CL-DMDF: Dynamic Multimodal Data Fusion Model Based on Contrastive Learning

Dong Li[1], **Lingling Zhang**[1], **Binghao Han**[1], **Linlin Ding**[1*], **Yue Kou**[2]

[1]School of Information, Liaoning University, Shenyang, China
[2]School of Computer Science and Engineering, Northeastern University, Shenyang, China
dongli@lnu.edu.cn, {4032532425, 4032232396}@smail.lnu.edu.cn, dinglinlin@lnu.edu.cn, kouyue@cse.neu.edu.cn

Abstract

Multimodal data fusion involves integrating and analyzing information from multiple modalities to uncover latent correlations and complementary patterns, thereby enhancing data processing and decision-making. While existing methods for structured multimodal inputs are typically designed around specific tasks and assume fully observed modalities, real-world applications often suffer from uncertain or missing modality inputs due to various factors. Some traditional models overly emphasize local interactions within missing modalities, neglecting the global complementary cues embedded in multimodal representations. To overcome these limitations, we propose a Dynamic Multimodal Data Fusion model based on Contrastive Learning (CL-DMDF). CL-DMDF introduces a novel attention mechanism that operates across both feature and modality dimensions to compute reliable attention scores, effectively reflecting importance at each level. The CL-DMDF further incorporates an entity-centroid contrastive learning module that constructs centroid-based positive samples from entity features to enhance discriminative learning. Additionally, an adaptive fusion module is employed to improve the efficiency and accuracy of dynamic fusion strategies. Extensive experiments conducted on three datasets demonstrate the effectiveness of the CL-DMDF across diverse multimodal fusion tasks.

Code — https://github.com/zoo-111-p/CL-DMDF

Introduction

In the real world, humans perceive the environment through sensory receptors such as eyes, ears, skin, nose, and tongue, enabling them to see objects, hear sounds, feel textures, smell scents, and taste flavors. The information obtained from each sensory source or medium can be regarded as a modality. Multimodal refers to the use of two or more heterogeneous modalities, such as text, vision, or audio, for joint learning and inference.

The human brain unconsciously integrates information from different sensory receptors, or fuses modalities, extracting complementary information to form predictions or decisions. In addition, machines are highly dependent on sensors such as RGB cameras, microphones, and other types of sensors. Each sensor maps the observed objects or activities into the machine's domain, enabling it to make predictions or decisions based on the collected data.

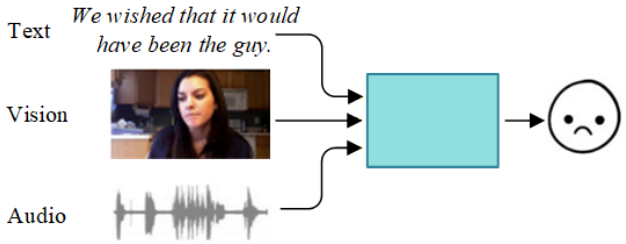

Figure 1: Relying solely on textual information is unlikely to accurately predict the current emotion during classification. However, audio and vision modalities can provide crucial cues for the multimodal network.

While existing methods for multimodal decision-making often rely on manual feature engineering, they often fail to capture cross-modal complementarity, leading to early information loss. Although models such as CNNs, LSTMs, transformers, and BERT can process multimodal inputs, the key challenge remains how to perform deep fusion to enhance machine intelligence.

Multimodal data fusion aims to combine information from multiple modalities—such as vision, audio, and textual—to enhance analytical accuracy. For example, in emotion recognition (Figure 1), integrating audio and visual signals improves classification by leveraging cross-modal complementarity.

For dynamic multimodal data, we propose a Dynamic Multimodal Data Fusion model based on Contrastive Learning (CL-DMDF). The model extracts modality-specific features and projects them into a shared embedding space. A dual-dimensional attention mechanism over feature and modality dimensions assigns weights to account for structural variability. Contrastive learning enhances feature discriminability, while an adaptive fusion module dynamically selects task-relevant strategies to generate a unified representation.

Multimodal fusion significantly enhances system perception by mitigating the limitations of single-modality inputs, which are often affected by factors such as lighting, noise,

*Corresponding author

and sensor failure. By integrating complementary cues, it improves accuracy and robustness in complex environments. For instance, in autonomous driving, combining camera images, radar data, and voice commands enhances safety and interaction reliability.

In this paper, we propose a novel Dynamic Multimodal Data Fusion model based on Contrastive Learning (CL-DMDF), to address the limitations of existing fusion models in terms of task adaptability and semantic expressiveness. Our main contributions are summarized as follows:

(1) We propose a dual-dimensional attention mechanism that jointly models feature-level and modality-level importance, enabling the computation of more reliable attention scores to guide effective multimodal integration.

(2) We introduce an entity-centroid contrastive learning module, which constructs positive and negative pairs based on attention-weighted modality features. This module enhances the discriminative power of representations and expands the embedding space.

(3) We present an adaptive fusion module that dynamically selects optimal fusion strategies based on the characteristics of input features. This allows the model to balance accuracy and computational efficiency across varying tasks.

