PLUM-Net: Prototype-Induced Label Structuring for Disentangled Multimodal Representation Network

Kehan Wang[1], Huan Zhao[1*], Yong Wei[1], Xupeng Zha[1], Guanghui Ye[1], Cheng Zhu[1], Yiming Liu[1], Zixing Zhang[1*]

[1]College of Computer Science and Electronic Engineering, Hunan University, China
hzhao@hnu.edu.cn, zixingzhang@hnu.edu.cn

Abstract

Existing multimodal representation learning approaches often rely on simple feature concatenation or unified transformations, which fail to effectively disentangle and leverage common and private information across different modalities in a progressive manner. Moreover, they typically lack adaptive modeling tailored to specific task requirements. To address these limitations, we propose a Prototype-Induced Label Structuring for Disentangled Multimodal Representation Network (PLUM-Net). It first employs a Multilevel Semantic Alignment module to synchronize global and local semantics across audio, visual and text streams. On this aligned foundation, a Prototype-based Single-modal Label Generation module derives modality-specific hard and soft labels that subtly steer the network toward a cleaner split between shared and private cues. Guided by these labels, the Task-conditioned Feature Bifurcator module channels information through the most beneficial common or private pathway for the given task, after which a Private Refinement module polishes and fuses each modality's idiosyncratic signals. Extensive experiments show that PLUM-Net delivers strong performance on datasets such as CMU-MOSI, CMU-MOSEI and UR-FUNNY, achieving an ACC-2 of 90.3% on CMU-MOSI, representing a 2%–4% improvement over previous SOTA models.

Introduction

In recent years, multimodal representation learning has become indispensable in intelligent human–computer interaction, cognitive computing and information retrieval (Cao et al. 2024; Yang et al. 2024b). Compared to unimodal paradigms, multimodal methodologies jointly exploit heterogeneous data streams (audio, visual and text) to encode complementary semantic cues, thereby improving model expressiveness and downstream task performance (Yang et al. 2025; Wang et al. 2025a).

However, current multimodal architectures exhibit two inherent weaknesses in representation extraction and fusion (Zhuang et al. 2025). First, **there exists latent entanglement between modality-common and modality-private information.** Each modality delivers both common information (cross-modal regularities) and private information (modality-exclusive details) (Zhang et al. 2024). However, most existing methods still rely on simple concatenation or linear fusion, without explicit disentanglement (Yang et al. 2025). As a result, common cues blur and private cues are suppressed, eroding generalization and discriminative power (Zhang et al. 2022). Second, **there are homogeneous supervisory signals that fail to accommodate modality heterogeneity.** Conventional frameworks impose a single task-level annotation on each modality and their joint representation, ignoring systematic distributional differences between unimodal and multimodal feature spaces (Xu et al. 2025; Wang et al. 2025b). Such uniform supervision fails to provide modality-aware feedback, hindering the alignment of unimodal embeddings with the multimodal fusion space, ultimately underutilizing modality-private signals (Yang et al. 2024a).

To address these limitations, we propose the Prototype-Induced Label Structuring for Disentangled Multimodal Representation Network (PLUM-Net). Guided by a modality-private prototype clustering strategy, PLUM-Net follows a progressive pipeline that begins with the explicit separation of common and private information, followed by their structured integration. First, the Multilevel Semantic Alignment module performs intra-modal, category-wise alignment by pulling features toward class prototypes and compacting centroids to minimize within-class variance; it then enforces cross-modal alignment to co-locate homologous centroids, thereby distilling high-fidelity common semantics while retaining modality-exclusive geometry. Next, the Prototype-based Single-modal Label Generator module applies adaptive clustering and distance-aware scoring to emit dynamic hard and soft labels for each modality, providing fine-grained, modality-aware supervision that enhances both complementarity and discriminability. Subsequently, the Task-conditioned Feature Bifurcator module uses these labels to disentangle task-relevant common cues from complementary private cues: its common branch extracts high-level semantics, whereas its private branch captures modality-private signals. Lastly, the Private Refinement module reconciles refined private features with distilled common features under soft labels guidance, yielding representations that are both globally coherent and locally discriminative.

Our main contributions are summarized as follows:

*These authors contributed equally.

- We propose PLUM-Net, a novel multimodal disentanglement framework that leverages prototype-based clustering and distance metrics to dynamically generate modality-private soft and hard labels, enabling the explicit separation and effective integration of common and private features.
- We design a prototype-based single-modal label generation mechanism that transforms global task labels into modality-private supervision signals, allowing for more targeted and discriminative unimodal feature learning.
- Extensive experiments on multiple public benchmarks demonstrate that PLUM-Net achieves superior performance in terms of classification accuracy and robustness.

Related Work

Multimodal semantic alignment and feature fusion have recently become core topics in multimodal learning. MULT (Bhattacharjee et al. 2022) performs modality fusion via gated mechanisms but does not explicitly align cross-modal semantics. MAG-BERT (Rahman et al. 2020) introduces modality-adaptive guidance to reinforce text-centric cooperation. MISA (Hazarika, Zimmermann, and Poria 2020) maps multimodal features into shared and private subspaces, laying the structural foundation for explicit alignment.

