comparative experiment against two direct alternatives: cosine-max and cosine-mean. The results are presented in Table 4.

Table 4. Ablation study on the aggregation strategy in the USD module on the MOSI dataset. The best results are in bold.

Aggregation Strategy	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE (↓)
min(Ours)	**46.48**	**85.52**	**85.48**	**0.718**
max	44.02	83.08	83.12	0.765
mean	45.61	84.54	83.75	0.781

The empirical data strongly corroborates our theoretical predictions. The aggregation strategy, which operates on a principle of strict consensus, consistently and significantly outperforms both the mean and max strategies across all key evaluation metrics.

The max strategy, being overly permissive by design, likely incorporates noisy or modality-pair-specific information into the "common" representation, thereby degrading performance.

The mean strategy, while a reasonable compromise, tends to blur the distinction between truly common and partially relevant features, leading to a suboptimal, "polluted" representation.

In contrast, our min strategy successfully identifies the purest set of modality-invariant features. By providing the downstream GIFT module with the cleanest possible inputs, it enables a more effective and robust fusion process, ultimately leading to superior model performance.

In summary, these ablation studies provide compelling evidence that the superior performance of DMBT arises from the synergistic contributions of each of its core architectural innovations.

4.5. Further Analysis

Performance on Sentiment Categories. To dissect the model's behavior, the 7-class confusion matrix and per-category accuracies on the MOSI test set are visualized in Figure 4. The model demonstrates high accuracy for the five core sentiment categories (N, WN, NT, WP, P). However, a notable performance drop is observed for the extreme categories (HN and HP), which the model tends to confuse with their less intense, adjacent counterparts. This difficulty in discerning subtle intensity differences is a primary reason for the model's performance on the Acc-7 metric.

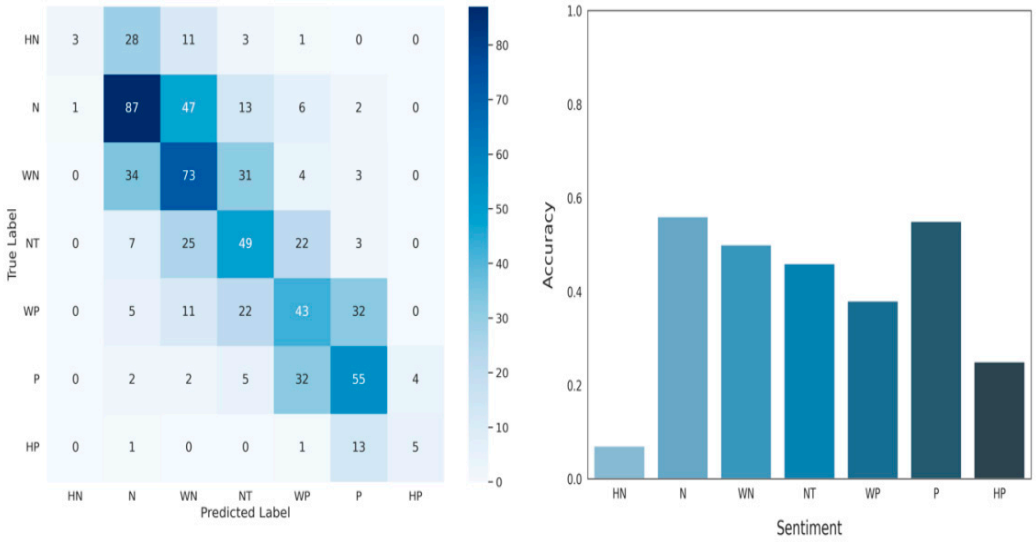


Figure 4. Performance analysis of DMBT on the CMU-MOSI test set. (**Left**) The 7-class confusion matrix. (**Right**) Per-category classification accuracy. The sentiment labels are abbreviated as follows: HN (Highly Negative), N (Negative), WN (Weakly Negative), NT (Neutral), WP (Weakly Positive), P (Positive), and HP (Highly Positive).

Visualization of Disentangled Representations. To investigate the quality of the learned representations, the disentangled feature spaces of DMBT and a strong baseline (DLF) are visualized using t-SNE in Figure 5. The visualization provides strong qualitative evidence for the superiority of the USD module. The feature clusters learned by DMBT exhibit significantly better intra-class compactness and inter-class separability compared to those learned by DLF, which shows considerable overlap between its "Text-Common" and "Text-Specific" representations.

