

---

# Theory-Driven Label-Specific Representation for Incomplete Multi-View Multi-Label Learning

---

**Quanjiang Li<sup>1†</sup>**

liquanjiang@nudt.edu.cn

**Tianxiang Xu<sup>2†</sup>**

xtx09@hotmail.com

**Tingjin Luo<sup>1\*</sup>**

tingjinluo@hotmail.com

**Yan Zhong<sup>2</sup>**

zhongyan@stu.pku.edu.cn

**Yang Li<sup>1</sup>**

liyang\_albert@foxmail.com

**Yiyun Zhou<sup>3</sup>**

yiyunzhou@zju.edu.cn

**Chenping Hou<sup>1</sup>**

hcpnudt@hotmail.com

<sup>1</sup>National University of Defense Technology, <sup>2</sup>Peking University, <sup>3</sup>Zhejiang University

## Abstract

Multi-view multi-label learning typically suffers from dual data incompleteness due to limitations in feature storage and annotation costs. The interplay of heterogeneous features, numerous labels, and missing information significantly degrades model performance. To tackle the complex yet highly practical challenges, we propose a Theory-Driven Label-Specific Representation (TDLSR) framework. Through constructing the view-specific sample topology and prototype association graph, we develop the proximity-aware imputation mechanism, while deriving class representatives that capture the label correlation semantics. To obtain semantically distinct view representations, we introduce principles of information shift, interaction and orthogonality, which promotes the disentanglement of representation information, and mitigates message distortion and redundancy. Besides, label-semantic-guided feature learning is employed to identify the discriminative shared and specific representations and refine the label preference across views. Moreover, we theoretically investigate the characteristics of representation learning and the generalization performance. Finally, extensive experiments on public datasets and real-world applications validate the effectiveness of TDLSR.

## 1 Introduction

The popularity of multi-view learning stems from its ability to provide comprehensive representations of samples [29, 9]. Integrating multi-source information effectively reveals latent semantic associations in multimodal data, thereby enriching the feature space and improving model accuracy and generalization [10]. With the advancement of information technology, single-label proves inadequate in meeting the object labeling and recognition demands [3]. In practice, objects frequently fall under multiple categories simultaneously, with a document in text classification [40] being assigned labels like topic, intended audience, sentiment and so on. Benefiting from the holistic characterization offered by multi-view approaches [41] and the capacity of multi-label methods [31] to capture all sample attributes, diverse views and labels are emerging as the dominant form of training data. Consequently, multi-view multi-label classification (MvMLC) has attracted significant research attention, as it delivers a refined understanding of the complexity and diversity inherent in real-world situations.

---

<sup>†</sup>Equal contribution

<sup>\*</sup>Corresponding Author

Existing MvMLC methods are capable of implementing both multivariate feature fusion and multi-objective joint discrimination. For example, LVSL [43] performed view-specific label learning and leveraged low-rank label structures to enhance performance.  $EF^2FS$  [12] proposed an embedded feature selection model that integrated feature aggregation and enhancement. However, these ideal methods assume the presence of both features and labels, whereas limitations in feature collection techniques and annotation complexity result in the unavailability of partial views and tags [37], motivating researchers to explore the incomplete multi-view multi-label classification (iMvMLC) approach. Early efforts like iMVWL [33] employed a joint learning strategy to refine the shared subspace and enhance the robustness of the weak label classifier. NAIM3L [21] combined matrix factorization with global high-rank and local low-rank constraints to obtain a shared label space. Given the ability to capture complex semantic information, deep learning based methods have demonstrated promising performance. The pioneering work DICNet [25] introduced an incomplete instance-level contrastive learning method for improved consensus representation, while LMVCAT [26] designed two transformer-based modules for cross-view feature aggregation and category awareness.

Despite the emergence of various effective iMvMLC methods, they remain deficient in feature information reconstruction, semantic distinctness of extracted representations and the coupling between label correlation semantics with feature connotation. (i) The absence of views severely undermines both the capture of cross-view dependencies and the stability of downstream modules. Approaches [4] that only mask missing samples tend to circumvent the challenge, rather than engaging in robust information recovery. AIMNet [23] attempted to perform missing imputation via cross-view global attention computation. However, the integration of a global weighting scheme may further exacerbate reconstruction noise since not all samples are strongly correlated. (ii) The essence of multi-view learning lies in extracting the complementarity and consistency across diverse views [36]. Additionally, increased attention should be devoted to feature separation in multi-label scenarios [42], as distinct labels exhibit varying sensitivities towards particular features. DIMC [38], TSIEN [34] and SIP [27] primarily focused on shared subspaces while overlooking the unique characteristics of individual views. MTD [24] attempted to obtain shared and specific representations through geometric distance constraints. However, this linear interaction pattern struggled to reflect the intricate inter-view relationships. (iii) Label relevance is fundamental to multi-label learning and distinguishes it from multi-class problem [44], which makes multiple independent binary classifications insufficient as done in some methods [24]. Besides, instead of being considered in isolation, mutually dependent label semantics should interact seamlessly with feature information to facilitate both the perception of label-specific features and the selection of the preferred labels corresponding to those features.

To tackle these issues, we propose a Theory-Driven Label-Specific Representation framework named TDLSR. The motivation behind TDLSR is to minimize view reconstruction errors, improve the semantic discriminability of extracted representations and strengthen the interaction between label-relevant semantics and feature information. We first construct view-specific instance relation graphs using the attention mechanism integrated with neighborhood-aware selection. Without any discrepancy arising from network parameter updates, the reconstruction risk is minimized by propagating highly relevant sample information across views. By introducing the principles of information shift, interaction and orthogonality, we develop a mutual information optimization model that aligns the core constituents of representations with their respective views, facilitates interaction among shared representations, suppresses cross-talk between private components and enforces orthogonality of shared and specific information from the same view. Label correlation information is transmitted via graph-induced relational network modeling, which results in the emergence of interdependent category representatives. Through discrete engagements between each class prototype and semantically distinct features, the most sensitive feature ensemble for each label across both shared and specific feature pools can be identified. We further establish the generalization error bound via error decomposition, showing that feature disentanglement maximizes the mutual information between representations and objects and reduces generalization error. The main contributions of our work are summarized as follows:

- We propose a general multi-view representation extraction model inspired by information theory. This model guarantees unbiased representations through the constraint of information shift, separates shared and specific semantic via the regulation of information interaction, and eliminates representation redundancy by strengthening information orthogonality.
- TDLSR enhances the propagation of feature dependencies and label correlation semantics by constructing relational graphs. Besides, it introduces label-specific shared and private feature

learning for the first time. Theoretically, we prove the discriminability and effectiveness of feature extraction and derive the generalization error bound.

- Extensive experimental results across diverse public available datasets, along with applications on real-world NBA data, validate the effectiveness and robustness of our method.

## 2 Method

In this section, we present the following critical components of our TDLSR as shown in Fig. 1: proximity-aware graph attention recovery mechanism, universal view extraction framework under mutual information constraints and multi-label semantic and label-specific representation learning.

### 2.1 Problem definition

We define  $\{\mathbf{X}^{(v)}\}_{v=1}^V$  as original multi-view setting, where  $\mathbf{X}^{(v)} = \{\mathbf{x}_i^{(v)}\}_{i=1}^N \in \mathbb{R}^{N \times d_v}$  represents  $d_v$  dimensional feature matrix of the  $v$ -th view. The label matrix  $\mathbf{Y} \in \{0, 1\}^{N \times C}$  corresponds to  $C$  categories, with  $\mathbf{Y}_{i,j} = 1$  if sample  $i$  is tagged as class  $j$ . Besides, we let  $\mathbf{W} \in \{0, 1\}^{N \times V}$  and  $\mathbf{G} \in \{0, 1\}^{N \times C}$  denote the missing indicator for views and labels, respectively. Specifically,  $\mathbf{W}_{i,j} = 1$  if the  $j$ -th view of the  $i$ -th sample is available, otherwise  $\mathbf{W}_{i,j} = 0$ . Similarly,  $\mathbf{G}_{i,j} = 1$  or 0 reflects the certainty of the corresponding label. Our goal is to train an end-to-end neural network capable of performing classification inference on incomplete multi-view weak multi-label data.

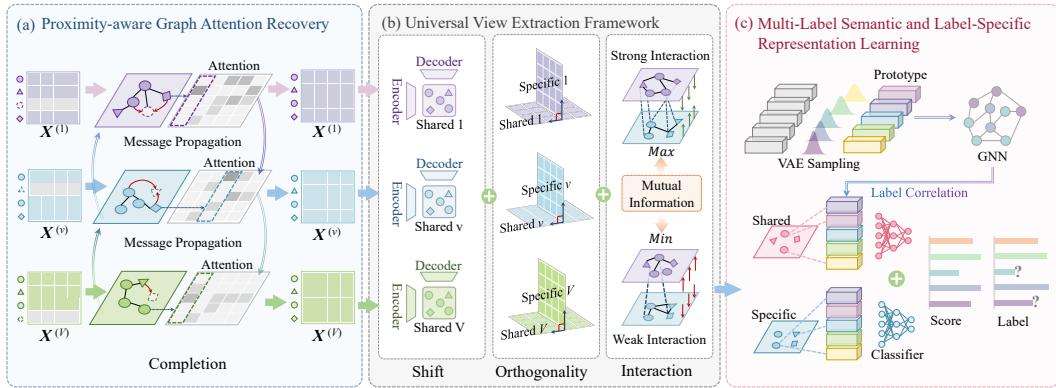


Figure 1: The main framework of our proposed TDLSR. Different shapes signify different samples.

### 2.2 Proximity-aware Graph Attention Recovery Mechanism

The lack of diverse views limits the capacity of deep neural models to extract high-level representations [20]. Consequently, data augmentation through reconstruction techniques is essential for improving model accuracy and stability. To mitigate information deficiency arising from incompleteness, we propose an attention-based relational graph construction strategy to propagate similarity signals across samples for missing imputation. For any view  $v$ , the attention score for an instance pair is computed by  $\mathbf{B}_{i,j}^{(v)} = e^{h(\mathbf{x}_i^{(v)})h(\mathbf{x}_j^{(v)T})/\tau}$ , where  $\tau$  is the temperature parameter and  $h(\cdot)$  denotes the normalization function. Each row of the matrix  $\mathbf{B}^{(v)}$  quantifies the degree of similarity between the corresponding sample with all other samples [23], which facilitates the identification of the  $k$ -nearest neighbors for each instance. Therefore, we construct the view-specific graph  $\hat{\mathbf{S}}^{(v)} \in \mathbb{R}^{N \times N}$  through attention-induced proximity awareness, where  $\hat{\mathbf{S}}_{i,j}^{(v)} = 1$  means  $\mathbf{W}_{i,v}\mathbf{W}_{j,v} = 1$  and  $\mathbf{B}_{i,j}^{(v)}$  is one of the top- $k$  largest elements in the  $i$ -th row, i.e.,  $\mathbf{x}_j^{(v)}$  is the neighbor of  $\mathbf{x}_i^{(v)}$ . Considering similarity relations between instances in existing views are applicable to the missing views, the transferred graph for finding the available instances related to the missing ones can be obtained:

$$\mathbf{K}^{(v)} = \sum_{k=1, k \neq v}^V \hat{\mathbf{S}}^{(k)} \text{diag}(\mathbf{W}_{:,v}), \quad (1)$$

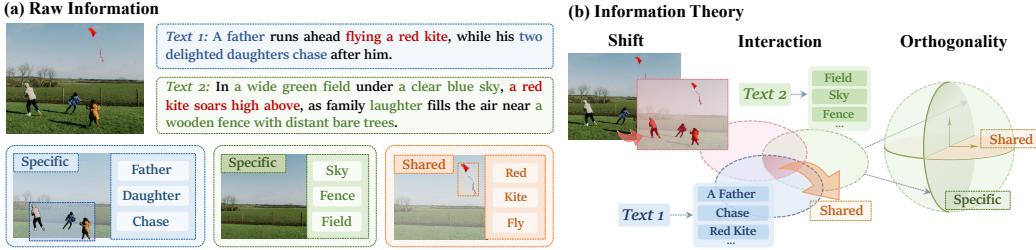


Figure 2: Depiction of information shift, interaction, and orthogonality.

where  $\text{diag}(\cdot)$  creates a diagonal matrix and  $\mathbf{W}_{:,v}$  is the  $v$ -th column. Besides, the contribution of available samples to reconstruct missing ones is governed by the maximal computable correlation from alternative views:

$$\bar{\mathbf{B}}_{i,j} = \max \left( \mathbf{B}_{i,j}^{(1)} \hat{\mathbf{S}}_{i,j}^{(1)}, \mathbf{B}_{i,j}^{(2)} \hat{\mathbf{S}}_{i,j}^{(2)}, \dots, \mathbf{B}_{i,j}^{(V)} \hat{\mathbf{S}}_{i,j}^{(V)} \right), \quad (2)$$

where  $\bar{\mathbf{B}}_{i,j}$  encapsulates the influence of the  $j$ -th sample in the recovery process of the  $i$ -th sample. By treating  $\mathbf{K}$  as the adjacency matrix and  $\bar{\mathbf{B}}$  as the edge weight, we can get the reconstructed data after message propagation:

$$\hat{\mathbf{x}}_i^{(v)} = \frac{\sum_{\mathbf{K}_{i,j}^{(v)} \geq 1} \mathbf{K}_{i,j}^{(v)} \bar{\mathbf{B}}_{i,j} \mathbf{x}_j^{(v)}}{\sum_{\mathbf{K}_{i,j}^{(v)} \geq 1} \bar{\mathbf{B}}_{i,j}}. \quad (3)$$

Since  $\hat{\mathbf{X}}^{(v)}$  serves as an approximate proxy of missing instances, we combine it with the original view to generate the final recovery matrix for downstream tasks:

$$\mathbf{Z}_{i,:}^{(v)} = \hat{\mathbf{X}}_{i,:}^{(v)} (1 - \mathbf{W}_{i,v}) + \mathbf{X}_{i,:}^{(v)} \mathbf{W}_{i,v}. \quad (4)$$

### 2.3 Universal View Extraction Framework under Mutual Information Constraints

Previous studies [27] have primarily concentrated on enabling networks to extract shared information, while overlooking the contributions of each unique view. Moreover, the reliance on linear geometric constraints in early view-specific representation learning [24] falls short of capturing complex feature interactions. To effectively assess both the consistency and the complementarity of distinct features, we propose a general model that directly constrains the mutual information between representations for precisely measuring their interaction degree. Two groups of multi-layer perceptrons  $\{\mathbf{M}_v^S\}_{v=1}^V$  and  $\{\mathbf{M}_v^O\}_{v=1}^V$  are employed as shared and view-private channel to extract the corresponding representations, i.e.,  $\{\mathbf{M}_v^S : \mathbf{Z}^{(v)} \rightarrow \mathbf{S}^{(v)}\}_{v=1}^V$  and  $\{\mathbf{M}_v^O : \mathbf{Z}^{(v)} \rightarrow \mathbf{O}^{(v)}\}_{v=1}^V$ . Taking the image-text retrieval task [32] for explanation, as illustrated in the Fig. 2, an image is typically paired with multiple textual descriptions that emphasize on individual relationships and environmental context, respectively. It requires the model to simultaneously extract the consistent shared representations (central actions) and unique modality-specific elements (e.g., distinct individual identities and background details) for holistic multimodal perception. To achieve this, maximizing  $I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)})$  is critical for guaranteeing that the representations stay consistent with the core semantics of raw modalities, such as people, kit, background, etc. Besides, comprehensive feature disentanglement is facilitated by maximizing the mutual information  $I(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{s}^{(v)}|\mathbf{z}^{(v)})$  and minimizing  $I(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)})$  between representations from any pair of views  $v$  and  $v^*$  ( $1 \leq v \neq v^* \leq V$ ), which focuses on capturing the central characteristics of the actions and preserving unique information embedded in each modality, respectively. Furthermore, establishing complete representation orthogonality from the same modality, as indicated by  $I(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) = 0$ , effectively minimizes semantic redundancy, such as the identification of the individual and their actions.