To alleviate semantic inconsistency across modalities, researchers have progressively incorporated contrastive and alignment objectives. CM-BERT (Yang, Xu, and Gao 2020) applies a cross-modal contrastive loss to encourage consistency, whereas UniMSE (Hu et al. 2022) achieves soft alignment with a shared-teacher distribution, alleviating bias introduced by hard labels. Foal-Net (Li et al. 2024a) first aligns audio–video affect and then corrects semantic drift through cross-modal emotion matching. GSIFN (Jin 2024) embeds graph structure and interleaved masking into a Transformer, enabling low-cost, multi-scale alignment.

Beyond feature-level objectives, Self-MM (Yu et al. 2021) employs multi-task self-supervision to generate pseudo-labels for each modality, thereby guiding gradient flow. MIMM (Han, Chen, and Poria 2021) maximizes mutual information to suppress redundancy; BBFN (Han et al. 2021) uses a "dual–dual modality" adversarial scheme to reinforce both relevance and diversity; MGHF (Sami et al. 2025) combines cross-modal attention with a gated recurrent hierarchy to capture long-range temporal dependencies.

Label modeling has likewise advanced. Foal-Net (Li et al. 2024a) proposes cross-label matching, while C-MIB (Mai, Zeng, and Hu 2022) incorporates an information-bottleneck regulariser to align affective semantics at the label level and reduce semantic noise.

Despite steady progress in alignment, soft-label supervision and redundancy suppression, real-world systems still confront label noise and mismatched semantic granularity. To tackle these challenges, we introduce PLUM-Net, which integrates a Multilevel Semantic Alignment module, a Prototype-based Single-modal Label Generation module, a Task-conditioned Feature Bifurcator module, and a private Refinement module, capturing cross-modal commonalities while preserving modality-specific signals and thereby improving the practical applicability of multimodal learning.

Method

In this section, we systematically detail the overall architecture of our proposed Prototype-Induced Label Structuring for Disentangled Multimodal Representation Network. As depicted in Figure 1, PLUM-Net consists of four essential modules: the Multilevel Semantic Alignment module, Prototype-based Single-modal Label Generation module, Task-conditioned Feature Bifurcator module and private refinement module. The theoretical foundations and implementation details of each module are discussed sequentially in the following subsections.

Definition

Multimodal representation learning aims to fuse heterogeneous signals (audio, video and text) so that complementary information from each modality can jointly enhance semantic understanding and improve generalization. Formally, each modality's data is represented by a sequence denoted by X_m, where $m \in \{a, v, t\}$ corresponds to audio, video and text modalities, respectively.

Multilevel Semantic Alignment Module

To achieve hierarchical discriminative optimization and globally coherent representations within the multimodal feature space, PLUM-Net introduces a Multilevel Semantic Alignment module at the input stage. The module adopts a unified contrastive learning paradigm to systematically align intra-modal and inter-modal representations.

First, an intra-modal contrastive loss $\mathcal{L}_{\text{intra}}^{(m)}$ is applied to each modality $m \in a, v, t$, encouraging embeddings from the same class to be pulled closer together while pushing those from different classes apart. This enhances the structural consistency and discriminative power of unimodal features:

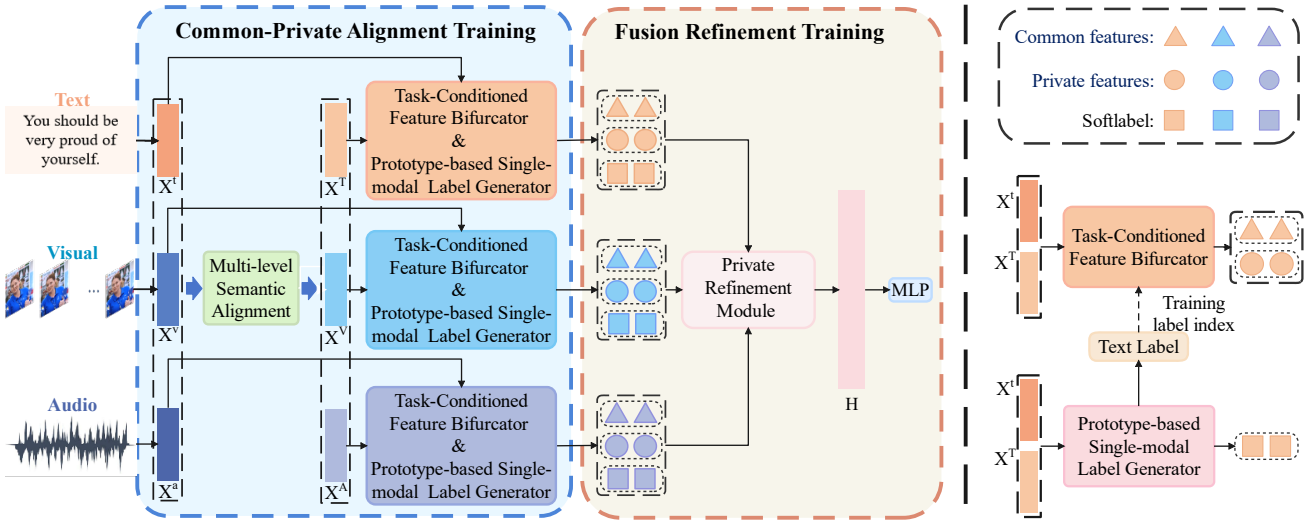
$$\mathcal{L}_{\text{intra}}^{(m)} = -\frac{1}{N_m} \sum_{i=1}^{N_m} \log \frac{\exp(\text{sim}(\mathbf{h}_i^{(m)}, \mathbf{h}_{i^+}^{(m)})/\tau)}{\sum_{j=1}^{N_m} \exp(\text{sim}(\mathbf{h}_i^{(m)}, \mathbf{h}_j^{(m)})/\tau)}, \tag{1}$$