(4) CL-DMDF's innovation lies in the collaborative design of its three components, optimizing a single objective within a unified architecture, representing an innovation at the framework level. We conduct comprehensive experiments on three representative datasets. The results demonstrate that CL-DMDF consistently outperforms strong baselines across diverse tasks, validating the effectiveness and generalizability of the model.

Related Work

Multimodal data fusion under dynamic conditions aligns with conventional fusion paradigms. This section reviews representative studies in this area.

Multimodal Data Fusion

Multimodal fusion has traditionally been categorized into data-level, feature-level, and decision-level approaches. Early data-level models (Camille, Clément, and Laurent 2013) concatenated RGB and depth images at the input, while SSR-CNN (Liu et al. 2019) employed a single-stream architecture to integrate modalities. Feature-level approaches (Li et al. 2018; Hu et al. 2020a) encoded modalities independently and fused them during decoding; FFN (Janani et al. 2021) improved this by concatenating outputs from modality-specific encoders.

For decision-level fusion, (Nihar, Kevin, and Peyman 2021) introduced a multimodal variational autoencoder (VAE) that learned a shared latent space from image features. Similarly, an end-to-end VAE framework (Dhruv et al. 2019) addressed fake news detection by encoding and reconstructing joint textual and visual embeddings.

Recent studies extended these paradigms. (Chen, Wang, and Zhang 2024) proposed a progressive cross-modal attention mechanism for adaptive fusion across abstraction levels. (Liu, Fan, and Li 2025) incorporated latent-variable modeling to capture modality-specific uncertainty. (Zhang, Hu, and Tan 2025) further explored lightweight transformer-based fusion under real-time constraints, improving the trade-off between efficiency and performance.

Dynamic Multimodal Data Fusion

Dynamic attention-based fusion methods were generally categorized into intra-modality self-attention, cross-modality cross-attention, and transformer-based approaches. (Gao et al. 2019) applied hard attention to generate spatial binary masks for selective feature propagation. (Mateusz et al. 2018) introduced bidirectional cross-modal attention for vision-language alignment, and (Hu et al. 2020b) proposed a dot-product cross-attention to capture audio-text correlations. MVAE integrated multiple modalities for tasks such as fake news detection.Transformer-based fusion methods leveraged cross-modal attention to model long-range dependencies. (Sun et al. 2021) designed a cross-modal transformer for MRI–acoustic signal alignment, while (Xu, Feng, and Huang 2022) adopted self-attention to capture inter-modal relations. (Yang, Tan, and Gao 2024) presented a multi-head sparse transformer with hierarchical fusion, demonstrating robustness on noisy video-text datasets. (Liu, Zhao, and Zhang 2025) introduced an adaptive sparse attention framework that pruned modality contributions based on semantic uncertainty. More recently, UniFM (Jiang, Zhang, and Tan 2024) and MM-TokenMixer (Li, Xu, and Liu 2025) optimized cross-modal token integration using shared representations and token-wise mixing, resulting in improved generalization between benchmarks.

Graph-based methods evolved from standard GCNs to spatio-temporal graph architectures. (Chih et al. 2022) utilized deep GCNs for emotion recognition, and (Hu et al. 2021) combined multimodal GATs with temporal convolutions to model temporal-spatial patterns. (Ding, Sun, and Zhao 2023) integrated multi-head attention into GNNs for scene graph embeddings, followed by cross-modal alignment in (Yang et al. 2023). (Wang, Lin, and Song 2025) proposed a reinforcement-guided GNN with modality-aware policy learning for dynamic social event detection. Furthermore, GNN-Adapter (Wu, Sun, and Huang 2024) incorporated lightweight graph modules into pre-trained multimodal models, enhancing efficiency without retraining the backbones.

Differences from Existing Work

Our work differs from traditional methods in several key aspects.

Most existing fusion models adopt static strategies across tasks, limiting adaptability. In contrast, CL-DMDF employs a dynamic fusion mechanism that selects task-relevant strategies based on modality reliability. Existing attention-based methods often yield unstable weights due to unsupervised design, while graph-based models struggle to generalize to unseen modalities. CL-DMDF addresses this via a dual-dimensional attention mechanism that jointly considers feature-level and modality-level importance. Unlike previous contrastive learning approaches with coarse alignment, our entity-centric contrastive module captures fine-grained