As mentioned above, feature separation is driven by the following three criteria: (i) prevent information shift, (ii) optimize information interaction and (iii) promote information orthogonality. Under

the condition of mutual information constraints, the universal model can be expressed as

$$\left\{ \begin{array}{l} \max \underbrace{\sum_{v=1}^V I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)})}_{\text{information shift}} + \underbrace{\sum_{v=1}^V \sum_{v^* \neq v} (I(\mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{s}^{(v)} | \mathbf{z}^{(v)}) - I(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}))}_{\text{information interaction}} \\ \text{s.t. } \underbrace{\sum_{v=1}^V I(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)})}_{\text{information orthogonality}} = 0. \end{array} \right. \quad (5)$$

By employing the Lagrange multiplier method, the equality constraint can be appropriately scaled. Direct computation of mutual information is impractical due to the challenges of distribution inference and the complexity of high-dimensional integrals. Thus, stability and feasibility are typically ensured by refining estimable mutual information bounds [35]. Regarding the information shift term, its lower bound is commonly represented through a reconstruction loss[15], where the representation  $\mathbf{s}^{(v)}$  is decoded via  $q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})$  to faithfully preserve the original view:

$$I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)}) \geq \mathbb{E}_{p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)})} \left[ \log q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) \right]. \quad (6)$$

For the second term, based on the definition expansion and the non-negativity of entropy, we can derive the following lower bound by introducing another variational distribution  $q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)})$ :

$$\begin{aligned} I(\mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{s}^{(v)} | \mathbf{z}^{(v)}) &\geq \int \int p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)} / \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) p(\mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ &+ \underbrace{\int \int p(\mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)} / \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) \log \frac{p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)} / \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)})}{q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)})} d\mathbf{s}^{(v)} d\mathbf{s}^{(v^*)}}_{\text{Kullback-Leibler divergence}}, \end{aligned} \quad (7)$$

where  $p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)} / \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)})$  refers to the conditional distribution between the shared representations from  $\mathbf{z}^{(v)}$  and  $\mathbf{z}^{(v^*)}$ . Leveraging the non-negativity of Kullback–Leibler divergence between the distributions  $p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)} / \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)})$  and  $q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)})$ , the tighter lower bound is further obtained. Similarly, upper bounds for the remaining two negative terms can be derived following the same rationale. Therefore, the problem (16) is transformed into optimizing the following objective:

$$\begin{aligned} \mathcal{L}_{IB} = & - \underbrace{\mathbb{E}_{p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)})} \left[ \log q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) \right]}_{\text{lower bound for information shift}} - \underbrace{\int \int p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}, \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}}_{\text{lower bound for shard information interaction}} \\ & + \underbrace{D_{KL}(p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}) q(\mathbf{o}^{(v)} | \mathbf{o}^{(v^*)}))}_{\text{upper bound for specific information interaction}} \\ & + \underbrace{\beta D_{KL}(p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v)} | \mathbf{z}^{(v)}) q(\mathbf{s}^{(v)} | \mathbf{o}^{(v)}))}_{\text{upper bound for information orthogonality}}, \end{aligned} \quad (8)$$

where  $\beta$  is the Lagrange multiplier. Minimizing  $\mathcal{L}_{IB}$  promotes effective distinction between shared and specific feature semantic, with implementation details provided in the Appendix.

## 2.4 Multi-Label Semantic and Label-Specific Representation Learning

In multi-label classification tasks, binary encoding struggles to capture the underlying label semantics [19]. To overcome this limitation, we employ a data-driven approach to learn label prototypes [27], which provides a clear insight into label structures and enhances effective semantic associations with features. After initializing the one-hot vector  $\mathbf{b}_i \in \mathbb{R}^C$  for each category, we utilize two stochastic encoders to model the prototype distribution i.e.,  $\mathbf{h}_i \sim \mathcal{N}(\mu_i, \sigma_i^2 \mathbf{I})$ , where the mean  $\mu_i$  and variance  $\sigma_i$  are determined by the encoders  $g_\mu(\mathbf{b}_i)$  and  $g_{\sigma^2}(\mathbf{b}_i)$ . Then, we sample from the distribution with the reparameterization trick to obtain the prototype representation  $\mathbf{h}_i^0 = \frac{1}{s} \sum_{d=1}^s (\mu_i + \sigma_i \odot \delta^d)$  [7], where  $\delta^d$  means the  $d$ -th individual sampling and  $\odot$  denotes the element-wise product.

The prototype representations are independently learned for each label, yet investigating intrinsic label correlation remains a core challenge in multi-label learning [3]. For bridging this gap, graph neural networks are leveraged to propagate prior correlation information and refine the label representations to encapsulate the inherent relationships of label semantics. Prototypes are regraded as a set of nodes positioned on the label relation graph, with edge weights reflecting the correlation between corresponding label pairs. Besides, the label correlation matrix  $\mathbf{A}$  quantitatively characterized on the training data serves as the appropriate substitute for the adjacency matrix. Rather than utilizing co-occurrence frequency to evaluate correlation degree, we use the Jaccard distance [30] calculated over positive labels, as we are only concerned with the categories assigned to each instance. By computing the intersection and union of two classes regarding positive values, we can obtain

$$A_{i,j} = \frac{\langle \mathbf{Y}_{:,i} \cdot \mathbf{Y}_{:,j} \rangle}{\sum_{k=1}^N (\mathbf{Y}_{k,i} + \mathbf{Y}_{k,j}) - \langle \mathbf{Y}_{:,i} \cdot \mathbf{Y}_{:,j} \rangle}, \quad (9)$$

where  $A_{ii}$  is set to 0 to eliminate self-dependency. Given an aggregated matrix  $\mathbf{H} \in \mathbb{R}^{C \times d}$ , the label embeddings corresponding to each row can be updated by passing through the GIN layer with propagated correlation information [11], i.e.,  $\mathbf{E} = f[(1 + \epsilon) \mathbf{H} + \mathbf{A}\mathbf{H}]$ , where  $f(\cdot)$  denotes a fully-connected layer followed by Batch Normalization and Leaky ReLU activation, and  $\epsilon$  is a learnable scalar that controls the influence of node's own features. To reinforce cohesion between relevant prototype representations and distinguish unrelated ones, we employ the following objective that aligns representation similarity with label correlation:

$$\mathcal{L}_{le} = -\frac{1}{C^2} \sum_{i=1}^C \sum_{j=1}^C \hat{A}_{ij} \log(\cos(\mathbf{E}_i, \mathbf{E}_j)) + (1 - \hat{A}_{ij}) \log(1 - \cos(\mathbf{E}_i, \mathbf{E}_j)), \quad (10)$$

where  $\cos(\mathbf{E}_i, \mathbf{E}_j)$  is the cosine similarity and  $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  with  $\mathbf{I}$  denoting an identity matrix. Multi-label classification tasks often involve labels with varying sensitivities to different feature subsets [14]. Consequently, label-specific feature learning has become a widely adopted technique to select the most relevant features tailored to classifying each label. However, label-specific disentangled feature learning remains underexplored, despite its potential to boost model performance. For instance, in image recognition, shared information capture general visual cues, while private features highlight other distinctive traits tied to each label, such as breed for "dog" and texture for "cat". Thus, we treat activated label prototypes as feature importance scores and engage them with both shared and private representations to discern the label-specific view embeddings:

$$\hat{\mathbf{P}}_i^{(v)} = \left[ \sigma_S(\mathbf{E}_1) \odot x_i^{(v)}; \sigma_S(\mathbf{E}_2) \odot x_i^{(v)}; \dots; \sigma_S(\mathbf{E}_C) \odot x_i^{(v)} \right], \quad (11)$$

where  $\sigma_S$  is the Sigmoid function. According to the Eq. (11), we can obtain the label-specific shared and private features, i.e.,  $\{\mathbf{S}^{(v)} \rightarrow \hat{\mathbf{U}}^{(v)} \in \mathbb{R}^{N \times C \times d}\}_{v=1}^V$  and  $\{\mathbf{O}^{(v)} \rightarrow \hat{\mathbf{V}}^{(v)} \in \mathbb{R}^{N \times C \times d}\}_{v=1}^V$ . The interaction between label semantics and view representations supports a bidirectional selection mechanism, in which discriminative views are assigned to specific labels, while information-related label subsets are uncovered associated with distinct views. Processing through the linear classifiers, view-specific predictions  $\mathbf{U}^{(v)} \in \mathbb{R}^{N \times C}$  and  $\mathbf{V}^{(v)} \in \mathbb{R}^{N \times C}$  are generated. Given the variability in reconstruction quality, views with higher recovery accuracy should be emphasized in the fusion process. For this aspect, certain reconstructed views characterized by low attention scores in relation to their associated samples naturally exhibit reduced confidence [23]. Thus, we calculate the maximum original attention for each instance with respect to other instances as its confidence score:

$$\mathbf{Q}_{i,v} = \max \left( \{\tau \log \bar{\mathbf{B}}_{i,j} \mathbf{W}_{j,v}\}_{j=1}^n \right), \quad (12)$$

where  $\bar{\mathbf{B}}$  is computed by Eq. (2) and  $\mathbf{Q} \in \mathbb{R}^{N \times V}$  stores the confidence score for individual instances. Since  $\mathbf{Q}$  is tailor-made for the missing samples, the confidence matrix is updated as  $\mathbf{Q}' = (1 - \mathbf{W}) \odot \mathbf{Q} + \mathbf{W}$ . During the late fusion, we combine the feature reconstruction efficiency to obtain the final prediction:

$$\bar{\mathbf{P}}_{i,:} = \sigma_S \left( \frac{\sum_{v=1}^V \mathbf{U}_{i,:}^{(v)} \mathbf{Q}'_{i,v}}{\sum_{v=1}^V \mathbf{Q}'_{i,v}} + \frac{\sum_{v=1}^V \mathbf{V}_{i,:}^{(v)} \mathbf{Q}'_{i,v}}{\sum_{v=1}^V \mathbf{Q}'_{i,v}} \right). \quad (13)$$

Then, we employ the weighted cross-entropy loss to mitigate the impact of missing labels:

$$\mathcal{L}_{bce} = -\frac{1}{\sum_{i,j} \mathbf{G}_{i,j}} \sum_{i=1}^N \sum_{j=1}^C (\mathbf{Y}_{i,j} \log(\bar{\mathbf{P}}_{i,j}) + (1 - \mathbf{Y}_{i,j}) \log(1 - \bar{\mathbf{P}}_{i,j})) \mathbf{G}_{i,j}. \quad (14)$$

By incorporating  $\lambda_1$  and  $\lambda_2$  to balance the loss effects, our total training loss can be expressed as

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{IB} + \lambda_2 \mathcal{L}_{le}. \quad (15)$$

## 2.5 Theoretical Results

In this subsection, we aim to theoretically explore the fundamental mechanisms that contribute to the model performance and the generalization capability of TDLSR. Through rigorous derivations (proofs in the Appendix), we obtain the following theorems:

**Theorem 1. (Discriminability of Label-specific Representation.)** For label prototypes  $\mathbf{E}_j$  and  $\mathbf{E}_k$  such that  $k \neq j$  for all  $k$ , and view representations  $\mathbf{X}^{(v)}$  and  $\mathbf{X}^{(v^*)}$  such that  $v^* \neq v$  for all  $v^*$ , the discriminability of  $\hat{\mathbf{P}}^{(v)}$  for class  $j$  necessitates that either of the following conditions be satisfied:

$$\begin{cases} \mathbb{E} \left[ \mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} \left[ \mathbf{X}^{(v)} \sigma_S(\mathbf{E}_k)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_k)^\top \\ \mathbb{E} \left[ \mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} \left[ \mathbf{X}^{(v^*)} \sigma_S(\mathbf{E}_j)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v^*)} \sigma_S(\mathbf{E}_j)^\top. \end{cases}$$

**Theorem 2. (Effectiveness of Disentangled Representation.)** Let the disentangled representation be denoted as  $\mathbf{R} = (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)})$ , where the information entropy of each representation is assumed to be fixed, i.e.,  $H(\mathbf{S}^{(v)}) = H(\mathbf{O}^{(v)}) = H^0$  ( $1 \leq v \leq V$ ). Then, in the case where each shared and specific representation is indispensable for prediction,  $I(R; Y)$  will attain its maximum when  $R = R_*$ , with  $R_*$  being the optimal solution of the problem (16).

**Theorem 3. (Generalization Error Bound.)** Our model is designed to learn a vector-valued function  $f = (f_1, \dots, f_C) : \mathcal{X} \mapsto \mathbb{R}^C$ . The expected risk and empirical risk w.r.t. the training dataset  $D$  are denoted as  $R(f) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})]$  and  $\hat{R}_D(f) = \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell(\sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c})$ , where  $f^{av}(\cdot)$  refers to the late fusion of multiple views. With probability at least  $1 - \delta$ , we have the following generalization error bound:

$$\begin{aligned} R(f) - \hat{R}_D(f) &\leq \frac{\tilde{\mathcal{K}}_1}{\sqrt{NV}} + \frac{\tilde{\mathcal{K}}_2}{N^{3/4}V^{1/4}} + \overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y}) \\ &\quad + \tilde{\mathcal{K}}_3 \sqrt{\frac{\sum_{c=1}^C \left( \sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) + \tilde{\mathcal{K}}_4 \right)}{NC}}, \end{aligned}$$

where  $\tilde{\mathcal{K}}_1 = \tilde{\mathcal{K}}_3 = \mathcal{O}(C)$ ,  $\tilde{\mathcal{K}}_2 = \mathcal{O}(\sqrt{C})$ ,  $\tilde{\mathcal{K}}_4$  is constant of order  $\tilde{\mathcal{O}}(1)$  as  $N, V \rightarrow \infty$ , and  $\overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y})$  is the generalization error related to the view reconstruction quality. Moreover, the generalization error bound becomes increasingly tighter during the optimization of the problem (16).

## 3 Experiments

### 3.1 Datasets and metrics

In our experiments, we utilize six popular multi-view multi-label datasets to validate the performance of our TDLSR, i.e., Corel 5k [5], ESPGame [1], IAPRTC12 [8], Mirflickr [16], Pascal07 [6], OBJECT [13]. In accordance with [25, 38], we select six metrics to construct a comprehensive evaluation system, i.e., Hamming Loss (HL), Ranking Loss (RL), OneError (OE), Coverage (Cov), Average Precision (AP), and Area Under Curve (AUC). To facilitate comparison, we present 1-HL, 1-OE, 1-Cov, and 1-RL values in the report, where higher values correspond to better performance.

### 3.2 Comparison methods

To measure the advancement of our TDLSR, nine state-of-the-art methods are selected for comparison experiments, i.e., AIMNet [23], DICNet [25], DIMC [38], iMVWL [33], LMVCAT [26], MTD [24], SIP [27], LVSL [43], DM2L [28]. Specifically, the first seven methods can simultaneously address the issues of missing views and labels. LVSL is a MvMLC method unable to handle missing data.

Thus, we use the mean of available instances to complete the missing views and fill the unknown labels with "0". DM2L is a kernel-based nonlinear method for incomplete multi-label learning. Then, we concatenate all views into a single-view representation for the execution of DM2L. All parameters of compared methods are configured as the recommended values in their original codes.

### 3.3 Implementation details

Each dataset is divided into training, validation and test sets in the ratio of 7:1:2. To simulate the partial view setting, a specified proportion of instances based on the Partial Example Ratio (PER), are randomly marked as unavailable in each view. Additionally, we ensure that each sample contains at least one complete view to avoid invalid cases. For weak supervision, we introduce label omissions for both positive and negative tags in each category applying the same proportion determined by the Label Missing Ratio (LMR). The process of constructing incomplete data is repeated multiple times to mitigate the impact of experimental randomness. Our model is implemented by PyTorch on one NVIDIA GeForce RTX 4090 GPU of 24GB memory.