where N_m is the number of training instances in modality m, $\mathbf{h}_i^{(m)}$ denotes the embedding of sample i in modality m, $\mathbf{h}_{i^+}^{(m)}$ represents a positive embedding from the same class, $\text{sim}(\cdot, \cdot)$ denotes a similarity function (e.g., cosine similarity) and τ is a temperature parameter.

Building upon this intra-modal alignment, the Multilevel Semantic Alignment module introduces an inter-modal contrastive loss $\mathcal{L}_{\text{inter-modal}}$ that encourages embeddings from the same class across different modalities to be closer, while pushing embeddings from different classes apart:

$$\mathcal{L}_{\text{inter-modal}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{h}_i^{\text{com}}, \mathbf{h}_{i^+}^{\text{com}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{h}_i^{\text{com}}, \mathbf{h}_j^{\text{com}})/\tau)}, \tag{2}$$

where $\mathbf{h}_i^{\text{com}}$ is the common embedding produced by a unified projection layer and N is the total number of samples. These



(a) Overall framework of PLUM-Net

(b) Task-Conditioned Feature Bifurcator & Prototype-based Single-modal Label Generator Specific Flowchart

Figure 1: (a) The overall framework of PLUM-Net. During training, PLUM-Net performs sentence-level feature extraction through a two-stage training paradigm. In the Common–Private Alignment Training (blue background), the Multilevel Semantic Alignment module (green block) is first trained, followed by the Prototype-based Single-modal Label Generator module and the Task-conditioned Feature Bifurcator module. In the Fusion Refinement Training (yellow background), based on the Common-Private Alignment Training, the Private Refinement module is used to optimize the single-modal and joint representations of each modality. (b) Task-conditioned Feature Bifurcator & Prototype-based Single-modal Label Generation Specific Flowchart.

inter-modally aligned features are then passed to the common expert branch, providing a unified semantic substrate across modalities as well as highly discriminative cues for subsequent modules.

The overall loss function for the Multilevel Semantic Alignment module is defined as:

$$\mathcal{L}_{\text{MSA}} = \mathcal{L}_{\text{inter-modal}} + \mathcal{L}_{\text{intra}}^{(m)}. \tag{3}$$

By jointly optimizing local (intra-modal) and global (inter-modal) contrastive objectives, the Multilevel Semantic Alignment module hierarchically extracts common semantic information, providing both a unified semantic substrate across modalities and highly discriminative cues for the Prototype-based Single-modal Label Generation module and Task-conditioned Feature Bifurcator module.

Prototype-based Single-modal Label Generator Module

Traditional multimodal learning frameworks usually supervise each modality with identical task labels, disregarding modality-specific differences in expressive form, semantic distribution, and discriminative salience. This uniform supervision blurs modality-private structure and weakens unimodal discriminability.

To recover and amplify these cues, PLUM-Net first obtains aligned features through the Multilevel Semantic Alignment module and then feeds them to a Prototype-based Single-modal Label Generator module. For a fixed modality m and a sample i, we compute distances from its aligned unimodal feature $\tilde{\mathbf{f}}^{(m)}(i)$ to *all* class prototypes of that modality; a distance-based softmax then yields per-modality soft labels, and the nearest prototype gives the hard labels. Given the aligned features $\tilde{\mathbf{f}}^{(m)}$, the module computes a modality-private prototype for each class k:

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \tilde{\mathbf{f}}^{(m)}(i), \tag{4}$$

where $\mathcal{S}_k$ represents the index set of samples belonging to class k in modality m. These prototypes characterize the intrinsic geometric structure of each modality's feature space.

For each sample $\tilde{\mathbf{f}}_{\text{single}}^{(m)}(i)$, the Euclidean distances $d_k = \|\tilde{\mathbf{f}}_{\text{single}}^{(m)}(i) - \mathbf{c}_k\|_2$ to all prototypes are measured and converted into a soft labels distribution using a distance-based softmax:

$$p_k = \frac{\exp(-d_k)}{\sum_j \exp(-d_j)}, \tag{5}$$

where p_k represents the probability that the sample belongs to class k. The accompanying hard labels is determined by assigning the sample to its nearest prototype.