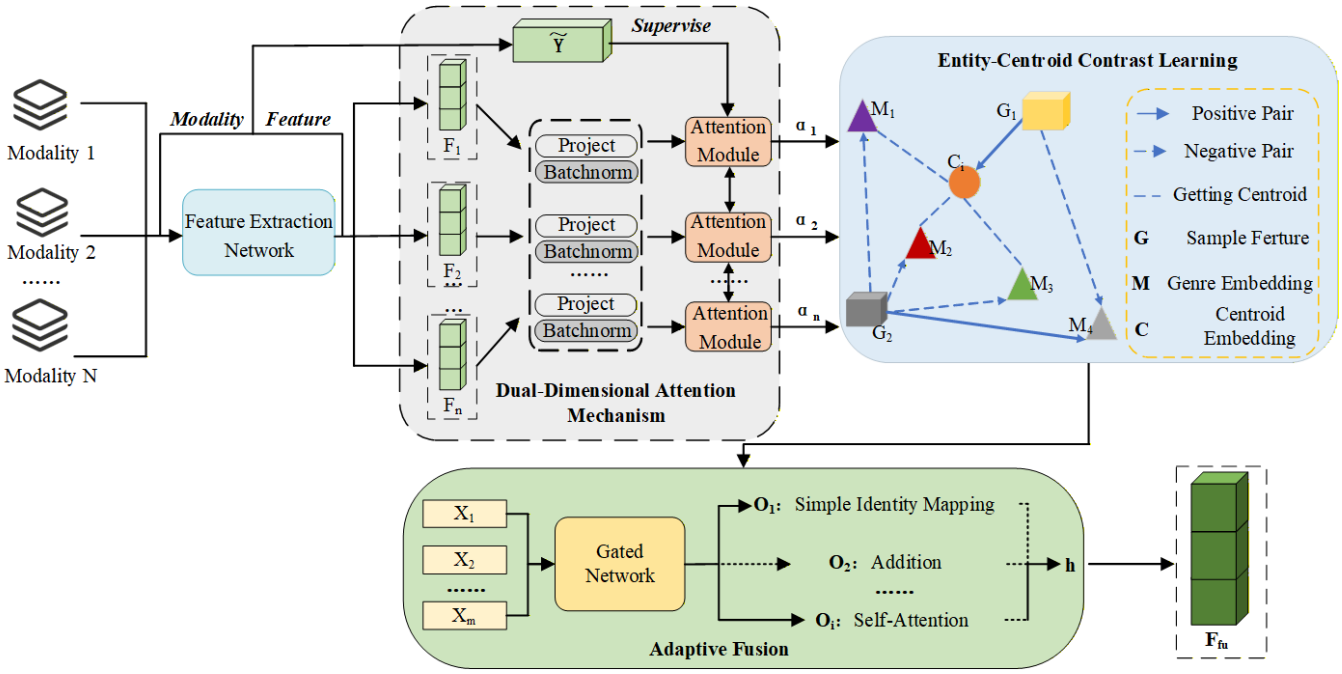

Figure 2: The overview of CL-DMDF. Features are first extracted from data of different modalities using a feature extraction network and projected into a unified dimensional space. A dual-dimensional attention mechanism is then employed to guide the allocation of attention. Subsequently, contrastive learning is applied to enhance the discriminability of the features. Finally, an adaptive fusion module selects the most appropriate fusion strategy based on the specific requirements of the task.

semantics, enhancing representation quality. Finally, to balance performance and efficiency, we introduce a resource-aware objective that guides the adaptive fusion module to avoid redundant computation under varying task complexities.

Method

We begin by presenting the proposed CL-DMDF model for dynamic multimodal fusion, followed by a formal definition of the problem and its implementation details.

Model Overview

This paper proposes a Dynamic Multimodal Data Fusion model based on Contrastive Learning, termed CL-DMDF. The model first extracts features from different modalities using dedicated feature extraction networks and projects them into a unified vector space. To account for the diversity of modality combinations and varying feature counts per entity, a dual-dimensional attention mechanism is introduced to guide attention allocation across modalities and features. This mechanism enhances the model's ability to concentrate on task-relevant information. To enhance feature discriminability, CL-DMDF integrates a contrastive learning module that sharpens representation boundaries by distinguishing similar from dissimilar samples. An adaptive fusion module further selects optimal strategies based on task demands and modality characteristics, enabling effective aggregation of heterogeneous features. The overall architecture is shown in Figure 2.

Dual-Dimensional Attention Mechanism

Different modalities may represent distinct entities with varying semantic features, complicating consistent and informative fusion due to modality-specific encoding differences.To address this, we propose a dual-dimensional attention mechanism that jointly considers both feature-level richness and modality-level presence when assigning attention weights to each entity.

Entities with broader feature coverage and presence across more modalities are assigned higher weights, enhancing their contribution to the final fused representation. This method prioritizes semantically important entities, enabling more effective cross-modal integration. Weighted features are used in contrastive learning, with an attention module, consisting of a shared linear layer and non-linear activation, further balancing modality contributions across samples.

For multimodal features, $F_i^1 \in R^{D_1}, F_i^2 \in R^{D_2}, \ldots F_i^n \in R^{D_n}$, where $\{D_i\}$ represents the features extracted from different modalities, we use batch normalization and linear projection functions to convert them into the same shape, is given by Equation 1:

$$h(x) = \text{project}\big(\text{batchnorm}(x)\big) \tag{1}$$

where $project(\cdot)$ and $batchnorm(\cdot)$ represent the linear projection function and batch normalization function, respectively. After alignment, all modality features are input into the module to obtain their attention scores a_i, which guide the contrastive learning process to generate centroids and expand the feature embedding space.