### 3.4 Experimental results and analysis

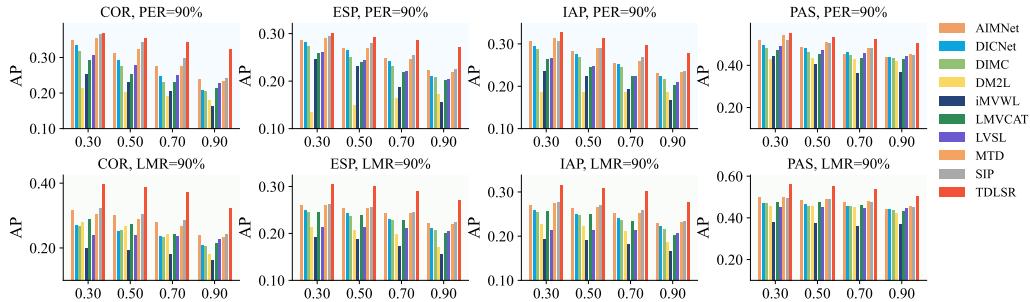


Figure 3: Experimental results on four datasets with PER and LMR changing from 30% to 90%.

To evaluate the effectiveness of our TDLSR in handling absent views and labels, we benchmark it against nine closely related algorithms across six datasets with varying levels of data sparsity. The proportion of missing views (PER) and labels (LMR) encompasses values of  $\{30\%, 50\%, 70\%, 90\%\}$ . The mean and standard deviation of the results with PER and LMR fixed at 50% are reported in Table 1. Besides, the average ranking of each algorithm based on the six metrics is calculated to perform a thorough assessment. Fig. 3 illustrates the variation in AP as PER and LMR changes from 30% to 90%. The other relevant results will be presented in the Appendix.

Drawing from the comparison results, we have the following observations: (i) Our method exhibits outstanding performance on almost all metrics across all six datasets. As shown in Table 1, despite the fluctuating rankings of other methods, TDLSR consistently holds the top position. Therefore, our method effectively addresses the iMVMLC problem and maintains stable outcomes. (ii) SIP and MTD are top-performing methods that always appear among the top three. The reason our method surpasses these leading approaches lies in the mutual information optimization, which constrains complex interactions between representations that are insufficiently addressed by MTD. It also overcomes the limitations of SIP by accounting for the impact of private features and transmitting label correlation information to refine label prototypes. Compared to AIMNet that similarly engages in view recovery, our method achieves an improvement in AP from 0.40 to 0.45 on Corel 5k, which demonstrates that our proximity-aware strategy can greatly suppress reconstruction noise. Achieving over a 10% performance gain against traditional multi-label methods like DM2L and deep learning frameworks such as DIMC and DICNet that disregard label correlation, our approach highlights its strength in capturing high-level view representations and leveraging label dependencies to enhance overall performance. (iii) As depicted in Fig. 3, our method exhibits remarkable performance and strong robustness across a wide range of missing ratios. Moreover, our method is particularly well-suited for highly incomplete settings. For instance, when PER reaches 90% on Corel5k and ESPGame, all baseline methods collapse, while our approach continues to deliver commendable results.

Table 1: Experimental results of nine methods on the six datasets with 50% PER and 50% LMR. ‘Ave.R’ refers to the mean ranking of the corresponding method across all six metrics.

DATA	METRIC	AIMNet	DICNet	DIMC	DM2L	iMVWL	LMVCAT	LVSL	MTD	SIP	TDLSR
COR	1-HL	<b>0.988</b> <sub>0.000</sub>	<b>0.987</b> <sub>0.000</sub>	<b>0.987</b> <sub>0.000</sub>	<b>0.987</b> <sub>0.000</sub>	0.978 <sub>0.000</sub>	0.986 <sub>0.000</sub>	<b>0.987</b> <sub>0.000</sub>	<b>0.988</b> <sub>0.000</sub>	<b>0.988</b> <sub>0.000</sub>	<b>0.988</b> <sub>0.000</sub>
	1-OE	0.478 <sub>0.011</sub>	0.460 <sub>0.012</sub>	0.446 <sub>0.009</sub>	0.378 <sub>0.014</sub>	0.308 <sub>0.017</sub>	0.448 <sub>0.011</sub>	0.353 <sub>0.017</sub>	<b>0.492</b> <sub>0.011</sub>	<b>0.492</b> <sub>0.014</sub>	<b>0.541</b> <sub>0.014</sub>
	1-Cov	0.766 <sub>0.004</sub>	0.726 <sub>0.007</sub>	0.709 <sub>0.008</sub>	0.640 <sub>0.007</sub>	0.701 <sub>0.003</sub>	0.720 <sub>0.006</sub>	0.720 <sub>0.005</sub>	0.754 <sub>0.005</sub>	<b>0.780</b> <sub>0.004</sub>	0.801 <sub>0.009</sub>
	1-RL	0.900 <sub>0.002</sub>	0.881 <sub>0.004</sub>	0.874 <sub>0.004</sub>	0.843 <sub>0.004</sub>	0.864 <sub>0.002</sub>	0.876 <sub>0.004</sub>	0.879 <sub>0.003</sub>	0.893 <sub>0.004</sub>	<b>0.908</b> <sub>0.003</sub>	<b>0.917</b> <sub>0.004</sub>
	AP	0.404 <sub>0.005</sub>	0.381 <sub>0.006</sub>	0.370 <sub>0.005</sub>	0.318 <sub>0.005</sub>	0.281 <sub>0.005</sub>	0.379 <sub>0.006</sub>	0.311 <sub>0.005</sub>	0.410 <sub>0.007</sub>	0.414 <sub>0.006</sub>	<b>0.450</b> <sub>0.006</sub>
	AUC	0.903 <sub>0.002</sub>	0.883 <sub>0.004</sub>	0.877 <sub>0.004</sub>	0.846 <sub>0.004</sub>	0.867 <sub>0.002</sub>	0.879 <sub>0.003</sub>	0.882 <sub>0.002</sub>	0.895 <sub>0.003</sub>	<b>0.910</b> <sub>0.002</sub>	<b>0.919</b> <sub>0.004</sub>
	AVE	3.5	5	7.333	9	9.5	6.833	7.333	3.167	2.333	1
ESP	1-HL	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	0.972 <sub>0.000</sub>	<b>0.982</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>	<b>0.983</b> <sub>0.000</sub>
	1-OE	0.442 <sub>0.006</sub>	0.440 <sub>0.009</sub>	0.431 <sub>0.009</sub>	0.302 <sub>0.008</sub>	0.343 <sub>0.010</sub>	0.431 <sub>0.006</sub>	0.365 <sub>0.000</sub>	<b>0.452</b> <sub>0.007</sub>	0.450 <sub>0.006</sub>	<b>0.477</b> <sub>0.007</sub>
	1-Cov	0.621 <sub>0.003</sub>	0.601 <sub>0.003</sub>	0.586 <sub>0.004</sub>	0.532 <sub>0.003</sub>	0.548 <sub>0.004</sub>	0.587 <sub>0.003</sub>	0.578 <sub>0.002</sub>	0.617 <sub>0.004</sub>	0.622 <sub>0.004</sub>	0.646 <sub>0.004</sub>
	1-RL	0.845 <sub>0.002</sub>	0.836 <sub>0.002</sub>	0.830 <sub>0.002</sub>	0.804 <sub>0.002</sub>	0.807 <sub>0.002</sub>	0.827 <sub>0.002</sub>	0.829 <sub>0.001</sub>	0.843 <sub>0.002</sub>	0.847 <sub>0.002</sub>	<b>0.859</b> <sub>0.002</sub>
	AP	0.305 <sub>0.003</sub>	0.300 <sub>0.003</sub>	0.294 <sub>0.003</sub>	0.229 <sub>0.003</sub>	0.243 <sub>0.004</sub>	0.293 <sub>0.003</sub>	0.266 <sub>0.003</sub>	<b>0.309</b> <sub>0.003</sub>	<b>0.309</b> <sub>0.004</sub>	0.328 <sub>0.004</sub>
	AUC	0.850 <sub>0.001</sub>	0.841 <sub>0.002</sub>	0.835 <sub>0.002</sub>	0.808 <sub>0.001</sub>	0.813 <sub>0.002</sub>	0.832 <sub>0.001</sub>	0.834 <sub>0.001</sub>	0.847 <sub>0.002</sub>	<b>0.851</b> <sub>0.002</sub>	<b>0.863</b> <sub>0.002</sub>
	AVE	3.833	4.5	5.833	9.667	9.167	7.333	7.167	3.833	2.667	1
IAP	1-HL	<b>0.981</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>	<b>0.980</b> <sub>0.000</sub>	0.969 <sub>0.000</sub>	<b>0.980</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>	<b>0.981</b> <sub>0.000</sub>
	1-OE	0.457 <sub>0.007</sub>	0.464 <sub>0.008</sub>	0.454 <sub>0.006</sub>	0.378 <sub>0.008</sub>	0.351 <sub>0.008</sub>	0.433 <sub>0.009</sub>	0.377 <sub>0.007</sub>	<b>0.479</b> <sub>0.007</sub>	0.459 <sub>0.005</sub>	<b>0.491</b> <sub>0.008</sub>
	1-Cov	0.675 <sub>0.004</sub>	0.649 <sub>0.005</sub>	0.630 <sub>0.005</sub>	0.555 <sub>0.003</sub>	0.565 <sub>0.004</sub>	0.646 <sub>0.004</sub>	0.605 <sub>0.004</sub>	0.670 <sub>0.004</sub>	<b>0.678</b> <sub>0.003</sub>	0.706 <sub>0.005</sub>
	1-RL	0.884 <sub>0.001</sub>	0.874 <sub>0.002</sub>	0.868 <sub>0.002</sub>	0.837 <sub>0.002</sub>	0.833 <sub>0.002</sub>	0.868 <sub>0.002</sub>	0.857 <sub>0.002</sub>	0.882 <sub>0.002</sub>	<b>0.886</b> <sub>0.001</sub>	<b>0.899</b> <sub>0.002</sub>
	AP	0.329 <sub>0.003</sub>	0.326 <sub>0.003</sub>	0.318 <sub>0.002</sub>	0.254 <sub>0.002</sub>	0.236 <sub>0.002</sub>	0.313 <sub>0.004</sub>	0.262 <sub>0.002</sub>	<b>0.340</b> <sub>0.002</sub>	0.331 <sub>0.003</sub>	<b>0.358</b> <sub>0.004</sub>
	AUC	0.885 <sub>0.001</sub>	0.876 <sub>0.002</sub>	0.870 <sub>0.001</sub>	0.838 <sub>0.001</sub>	0.835 <sub>0.001</sub>	0.870 <sub>0.002</sub>	0.859 <sub>0.001</sub>	0.883 <sub>0.002</sub>	<b>0.887</b> <sub>0.001</sub>	<b>0.899</b> <sub>0.002</sub>
	AVE	4	4.333	6	8.833	9.833	6.833	8	2.833	2.833	1
MIR	1-HL	0.890 <sub>0.001</sub>	0.890 <sub>0.001</sub>	0.890 <sub>0.001</sub>	0.876 <sub>0.001</sub>	0.840 <sub>0.004</sub>	0.880 <sub>0.004</sub>	0.877 <sub>0.001</sub>	<b>0.893</b> <sub>0.001</sub>	0.890 <sub>0.001</sub>	<b>0.896</b> <sub>0.001</sub>
	1-OE	0.646 <sub>0.009</sub>	0.647 <sub>0.010</sub>	0.645 <sub>0.008</sub>	0.533 <sub>0.008</sub>	0.511 <sub>0.016</sub>	0.639 <sub>0.009</sub>	0.609 <sub>0.007</sub>	<b>0.667</b> <sub>0.006</sub>	0.654 <sub>0.007</sub>	<b>0.690</b> <sub>0.009</sub>
	1-Cov	0.673 <sub>0.003</sub>	0.661 <sub>0.004</sub>	0.657 <sub>0.003</sub>	0.615 <sub>0.002</sub>	0.588 <sub>0.013</sub>	0.665 <sub>0.002</sub>	0.624 <sub>0.002</sub>	<b>0.681</b> <sub>0.002</sub>	0.668 <sub>0.006</sub>	<b>0.694</b> <sub>0.003</sub>
	1-RL	0.874 <sub>0.002</sub>	0.869 <sub>0.003</sub>	0.867 <sub>0.003</sub>	0.835 <sub>0.001</sub>	0.809 <sub>0.014</sub>	0.862 <sub>0.003</sub>	0.847 <sub>0.001</sub>	<b>0.878</b> <sub>0.001</sub>	0.873 <sub>0.002</sub>	<b>0.888</b> <sub>0.002</sub>
	AP	0.599 <sub>0.003</sub>	0.595 <sub>0.007</sub>	0.592 <sub>0.006</sub>	0.519 <sub>0.003</sub>	0.494 <sub>0.017</sub>	0.589 <sub>0.004</sub>	0.548 <sub>0.003</sub>	<b>0.614</b> <sub>0.004</sub>	0.603 <sub>0.005</sub>	<b>0.631</b> <sub>0.004</sub>
	AUC	0.861 <sub>0.001</sub>	0.855 <sub>0.002</sub>	0.854 <sub>0.002</sub>	0.828 <sub>0.001</sub>	0.801 <sub>0.017</sub>	0.852 <sub>0.003</sub>	0.839 <sub>0.001</sub>	<b>0.864</b> <sub>0.001</sub>	0.859 <sub>0.002</sub>	<b>0.875</b> <sub>0.001</sub>
	AVE	3.833	4.667	6.167	9	10	6.667	8	2	3.667	1
OBJ	1-HL	0.948 <sub>0.001</sub>	0.948 <sub>0.001</sub>	0.947 <sub>0.001</sub>	0.935 <sub>0.000</sub>	0.899 <sub>0.002</sub>	0.940 <sub>0.003</sub>	0.935 <sub>0.001</sub>	<b>0.949</b> <sub>0.001</sub>	0.948 <sub>0.001</sub>	<b>0.953</b> <sub>0.001</sub>
	1-OE	0.619 <sub>0.015</sub>	0.601 <sub>0.011</sub>	0.594 <sub>0.012</sub>	0.537 <sub>0.011</sub>	0.465 <sub>0.018</sub>	0.604 <sub>0.016</sub>	0.450 <sub>0.008</sub>	<b>0.627</b> <sub>0.011</sub>	0.626 <sub>0.009</sub>	<b>0.685</b> <sub>0.011</sub>
	1-Cov	0.806 <sub>0.006</sub>	0.794 <sub>0.006</sub>	0.793 <sub>0.006</sub>	0.768 <sub>0.005</sub>	0.744 <sub>0.008</sub>	0.796 <sub>0.008</sub>	0.759 <sub>0.006</sub>	<b>0.812</b> <sub>0.006</sub>	0.809 <sub>0.006</sub>	<b>0.834</b> <sub>0.007</sub>
	1-RL	0.888 <sub>0.005</sub>	0.876 <sub>0.004</sub>	0.875 <sub>0.004</sub>	0.860 <sub>0.004</sub>	0.833 <sub>0.006</sub>	0.878 <sub>0.006</sub>	0.850 <sub>0.004</sub>	<b>0.890</b> <sub>0.005</sub>	0.889 <sub>0.004</sub>	<b>0.910</b> <sub>0.004</sub>
	AP	0.639 <sub>0.010</sub>	0.627 <sub>0.009</sub>	0.623 <sub>0.010</sub>	0.577 <sub>0.009</sub>	0.512 <sub>0.014</sub>	0.630 <sub>0.012</sub>	0.537 <sub>0.008</sub>	0.649 <sub>0.009</sub>	<b>0.649</b> <sub>0.009</sub>	0.692 <sub>0.009</sub>
	AUC	0.897 <sub>0.004</sub>	0.886 <sub>0.004</sub>	0.885 <sub>0.004</sub>	0.872 <sub>0.004</sub>	0.846 <sub>0.006</sub>	0.888 <sub>0.006</sub>	0.864 <sub>0.004</sub>	<b>0.900</b> <sub>0.005</sub>	0.898 <sub>0.004</sub>	<b>0.918</b> <sub>0.004</sub>
	AVE	4	5.833	6.833	8.167	9.833	5.333	9	2	3	1
PAS	1-HL	0.931 <sub>0.001</sub>	0.931 <sub>0.000</sub>	0.931 <sub>0.001</sub>	0.927 <sub>0.001</sub>	0.882 <sub>0.004</sub>	0.915 <sub>0.005</sub>	0.928 <sub>0.001</sub>	<b>0.933</b> <sub>0.001</sub>	<b>0.932</b> <sub>0.001</sub>	<b>0.933</b> <sub>0.001</sub>
	1-OE	0.462 <sub>0.009</sub>	0.443 <sub>0.007</sub>	0.435 <sub>0.010</sub>	0.419 <sub>0.006</sub>	0.366 <sub>0.039</sub>	0.433 <sub>0.017</sub>	0.418 <sub>0.008</sub>	0.473 <sub>0.008</sub>	0.468 <sub>0.008</sub>	0.495 <sub>0.013</sub>
	1-Cov	0.781 <sub>0.007</sub>	0.749 <sub>0.003</sub>	0.738 <sub>0.004</sub>	0.720 <sub>0.004</sub>	0.674 <sub>0.011</sub>	0.759 <sub>0.006</sub>	0.738 <sub>0.003</sub>	<b>0.790</b> <sub>0.006</sub>	0.778 <sub>0.004</sub>	<b>0.817</b> <sub>0.004</sub>
	1-RL	0.830 <sub>0.006</sub>	0.803 <sub>0.002</sub>	0.792 <sub>0.008</sub>	0.778 <sub>0.003</sub>	0.736 <sub>0.011</sub>	0.808 <sub>0.006</sub>	0.797 <sub>0.002</sub>	<b>0.836</b> <sub>0.005</sub>	0.828 <sub>0.004</sub>	<b>0.862</b> <sub>0.004</sub>
	AP	0.548 <sub>0.007</sub>	0.517 <sub>0.004</sub>	0.510 <sub>0.008</sub>	0.482 <sub>0.005</sub>	0.438 <sub>0.022</sub>	0.524 <sub>0.009</sub>	0.486 <sub>0.005</sub>	<b>0.562</b> <sub>0.005</sub>	0.552 <sub>0.006</sub>	<b>0.590</b> <sub>0.008</sub>
	AUC	0.851 <sub>0.005</sub>	0.827 <sub>0.002</sub>	0.817 <sub>0.008</sub>	0.806 <sub>0.003</sub>	0.767 <sub>0.011</sub>	0.830 <sub>0.006</sub>	0.823 <sub>0.002</sub>	<b>0.855</b> <sub>0.005</sub>	0.848 <sub>0.005</sub>	<b>0.880</b> <sub>0.003</sub>
	AVE	3.5	5.667	7.167	8.667	10	6	7.5	2	3.5	1