By providing modality-aware soft and hard labels, the prototype-based single-modal label generator preserves modality-private structure and significantly improves the discriminability and robustness of PLUM-Net during multimodal fusion and downstream prediction.

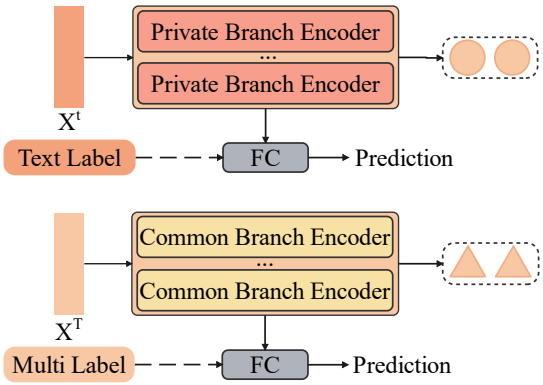


Figure 2: Architecture of the Task-conditioned Feature Bifurcator module

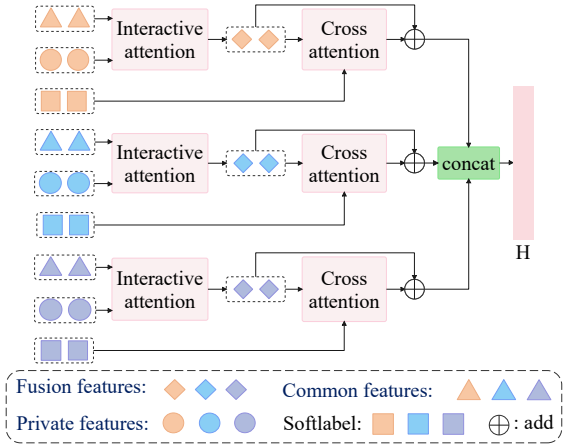


Figure 3: Architecture of the Private Refinement module.

Task-conditioned Feature Bifurcator Module

In multimodal learning, an ideal semantic representation must encode both cross-modal common semantics and unimodal, fine-grained discriminative cues (Wang et al. 2024). To meet this requirement, as depicted in Figure 2, PLUM-Net equips each modality (audio a, visual v and text t) with a Task-conditioned Feature Bifurcator module. Within this module, each modality is split into: (1) a common branch, which models task-relevant cross-modal semantics and (2) a private branch, which captures modality-private details.

For each branch, we employ a Transformer encoder structure. The common branch takes the common representations generated by the Multilevel Semantic Alignment module and refines them through task-oriented layers, thereby enhancing the semantic expressiveness and discriminative capability of high-level common features. Conversely, the private branch operates directly on the original modality-private features $\tilde{\mathbf{f}}_{\text{single}}$, extracting an initial set of private cues as the basis for subsequent fine-grained modeling. This hierarchical bifurcation allows PLUM-Net to represent common and private information at multiple granular levels.

Training proceeds with two cross-entropy loss. The common branch is supervised using multimodal ground-truth labels:

$$\mathcal{L}_{\text{multi}}^{(m)} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,k}\log\left(\hat{y}_{i,k}^{(m)}\right), \tag{6}$$

where $y_{i,k}$ is the ground-truth label for i in class k and $\hat{y}_{i,k}^{(m)}$ is the probability predicted by the common branch of modality m.

The private branch is supervised using modality-private labels generated by the Prototype-based Single-modal Label Generation module:

$$\mathcal{L}_{\text{single}}^{(m)} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,k}^{(m)}\log\left(\hat{y}_{i,k}^{(m,\text{priv})}\right), \tag{7}$$

where $\hat{y}_{i,k}^{(m,\text{priv})}$ denotes the probability predicted by the private branch.

The overall loss for Task-conditioned Feature Bifurcator module is the sum of common and private-branch losses over all three modalities:

$$\mathcal{L}_{TFB} = \sum_{m\in\{a,v,t\}}\left[\mathcal{L}_{\text{multi}}^{(m)} + \mathcal{L}_{\text{single}}^{(m)}\right], \tag{8}$$

where a, v, and t represent the audio, visual, and text modalities, respectively. By explicitly disentangling task-relevant shared semantics from modality-specific information, the Task-conditioned Feature Bifurcator module achieves a dual goal: it provides globally coherent signals essential for cross-modal reasoning while preserving the rich, discriminative cues unique to each modality, thereby reducing negative transfer and enhancing multimodal complementarity.

Private Refinement Module

After dual-branch extraction, PLUM-Net yields two complementary vectors: (1) the common representation, capturing cross-modal common semantics, and (2) the modality-private representation. Naively concatenating these vectors expands the feature space but introduces redundancy and imbalance, which may dilute task-relevant cues and weaken discriminative power (Li et al. 2024b). To address this issue, we introduce the Private Refinement Module, which refines private cues through bidirectional cross-feature attention and label-conditioned enhancement without relying on early fusion.