To improve attention reliability, we introduce a self-supervised dual-dimensional mechanism that assigns attention scores based on pseudo-labels $\tilde{Y}_n$, reflecting both feature count and modality coverage. Entities with limited features or modalities receive lower weights due to reduced semantic value. The pseudo-label $\tilde{Y}_n$ is defined over F as shown in Equations 2-4.

$$\overline{N} = \frac{1}{W}\sum_{i=1}^{W} F^i \tag{2}$$

$$\overline{M} = \frac{1}{W}\sum_{i=1}^{W}\sum_{E_i \in M_j} M_{E^i} \tag{3}$$

$$\tilde{Y} = \frac{N_i \sum_{E^i \in M_j} M_{E^i}}{N_i + \overline{N}\left(\sum_{E^i \in M_j} M_{E^i} + \overline{M}\right)} \tag{4}$$

Let W denote the number of entities and M_{E^i} the number of modalities containing entity E^i. Entities with feature and modality counts above the averages $\overline{N}$ and $\overline{M}$ are assigned higher pseudo-labels $\tilde{Y}_n$, while others receive lower values. To align with the activation-constrained attention range, $\tilde{Y} \in \{0, 1\}$ is used.

To ensure stable convergence of self-supervised attention scores, we incorporate an appropriate loss function as shown in Equation 5.

$$\Gamma_{\text{atten}} = -\sum_{i=1}^{B} \log\left(1 - |\alpha_{M_i} - \overline{Y_i}|\right) \tag{5}$$

Since α_{M_i} is a vector and $\tilde{Y}_n$ is a scalar, we first average α_{M_i} to obtain a scalar for regression. As both values lie within $(0, 1)$, directly applying standard L_1 or L_2 losses can lead to weak gradients. To address this, we adopt a logarithmic transformation in the loss function to maintain sufficiently strong gradients and ensure monotonicity during optimization.

We align features using batch normalization and linear projection to stabilize attention weights. CL-DMDF demonstrates robustness under noisy and low-sample conditions by focusing on key entities and enhancing features via two-dimensional attention and contrastive learning.

Entity-Centroid Contrastive Learning

Figure 3 illustrates the Entity-Centroid Contrastive Learning framework. To enhance discriminability, representations are projected into a shared embedding space. Unlike conventional methods, our approach captures multiple semantics per instance. For a batch of fused features F^{Mi}, each entity aggregates cross-modal information, with its embedding set defined in Equation 6.

$$G_i = F^{M_1}, F^{M_2}, \ldots, F^{M_n} \tag{6}$$

Let n denote the number of modalities in the fused feature set. The centroid C_i of group G_i expands the embedding space. Dual-dimensional attention scores guide weighted

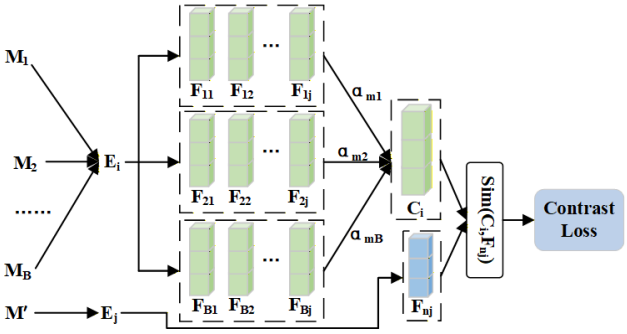

Figure 3: Entity-Centroid Contrastive Learning implementation details. Embedding features in vector space to capture diverse semantics and prevent over-alignment in contrastive learning.

feature aggregation, integrating modality-specific information, with features treated as positive samples (Equation 7).

$$C_i = \sum_{F^{M_i} \in G_i} \alpha_{M_i} F^{M_i} \tag{7}$$

Here, α_{M_i} denotes the attention scores generated by the dual-dimensional attention mechanism. After expanding the embedding space, features within this space are treated as positive samples, while those outside it are considered negative. Accordingly, the complete embedding set for all samples is defined as in Equation 8.

$$U_{i=1}^{B} G_i = \{F^{M_1}, F^{M_2}, \ldots, F^{M_j}\} \tag{8}$$

B denotes the number of entities, and j the total number of modality features. This yields the feature embedding set for all entities in the current batch. Samples outside this embedding space—i.e., the complement of $U_{i=1}^{B} G_i$ is considered a negative sample during reinforcement learning, as in Equation 9.

$$S_i = U_{i=1}^{B} G_i - C_i \tag{9}$$

In the loss design, since positive samples define an embedding space, we compute the similarity between each negative sample and all features within this space and sum the results to enhance feature discriminability, which can be formulated as follows:

$$q_i = \sum_{F^{M_i} \in S_i} \exp\left(\frac{F_i f_{E^{M_i}}}{\tau}\right) \tag{10}$$

$$\Gamma_{\text{contra}} = -\sum_{i=1}^{B} \log \frac{\exp\left(\frac{F_i f_{C_i}}{\tau}\right)}{\exp\left(\frac{F_i f_{C_i}}{\tau}\right) + q_i} \tag{11}$$

The temperature coefficient τ adjusts the model's sensitivity to negative samples. Lower τ sharpens similarity scores, assigning higher penalties to hard negatives. During training, negatives are repelled proportionally to their similarity. q_i denotes the total similarity between sample F_i and

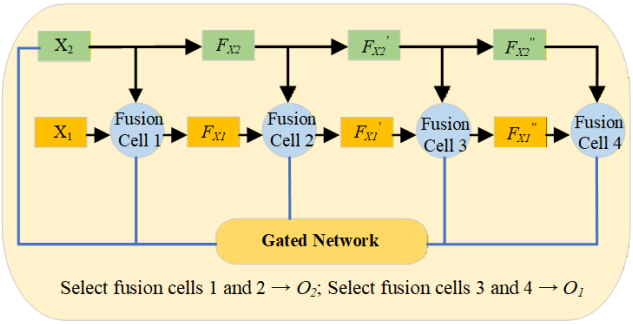

Figure 4: The adaptive fusion module enables finer-grained and more flexible decision-making by stacking multiple fusion units to construct a fusion network.

its positives S_i, with the objective of attracting positives and repelling negatives.