### 3.5 Ablation Study

The ablation experiments are conducted to deeply investigate the effect of the three crucial modules of TDLSR, i.e., proximity-aware graph attention recovery mechanism ( $S_1$ ), information theory-driven representation extraction framework ( $S_2$ ), multi-label semantic and label-specific representation learning ( $S_3$ ). After individually removing  $S_1$ ,  $S_2$  and  $S_3$ , we use mean imputation for missing samples, rely solely on a single multilayer perceptron (MLP) for feature extraction while discarding  $\mathcal{L}_{IB}$ , and directly employ a classifier based on fully connected layers without exploring label semantics, respectively. Based on the ablation results provided in Table 2, we have the following observations: (i) When either module is removed, the performance declines, which indicates the effectiveness and thoughtful design of our TDLSR. (ii) The recovery mechanism is crucial for enhancing performance, as it provides downstream modules with rich feature information. Moreover, feature separation outperforms the single-channel representations and incorporating category semantic learning enhances performance beyond that of classifier-only approaches. It demonstrates our thorough consideration of feature extraction and label associations.

## 4 Application to Comprehensive Potential Prediction of Players

To validate the practical applicability of our TDLSR, we evaluate its ability to predict multiple attributes of NBA players under partial data missingness. The NBA dataset was collected from Basketball-Reference [2], which contains 16,992 player-season records from the 2002-2022 seasons. Each sample is structured across six principal statistical views including scoring efficiency, rebounding and physical metrics, technical statistics, advanced efficiency metrics, player background, and season context. The prediction tasks comprise career stage classification, positional identification and awards prediction. Career stages are partitioned into early (first 25%), peak (middle 50%), and late (final

Table 2: Ablation study on Pascal07, OBJECT and Mirflickr with PER=50% and LMR=50%. ‘✓’ and ‘✗’ represent the used and not used corresponding item, respectively.

$S_1$	$S_2$	$S_3$	Pascal07				OBJECT				Mirflickr			
			AP	AUC	1-RL	1-OE	AP	AUC	1-RL	1-OE	AP	AUC	1-RL	1-OE
✗	✓	✓	0.546	0.852	0.830	0.455	0.650	0.903	0.894	0.633	0.594	0.859	0.872	0.649
✓	✗	✓	0.576	0.874	0.853	0.478	0.687	0.914	0.906	0.678	0.614	0.872	0.881	0.652
✓	✓	✗	0.582	0.874	0.857	0.486	0.690	0.912	0.904	0.680	0.616	0.870	0.882	0.659
✓	✓	✓	<b>0.599</b>	<b>0.882</b>	<b>0.864</b>	<b>0.519</b>	<b>0.702</b>	<b>0.924</b>	<b>0.916</b>	<b>0.688</b>	<b>0.631</b>	<b>0.875</b>	<b>0.889</b>	<b>0.687</b>

25%) phases according to each player’s professional timeline. Player positions (PG, SG, SF, PF, C) are represented using one-hot encoding, and multiple binary indicators corresponding to honors such as MVP awards and Defensive Player of the Year, are included to provide multi-task objectives and comprehensive modeling of player achievements.

Across varying levels of data incompleteness, with PER and LMR ranging from 50% to 90%, our method consistently surpasses baseline approaches, demonstrating superior robustness and reliability in attribute prediction. Moreover, all comparison methods fail to surpass an AP of 0.6 at 90% missing ratio, whereas our TDLSR achieves 0.668. Despite incomplete technical statistics and constrained annotation resources, our method remains effective in predicting player potential, including career development and honor attainment, which offers considerable promise for real-world applications.

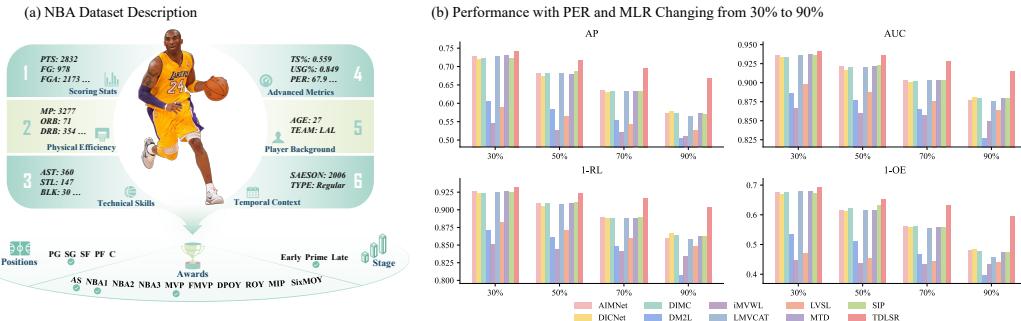


Figure 4: Overview of the NBA dataset and performance with missing views and labels.

## 5 Conclusion

In this paper, we propose a theory-driven label-specific representation (TDLSR) framework for addressing the iMvMLC problem. Specifically, structural dependencies are modeled through graph attention mechanism inside each view for recovery, with the reconstruction fidelity adaptively tailored to enhance classification efficacy. Meanwhile, we construct a universal feature extraction model, where mutual information optimization serves to regulate information shift, interaction and orthogonality between representations. On the basis of complete feature semantic separation, we independently interact each representation with label prototypes that encode correlation semantic, aiming to extract label-specific discriminative features and uncover representation-sensitive label subsets. Moreover, we theoretically validate the effectiveness of representation learning and its influence on the generalization performance. Finally, the superiority of TDLSR are validated through extensive experiments and application to the NBA dataset. In the future, we will explore leveraging the prior knowledge embedded in LLM to facilitate label semantic learning.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62376281, the NSF for Distinguished Young Scholars under Grant No. 62425607, and the Key NSF of China under Grant No. 62136005.

## References

[1] Ahn, L. V.; and Dabbish, L. 2004. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*, 319–326.

[2] Basketball Reference. 2025. NBA Player Totals Statistics. <https://www.basketball-reference.com>. Accessed: 2025-03-03.

[3] Chen, T.; Wang, W.; Pu, T.; Qin, J.; Yang, Z.; Liu, J.; and Lin, L. 2024. Dynamic correlation learning and regularization for multi-label confidence calibration. *IEEE Transactions on Image Processing*, 33: 4811–4823.

[4] Duan, K.; Cui, S.; Shinnou, H.; and Bao, S. 2025. View-Channel Mixer Network for Double Incomplete Multi-View Multi-Label learning. *Neurocomputing*, 617: 129013.

[5] Duygulu, P.; Barnard, K.; de Freitas, J. F.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of European Conference on Computer Vision*, 97–112.

[6] Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

[7] Federici, M.; Dutta, A.; Forré, P.; Kushman, N.; and Akata, Z. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *International Conference on Learning Representations*.

[8] Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop onto Image*, volume 2, 1–11.

[9] Guan, R.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; and Liu, X. 2025. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

[10] Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025. Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

[11] Hang, J.-Y.; and Zhang, M.-L. 2021. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9860–9871.

[12] Hao, P.; Gao, W.; and Hu, L. 2025. Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, 157: 110888.

[13] Hao, P.; Liu, K.; and Gao, W. 2024. Anchor-guided global view reconstruction for multi-view multi-label feature selection. *Information Sciences*, 679: 121124.

[14] Huang, J.; Li, G.; Huang, Q.; and Wu, X. 2016. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 28(12): 3309–3323.

[15] Huang, W.; Yang, S.; and Cai, H. 2023. Generalized information-theoretic multi-view clustering. *Advances in Neural Information Processing Systems*, 36: 58752–58764.

[16] Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of ACM International Conference on Multimedia Information Retrieval*, 39–43.

[17] Jia, X.; Jing, X.-Y.; Zhu, X.; Chen, S.; Du, B.; Cai, Z.; He, Z.; and Yue, D. 2020. Semi-supervised multi-view deep discriminant representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2496–2509.

[18] Kawaguchi, K.; Deng, Z.; Luh, K.; and Huang, J. 2022. Robustness implies generalization via data-dependent generalization bounds. In *International conference on machine learning*, 10866–10894. PMLR.

[19] Li, Q.; Luo, T.; Jiang, M.; Jiang, Z.; Hou, C.; and Li, F. 2025. Semi-Supervised Multi-View Multi-Label Learning with View-Specific Transformer and Enhanced Pseudo-Label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18430–18438.

[20] Li, Q.; Luo, T.; Jiang, M.; Liao, J.; and Jiang, Z. 2024. Deep Incomplete Multi-View Network Semi-Supervised Multi-Label Learning with Unbiased Loss. In *Proceedings of ACM International Conference on Multimedia*, 9048–9056.

[21] Li, X.; and Chen, S. 2021. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5918–5932.

[22] Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

[23] Liu, C.; Jia, J.; Wen, J.; Liu, Y.; Luo, X.; Huang, C.; and Xu, Y. 2024. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13864–13872.

[24] Liu, C.; Wen, J.; Liu, Y.; Huang, C.; Wu, Z.; Luo, X.; and Xu, Y. 2023. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. *Advances in Neural Information Processing Systems*, 36: 32387–32400.

[25] Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8807–8815.

[26] Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8816–8824.

[27] Liu, C.; Xu, G.; Wen, J.; Liu, Y.; Huang, C.; and Xu, Y. 2024. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *Forty-first International Conference on Machine Learning*.

[28] Ma, Z.; and Chen, S. 2021. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. *Pattern Recognition*, 111: 107675.

[29] Pan, E.; and Kang, Z. 2021. Multi-view contrastive graph clustering. *Advances in Neural Information Processing Systems*, 34: 2148–2159.

[30] Park, L. A.; and Read, J. 2019. A blended metric for multi-label optimisation and evaluation. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, 719–734. Springer.

[31] Si, C.; Jia, Y.; Wang, R.; Zhang, M.-L.; Feng, Y.; and Qu, C. 2023. Multi-label classification with high-rank and high-order label correlations. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4076–4088.

[32] Song, Q.; Gong, T.; Gao, S.; Zhou, H.; and Li, J. 2024. QUEST: Quadruple Multimodal Contrastive Learning with Constraints and Self-Penalization. *Advances in Neural Information Processing Systems*, 37: 28889–28919.

[33] Tan, Q.; Yu, G.; Domeniconi, C.; Wang, J.; and Zhang, Z. 2018. Incomplete multi-view weak-label learning. In *International Joint Conference on Artificial Intelligence*, 2703–2709.

[34] Tan, X.; Zhao, C.; Liu, C.; Wen, J.; and Tang, Z. 2024. A two-stage information extraction network for incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15249–15257.

[35] Tian, X.; Zhang, Z.; Wang, C.; Zhang, W.; Qu, Y.; Ma, L.; Wu, Z.; Xie, Y.; and Tao, D. 2023. Variational distillation for multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4551–4566.

[36] Wang, S.; Liu, X.; Liao, Q.; Wen, Y.; Zhu, E.; and He, K. 2025. Scalable Multi-View Graph Clustering With Cross-View Corresponding Anchor Alignment. *IEEE Transactions on Knowledge and Data Engineering*, 2932–2945.

[37] Wang, Y.; Li, Q.; Chang, D.; Wen, J.; Xiao, F.; and Zhao, Y. 2025. A Category-Driven Contrastive Recovery Network for Double Incomplete Multi-view Multi-label Classification. *IEEE Transactions on Multimedia*, 1–10.

[38] Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 11396–11408.

[39] Wen, W.; Gong, T.; Dong, Y.; Yu, S.; and Zhang, W. 2025. Towards the Generalization of Multi-view Learning: An Information-theoretical Analysis. *arXiv preprint arXiv:2501.16768*.

- [40] Ye, H.; Sunderraman, R.; and Ji, S. 2024. MatchXML: an efficient text-label matching framework for extreme multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, 36: 4781–4793.
- [41] Yuan, H.; Sun, Y.; Zhou, F.; Wen, J.; Yuan, S.; You, X.; and Ren, Z. 2025. Prototype Matching Learning for Incomplete Multi-view Clustering. *IEEE Transactions on Image Processing*, 34: 828–841.
- [42] Zhang, X.; Wang, R.; Chen, S.; Jia, Y.; and Wang, D. D. 2024. AME-LSIFT: Attention-Aware Multi-Label Ensemble With Label Subset-Specific Features. *IEEE Transactions on Knowledge and Data Engineering*, 7627–7642.
- [43] Zhao, D.; Gao, Q.; Lu, Y.; and Sun, D. 2022. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 25: 7235–7247.
- [44] Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2017. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1081–1094.