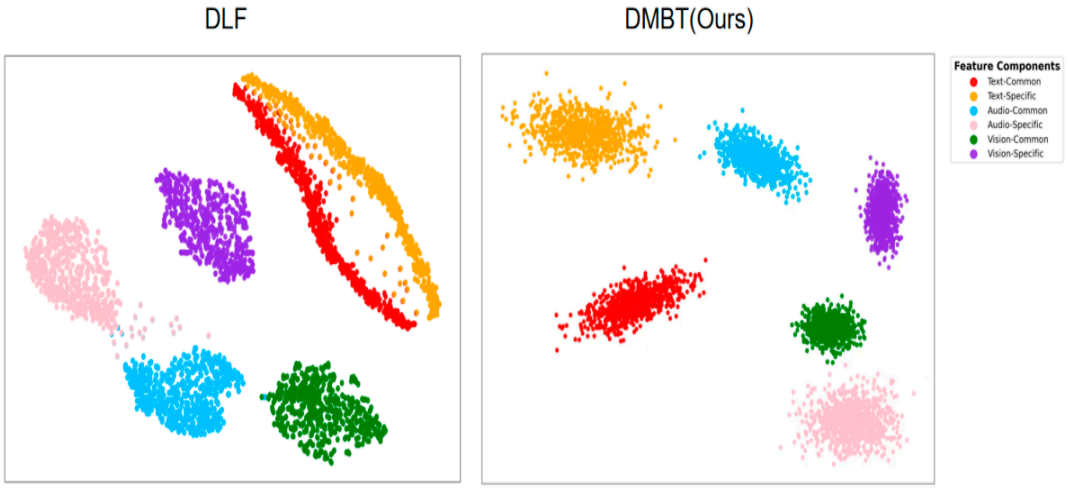


Figure 5. t-SNE visualization of the disentangled feature representations on the CMU-MOSI benchmark. The plot compares the feature space learned by a strong baseline, DLF (**left**), with that of the proposed DMBT (**right**). Different colors represent the common and specific features from each modality.

Data-Driven Rationale for Performance on Extremes. The clearer feature separation in Figure 5 also offers a deeper, data-driven explanation for the lower accuracy on extreme sentiment classes. The issue appears to be rooted in the dataset's intrinsic properties rather than architectural limitations.

Data Imbalance: The "HN" and "HP" classes suffer from a long-tail distribution, with far fewer samples than other categories, which hinders the model's ability to learn their distinguishing features.

Inherent Semantic Ambiguity: The perceptual difference between "negative" and "highly negative" is inherently subtle across language, visual, and acoustic cues. Consequently, their feature representations are naturally closer in the semantic space. A model with a superior representational capacity, such as DMBT, will more faithfully reflect this intrinsic data ambiguity, which can paradoxically lead to confusion during classification between semantically adjacent but distinct classes.

5. Conclusions

In this paper, we proposed a novel, end-to-end DMBT framework to improve the performance of MSA. The core of DMBT is twofold: The USD module and GIFT with MBTA engine and DFG. Extensive experiments and comprehensive ablation studies demonstrated the superiority of DMBT.

Despite these promising results, we acknowledge several limitations that pave the way for future work. As our analysis revealed, the model's performance is noticeably lower on extreme sentiment categories, largely due to the long-tail distribution of the training data. Moreover, our current framework assumes the presence of complete modal data, which may not hold true in real-world scenarios where data can be missing or corrupted. Additionally, we lack a formal analysis of the Dynamic Fusion Gate's stability under perturbation, leaving its potential to overfit to sample-specific noise an open question.

Looking ahead, several tangible directions are planned. To address the challenge of recognizing extreme emotions, we will explore integrating advanced data augmentation techniques or specialized loss functions designed for imbalanced data. To enhance the model's robustness for real-world deployment, we will focus on extending the DMBT framework to dynamically adapt to missing modalities, ensuring stable performance even with incomplete inputs. Finally, we intend to investigate the generalizability of our

core USD and GIFT modules to a broader range of multimodal tasks, with the goal of contributing a more versatile solution to the wider field of multimodal understanding.

Author Contributions: Conceptualization, R.G.; Methodology, R.G.; Software, R.G.; Validation, R.G.; Formal analysis, R.G.; Investigation, R.G.; Resources, G.G. and F.J.; Data curation, R.G.; Writing—original draft, R.G.; Writing—review and editing, G.G. and F.J.; Visualization, R.G.; Supervision, G.G.; Project administration, G.G.; Funding acquisition, F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Chinese Key Project for Innovation and Entrepreneurship of College Students (202510320054) and in part by the National Natural Science Foundation of China (62401235).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in thearticle, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors are grateful to the editor and the anonymous reviewers for their useful comments and advice, which were vital for improving the quality of this paper.