Given multimodal inputs $X = \{x_1, x_2, ..., x_M\}$, we define a set of candidate fusion operations $\{O_i\}_{i=1}^{B}$, such as element-wise addition, concatenation, or attention-based fusion. A gating network $G(\cdot)$ processes X and outputs a weight vector of B. The final fused representation h is computed as the weighted sum over all fusion operations, as in Equation 12.

$$h = \sum_{i=1}^{B} g_i o_i(x) \tag{12}$$

The Entity-Centric Contrastive Learning module tackles class imbalance by using attention-weighted features for positives and out-of-embedding samples for negatives.

Adaptive Fusion

The adaptive fusion module integrates multiple fusion operations with a gating mechanism that dynamically selects task-relevant strategies based on contrastively enhanced, weighted features, enabling flexible multimodal fusion.

As shown in Figure 4, X_1 and X_2 are inputs from two modalities. The module includes four fusion blocks and a global gating network, where F_r, F_t are shallow features, and F_r', F_t', F_r'', F_t'' are deeper features. The gating network selects fusion units based on task requirements.

To enhance efficiency, deeper operations are skipped when shallow features suffice. The gating network selects operations based on task complexity. A resource-aware loss is introduced, with $C(O_{i,j})$ representing the cost of operation j in fusion unit i. During inference, the adaptive module dynamically adjusts fusion strategies based on input features, managing modality uncertainty through resource-aware optimization (Eq. 13).

$$\Gamma_{\text{fusion}} = \Gamma_{\text{task}} + \sum_{j=1}^{F} \sum_{i=1}^{B} g_i^{(j)} C(O_{i,j}) \tag{13}$$

Here, Γ_{task} denotes the task loss, $g_i^{(j)}$ is the policy weight assigned to the j-th operation by the i-th fusion cell, B is the total number of candidate operations, and F is the number of fusion cells. The final training objective of the CL-DMDF model combines the attention loss Γ_{atten} (Equation 5), contrastive loss Γ_{contra} (Equation 11), and fusion loss Γ_{fusion} (Equation 13), formally defined in Equation 14.

$$\Gamma = \Gamma_{\text{atten}} + \Gamma_{\text{contra}} + \Gamma_{\text{fusion}} \tag{14}$$

Dynamic Multimodal Fusion Algorithm

This section outlines the general CL-DMDF algorithm. Given a multimodal data set D, hyperparameter τ, and iteration count num, the model initializes and outputs the fused feature representations. The computation is defined as follows.

- **Step 1**: Initialize all modalities and their corresponding feature embeddings, then extract and embed modality-specific features from the dataset.
- **Step 2**: Extract modality features and enhance embeddings by minimizing a temperature-scaled contrastive loss between positive and negative pairs.
- **Step 3**: Compute attention scores based on feature and modality averages, and fuse features via the adaptive module to obtain the final representation.

For full implementation details, please refer to the **supplementary material**.

Experiment

We evaluated CL-DMDF on three public datasets, with experiments including setup details, comparisons, and ablations.

Experimental Setup

Datasets. *MM-IMDB* contains approximately 25,000 training and 25,000 testing movie reviews, each consisting of textual content and a corresponding movie poster image, labeled with either positive or negative sentiment. *NYU Depth V2* includes 1,449 images captured from more than 400 indoor scenes, each with pixel-level depth annotations. *CMU-MOSEI* consists of more than 23,000 video clips, each annotated with sentiment categories and containing data from multiple modalities.

To improve the accuracy of the model, we progressively adjust key hyperparameters during training. Selected experimental settings for CL-DMDF are summarized in Table 1. Please refer to the **supplementary material** for specific extractor configurations.

Baseline Methods. To ensure fair evaluation across heterogeneous tasks, we adopt dataset-specific metrics. *MM-IMDB*, a multi-label text-image classification dataset, uses *Micro-F1* and *Macro-F1*, capturing global and per-class performance, respectively. *NYU Depth V2*, for semantic segmentation, is evaluated via *Mean Intersection over Union (MIoU)*. *CMU-MOSEI*, a multimodal sentiment benchmark, reports *Accuracy (Acc)* and *Mean Absolute Error (MAE)* for classification and regression. This metric selection supports a comprehensive and task-aligned assessment of CL-DMDF.

Parameter	MM-IMDB	NYU-V2	CMU-MOSEI
Epoch	500	500	500
Batch size	32	32	32
Learning rate	0.001	0.001	0.001
Vector dimension	200	200	200

Table 1: Hyperparameter settings of CL-DMDF across different datasets.