## A Derivation and Implementation of the Loss Function for the Universal View Extraction Model

### A.1 Derivation of the Loss Function

The universal view extraction model under mutual information constraints is

$$\left\{ \begin{array}{l} \max \underbrace{\sum_{v=1}^V I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)})}_{\text{information shift}} + \underbrace{\sum_{v=1}^V \sum_{v^* \neq v} (I(\mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{s}^{(v)} | \mathbf{z}^{(v)}) - I(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}))}_{\text{information interaction}} \\ \text{s.t. } \underbrace{\sum_{v=1}^V I(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)})}_{\text{information orthogonality}} = 0. \end{array} \right. \quad (16)$$

By incorporating variational derivations from the information theory, we arrive at the following objective function for optimization:

$$\begin{aligned} \mathcal{L}_{IB} = & - \underbrace{\mathbb{E}_{p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)})} \left[ \log q(\mathbf{s}^{(u)} | \mathbf{z}^{(v)}) \right]}_{\text{lower bound for information shift}} - \underbrace{\int \int p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}, \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}}_{\text{lower bound for shard information interaction}} \\ & + \underbrace{D_{KL}(p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}) q(\mathbf{o}^{(v)} | \mathbf{o}^{(v^*)}))}_{\text{upper bound for specific information interaction}} \\ & + \underbrace{\beta D_{KL}(p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v)} | \mathbf{z}^{(v)}) q(\mathbf{s}^{(v)} | \mathbf{o}^{(v)}))}_{\text{upper bound for information orthogonality}}. \end{aligned} \quad (17)$$

*Proof.* Due to the challenge of distribution inference and the complexity of high-dimensional integration, directly computing mutual information is impractical. Therefore, it is common to ensure stability and feasibility by optimizing an estimable bound of mutual information. For the information shift term, we have the following transformation based on the definitions of mutual information and information entropy:

$$\begin{aligned} & I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)}) \\ &= \iint p(\mathbf{s}^{(v)}, \mathbf{z}^{(v)}) \log \frac{p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})}{p(\mathbf{z}^{(v)})} d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &= \iint p(\mathbf{s}^{(v)}, \mathbf{z}^{(v)}) \log p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} + \int p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) \int p(\mathbf{z}^{(v)}) \log \frac{1}{p(\mathbf{z}^{(v)})} d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &= \iint p(\mathbf{s}^{(v)}, \mathbf{z}^{(v)}) \log p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} + H(\mathbf{z}^{(v)}). \end{aligned} \quad (18)$$

Since Eq. (18) is intractable, we approximate  $p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})$  using a stochastic variational distribution  $q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})$ , which can be reasonably estimated. Combing the fact that entropy  $H(\mathbf{z}^{(v)}) \geq 0$ , we can further obtain the lower bound:

$$\begin{aligned} & I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)}) \\ &\geq \iint p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)}) \log p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &= \int p(\mathbf{z}^{(v)}) \int p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}) \log q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &\quad + \int p(\mathbf{s}^{(v)}) \int p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)}) \log \frac{p(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})}{q(\mathbf{z}^{(v)} | \mathbf{s}^{(v)})} d\mathbf{s}^{(v)} d\mathbf{z}^{(v)}. \end{aligned} \quad (19)$$

The Kullback-Leibler divergence is denoted as

$$D_{KL}(p(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})\|q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})) = \int p(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) \log \frac{p(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})}{q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})} d\mathbf{z}^{(v)} \geq 0 \quad (20)$$

Thus, we have

$$\begin{aligned} & I(\mathbf{s}^{(v)}; \mathbf{z}^{(v)}) \\ &= \int p(\mathbf{z}^{(v)}) \int p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \log q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ & \quad + \int p(\mathbf{s}^{(v)}) D_{KL}(p(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})\|q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)})) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &\geq \int p(\mathbf{z}^{(v)}) \int p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \log q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &= \int p(\mathbf{s}^{(v)}, \mathbf{z}^{(v)}) \log q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) d\mathbf{s}^{(v)} d\mathbf{z}^{(v)} \\ &= \mathbb{E}_{p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)})} \left[ \log q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) \right]. \end{aligned} \quad (21)$$

For the second term, based on the definition and non-negativity of entropy, a lower bound can be derived by introducing another estimable variational distribution  $q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})$ . By expanding the mutual information in its integral form, we have

$$I(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{s}^{(v)}|\mathbf{z}^{(v)}) = \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}, \mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log \frac{p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}, \mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})}{p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)})} d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}, \quad (22)$$

Considering  $p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}, \mathbf{s}^{(v)}|\mathbf{z}^{(v)}) = p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})$ , we can get

$$\begin{aligned} & I(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \\ &= \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log \frac{p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})}{p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v)}/\mathbf{z}^{(v)})} d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ &= \iint p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}/\mathbf{s}^{(v)}|\mathbf{z}^{(v)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ & \quad + \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \log \frac{1}{p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)})} d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}. \end{aligned} \quad (23)$$

Since  $H(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) = - \int p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \log p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}) d\mathbf{s}^{(v)} \geq 0$ , we obtain

$$\begin{aligned} & I(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{s}^{(v)}|\mathbf{z}^{(v)}) \\ &\geq \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ &= \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ & \quad + \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log \frac{p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})}{q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})} d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \end{aligned} \quad (24)$$

Based on the definition of Kullback-Leibler divergence, we have

$$\begin{aligned} & D_{KL}(p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})\|q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})) \\ &= \int p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log \frac{p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})}{q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})} d\mathbf{s}^{(v)}. \end{aligned} \quad (25)$$

Since  $D_{KL}(p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})\|q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})) \geq 0$ , we have

$$\begin{aligned} I(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{s}^{(v)}|\mathbf{z}^{(v)}) &\geq \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}/\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} \\ &= \iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}, \mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}. \end{aligned} \quad (26)$$

For the third term  $I(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)})$ , we derive the following upper bound:

$$\begin{aligned}
& I(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \\
&= \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)})}{p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&= \iint p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)}/\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})}{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&= \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)}/\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})}{q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&\quad + \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})}{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&= \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)}/\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})}{q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&\quad - \int p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)}/\mathbf{o}^{(v)}|\mathbf{z}^{(v)}) D_{KL}(p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \| q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})) d\mathbf{o}^{(v^*)} \\
&\leq \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)}/\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})}{q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&= \iint p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \log \frac{p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)})}{p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})} d\mathbf{o}^{(v^*)} d\mathbf{o}^{(v)} \\
&= D_{KL}(p(\mathbf{o}^{v^*}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)})). \tag{27}
\end{aligned}$$

Similarly, we can derive the upper bound for the last term, which exhibit structure analogous to that of Eq. (27):

$$I(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \leq D_{KL}(p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)})q(\mathbf{s}^{(v)}|\mathbf{o}^{(v)})). \tag{28}$$

Combining Eqs. (21), (25), (27) and (28), the objective is naturally transformed into minimizing its upper bound:

$$\begin{aligned}
\mathcal{L}_{IB} &= - \underbrace{\mathbb{E}_{p(\mathbf{z}^{(v)}, \mathbf{s}^{(v)})} \left[ \log q(\mathbf{z}^{(v)}|\mathbf{s}^{(v)}) \right]}_{\text{lower bound for information shift}} - \underbrace{\iint p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}, \mathbf{s}^{(v^*)}|\mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)}}_{\text{lower bound for shard information interaction}} \\
&\quad + \underbrace{D_{KL}(p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v^*)}|\mathbf{z}^{(v^*)})q(\mathbf{o}^{(v)}|\mathbf{o}^{(v^*)}))}_{\text{upper bound for specific information interaction}} \\
&\quad + \underbrace{\beta D_{KL}(p(\mathbf{s}^{(v)}|\mathbf{z}^{(v)}; \mathbf{o}^{(v)}|\mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v)}|\mathbf{z}^{(v)})q(\mathbf{s}^{(v)}|\mathbf{o}^{(v)}))}_{\text{upper bound for information orthogonality}}. \tag{29}
\end{aligned}$$

□

## A.2 Implementation of the Loss Function

Data-driven contrastive learning [22] is used to compute various complex distributions. Specifically, representations are treated as probability vectors over  $D$  classes (corresponding to  $d$  dimension) via a Softmax activation function, and the joint distribution matrix is obtained by

$$\mathbf{P}^{(v, v^*)} = \sum_{i=1}^N \left( \mathbf{s}_i^{(v)} \right)^T \mathbf{s}_i^{(v^*)}, \quad \mathbf{Q}^{(v, v^*)} = \sum_{i=1}^N \left( \mathbf{o}_i^{(v)} \right)^T \mathbf{o}_i^{(v^*)}, \quad \mathbf{M}^{(s^v, o^v)} = \sum_{i=1}^N \left( \mathbf{s}_i^{(v)} \right)^T \mathbf{o}_i^{(v^*)}.$$

Due to the strong coupling between  $\mathbf{s}^{(v)}$  and  $\mathbf{s}^{(v^*)}$ , the variational distribution  $q(\mathbf{s}^{(v)}|\mathbf{s}^{(v^*)})$  can be effectively estimated by the obtained conditional distribution. In contrast, owing to the extremely

weak correlation between  $\mathbf{o}^{(v)}$  and  $\mathbf{o}^{(v^*)}$ , as well as  $\mathbf{o}^{(v)}$  and  $\mathbf{s}^{(v)}$ , the distributions  $q(\mathbf{o}^{(v)}/\mathbf{o}^{(v^*)})$  and  $q(\mathbf{s}^{(v)}/\mathbf{o}^{(v)})$  can be approximated by the corresponding marginal distribution. Since the first information shift term is equivalent to the reconstruction loss, we construct a decoder for each view  $v$  to obtain  $\hat{\mathbf{z}}^v$ , which is used to approximate the original view. Subsequently, by converting the integral to a summation form, each term in Eq. (29) can be expressed as

$$\begin{cases} \int \int p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}, \mathbf{s}^{(v^*)} | \mathbf{z}^{(v^*)}) \log q(\mathbf{s}^{(v)} | \mathbf{s}^{(v^*)}) d\mathbf{s}^{(v^*)} d\mathbf{s}^{(v)} = \sum_{d=1}^D \sum_{d'=1}^D \mathbf{P}_{d,d'}^{(v,v^*)} \log \frac{\mathbf{P}_{d,d'}^{(v,v^*)}}{\mathbf{P}_{d'}^{(v^*)}} \\ D_{KL}(p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v^*)} | \mathbf{z}^{(v^*)}) q(\mathbf{o}^{(v)} | \mathbf{o}^{(v^*)})) = \sum_{d=1}^D \sum_{d'=1}^D \mathbf{Q}_{d,d'}^{(v,v^*)} \log \frac{\mathbf{Q}_{d,d'}^{(v,v^*)}}{(\mathbf{Q}_d^{(v)})^\alpha \mathbf{Q}_{d'}^{(v^*)}} \\ D_{KL}(p(\mathbf{s}^{(v)} | \mathbf{z}^{(v)}; \mathbf{o}^{(v)} | \mathbf{z}^{(v)}) \| p(\mathbf{o}^{(v)} | \mathbf{z}^{(v)}) q(\mathbf{s}^{(v)} | \mathbf{o}^{(v)})) = \sum_{d=1}^D \sum_{d'=1}^D \mathbf{M}_{d,d'}^{(s^v, o^v)} \log \frac{\mathbf{M}_{d,d'}^{(s^v, o^v)}}{(\mathbf{M}_d^{(s^v)})^\alpha \mathbf{M}_{d'}^{(o^v)}}, \end{cases} \quad (30)$$

where the Lagrange multiplie  $\beta$  is fixed to 1, and the marginal distributions  $\mathbf{P}_d^{(v)}$  and  $\mathbf{P}_{d'}^{(v^*)}$  can be obtained by summing over the  $d$ -th row and the  $d'$ -th column of the joint distribution matrix  $\mathbf{P}^{(v,v^*)}$ , with other symbols defined as the same. In this conversion, we let  $q(\mathbf{s}^v / \mathbf{s}^{v^*}) = (\psi(\mathbf{Q}^{(v)}))^\alpha$  and  $q(\mathbf{z}^v / \mathbf{s}^v) = (\psi(\mathbf{M}^{(s^v)}))^\alpha$ , where  $\psi$  is a fully connected layer, and  $\alpha$  is a balance factor to preserve crucial information and ensure model stability [17]. We set the value of  $\alpha$  to 10 in our experiments.

## B Detailed Proof for the Theoretical Results

In this section, we will provide a rigorous proof of the theoretical results mentioned in the main text.

### B.1 Proof of the Theorem 1

**Theorem 4. (Discriminability of Label-specific Representation.)** For label prototypes  $\mathbf{E}_j$  and  $\mathbf{E}_k$  such that  $k \neq j$  for all  $k$ , and view representations  $\mathbf{X}^{(v)}$  and  $\mathbf{X}^{(v^*)}$  such that  $v^* \neq v$  for all  $v^*$ , the discriminability of  $\hat{\mathbf{P}}^{(v)}$  for class  $j$  necessitates that either of the following conditions be satisfied:

$$\begin{cases} \mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_k)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_k)^\top \\ \mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} [\mathbf{X}^{(v^*)} \sigma_S(\mathbf{E}_j)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v^*)} \sigma_S(\mathbf{E}_j)^\top. \end{cases}$$

*Proof.* Since our training data is multi-view and multi-label, the discriminative power of view  $v$  with respect to label  $j$  involves two key considerations. First, among all labels, view  $v$  yields the most accurate prediction for label  $j$ ; second, across all views, view  $v$  provides the prediction for label  $j$  with the highest confidence. Thus, we proceed with the proof from the following two aspects:

(i) For  $\mathbf{E}_j$  and  $\mathbf{E}_k$  such that  $k \neq j$  for all  $k$ , the following inequality regarding view  $v$  holds:

$$\mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_k)^\top] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_k)^\top. \quad (31)$$

The  $j$ -th component of  $\hat{\mathbf{P}}_i^{(v)}$  (corresponding to label  $j$ ) is

$$\hat{\mathbf{P}}_{i,j}^{(v)} = \sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)},$$

where  $\hat{\mathbf{P}}_{i,j}^{(v)}$  is a vector where each element is the dot product of the corresponding elements of  $\sigma_S(\mathbf{E}_j)$  and  $\mathbf{x}_i^{(v)}$ . Since a linear classifier (e.g., fully connected layers) is used for classification, the prediction score for sample  $i$  is

$$\mathbf{U}_i^{(v)} = \mathbf{W} \hat{\mathbf{P}}_i^{(v)} + \mathbf{b},$$

where  $\mathbf{W} \in \mathbb{R}^{C \times (C \times d)}$  is the weight matrix and  $\mathbf{b} \in \mathbb{R}^C$  is the bias term. Given that  $\hat{\mathbf{P}}_i^{(v)} \in \mathbb{R}^{C \times d}$ , we can flatten it into  $C$  individual vectors for each label. Besides,  $\mathbf{W}$  has a block-diagonal structure, i.e., the weights for each label  $j$  only act on  $\hat{\mathbf{P}}_{i,j}^{(v)}$ . Then, the prediction score for label  $j$  simplifies to

$$U_{i,j}^{(v)} = \mathbf{w}_j^\top \hat{\mathbf{P}}_{i,j}^{(v)} + b_j = \mathbf{w}_j^\top (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) + b_j,$$

where  $\mathbf{w}_j \in \mathbb{R}^d$  is the classifier weight for label  $j$ . To express discriminability, we need to show

$$\mathbb{E} [U_{:,j}^{(v)}] > \mathbb{E} [U_{:,k}^{(v)}], \quad \forall k \neq j.$$

The expectation is obtained by averaging over all empirical samples:

$$\begin{aligned} \mathbb{E} [U_{:,j}^{(v)}] &= \frac{1}{N} \sum_{i=1}^N \mathbf{w}_j^\top (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) + b_j, \\ \mathbb{E} [U_{:,k}^{(v)}] &= \frac{1}{N} \sum_{i=1}^N \mathbf{w}_k^\top (\sigma_S(\mathbf{E}_k) \odot \mathbf{x}_i^{(v)}) + b_k. \end{aligned}$$

Assuming classifier weights  $\mathbf{w}_k$  are independent of input data (e.g., optimized during training) and biases  $b_k$  are constants, the core comparison reduces to

$$\sum_{i=1}^N \mathbf{w}_j^\top (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) \quad \text{vs.} \quad \sum_{i=1}^N \mathbf{w}_k^\top (\sigma_S(\mathbf{E}_k) \odot \mathbf{x}_i^{(v)}). \quad (32)$$

During training, the classifier weights  $\mathbf{w}_j$  tend to align with label prototypes  $\mathbf{E}_j$  since  $\mathbf{E}_j$  captures the semantic meaning of label  $j$ . Thus, we approximate  $\mathbf{w}_k \approx \mathbf{E}_k^\top$ . Substituting it into Eq. (32), the comparison terms simplify to

$$\sum_{i=1}^N \mathbf{E}_j = (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) \quad \text{vs.} \quad \sum_{i=1}^N \mathbf{E}_k (\sigma_S(\mathbf{E}_k) \odot \mathbf{x}_i^{(v)}).$$

Using properties of dot products:

$$\begin{aligned} \mathbf{E}_j (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) &= \sum_{l=1}^d E_{j,l} \cdot \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v)}, \\ \mathbf{E}_k (\sigma_S(\mathbf{E}_k) \odot \mathbf{x}_i^{(v)}) &= \sum_{l=1}^d E_{k,l} \cdot \sigma_S(E_{k,l}) \cdot x_{i,l}^{(v)}. \end{aligned}$$

Then, the comparison is further refined into

$$\sum_{i=1}^N \sum_{l=1}^d E_{j,l} \cdot \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v)} \quad \text{vs.} \quad \sum_{i=1}^N \sum_{l=1}^d E_{k,l} \cdot \sigma_S(E_{k,l}) \cdot x_{i,l}^{(v)}.$$

Note that  $\sigma_S(E_{j,l}) \in (0, 1)$  and is monotonically increasing. For positive samples of label  $i$ ,  $\mathbf{E}_j$  and  $\mathbf{x}_i^{(v)}$  are better aligned, meaning  $E_{j,l} \cdot x_{i,l}^{(v)}$  has a higher expectation. Thus, if the condition is

$$\mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top] > \mathbb{E} [\mathbf{X}^{(v)} \sigma_S(\mathbf{E}_k)^\top].$$

Then, the following inequality holds:

$$\sum_{i=1}^N \sum_{l=1}^d \sigma_S(E_{k,l}) \cdot x_{i,l}^{(v)} > \sum_{i=1}^N \sum_{l=1}^d \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v)}.$$

Assuming  $\mathbf{E}_j$  and  $\mathbf{E}_k$  have similar scales ( $\|\mathbf{E}_j\|_2^2 \approx \|\mathbf{E}_k\|_2^2$ ), we conclude

$$\sum_{i=1}^N \mathbf{E}_j (\sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)}) > \sum_{i=1}^N \mathbf{E}_k (\sigma_S(\mathbf{E}_k) \odot \mathbf{x}_i^{(v)}). \quad (33)$$

Furthermore, we have

$$\mathbb{E} \left[ U_{:,j}^{(v)} \right] > \mathbb{E} \left[ U_{:,k}^{(v)} \right], \quad \forall k \neq j.$$

Therefore, the prediction score for label  $j$  exceeds that of all other labels, demonstrating the discriminative capability of view  $v$  for class  $j$ .