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

References

1. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef] [PubMed]
2. Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; et al. Context de-confounded emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19005–19015.
3. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. *Proc. Conf. Assoc. Comput. Linguist. Meet.* **2019**, *2019*, 6558. [PubMed]
4. Liu, S.; Ma, X.; Deng, S.; Suo, Y.; Zhang, J.; Ng, W.W. Lightweight multimodal Cycle-Attention Transformer towards cancer diagnosis. Expert Syst. *Appl.* **2024**, *255*, 124616.
5. Hang, C.N.; Ho, S.M. Personalized Vocabulary Learning through Images: Harnessing Multimodal Large Language Models for Early Childhood Education. In Proceedings of the 2025 IEEE Integrated STEM Education Conference (ISEC), Princeton, NJ, USA, 15 March 2025; pp. 1–7.
6. Marquez-Carpintero, L.; Viejo, D.; Cazorla, M. Enhancing engineering and STEM education with vision and multimodal large language models to predict student attention. *IEEE Access* **2025**, *13*, 114681–114695. [CrossRef]
7. Ali, K.; Hughes, C.E. A unified transformer-based network for multimodal emotion recognition. *arXiv* **2023**, arXiv:2308.14160. [CrossRef]
8. Ezzameli, K.; Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Inf. Fusion* **2023**, *99*, 101847. [CrossRef]
9. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
10. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250. [CrossRef]
11. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
12. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
13. Wang, Z.; Wan, Z.; Wan, X. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2514–2520.
14. Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; Xu, B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Trans. Multimed.* **2023**, *26*, 776–788. [CrossRef]

15. Gan, C.; Fu, X.; Feng, Q.; Zhu, Q.; Cao, Y.; Zhu, Y. A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Syst. Appl.* **2024**, *242*, 122731. [CrossRef]
16. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv* **2023**, arXiv:2310.05804.
17. Lei, Y.; Yang, D.; Li, M.; Wang, S.; Chen, J.; Zhang, L. Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In Proceedings of the CAAI International Conference on Artificial Intelligence, Fuzhou, China, 22–23 July 2023; Springer: Singapore, 2023; pp. 189–200.
18. Hasan, M.K.; Islam, M.S.; Lee, S.; Rahman, W.; Naim, I.; Khan, M.I.; Hoque, E. Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models. *arXiv* **2023**, arXiv:2303.15430.
19. Xie, X.; Zhang, X.; Tang, X.; Zhao, J.; Xiong, D.; Ouyang, L.; Yang, B.; Zhou, H.; Ling, B.W.K.; Teo, K.L. MACTFusion: Lightweight cross transformer for adaptive multimodal medical image fusion. *IEEE J. Biomed. Health Inform.* **2024**, *29*, 3317–3328. [CrossRef] [PubMed]
20. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv* **2018**, arXiv:1806.06176.
21. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
22. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1642–1651.
23. Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; Hu, J. DLF: Disentangled-language-focused multimodal sentiment analysis. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 21180–21188. [CrossRef]
24. Wei, Y.; Zhu, Y.; Bai, T.; Wu, B. A cross-temporal contrastive disentangled model for ancient Chinese understanding. *Neural Netw.* **2024**, *179*, 106559. [CrossRef] [PubMed]
25. Liu, Z.; Pan, L.; Guo, X.; Zhao, J. CoDeGAN: Contrastive Disentanglement for Generative Adversarial Network. *Neurocomputing* **2025**, *648*, 130478. [CrossRef]
26. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
27. Zadeh, A.A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, VIC, Australia, 15–20 July 2018; pp. 2236–2246.
28. Williams, J.; Kleinegesse, S.; Comanescu, R.; Radu, O. Recognizing emotions in video using multimodal DNN feature fusion. In Proceedings of the Grand Challenge and Workshop on Human Multimodal Language, Melbourne, VIC, Australia, 20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 11–19.
29. Williams, J.; Comanescu, R.; Radu, O.; Tian, L. Dnn multimodal fusion techniques for predicting video sentiment. In Proceedings of the Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Melbourne, VIC, Australia, 20 July 2018; pp. 64–72.
30. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
31. Lv, F.; Chen, X.; Huang, Y.; Duan, L.; Lin, G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2554–2562.
32. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. *Proc. Conf. Assoc. Comput. Linguist. Meet.* **2020**, *2020*, 2359. [PubMed]
33. Li, Y.; Wang, Y.; Cui, Z. Decoupled multimodal distilling for emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6631–6640.

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