As shown in Figure 3, the module first applies a bidirectional cross-feature attention mechanism. In one direction, the common representation serves as the query, and the private representation acts as both key and value. In the reverse direction, the private representation becomes the query, while the common representation serves as key and value. This symmetric attention structure facilitates mutual feature interaction and explicit modeling of complementary information. Following the standard Transformer formula-

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr	ACC-2[%]	F1[%]	ACC-7[%]	MAE↓	Corr	ACC-2[%]	F1[%]	ACC-7[%]
MULT	0.767	0.799	-/83.7	-/83.7	41.5	0.625	0.775	-/84.7	-/84.7	50.7
ICCN	0.860	0.710	-/85.1	-/85.0	40.0	0.579	0.713	-/84.4	-/84.3	51.4
CubeMLP	0.770	0.767	-/85.6	-/85.5	45.5	0.529	0.789	-/86.2	-/86.2	54.9
BBFN	0.776	0.756	-/84.4	-/84.3	46.0	0.565	0.767	-/86.2	-/86.2	54.8
MIMM	0.700	0.800	84.1/86.0	84.0/86.0	46.7	0.526	0.772	82.24/86.0	82.7/85.9	54.2
UniMSE	0.691	0.809	85.9/86.9	85.8/86.4	48.7	0.523	0.773	85.9/87.5	85.8/87.5	54.4
GLoMo	0.718	0.782	84.1/86.7	83.9/86.6	48.3	0.539	0.771	83.7/86.5	84.0/86.4	55.0
Self-MM	0.713	0.798	84.0/86.0	84.4/86.0	-	0.530	0.765	82.8/85.2	82.5/85.3	-
DTMD	0.705	0.799	84.0/86.0	83.9/86.0	47.5	0.531	0.767	84.8/86.1	84.9/85.9	53.7
DMD	0.710	0.792	-/86.0	-/86.0	45.6	0.537	0.771	-/86.6	-/86.6	54.5
MM-CoT	0.647	0.798	-/88.3	-/88.2	48.1	0.486	0.798	-/88.3	-/88.4	56.0
PLUM-Net (Ours)	**0.448**	**0.818**	**88.2/90.3**	**88.1/90.3**	**66.4**	**0.372**	**0.812**	**87.2/89.6**	**87.1/89.7**	**66.8**

Table 1: Performance comparison between the proposed PLUM-Net and other SOTA multimodal learning methods on the CMU-MOSI and CMU-MOSEI datasets.

tion (Vaswani et al. 2017), this produces the fused feature $\mathbf{H}^{\mathbf{fusion}}$.

To further highlight task-relevant discriminative signals, the Private Refinement module incorporates a label-conditioned attention enhancement mechanism. Concretely, we project the fused features to serve as the Query, with the soft labels as the Key and Value. Again, following the formulation (Vaswani et al. 2017), this produces a feature with enhanced private information $\mathbf{H}^{\mathbf{private}}$.

Finally, residual enhancement generates the refined feature representation:

$$\mathbf{H} = \mathbf{H}^{\mathbf{fusion}} + \mathbf{H}^{\mathbf{private}}. \tag{9}$$

This structure enables label-aware channel reweighting, effectively amplifying high-quality discriminative signals in the private space while maintaining semantic consistency with shared representations.

Optimization Objective

The training procedure of PLUM-Net follows a two-stage optimization strategy, progressively refining both common and private representations across modalities.

Stage 1: Common-Private Alignment Training. In this stage, the Multilevel Semantic Alignment module produces semantically consistent shared features and structurally aligned unimodal features. These are passed to the Task-conditioned Feature Bifurcator module, where the common branch is supervised with multimodal ground-truth labels and the private branch is supervised with prototype-based unimodal labels. The total loss for this stage is defined as:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{TFB} + \lambda_1 \mathcal{L}_{\text{MSA}}, \tag{10}$$

where $\mathcal{L}_{TFB}$ denotes the overall loss for the Task-conditioned Feature Bifurcator module, and $\mathcal{L}_{\text{MSA}}$ represents the contrastive loss from multilevel semantic alignment. λ_1 is a hyperparameter.

Stage 2: Fusion Refinement Training. Based on representations learned in Stage 1, this stage further enhances the fused features for final task predictions.

For classification tasks:

$$\mathcal{L}_{\text{stage2}} = -\frac{1}{n}\sum_{i=1}^{n} y_i^{\text{final}} \log(\hat{y}_i^{\text{final}}). \tag{11}$$

For regression tasks:

$$\mathcal{L}_{\text{stage2}} = \frac{1}{n}\sum_{i=1}^{n} \left| y_i^{\text{final}} - \hat{y}_i^{\text{final}} \right|. \tag{12}$$

Overall Objective. The complete loss for PLUM-Net is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}_{\text{stage2}}. \tag{13}$$

This staged optimization procedure first disentangles and aligns modality-specific and shared semantics, then selectively fuses them for downstream tasks, thereby promoting robustness and task-specific generalization.