	Model	MicroF1	MacroF1
Unimodal	Unimodal Text	59.37	47.59
	Unimodal Image	40.31	25.76
Encoder-Decoder	LRMF	58.95	50.73
	CCA	60.31	50.45
	MFM	56.44	48.53
	ReFNet	59.45	51.51
Attention-Based	RMFE	58.67	49.82
	DynMM	60.35	51.60
Graph-Based	MI-Matrix	55.87	46.77
	CL-DMDF(ours)	**63.25**	**53.28**

Table 2: Comparison of CL-DMDF with baseline methods on MM-IMDB dataset.

	Model	MIoU (%)	MAdds
Encoder-Decoder	ACNet	48.3	126.2
	CEN	51.1	618.3
Attention-Based	ESANet	50.6	56.9
	LW-RefineNet	43.6	**38.5**
Graph-Based	MLTC	50.4	147.6
	CL-DMDF (ours)	**52.3**	**43.4**

Table 3: Comparison of CL-DMDF with baseline methods on NYU Depth V2 dataset.

	Model	Acc (%)	MAE
Encoder-Decoder	MFM	78.1	0.951
	CM-BERT	84.5	**0.729**
	LRMF	76.4	0.912
	MUIT	83.0	0.871
Attention-Based	MARN	77.1	0.968
Graph-Based	MLTC	78.4	0.922
	CL-DMDF (ours)	**85.4**	**0.737**

Table 4: Comparison of CL-DMDF with baseline methods on the CMU-MOSEI dataset.

We evaluate CL-DMDF on *MM-IMDB*, *NYU Depth V2*, and *CMU-MOSEI*, comparing it against three fusion model categories and unimodal baselines.

(1)**Unimodal Models**: Text-only and image-only baselines are included to assess the standalone discriminative power of individual modalities.

(2) **Encoder-Decoder Models**: CCA (Sun et al. 2020), MFM (Braz et al. 2021), ReFNet (Sudharsan, Diyi, and Soon 2021), MUIT (Bai et al. 2019), CM-BERT (Yang, Xu, and Gao 2020), and CEN (Wang et al. 2020);

(3) **Attention-Based Models**: MulCon (Chih et al. 2022), MARN (Amir et al. 2018), LW-RefineNet (Nekrasov, Shuai-Hua, and Reid 2018), and ESANet (Schlegel et al. 2021);

(4) **Graph-Based Models**: MLTC (Wang, Dai et al. 2022) and MI-Matrix (Muralidhar et al. 2019), which take advantage of graph reasoning for modality interaction.

The selected benchmarks cover diverse modalities (discrete, continuous, temporal), validating our model's ability to handle dynamic multimodal interactions and supporting future adaptation to unseen modalities.

Experiment Results

Comparative Experiments. To thoroughly evaluate the performance of CL-DMDF, we performed comparative experiments on the MM-IMDB, NYU Depth V2 and CMU-MOSEI datasets. The results are presented in Tables 2-4.

Table 2 summarizes the results of CL-DMDF on the MM-IMDB dataset. The unimodal text model outperforms the image-only model, highlighting the advantage of text features. Most baseline fusion methods lacking modality-aware adaptation offer limited improvement. DynMM achieves Micro-F1 and Macro-F1 scores of 60.35% and 51.60%, respectively, while CL-DMDF's adaptive fusion and contrastive learning methods improve by 2.9% and 1.68%, respectively, with Micro-F1 and Macro-F1 scores reaching 63.25% and 53.28%, respectively.

Table 3 summarizes CL-DMDF's results on the NYU Depth V2 dataset, comparing it with state-of-the-art semantic segmentation methods. CL-DMDF achieves the highest MIoU (52.3%), surpassing CEN by 1.2% while using only 1/14 of its computational cost. While CEN's dynamic fusion strategy incurs high resource overhead, CL-DMDF balances performance and efficiency through task-adaptive fusion, outperforming all baselines in both accuracy and resource use.

Table 4 shows CL-DMDF's results on the CMU-MOSEI dataset, comparing it with baselines, including CM-BERT. While CM-BERT improves performance through cross-modal fine-tuning, CL-DMDF outperforms with higher accuracy (85.4%, +0.9%) and competitive MAE (0.737) by leveraging adaptive feature weighting and contrastive centroid embeddings, demonstrating superior representation learning.

Ablation Study. Training settings: To assess the effectiveness of the proposed training strategy, we conduct ablation studies on the MM-IMDB dataset under the following three settings:

- **-D (Dual-Dimensional Attention Mechanism)**: Removes the dual-dimensional attent ion mechanism.
- **-G (Entity-Centroid Contrastive Learning)**: Removes the contrastive learning module; the fused features are not enhanced via contrastive learning.

Model	MicroF1 (%)	MacroF1 (%)
CL-DMDF(ours)	63.25	53.28
CL-DMDF-D	63.21	53.25
CL-DMDF-D-G	62.56	52.45
CL-DMDF-D-G-F	59.87	50.41

Table 5: Ablation study results on the MM-IMDB dataset.

Model	MicroF1 (%)	MacroF1 (%)
MulCon	62.14	52.03
MCL	62.38	52.26
MLTC	62.76	52.45
C-GMVAE	63.03	52.87
CL-DMDF(ours)	**63.25**	**53.28**

Table 6: Comparison of CL-DMDF with contrastive learning baselines on the MM-IMDB dataset.

- **-F (Adaptive Fusion)**: Replaces the adaptive fusion module with a static fusion strategy, where modality features are fused using a fixed rule.