(ii) For view  $v$  and  $v^*$  such that  $v^* \neq v$  for all  $v^*$ , the following inequality regarding label  $j$  holds:

$$\mathbb{E} \left[ \mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v)} \sigma_S(\mathbf{E}_j)^\top > \mathbb{E} \left[ \mathbf{X}^{(v^*)} \sigma_S(\mathbf{E}_j)^\top \right] = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^{(v^*)} \sigma_S(\mathbf{E}_j)^\top. \quad (34)$$

The prediction score for label  $j$  assigned by view  $v$  and  $v^*$ , respectively, is

$$\begin{aligned} U_{i,j}^{(v)} &= \mathbf{w}_j^\top \hat{\mathbf{P}}_{i,j}^{(v)} + b_j = \mathbf{w}_j^\top \left( \sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)} \right) + b_j, \\ U_{i,j}^{(v^*)} &= \mathbf{w}_j^\top \hat{\mathbf{P}}_{i,j}^{(v^*)} + b_j = \mathbf{w}_j^\top \left( \sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v^*)} \right) + b_j, \end{aligned}$$

To express discriminability, we need to show

$$\mathbb{E} \left[ U_{:,j}^{(v)} \right] > \mathbb{E} \left[ U_{:,j}^{(v^*)} \right], \quad \forall v \neq v^*.$$

Following the same rationale, we can transform the comparison into

$$\sum_{i=1}^N \sum_{l=1}^d E_{j,l} \cdot \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v)} \quad \text{vs.} \quad \sum_{i=1}^N \sum_{l=1}^d E_{j,l} \cdot \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v^*)}.$$

Thus, if the condition is

$$\mathbb{E} \left[ \mathbf{X}^{(v)} \sigma_S(\mathbf{E}_j)^\top \right] > \mathbb{E} \left[ \mathbf{X}^{(v^*)} \sigma_S(\mathbf{E}_j)^\top \right].$$

Then, the following inequality holds:

$$\sum_{i=1}^N \sum_{l=1}^d \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v)} > \sum_{i=1}^N \sum_{l=1}^d \sigma_S(E_{j,l}) \cdot x_{i,l}^{(v^*)}.$$

Similarly, we conclude

$$\sum_{i=1}^N \mathbf{E}_j \left( \sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v)} \right) > \sum_{i=1}^N \mathbf{E}_j \left( \sigma_S(\mathbf{E}_j) \odot \mathbf{x}_i^{(v^*)} \right). \quad (35)$$

Furthermore, we have

$$\mathbb{E} \left[ U_{:,j}^{(v)} \right] > \mathbb{E} \left[ U_{:,k}^{(v^*)} \right], \quad \forall v \neq v^*.$$

Therefore, the prediction score of the  $v$ -th view for class  $j$  is higher than that of all other views, indicating that view  $v$  exhibits discriminative power for label  $j$ . Combining (i) and (ii), we complete the proof.  $\square$

## B.2 Proof of the Theorem 2

**Theorem 5. (Effectiveness of Disentangled Representation.)** Let the disentangled representation be denoted as  $\mathbf{R} = (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)})$ , where the information entropy of each representation is assumed to be fixed, i.e.,  $H(\mathbf{S}^{(v)}) = H(\mathbf{O}^{(v)}) = H^0$  ( $1 \leq v \leq V$ ). Then, in the case where each shared and specific representation is indispensable for prediction,  $I(\mathbf{R}; \mathbf{Y})$  will attain its maximum when  $\mathbf{R} = \mathbf{R}_*$ , with  $\mathbf{R}_*$  being the optimal solution of the problem (16).

*Proof.* By integrating core features from the raw data to construct shared representations and leveraging classification loss to derive view-specific representations aligned with particular labels, our approach enables the shared features to encapsulate primary generalization information across multiple views, while the specific features are designed to capture discriminative information that is

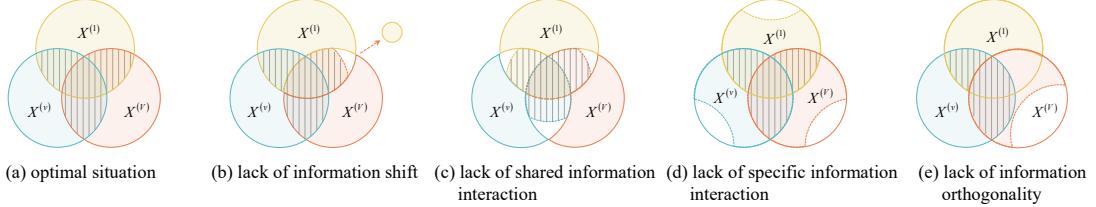


Figure 5: Feature extraction visualization. Fig. (a) illustrates the optimal solution obtained by our model, while the remaining four figures display cases where the constraints of information shift, interaction, and orthogonality are successively removed. The shaded areas represent the shared representations, whereas the other colored regions indicate the view-specific representations.

finely attuned to task-specific characteristics. Thus, the assumption that each shared and specific representation is indispensable for prediction aligns well with our model. Besides, as all representations are extracted into the same dimensional space, it is reasonable to posit that they encode equivalent quantities of information, i.e.,  $H(\mathbf{S}^{(v)}) = H(\mathbf{O}^{(v)}) = H^0$  ( $1 \leq v \leq V$ ). Let define the optimal solution of the problem (16) as  $\mathbf{R}_* = (\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)})$ . Then, we specify that identifying each class  $\mathbf{Y}^c$  requires a corresponding representation setting  $\mathbf{R}^{\mathbf{Y}^c}$  to be mined, i.e.,  $H(\mathbf{Y}^c) = H(\mathbf{R}^{\mathbf{Y}^c})$ , which indicates that the successful extraction of  $\mathbf{R}^{\mathbf{Y}^c}$  is sufficient for predicting  $\mathbf{Y}^c$ . Based on this, we establish the following formulation:

$$\begin{cases} H(\mathbf{R}^{\mathbf{Y}^c}) = H(\mathbf{S}^{\mathbf{Y}^c}) + H(\mathbf{O}^{\mathbf{Y}^c}), H(\mathbf{S}^{\mathbf{Y}^c}) = \sum_{v=1}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}), H(\mathbf{O}^{\mathbf{Y}^c}) = \sum_{v=1}^V g_c^v \cdot H(\mathbf{O}_*^{(v)}) \\ \sum_{c=1}^C \sum_{v=1}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) + \sum_{c=1}^C \sum_{v=1}^V g_c^v \cdot H(\mathbf{O}_*^{(v)}) = H(\mathbf{R}_*). \end{cases} \quad (36)$$

where  $f_c^v$  and  $g_c^v$  are indicator functions representing the shared and specific information components that constitute  $\mathbf{R}^{\mathbf{Y}^c}$ . According to Eq. (36), all information contained in the extracted representations is essential for predicting the target labels. Next, we demonstrate from three perspectives that our feature extraction model is capable of maximizing the mutual information between the representations and the labels, i.e., the representations contain insufficient information for prediction, just enough information, or redundant information beyond what is necessary for prediction.

$$(i) (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}) \subset \mathbf{R}^{\mathbf{Y}^1} \cup \mathbf{R}^{\mathbf{Y}^2} \dots \cup \mathbf{R}^{\mathbf{Y}^C}.$$

In this case, we denote the mutual information between  $\mathbf{R}_*$  and  $\mathbf{Y}$  as  $I^0$ , i.e.,

$$\begin{aligned} I(\mathbf{R}_*; \mathbf{Y}) &= \sum_{c=1}^C I((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)}); \mathbf{Y}^c) \\ &= \sum_{c \in \mathcal{L}} I(\mathbf{R}^{\mathbf{Y}^c}; \mathbf{Y}^c) + \sum_{c \in \mathcal{U}} I((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)}); \mathbf{Y}^c) \\ &= \sum_{c \in \mathcal{L}} H(\mathbf{Y}^c) - H(\mathbf{Y}^c | \mathbf{R}^{\mathbf{Y}^c}) + \sum_{c \in \mathcal{U}} H((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)})) \\ &\quad - H((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)}) | \mathbf{Y}^c) \\ &= \sum_{c \in \mathcal{L}} H(\mathbf{Y}^c) + \sum_{c \in \mathcal{U}} H((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)})) \\ &= \sum_{c \in \mathcal{L}} H(\mathbf{Y}^c) + 2V|\mathcal{U}|H^0 - (V-1)|\mathcal{U}|I(\mathbf{S}_*^{(1)}; \dots; \mathbf{S}_*^{(V)}) = I^0. \end{aligned}$$

where  $\mathcal{L}$  and  $\mathcal{U}$  denote the subsets of labels that are fully predictable and not adequately predictable, respectively, and  $|\mathcal{L}| + |\mathcal{U}| = C$ . When the core shared feature information is essential to all labels, reducing  $I(\mathbf{S}_*^{(1)}; \dots; \mathbf{S}_*^{(V)})$  to enhance label-specific information associated with certain labels will not lead to an increase in  $I(\mathbf{R}_*; \mathbf{Y})$ . Since  $\sum_{c \in \mathcal{L}} H(\mathbf{Y}^c)$  is fixed, postulating the existence of an alternative representation  $\mathbf{R}$  satisfying  $I(\mathbf{R}; \mathbf{Y}) > I(\mathbf{R}_*; \mathbf{Y})$  would inevitably contradict the

assumption that the information entropy of each representation is fixed at  $H^0$ . Thus, we can obtain  $I(\mathbf{R}_*; \mathbf{Y}) \geq I(\mathbf{R}; \mathbf{Y})$ . Given the insufficiency of prediction information and a fixed amount of information encoded in the representations, enhancing performance necessitates incorporating usable content beyond raw information and introducing additional constraints offers no further gain in the mutual information between representations and labels.

$$(ii) (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}) = \mathbf{R}^{\mathbf{Y}^1} \cup \mathbf{R}^{\mathbf{Y}^2} \dots \cup \mathbf{R}^{\mathbf{Y}^C}.$$

In this case, similarly, we have the following equality:

$$\begin{aligned} I(\mathbf{R}_*; \mathbf{Y}) &= \sum_{c=1}^C I((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)}); \mathbf{Y}^c) \\ &= \sum_{c=1}^C I(\mathbf{R}^{\mathbf{Y}^c}; \mathbf{Y}^c) \\ &= \sum_{c=1}^C H(\mathbf{R}^{\mathbf{Y}^c}) - H(\mathbf{R}^{\mathbf{Y}^c} | \mathbf{Y}^c) \\ &= \sum_{c=1}^C \sum_{v=1}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) + \sum_{c=1}^C \sum_{v=1}^V g_c^v \cdot H(\mathbf{O}_*^{(v)}). \end{aligned}$$

If the information shift term is not sufficiently optimized as shown in Fig. 5.(b), there will exist a shared representation  $\tilde{\mathbf{S}}^{(m)}$  in the setting  $\mathbf{R}$  such that

$$\begin{aligned} I(\mathbf{R}; \mathbf{Y}) &= \sum_{c \in T_m^n} \sum_{v=1}^V \tilde{f}_c^v \cdot H(\mathbf{S}^{(v)}) + \sum_{c \in T_m} \sum_{v=1, \neq m}^V \tilde{f}_c^v \cdot H(\mathbf{S}^{(v)}) \\ &\quad + \sum_{c \in T_m} \tilde{f}_c^m \cdot H(\tilde{\mathbf{S}}^{(m)}) + \sum_{c=1}^C \sum_{v=1}^V g_c^v \cdot H(\mathbf{O}_*^{(v)}). \end{aligned}$$

where  $T_m$  denotes the set of labels whose prediction processes involve contributions from  $\tilde{\mathbf{S}}^{(m)}$ , while  $T_m^n$  is the set unaffected by  $\tilde{\mathbf{S}}^{(m)}$ . Since the change in information entropy occurs exclusively in  $\tilde{\mathbf{S}}^{(m)}$ , we have

$$\begin{cases} \sum_{c \in T_m^n} \sum_{v=1}^V \tilde{f}_c^v \cdot H(\mathbf{S}^{(v)}) = \sum_{c \in T_m^n} \sum_{v=1}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) \\ \sum_{c \in T_m} \sum_{v=1, \neq m}^V \tilde{f}_c^v \cdot H(\mathbf{S}^{(v)}) = \sum_{c \in T_m} \sum_{v=1, \neq m}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) \\ \sum_{c \in T_m} \tilde{f}_c^m \cdot H(\tilde{\mathbf{S}}^{(m)}) \leq \sum_{c \in T_m} f_c^m \cdot H(\mathbf{S}_*^{(m)}). \end{cases}$$

Furthermore, we have the following comparison:

$$\begin{aligned} I(\mathbf{R}; \mathbf{Y}) &\leq \sum_{c \in T_m^n} \sum_{v=1}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) + \sum_{c \in T_m} \sum_{v=1, \neq m}^V f_c^v \cdot H(\mathbf{S}_*^{(v)}) \\ &\quad + \sum_{c \in T_m} f_c^m \cdot H(\mathbf{S}_*^{(m)}) + \sum_{c=1}^C \sum_{v=1}^V g_c^v \cdot H(\mathbf{O}_*^{(v)}) \\ &= I(\mathbf{R}_*; \mathbf{Y}). \end{aligned}$$

If the shared information interaction term is not sufficiently optimized (Fig. 5.(c)), there will exist a set of shared representations whose task-relevant information is significantly reduced due to ineffective interactions. Under such circumstances, it similarly follows that  $I(\mathbf{R}; \mathbf{Y}) \leq I(\mathbf{R}_*; \mathbf{Y})$ . Moreover, insufficient suppression of interactions among specific representations causes overlap, thereby reducing the useful information entropy of each representation (Fig. 5.(d)). Additionally, inadequate optimization of the orthogonality constraint can introduce redundancy between shared and specific representations, which in turn hinders the extraction of the specific representations (Fig. 5.(e)). For this two aspects, we also have  $I(\mathbf{R}; \mathbf{Y}) \leq I(\mathbf{R}_*; \mathbf{Y})$ .

$$(iii) \mathbf{R}^{\mathbf{Y}^1} \cup \mathbf{R}^{\mathbf{Y}^2} \dots \cup \mathbf{R}^{\mathbf{Y}^C} \subset (\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}).$$

In this case, we can obtain

$$\begin{aligned}
I(\mathbf{R}_*; \mathbf{Y}) &= \sum_{c=1}^C I((\mathbf{S}_*^{(1)}, \dots, \mathbf{S}_*^{(V)}, \mathbf{O}_*^{(1)}, \dots, \mathbf{O}_*^{(V)}); \mathbf{Y}^c) \\
&= \sum_{c=1}^C I(\mathbf{R}^{\mathbf{Y}^c}; \mathbf{Y}^c) \\
&= \sum_{c=1}^C H(\mathbf{Y}^c) - H(\mathbf{Y}^c | \mathbf{R}^{\mathbf{Y}^c}) \\
&= \sum_{c=1}^C H(\mathbf{Y}^c).
\end{aligned}$$

Given that  $\sum_{c=1}^C H(\mathbf{Y}^c)$  quantifies the total label information, the mutual information  $I(\mathbf{R}; \mathbf{Y})$  obtained from any alternative representation is inherently constrained by this upper bound. Therefore, we conclude that  $I(\mathbf{R}; \mathbf{Y}) \leq I(\mathbf{R}_*; \mathbf{Y})$ .