Experiments

Datasets

To evaluate the effectiveness of our proposed PLUM-Net method, we conducted experiments on four widely used multimodal emotion recognition datasets: CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018), UR-FUNNY (Hasan et al. 2019) and MUStARD (Castro et al. 2019).

Implementation Details

In the experiments, the representation dimensions are set to 256 for CMU-MOSI and 128 for CMU-MOSEI, respectively. The maximum number of training epochs for the Common-Private Alignment Training phase is 160 for CMU-MOSI and 50 for CMU-MOSEI, while the maximum number of epochs in the Fusion Refinement Training phase is 140 and 50, respectively. We use the Adam optimizer with a learning rate of 0.00005 and a dropout rate of 0.6 for CMU-MOSI.

Model	UR-FUNNY[%]	MUStARD[%]
MISA (BERT)	69.62	66.18
MISA (ALBERT)	69.82	66.18
MAG-BERT (ALBERT)	67.20	69.12
MAG-BERT (XLNet)	72.43	76.47
GLoMo	74.75	79.36
PLUM-Net (Ours)	**77.26**	**83.59**

Table 2: The comparison with baselines on UR-FUNNY and MUStARD, in terms of ACC-2. Models in parentheses indicate the text features used.

MSA	PSL&TFB	PRM	CMU-MOSI	
			ACC-2[%]	F1[%]
			83.6/86.2	83.4/86.1
✓			86.0/88.4	85.9/88.4
✓	✓		87.9/89.9	87.8/89.9
✓	✓	✓	**88.2/90.3**	**88.1/90.3**

Table 3: Ablation studies of the PLUM-Net on the MOSI dataset. MSA: Multilevel Semantic Alignment module; PSL: Prototype-based Single-modal Label Generation module; TFB: Task-conditioned Feature Bifurcator module; PRM: Private Refinement module.

Quantitative Results

Multimodal Sentiment Analysis Table 1 presents the performance of PLUM-Net on CMU-MOSI and CMU-MOSEI, compared against strong baselines and recent state-of-the-art (SOTA) models. Across every metric—MAE, Corr, ACC-2, F1 and ACC-7, PLUM-Net consistently delivers superior results. On CMU-MOSI, PLUM-Net achieves the lowest MAE (0.448) and the highest Pearson correlation (0.818), along with an ACC-2 of 88.2%, F1 of 90.3% and ACC-7 of 66.4%, all of which significantly outperform existing models such as MM-CoT and UniMSE. On CMU-MOSEI, PLUM-Net continues to improve performance, yielding an MAE of 0.372, a Corr of 0.812 and a new best ACC-7 of 66.8%. These gains highlight the effectiveness of the proposed prototype-guided supervision and task-conditioned feature bifurcation, which together enhance overall prediction accuracy and enable fine-grained sentiment modeling.

Multimodal Humor Detection and Multimodal Emotion Recognition We further evaluate PLUM-Net on two multimodal humor and sarcasm detection benchmarks: UR-FUNNY and MUStARD. As shown in Table 2, PLUM-Net achieves the highest ACC-2 score on UR-FUNNY (77.26%) and competitive performance on MUStARD (83.59%), outperforming prior strong baselines such as GLoMo (74.75% / 79.36%) and MAG-BERT (72.43% / 76.47%) under various text backbones. These results highlight the generalization ability of PLUM-Net across tasks involving subtle and implicit multimodal semantics, such as humor and sarcasm.

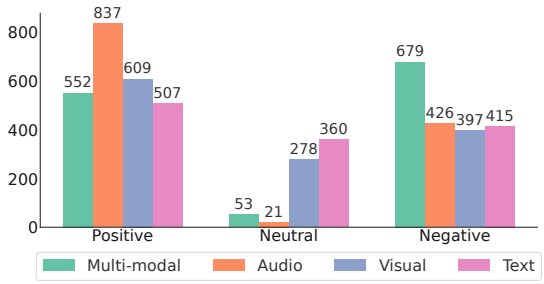


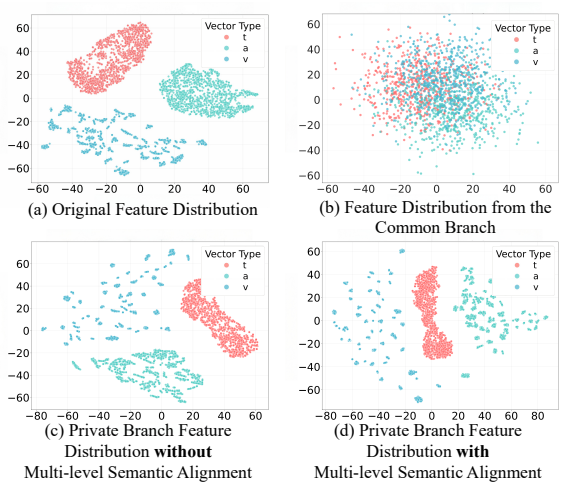
Figure 4: Distribution of multimodal and single-modal Labels