Table 5 shows that replacing the adaptive fusion module with a static one causes a performance drop of over 5% across all metrics, underscoring the importance of dynamic fusion. Ablation experiments also validated CL-DMDF's effective adaptation under noisy or missing modes.

Removing the dual-dimensional attention or contrastive learning module led to consistent performance drops, confirming their effectiveness in weighting and feature discrimination.

Contrastive Learning Method Experiments. To evaluate the contrastive learning module, we conducted ablation studies by replacing it with existing methods, including MulCon (Chih et al. 2022), MCL (Xu, Feng, and Huang 2022), MLTC (Wang, Dai et al. 2022), and C-GMVAE (Bai, Kong, and Gomes 2022). As shown in Table 6, our module consistently outperforms all baselines. Centroid-based type embeddings also outperform random initialization, capturing inter-class semantics and improving feature discrimination.

Parameter Analysis. In the entity-centroid contrastive module, the temperature parameter τ controls similarity scaling between positive and negative pairs, affecting embedding separation. A small τ overemphasizes hard negatives and distorts embeddings. Figure 5 shows that $\tau = 0.1$ achieves optimal results on the MM-IMDB dataset, with 83.2% Macro-F1 and 84.9% Micro-F1, enabling effective multimodal integration and improving downstream performance.

Analysis of Dual-Dimensional Attention Mechanism. To assess the effectiveness of the dual-dimensional attention mechanism, we conduct a case study (Table 7) using three test samples A_1^r, A_2^r, A_3^r, in which semantically salient modalities are marked in red. The attention module assigns higher scores to informative modalities, validating its ability to identify modality relevance.

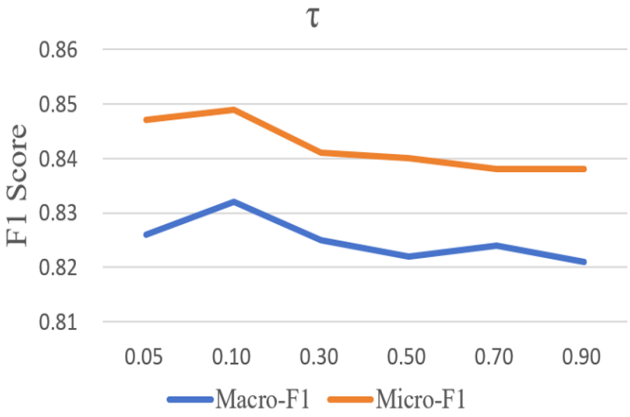

Figure 5: Effect of parameter τ on MM-IMDB.

Poster	A^I	Plot	A^T	Number	A^K
	**0.824**	The desert can be a lonely place for the ...	0.370	29	**0.822**
	**0.373**	A young boy struggles on his own in a run ...	**0.708**	38	**0.816**
	**0.659**	A documentary directed by one of their own ...	**0.793**	5	0.307

Table 7: Case study on dual-dimensional attention module.

As shown in Table 7, key modalities received high reliability scores, while less informative cues (e.g., emotion) scored lower, confirming the dual-dimensional attention effectively highlights salient modalities.

Conclusion

We propose Dynamic Multimodal Data Fusion model based on Contrastive Learning(CL-DMDF) that addresses semantic inconsistency and task-adaptive fusion. CL-DMDF integrates dual-dimensional attention for reliable weighting, centroid-based contrastive learning for discriminative representations, and an adaptive fusion module to balance accuracy and efficiency. Experiments on three benchmarks demonstrate consistent improvements over state-of-the-art methods. Future work will explore extensions to incomplete modalities and real-time scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62472204, 52574191, 62072220), the youth talent support program of 'Xing Liao Talent Program' (XLYC2203003), the Basic Scientific Research Project of the Department of Education of Liaoning Province (LJ232510140001).

References

Amir, Z.; Liang, P. P.; Poria, S.; et al. 2018. Multi-Attention Recurrent Network for Human Communication Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5642.

Bai, J. W.; Kong, S. F.; and Gomes, C. P. 2022. Gaussian Mixture Variational Autoencoder with Contrastive Learning for Multi-Label Classification. In *Proceedings of the International Conference on Machine Learning*, 1383–1398.

Bai, S.; Liang, P. P.; Salakhutdinov, R.; et al. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the Conference of the Association for Computational Linguistics*, 6558.

Braz, L.; Teixeira, V.; Pedrini, H.; et al. 2021. Image-text Integration Using a Multimodal Fusion Network Module for Movie Genre Classification. In *Proceedings of the 11th International Conference of Pattern Recognition Systems*, 200–205.

Camille, C.; Clément, F.; and Laurent, N. 2013. Indoor Semantic Segmentation Using Depth Information. *arXiv*. ArXiv:1301.3572.

Chen, Y.; Wang, H.; and Zhang, J. 2024. Progressive Cross-Modal Attention Mechanism for Hierarchical Multimodal Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11245–11254.

Chih, C. H.; Pi, J. T.; Ting, C. Y.; et al. 2022. A Comprehensive Study of Spatiotemporal Feature Learning for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7130–7134.

Dhruv, K.; Jaipal, S. G.; Manish, G.; et al. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of the World Wide Web Conference*, 2915–2921.