Combining (i) (ii) and (iii), we deduce that the maximization of  $I(\mathbf{R}; \mathbf{Y})$  is attained when  $\mathbf{R}$  serves as the optimal solution  $\mathbf{R}_*$  of the problem (16). Moreover, the principles of information shift, interaction, and orthogonality are all indispensable for achieving this.  $\square$

### B.3 Proof of the Theorem 3

**Theorem 6. (Generalization Error Bound.)** *Our model is designed to learn a vector-valued function  $f = (f_1, \dots, f_C) : \mathcal{X} \mapsto \mathbb{R}^C$ . The expected risk and empirical risk w.r.t. the training dataset  $D$  are denoted as  $R(f) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})]$  and  $\widehat{R}_D(f) = \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell(\sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c})$ , where  $f^{av}(\cdot)$  refers to the late fusion of multiple views. With probability at least  $1 - \delta$ , we have the following generalization error bound:*

$$\begin{aligned}
R(f) - \widehat{R}_D(f) &\leq \frac{\tilde{\mathcal{K}}_1}{\sqrt{NV}} + \frac{\tilde{\mathcal{K}}_2}{N^{3/4}V^{1/4}} + \overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y}) \\
&\quad + \tilde{\mathcal{K}}_3 \sqrt{\frac{\sum_{c=1}^C \left( \sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) + \tilde{\mathcal{K}}_4 \right)}{NC}},
\end{aligned}$$

where  $\tilde{\mathcal{K}}_1 = \tilde{\mathcal{K}}_3 = \mathcal{O}(C)$ ,  $\tilde{\mathcal{K}}_2 = \mathcal{O}(\sqrt{C})$ ,  $\tilde{\mathcal{K}}_4$  is constant of order  $\tilde{\mathcal{O}}(1)$  as  $N, V \rightarrow \infty$ , and  $\overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y})$  is the generalization error related to the view reconstruction quality. Moreover, the generalization error bound becomes increasingly tighter during the optimization of the problem (16).

*Proof.* The training dataset denoted as  $D = \{(\mathbf{x}_i, \mathbf{y}_i) : i \in [N]\}$  is drawn from a probability distribution over  $\mathcal{X} \times \mathcal{Y}$ . Each  $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(V)})$  consists of  $V$  views and  $\mathbf{y}_i = (Y_{i,1}, \dots, Y_{i,C})$ . Our strategy is to learn a vector-valued function  $\mathbf{f} = (f_1, \dots, f_C) : \mathcal{X} \mapsto \mathbb{R}^C$  and determine relevant labels by applying a thresholding criterion. The goal of learning is to find a hypothesis  $f \in \mathcal{F}$  with good generalization performance by minimizing the loss  $\ell$  on the dataset  $D$ . As missing labels inevitably degrade generalization, we focus on deriving the tightest generalization error bound under complete labeling. Thus, the expected risk and empirical risk w.r.t. the training dataset  $D$  are denoted as  $R(f) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})]$  and  $\widehat{R}_D(f) = \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell(\sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c})$ , respectively, where  $f^{av}(\cdot)$  refers to the result fusion of multiple views,  $\mathbf{Q}$  is the reconstruction quality score serving as the fusion weights. We define the function class of TDLSR as follows:

$$\begin{aligned}
\mathcal{F} &= \{\mathbf{X}^{(v)} \mapsto \mathbf{f}(\mathbf{X}^{(v)}) : f(\mathbf{X}^{(v)}) = (f_1(\mathbf{X}^{(v)}), \dots, f_C(\mathbf{X}^{(v)})), f_c(\mathbf{X}^{(v)}) = \psi_c \phi^{(v)}(\mathbf{X}^{(v)}), \\
&\quad \phi^{(v)} = (\phi_v^s, \phi_v^o), \psi_c = (\mathbf{w}_c^s \zeta_c, \mathbf{w}_c^o \zeta_c), \mathbf{w} \in \mathbb{R}^{1 \times d}\}, \\
&\quad (37)
\end{aligned}$$

where  $\phi^s$  and  $\phi^o$  denote the view-common and view-specific representation extractors,  $\zeta_j$  is a nonlinear mapping induced by the label-specific representation construction and  $\mathbf{w}$  is the classification head. Besides, the disentangled representation  $\mathbf{R}$  is expressed by  $\mathbf{R} = \phi(\mathbf{X}) = (\phi^s(\mathbf{X}), \phi^o(\mathbf{X})) =$

$(\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)})$ . Let multi-view data  $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^V$  be produced with a hidden label-specific function  $\theta^c$  by  $\mathbf{X} = \theta^c(\mathbf{Y}^c, \mathbf{V})$ , where  $\mathbf{Y}^c$  is the randomly generated single label and  $\mathbf{V} = \{\mathbf{V}^{(v)} = (\mathbf{V}_1^{(v)}, \dots, \mathbf{V}_d^{(v)})\}_{v=1}^V \in \mathbb{R}^{m \times d}$  are nuisance variables. Denote the conditional random variables of  $\mathbf{X}$  and  $\mathbf{R}$  given the category  $\mathbf{Y}^c = \mathbf{y}^c$  as  $\mathbf{X}_{\mathbf{y}^c}$  and  $\mathbf{R}_{\mathbf{y}^c}$ , respectively. For any  $\mathbf{y}^c \in \mathcal{Y}^* = \{-1, 1\}$ , the sensitivity  $c_\phi^{\mathbf{y}^c}$  of the representation function  $\phi = \{\phi^{(v)}\}_{v=1}^V$  w.r.t the nuisance variable  $\mathbf{V}_i^{(v)}$  is defined as

$$c_\phi^{\mathbf{y}^c} = \sup_{v \in [V]} \sup_{\mathbf{v}_1^{(v)}, \dots, \hat{\mathbf{v}}_i^{(v)}, \dots, \mathbf{v}_d^{(v)}} |\log(p_{\mathbf{r}}(\phi^{(v)} \circ \theta_{\mathbf{y}^c}(\mathbf{v}_1^{(v)}, \dots, \mathbf{v}_i^{(v)}, \dots, \mathbf{v}_d^{(v)}))) - \log(p_{\mathbf{r}}(\phi^{(v)} \circ \theta_{\mathbf{y}^c}(\mathbf{v}_1^{(v)}, \dots, \hat{\mathbf{v}}_i^{(v)}, \dots, \mathbf{v}_d^{(v)})))|, \quad (38)$$

where  $\theta_{\mathbf{y}^c}(\mathbf{v}^{(v)}) = \theta(\mathbf{y}^c, \mathbf{v}^{(v)})$  and  $p_{\mathbf{r}}(\mathbf{r}) = \mathbb{P}(\mathbf{R} = \mathbf{r})$ . Based on Eq. (38), we set the global sensitivity of  $\phi$  as  $c_\phi = \sup_{c \in [C]} c_\phi^{\mathbf{y}^c}$ . Let  $\mathcal{R}_c^x = \{\theta_{\mathbf{y}^c}(\mathbf{v}), \phi \circ \theta_{\mathbf{y}^c}(\mathbf{v}) : \mathbf{v} \in \mathcal{V}, \mathbf{y}^c \in \mathcal{Y}^*\}$  denote the complete set of multi-view data and their corresponding representations. For any  $\gamma > 0$ , we construct the following typical representation subset for each class [39]:

$$\mathcal{R}_{c,\gamma}^x = \left\{ \mathbf{X}, \mathbf{R} \in \mathcal{R}_c^x : -\log p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}) - H(\mathbf{R}_{\mathbf{Y}^c}) \leq c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} \right\}. \quad (39)$$

Define the function  $h'_{\mathbf{y}^c}(\mathbf{v}) = -\log p_{\mathbf{r}|\mathbf{y}^c}(h''_{\mathbf{y}^c}(\mathbf{v}))$ , where  $h''_{\mathbf{y}^c}(\mathbf{v}) = \phi \circ (\theta_{\mathbf{y}^c}(\mathbf{v}))$ . Let  $p_{\mathbf{v}}(\mathbf{v}) = P(\mathbf{V} = \mathbf{v})$  and  $h_{\mathbf{y}^c}^{-1}(\mathbf{r}) = \{\mathbf{v} \in \mathcal{V} : h_{\mathbf{y}^c}(\mathbf{v}) = \mathbf{r}\}$ , we have

$$\begin{aligned} \mathbb{E}_{\mathbf{V}} [h'_{\mathbf{y}^c}(\mathbf{V})] &= - \sum_{\mathbf{v} \in \mathcal{V}} p_{\mathbf{v}}(\mathbf{v}) \log p_{\mathbf{r}|\mathbf{y}^c}(h''_{\mathbf{y}^c}(\mathbf{v})) \\ &= - \sum_{\mathbf{r} \in \mathcal{R}_c^x} \sum_{\mathbf{v} \in h_{\mathbf{y}^c}^{-1}(\mathbf{r})} p_{\mathbf{v}}(\mathbf{v}) \log p_{\mathbf{r}|\mathbf{y}^c}(h''_{\mathbf{y}^c}(\mathbf{v})) \\ &= - \sum_{\mathbf{r} \in \mathcal{R}_c^x} \left( \sum_{\mathbf{v} \in h_{\mathbf{y}^c}^{-1}(\mathbf{r})} p_{\mathbf{v}}(\mathbf{v}) \right) \log p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}) \\ &= - \sum_{\mathbf{r} \in \mathcal{R}_c^x} p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}) \log p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}) \\ &= H(R_{\mathbf{y}^c}). \end{aligned}$$

By applying McDiarmid's inequality on  $h'_{\mathbf{y}^c}(\mathbf{V})$ , we get

$$\mathbb{P}(-\log p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}) - H(\mathbf{R}_{\mathbf{Y}^c}) \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{d(c_\phi)^2}\right).$$

Taking  $\delta = \exp\left(-\frac{2\epsilon^2}{d(c_\phi)^2}\right)$ , we have

$$\epsilon = c_\phi \sqrt{\frac{d \log(1/\delta)}{2}}.$$

After setting  $\delta = \gamma/\sqrt{NV}$ , i.e.,  $\mathbb{P}(\mathbf{X}, \mathbf{R} \notin \mathcal{R}_{c,\gamma}^x) \leq \delta = \frac{\gamma}{\sqrt{NV}}$ , we can obtain

$$\exp(-H(\mathbf{R}_{\mathbf{Y}^c}) - \epsilon) \leq \log p_{\mathbf{r}|\mathbf{y}^c}(\mathbf{r}).$$

Then, we further derive the following transformation:

$$|\mathcal{R}_{c,\gamma}^x| \exp(-H(\mathbf{R}_{\mathbf{Y}^c}) - \epsilon) = \sum_{\mathbf{r} \in \mathcal{R}_{c,\gamma}^x} \exp(-H(\mathbf{R}_{\mathbf{Y}^c}) - \epsilon) \leq \sum_{\mathbf{r} \in \mathcal{R}_{c,\gamma}^x} p_{\mathbf{r}}(\mathbf{r}) = 1,$$

which implies

$$|\mathcal{R}_{c,\gamma}^x| \leq \exp\left(H(\mathbf{R}_{\mathbf{Y}^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}}\right).$$

Therefore, regarding the property of the typical subset, the following lemma can be derived:

**Lemma 1.** For any  $\gamma > 0$  and all  $v \in [V]$ , we have

$$\mathbb{P}(\mathbf{X}, \mathbf{R} \notin \mathcal{R}_{c,\gamma}^x) \leq \frac{\gamma}{\sqrt{NV}}, \quad |\mathcal{R}_{c,\gamma}^x| \leq \exp \left( H(\mathbf{R}_{Y^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} \right).$$

Besides, we need the following lemma for subsequent proof:

**Lemma 2.** [18] The vector  $\mathbf{X} = (X_1, \dots, X_k)$  is defined to follow the multinomial distribution with parameters  $p = (p_1, \dots, p_k)$ . Let  $\bar{a}_1, \dots, \bar{a}_k \geq 0$  be fixed such that  $\sum_{i=1}^k \bar{a}_i p_i \neq 0$ . Then, for any  $\epsilon > 0$ , the following inequality holds:

$$\mathbb{P} \left( \sum_{i=1}^k \bar{a}_i \left( p_i - \frac{X_i}{m} \right) > \epsilon \right) \leq \exp \left( -\frac{m\epsilon^2}{\beta} \right),$$

where  $\beta = 2 \sum_{i=1}^k \bar{a}_i^2 p_i$ .

We further define  $T^{y^c} = |\mathcal{R}_{c,\gamma}^x|$ , where the typical subset  $\mathcal{R}_{c,\gamma}^x$  consists of elements  $\mathcal{R}_{c,\gamma}^x = \{(a_{c,1}^x, a_{c,1}^s, a_{c,1}^o), \dots, (a_{c,T}^x, a_{c,T}^s, a_{c,T}^o)\}$ . Besides, we introduce the following label-specific sets:

$$\begin{aligned} \mathcal{U}^{y^c} &= \left\{ i \in [N], v \in [V] : \mathbf{x}_i^{(v)}, \phi^{(v)}(\mathbf{x}_i^{(v)}) \notin \mathcal{R}_{c,\gamma}^x, \mathbf{Y}_{i,c} = \mathbf{y}^c \right\}, \\ \mathcal{U}_k^{y^c} &= \left\{ i \in [N], v \in [V] : \mathbf{x}_i^{(v)} = a_k^{c,x}, \phi^{(v)}(\mathbf{x}_i^{(v)}) = (a_k^{c,s}, a_k^{c,o}), \mathbf{Y}_{i,c} = \mathbf{y}^c \right\}. \end{aligned} \quad (40)$$

Then, we conduct an analysis of the classification generalization error:

$$R(f) - \widehat{R}_D(f) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})] - \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right). \quad (41)$$

Since  $\ell$  is a convex function, we can apply Jensen's inequality to obtain

$$\ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right) \leq \mathbf{Q}_{i,v} \sum_{v=1}^V \ell(f_c(\mathbf{x}_i^{(v)}), \mathbf{Y}_{i,c}). \quad (42)$$

Thus, we define a instance-level loss difference term  $\Delta \ell_i^c(\mathbf{Q}_{i,:}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,c})$  such that

$$\Delta \ell_i^c = \mathbf{Q}_{i,v} \sum_{v=1}^V \ell(f_c(\mathbf{x}_i^{(v)}), \mathbf{Y}_{i,c}) - \ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right). \quad (43)$$

Define  $q = \sup_{i \in [N]} \sup_{v \in [V]} Q_{i,v}$  as the maximum view reconstruction quality among all views. Regarding the cumulative loss calculation for individual views, we have

$$\begin{aligned} & \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \left( \mathbf{Q}_{i,v} \sum_{v=1}^V \ell(f_c(\mathbf{x}_i^{(v)}), \mathbf{Y}_{i,c}) \right) \\ &= \frac{\mathbf{Q}_{i,v}}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \left( \sum_{i,v \in \mathcal{U}^{y^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) + \sum_{k=1}^{T^{y^c}} \sum_{i,v \in \mathcal{U}_k^{y^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \right) \\ &= \frac{\mathbf{Q}_{i,v}}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{i,v \in \mathcal{U}^{y^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) + \frac{\mathbf{Q}_{i,v}}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{y^c}} \sum_{i,v \in \mathcal{U}_k^{y^c}} \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), \mathbf{y}^c) \\ &\geq \frac{q}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{i,v \in \mathcal{U}^{y^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) + \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{y^c}} \frac{q |\mathcal{U}_k^{y^c}|}{NC} \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), \mathbf{y}^c). \end{aligned} \quad (44)$$