Ablation Study

Table 3 reports the results of an ablation study conducted on the CMU-MOSI dataset. We first add the MSA; this single addition already raises ACC-2 and F1, confirming that deep hierarchical alignment is far more effective than shallow fusion. We then activate PSL together with TFB so that their combined effect can be isolated. Under task-aware control, explicitly disentangling common and private features reduces cross-modal representational conflict and enables more adaptive information routing, producing another clear jump in accuracy. Finally, the PRM is introduced; by amplifying under-utilised modality-private cues, PRM delivers the strongest overall results, particularly on fine-grained sentiment recognition. With each component in place, PLUM-Net improves ACC-2 by 4.6% and F1 by 4.2% on CMU-MOSI relative to baseline, demonstrating that each module is effective on its own and synergistic in combination.

Label Distribution Analysis

To clarify why PLUM-Net employs the Prototype-based Single-modal Label Generator module, we first examine the sentiment-label distributions derived from individual modalities (audio, visual and text), which differ significantly from the original multimodal annotations. Figure 4 reveals substantial discrepancies. For example, the audio modality labels significantly more samples as positive (837 vs. 552 in the multimodal ground truth) while nearly neglecting the neutral class (21 vs. 53). Conversely, text annotations tend to favor neutral and negative emotions, and the visual modality shows a more balanced yet still distinct distribution from the fused multimodal labels. These modality-private biases indicate that each modality perceives affective cues differently. Consequently, relying solely on hard one-hot labels can obscure ambiguity, misdirect optimization, and amplify noise from dominant modalities. PLUM-Net circumvents these issues by generating prototype-guided labels for each modality, preserving modality-private disagreements and enabling the network to leverage richer distributions of emotional signals rather than prematurely collapsing them. This approach allows the model to retain modality-level nuances while benefiting from joint optimization, which is reflected clearly in our performance gains and qualitative evaluations.



(a) Original Feature Distribution

(b) Feature Distribution from the Common Branch

(c) Private Branch Feature Distribution **without** Multi-level Semantic Alignment

(d) Private Branch Feature Distribution **with** Multi-level Semantic Alignment

Figure 5: t-SNE plot of common and private features learned by PLUM-Net on MOSI.

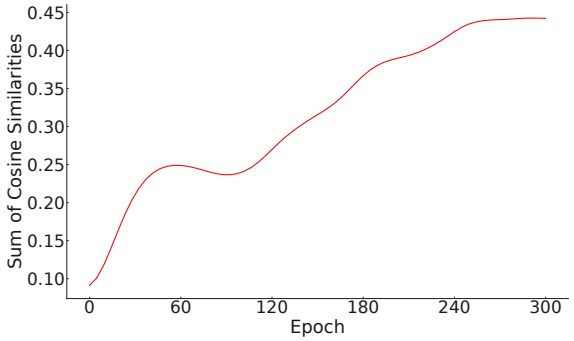


Figure 6: Semantic alignment effect of common branch without using multilevel semantic alignment.

Visualization of Feature Distributions

To illustrate how multilevel semantic alignment and the dual-branch architecture reshape the feature space, Figure 5 visualizes representations at several stages of PLUM-Net. Figure 5(a) shows raw features before modality-private encoding: the space is entangled and unstructured. After processing through the common branch, Figure 5(b) displays clear modality-invariant clusters, indicating that the common encoder captures high-level semantics even in the absence of an explicit alignment loss. Figure 5(c) and 5(d) examine the private branches. Without multilevel semantic alignment (Figure 5(c)), private features remain fragmented and modality-biased; with MSA (Figure 5(d)), they become more coherent and semantically meaningful while still preserving modality-private variance. Thus, MSA not only tightens global alignment in the common space but also improves cross-modality awareness in the private streams, preventing useful diversity from collapsing.

Together, these visual and statistical analyses validate the design philosophy of PLUM-Net: the common-private architecture facilitates disentangled yet cooperative encoding, while the Multilevel Semantic Alignment module ensures both global consistency and local specialization.

Analysis of Common Branch Consistency without Explicit Semantic Alignment Module

This subsection investigates how the common branch of the Task-conditioned Feature Bifurcator module (driven solely by the classification objective) achieves semantic consistency in the absence of the Multilevel Semantic Alignment module. Figure 6 illustrates a three-phase, U-shaped trajectory in the aggregated cosine similarity. Epoch 0–40: Similarity rises sharply as modality encoders optimize toward the same classification label, causing their common representations to converge along a task-relevant semantic axis. Epoch 40–100: Similarity declines by approximately 5% as encoders begin injecting modality-specific cues to enhance individual accuracy, increasing cross-modal divergence. Epoch 100 onward: Similarity gradually increases and stabilizes just above its peak. The common classification label continues to act as a semantic anchor, pruning away task-irrelevant dimensions and restoring most of the semantic consensus. Overall, the common branch quickly establishes task-oriented alignment. Because this alignment is exclusively driven by label semantics, it inherently prioritizes features that are most beneficial for classification. As a result, even without a dedicated alignment module, the common representations remain highly task-specific, providing a robust foundation for downstream performance.