Ding, C. Y.; Sun, S. L.; and Zhao, J. 2023. MST-GAT: A Multimodal Spatial–Temporal Graph Attention Network for Time Series Anomaly Detection. *Information Fusion*, 89: 527–536.

Gao, P.; Jiang, Z. K.; You, H. X.; et al. 2019. Dynamic Fusion with Intra-and Inter-Modality Attention Flow for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6639–6648.

Hu, J. W.; Liu, Y. C.; Zhao, J. M.; et al. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 5666–5675.

Hu, Z. W.; Feng, G.; Sun, J. Y.; et al. 2020a. Bi-Directional Relationship Inferring Network for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4424–4433.

Hu, Z. W.; Feng, G.; Sun, J. Y.; et al. 2020b. Bi-Directional Relationship Inferring Network for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4424–4433.

Janani, V.; Li, T.; Hamid, R. H.; et al. 2021. Multimodal Deep Learning Models for Early Detection of Alzheimer's Disease Stage. *Scientific Reports*, 11: 1–13.

Jiang, Y.; Zhang, R.; and Tan, J. 2024. UniFM: Unified Multimodal Fusion with Cross-modal Token Sharing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, R. Y.; Li, K. C.; Kuo, Y. C.; Shen, X. Y.; and Jia, J. 2018. Referring Image Segmentation via Recurrent Refinement Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5745–5753.

Li, Z.; Xu, C.; and Liu, H. 2025. MM-TokenMixer: Token Mixing for Efficient Multimodal Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, Q.; Zhao, Y.; and Zhang, W. 2025. Exploring Adaptive Sparse Attention for Multimodal Fusion under Uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6993–7001.

Liu, S.; Fan, Z.; and Li, M. 2025. LVMF: Latent-Variable Multimodal Fusion with Uncertainty-Aware Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. To appear.

Liu, Z. Y.; Shi, S.; Duan, Q. T.; et al. 2019. Salient Object Detection for RGB-D Image by Single Stream Recurrent Convolution Neural Network. *Neurocomputing*, 363: 46–57.

Mateusz, M.; Carl, D.; Adam, S.; et al. 2018. Learning Visual Question Answering by Bootstrapping Hard Attention. In *Proceedings of the European Conference on Computer Vision*, 3–20.

Muralidhar, S.; Jayakumar; Menick, J.; et al. 2019. Multiplicative Interactions and Where to Find Them. https://openreview.net/pdf?id=rylnK6VtDH. Accessed: 2025-05-10.

Nekrasov, V.; Shuai-Hua, S.; and Reid, I. 2018. Lightweight RefineNet for Real-Time Semantic Segmentation. In *Proceedings of the British Machine Vision Conference*, 125.

Nihar, B.; Kevin, D.; and Peyman, N. 2021. Generalized Zero-Shot Learning Using Multimodal Variational Auto-Encoder with Semantic Concepts. In *Proceedings of the 2021 IEEE International Conference on Image Processing*, 1284–1288.

Schlegel, D.; Lengerich, B.; Weninger, T.; et al. 2021. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis. In *Proceedings of the International Conference on Robotics and Automation*, 13525–13531.

Sudharsan, S.; Diyi, Y.; and Soon, L. 2021. Refining Multimodal Representations Using a Modality-Centric Self-Supervised Module. https://openreview.net/pdf?id=hB2HIO39r8G. Accessed: 2025-05-10.

Sun, L. C.; Liu, B.; Tao, J. H.; et al. 2021. Multimodal Cross- And Self-Attention Network for Speech Emotion Recognition. In *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4275–4279.

Sun, Z.; Sarma, P.; Scherlis, W.; et al. 2020. Learning Relationships Between Text, Audio, And Video Via Deep Canonical Correlation for Multimodal Language Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8992–8999.

Wang, R.; Dai, X.; et al. 2022. Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 672–679.

Wang, Y.; Huang, W.; Sun, F.; et al. 2020. Deep Multimodal Fusion by Channel Exchanging. *Advances in Neural Information Processing Systems*, 33: 4835–4845.

Wang, Y.; Lin, J.; and Song, J. 2025. Modality-Aware Reinforced Graph Fusion for Social Event Detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 234–245.

Wu, K.; Sun, X.; and Huang, Y. 2024. GNN-Adapter: Plug-and-Play Graph Modules for Pretrained Multimodal Transformers. In *Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xu, K. L.; Feng, M.; and Huang, W. Q. 2022. Seeing Speech: Magnetic Resonance Imaging-Based Vocal Tract Deformation Visualization Using Cross-Modal Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6947–6949.

Yang, K.-C.; Xu, H.; and Gao, K. 2020. CM-BERT: Cross-modal BERT for Text-Audio Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 521–528.

Yang, M.; Tan, Z.; and Gao, L. 2024. Dynamic Sparse Transformer for Robust Video-Text Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11720–11730.

Yang, X.; Peng, J. W.; Wang, Z. H.; et al. 2023. Transforming Visual Scene Graphs to Image Captions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 12427–12440.

Zhang, W.; Hu, Z.; and Tan, F. 2025. LightFusion: Efficient Transformer-based Multimodal Fusion for Real-Time Applications. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. To appear.