Conclusion

PLUM-Net is designed to tackle two persistent challenges in multimodal learning: the entanglement of modality-shared and modality-private representations, and the homogenized supervision that overlooks modality heterogeneity. The framework first uses a Multilevel Semantic Alignment module to progressively shrink cross-modal gaps. It then introduces a Prototype-based Single-modal Label Generator module, which produces dynamic hard and soft labels for each modality, injecting differentiated supervision from the source. Guided by these labels, the Task-conditioned Feature Bifurcator module disentangles and optimizes common and private information in parallel. Finally, the Private Refinement module further enhances modality-specific signals through label-aware adjustment and fusion. Working together, these four modules capture rich cross-modal commonality while leveraging unimodal nuances. Benchmark experiments show that PLUM-Net surpasses existing methods in accuracy, robustness and interpretability. PLUM-Net's main limitation is that severe class imbalance or noisy label distributions can dilute soft labels quality. Moreover, its performance can degrade sharply when one or more modalities are missing. Future work will pursue imbalance-aware clustering, noise-tolerant label smoothing, and self-supervised consistency checks to bolster resilience under modality dropout, label noise, and long-tailed data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62076092, the National Natural Science Foundation of China under Grant No. 62571184, the Science and Technology Innovation Program of Hunan Province under Grant No. 2025RC6003, and the Changsha Science and Technology Bureau Foundation under Grant No. kq2402082.

References

Bhattacharjee, D.; Zhang, T.; Süsstrunk, S.; and Salzmann, M. 2022. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 12031–12041.

Cao, R.; Lei, F.; Wu, H.; Chen, J.; Fu, Y.; Gao, H.; Xiong, X.; Zhang, H.; Hu, W.; Mao, Y.; et al. 2024. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 107703–107744.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards Multimodal Sarcasm Detection. In *Proceedings of the Association for computational linguistics (ACL)*, 4619–4629.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the international conference on multimodal interaction (ICMI)*, 6–15.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hasan, M. K.; Rahman, W.; Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the ACM international conference on multimedia (MM)*, 1122–1131.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Jin, Y. 2024. GSIFN: A Graph-Structured and Interlaced-Masked Multimodal Transformer-based Fusion Network for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2408.14809*.

Li, Q.; Gao, Y.; Wen, Y.; Wang, C.; and Li, Y. 2024a. Enhancing modal fusion by alignment and label matching for multimodal emotion recognition. *arXiv preprint arXiv:2408.09438*.

Li, W.; Zhou, H.; Yu, J.; Song, Z.; and Yang, W. 2024b. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 59808–59832.

Mai, S.; Zeng, Y.; and Hu, H. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25(14): 4121–4134.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Association for computational linguistics (ACL)*, volume 2020, 2359.

Sami, S. M.; Hasan, M. M.; Saadabadi, M. S. E.; Dawson, J.; Nasrabadi, N.; and Rao, R. 2025. MGHF: Multi-Granular High-Frequency Perceptual Loss for Image Super-Resolution. *arXiv preprint arXiv:2411.13548*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30.

Wang, C.; Qi, Q.; Wang, J.; Sun, H.; Zhuang, Z.; Wu, J.; Zhang, L.; and Liao, J. 2025a. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 12694–12702.

Wang, M.; Xing, J.; Jiang, B.; Chen, J.; Mei, J.; Zuo, X.; Dai, G.; Wang, J.; and Liu, Y. 2024. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 5517–5525.

Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; and Hu, J. 2025b. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 21180–21188.

Xu, J.; Chen, Z.; Yang, S.; Li, J.; Wang, H.; and Ngai, E. C. 2025. Mentor: multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 12908–12917.

Yang, K.; Xu, H.; and Gao, K. 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the ACM international conference on multimedia (MM)*, 521–528.

Yang, P.; Liu, N.; Liu, X.; Shu, Y.; Ji, W.; Ren, Z.; Sheng, J.; Yu, M.; Yi, R.; Zhang, D.; et al. 2024a. A multimodal dataset for mixed emotion recognition. *Scientific Data*, 11(1): 847.

Yang, Y.; Ma, H.; Meng, L.; Xu, S.; Xie, R.; and Meng, X. 2025. Curriculum conditioned diffusion for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 13035–13043.

Yang, Y.; Wan, F.; Jiang, Q.-Y.; and Xu, Y. 2024b. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 62108–62122.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, 10790–10797.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the Association for computational linguistics (ACL)*, volume 1, 2236–2246.

Zhang, L.; Jin, L.; Xu, G.; Li, X.; Xu, C.; Wei, K.; Liu, N.; and Liu, H. 2024. CAMEL: capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 9341–9349.

Zhang, Y.; Chen, M.; Shen, J.; and Wang, C. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 9100–9108.

Zhuang, W.; Huang, X.; Zhang, X.; and Zeng, J. 2025. Mathpuma: Progressive upward multimodal alignment to enhance mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 26183–26191.