Furthermore, we can obtain the upper bound for the empirical risk:

$$\begin{aligned} \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right) &\geq \frac{q}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \\ &+ \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{\mathbf{y}^c}} \frac{q|\mathcal{U}_k^{\mathbf{y}^c}|}{NC} \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), \mathbf{y}^c) + \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \Delta \ell_i^c(\mathbf{Q}_{i,:}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,c}). \end{aligned} \quad (45)$$

The random variables of the quality score of the  $v$ -th view and the  $c$ -th category are denoted as  $\mathbf{Q}^{(v)}$  and  $\mathbf{Y}^c$ . Similarly, the population risk can be decomposed as

$$\begin{aligned} &\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})] \\ &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c] \\ &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(f_c^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x] \\ &+ \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \in \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(f_c^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \in \mathcal{R}_{c,\gamma}^x]. \end{aligned} \quad (46)$$

Since  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(f_c^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y}^c)] = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(\sum_{v=1}^V \mathbf{Q}^{(v)} f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c)]$ , we can further decompose it based on the convexity of  $\ell$ :

$$\begin{aligned} \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(f_c^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y}^c)] &= \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\ell(\sum_{v=1}^V \mathbf{Q}^{(v)} f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c)] \\ &= \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\sum_{v=1}^V \mathbf{Q}^{(v)} \ell(f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c)] + \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\Delta \ell^c(\mathbf{Q}, \mathbf{X}, \mathbf{Y}^c)], \end{aligned} \quad (47)$$

where  $\ell^c(\mathbf{Q}, \mathbf{Y}^c)$  represents the change in loss  $\ell$  associated with class  $c$  induced by view fusion. Then, we can obtain the following transformation combining Eqs. (46) and (47):

$$\begin{aligned} &\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})] \\ &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\sum_{v=1}^V \mathbf{Q}^{(v)} \ell(f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x] \\ &+ \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \in \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\sum_{v=1}^V \mathbf{Q}^{(v)} \ell(f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \in \mathcal{R}_{c,\gamma}^x] \\ &+ \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\Delta \ell^c(\mathbf{Q}, \mathbf{X}, \mathbf{Y}^c) | \mathbf{Y}^c = y^c] \\ &\leq \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*, \mathbf{X}^{(j)} \sim \mathbf{X}} [\ell(\psi_c \circ \phi^j(\mathbf{X}^{(j)}), y^c) | \mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x] \\ &+ \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\Delta \ell^c(\mathbf{Q}, \mathbf{X}, \mathbf{Y}^c) | \mathbf{Y}^c = y^c]. \end{aligned} \quad (48)$$

Putting Eqs. (58) and (48) back into Eq. (41), we have

$$\begin{aligned}
& \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})] - \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right) \\
& \leq \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*, \mathbf{X}^j \sim \mathbf{X}} [\ell(\psi_c \circ \phi^j(\mathbf{X}^j), y^c) | \mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \\
& \quad \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x] - \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x) \frac{1}{|\mathcal{U}^{\mathbf{y}^c}|} \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \\
& \quad + \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x) \frac{1}{|\mathcal{U}^{\mathbf{y}^c}|} \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \\
& \quad - \frac{q}{NC} \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \\
& \quad + \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{\mathbf{y}^c}} \mathbb{P}(\mathbf{Y}^c = y^c, \mathbf{X}^{(j)} = a_k^{c,x}, \phi^j(\mathbf{X}^{(j)}) = (a_k^{c,s}, a_k^{c,o})) \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), y^c) \\
& \quad - \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{\mathbf{y}^c}} \frac{q|\mathcal{U}_k^{\mathbf{y}^c}|}{NC} \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), \mathbf{y}^c) \\
& \quad + \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\Delta \ell^c(\mathbf{Q}, \mathbf{X}, \mathbf{Y}^c) | \mathbf{Y}^c = y^c] - \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \Delta \ell_i^c(\mathbf{Q}_{i,:}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,c} = \mathbf{y}^c). \tag{49}
\end{aligned}$$

Let  $\mathcal{S}_{x,y} = \sup_{c \in [C]} \sup_{(\mathbf{X}, \mathbf{Y}^c) \in \mathcal{X} \times \mathcal{Y}^*} \sum_{v=1}^V \mathbf{Q}^{(v)} \ell(f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c)$  denote the maximum attainable losses among all classes and  $\mathcal{S}_{x,y}^s = \sup_{c \in [C]} \sup_{i \in [n]} \mathbf{Q}_{i,v} \sum_{v=1}^V \ell(f_c(\mathbf{x}_i^{(v)}), \mathbf{Y}_{i,c})$  represent the maximum instance-level losses. We consider the results of Eq. (49) as four terms, with the upper bound of the first term as follows:

$$\begin{aligned}
\mathcal{P}_1 &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x) (\mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*, \mathbf{X}^j \sim \mathbf{X}} [q \ell(\psi_c \circ \phi^j(\mathbf{X}^j), y^c) | \mathbf{Y}^c = y^c, \\
& \quad (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x] - \frac{1}{|\mathcal{U}^{\mathbf{y}^c}|} \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c)) \\
&\leq \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*, \mathbf{X}^j \sim \mathbf{X}} [\sum_{v=1}^V \mathbf{Q}^{(v)} \ell(f_c(\mathbf{X}^{(v)}), \mathbf{Y}^c) | \mathbf{Y}^c = y^c, (\mathbf{X}, \mathbf{R}) \notin \mathcal{R}_{c,\gamma}^x] \\
&\leq \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c) \frac{\gamma}{\sqrt{NV}} \mathcal{S}_{x,y} = \frac{C\gamma}{\sqrt{NV}} \mathcal{S}_{x,y}. \tag{50}
\end{aligned}$$

Define  $p_k^c = \mathbb{P}(\mathbf{Y}^c = y^c, \mathbf{X}^{(j)} = a_k^{c,x}, \phi^j(\mathbf{X}^{(j)}) = (a_k^{c,s}, a_k^{c,o}))$  for  $k \in [T^{\mathbf{y}^c}]$ ,  $p_{T^{\mathbf{y}^c}+1}^c = \mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x)$  and  $b_k^c = \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), y^c)$  for  $k \in [T^{\mathbf{y}^c} + 1]$ . Then, we can obtain the following term:

$$\mathcal{P}_{3,k}^c = \sum_{t=1}^{T^{\mathbf{y}^c}} (p_t^c - \frac{|\mathcal{U}_t^{\mathbf{y}^c}|}{NCq^{-1}}) b_t^c - (p_k^c - \frac{|\mathcal{U}_k^{\mathbf{y}^c}|}{NCq^{-1}}) b_k^c. \tag{51}$$

Applying Lemma 2, for any  $\epsilon > 0$  and  $k \in [T^{\mathbf{y}^c}]$ , we have

$$\mathbb{P}(\mathcal{P}_{3,k}^c \geq \epsilon) \leq \exp \left( -\frac{NCq^{-1}\epsilon^2}{2 \left( \sum_{t=1}^{T^{\mathbf{y}^c}} (p_t^c - \frac{|\mathcal{U}_t^{\mathbf{y}^c}|}{NCq^{-1}})^2 - p_k^c (b_k^c)^2 \right)} \right). \tag{52}$$

Similarly, we get the following inequality regarding  $k = T^{\mathbf{y}^c} + 1$ :

$$\mathbb{P} \left( p_{T^{\mathbf{y}^c}+1}^c - \frac{|\mathcal{U}^{\mathbf{y}^c}|}{NCq^{-1}} \geq \epsilon \right) \leq \exp \left( -\frac{NCq^{-1}\epsilon^2}{2p_{T^{\mathbf{y}^c}+1}^c} \right). \quad (53)$$

Let  $\delta$  be the right-hand side of Eq. (52) and (53), respectively, we can obtain the following variants:

$$\left\{ \begin{array}{l} \mathbb{P} \left( \mathcal{P}_{3,k}^c \geq \sqrt{\sum_{t=1}^{T^{\mathbf{y}^c}} p_t^c (b_t^c)^2 - p_k^c (b_k^c)^2} \sqrt{\frac{2 \log(1/\delta)}{NCq^{-1}}} \right) \leq \delta, \forall k \in [T^{\mathbf{y}^c}] \\ \mathbb{P} \left( p_{T^{\mathbf{y}^c}+1}^c - \frac{|\mathcal{U}^{\mathbf{y}^c}|}{NCq^{-1}} \geq \sqrt{\frac{2p_{T^{\mathbf{y}^c}+1}^c \log(1/\delta)}{NCq^{-1}}} \right) \leq \delta. \end{array} \right. \quad (54)$$

Taking union bounds over all  $\mathbf{y}^c \in \mathcal{Y}^*$  ( $|\mathcal{Y}^*| = 2$ ) and  $k \in [T^{\mathbf{y}^c}]$ , we have

$$\left\{ \begin{array}{l} \mathbb{P} \left( \mathcal{P}_{3,k}^c \geq \sqrt{\sum_{t=1}^{T^{\mathbf{y}^c}} p_t^c (b_t^c)^2 - p_k^c (b_k^c)^2} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}} \right) \leq \delta, \forall k \in [T^{\mathbf{y}^c}] \\ \mathbb{P} \left( p_{T^{\mathbf{y}^c}+1}^c - \frac{|\mathcal{U}^{\mathbf{y}^c}|}{NCq^{-1}} \geq \sqrt{\frac{2p_{T^{\mathbf{y}^c}+1}^c \log(2/\delta)}{NCq^{-1}}} \right) \leq \delta. \end{array} \right. \quad (55)$$

Then, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , we have the following scaling regarding the second term of Eq. (49) by using Jensen's inequality:

$$\begin{aligned} \mathcal{P}_2 &= \frac{1}{|\mathcal{U}^{\mathbf{y}^c}|} \sum_{c=1}^C \sum_{y^c \in \mathcal{Y}^*} \left( \mathbb{P} \left( \mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x \right) - \frac{q|\mathcal{U}^{\mathbf{y}^c}|}{NC} \right) \sum_{i,v \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c) \\ &\leq \sum_{c=1}^C \sum_{y^c \in \mathcal{Y}^*} \sqrt{\mathbb{P}(\mathbf{Y}^c = y^c, (\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x)} \frac{\sum_{i,j \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c)}{|\mathcal{U}^{\mathbf{y}^c}|} \sqrt{\frac{2 \log(2/\delta)}{NCq^{-1}}} \\ &= \sum_{c=1}^C \sum_{y^c \in \mathcal{Y}^*} \sqrt{\mathbb{P}(\mathbf{Y}^c = y^c)} \sqrt{\mathbb{P}(\mathbf{X}^{(j)}, \phi^j(\mathbf{X}^{(j)})) \notin \mathcal{R}_{c,\gamma}^x} \frac{\sum_{i,j \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c)}{|\mathcal{U}^{\mathbf{y}^c}|} \sqrt{\frac{2 \log(2/\delta)}{NCq^{-1}}} \\ &\leq \sum_{c=1}^C \frac{\sqrt{\gamma}}{(NV)^{1/4}} \frac{\sum_{i,j \in \mathcal{U}^{\mathbf{y}^c}} \ell(\psi_c \circ \phi^{(v)}(\mathbf{x}_i^{(v)}), \mathbf{y}^c)}{|\mathcal{U}^{\mathbf{y}^c}|} \sqrt{|\mathcal{Y}^*|} \sqrt{\sum_{y^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = y^c)} \sqrt{\frac{2 \log(2/\delta)}{NCq^{-1}}} \\ &\leq \mathcal{S}_{x,y}^s \frac{2\sqrt{\gamma}}{(NV)^{1/4}} \sqrt{\frac{C \log(2/\delta)}{Nq^{-1}}}. \end{aligned} \quad (56)$$

Similarly, for  $\mathcal{P}_{3,k}^c$ , with probability at least  $1 - \delta$ , we have

$$\begin{aligned} \mathcal{P}_{3,k}^c &\leq \sqrt{\sum_{t=1}^{T^{\mathbf{y}^c}} p_t^y (b_t^y)^2 - p_k^y (b_k^y)^2} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}} \\ &\leq \mathcal{S}_{x,y} \sqrt{\sum_{t=1}^{T^{\mathbf{y}^c}} p_t^y - p_k^y} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}} \\ &= \mathcal{S}_{x,y} \sqrt{\mathbb{P} \left( \mathbf{Y}^c = \mathbf{y}^c \bigcap (\mathbf{X}, \mathbf{R}) \in \mathcal{R}_{c,\gamma}^x \bigcap (\mathbf{X}^{(j)}, \phi^{(j)}(\mathbf{X}^{(j)})) \neq (a_k^{c,x}, a_k^{c,s}, a_k^{c,o}) \right)} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}} \\ &\leq \mathcal{S}_{x,y} \sqrt{\mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c)} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}}. \end{aligned} \quad (57)$$

Next, we can derive the constraint result for the third term:

$$\begin{aligned}
\mathcal{P}_3 &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sum_{k=1}^{T^{\mathbf{y}^c}} \left( \mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c, \mathbf{X}^{(j)} = a_k^{c,x}, \phi^j(\mathbf{X}^{(j)}) = (a_k^{c,s}, a_k^{c,o})) - \frac{q|\mathcal{U}_k^{\mathbf{y}^c}|}{NC} \right) \ell(\psi_c(a_k^{c,s}, a_k^{c,o}), \mathbf{y}^c) \\
&= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \frac{1}{T^{\mathbf{y}^c} - 1} \sum_{k=1}^{T^{\mathbf{y}^c}} \mathcal{P}_{3,k}^c \\
&\leq \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \frac{T^{\mathbf{y}^c}}{T^{\mathbf{y}^c} - 1} \mathcal{S}_{x,y} \sqrt{\mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c)} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}} \\
&\leq 2 \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathcal{S}_{x,y} \sqrt{\mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c)} \sqrt{\frac{2 \log(2T^{\mathbf{y}^c}/\delta)}{NCq^{-1}}}.
\end{aligned} \tag{58}$$

The property of  $T^{\mathbf{y}^c}$  is as follows:

$$T^{\mathbf{y}^c} = |\mathcal{R}_{c,\gamma}^x| \leq \exp \left( H(\mathbf{R}_{\mathbf{Y}^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} \right). \tag{59}$$

Combining Eqs. (58) and (59), we can get the following bound by using Jensen's inequality:

$$\begin{aligned}
\mathcal{P}_3 &\leq 2 \sum_{c=1}^C \mathcal{S}_{x,y} \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \sqrt{\mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c)} \sqrt{\frac{2 \left( H(\mathbf{R}_{\mathbf{Y}^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} \right) + 2 \log(2/\delta)}{NCq^{-1}}} \\
&\leq 2 \sum_{c=1}^C \mathcal{S}_{x,y} \sqrt{|\mathcal{Y}^*|} \sqrt{\sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c)} \sqrt{\frac{2 \left( H(\mathbf{R}_{\mathbf{Y}^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} \right) + 2 \log(2/\delta)}{NCq^{-1}}} \\
&= 4 \sum_{c=1}^C \mathcal{S}_{x,y} \sqrt{\frac{H(\mathbf{R}_{\mathbf{Y}^c}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} + \log(2/\delta)}{NCq^{-1}}} \\
&\leq 4C \mathcal{S}_{x,y} \sqrt{\sum_{c=1}^C \frac{H(\mathbf{R}|\mathbf{Y}^c) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} + \log(2/\delta)}{NCq^{-1}}}.
\end{aligned} \tag{60}$$

By applying the chain rule, we obtain the following conclusion:

$$\begin{aligned}
H(\mathbf{R}|\mathbf{Y}^c) &= I(\mathbf{X}^{(1)}; \mathbf{R}|\mathbf{Y}^c) + H(\mathbf{R}|\mathbf{Y}^c, \mathbf{X}^{(1)}) \\
&= I(\mathbf{X}^{(1)}; \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}|\mathbf{Y}^c) + H(\mathbf{R}|\mathbf{Y}^c, \mathbf{X}^{(1)}) \\
&= I(\mathbf{X}^{(1)}; \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}|\mathbf{Y}^c) + I(\mathbf{X}^{(1)}; \{\mathbf{O}^{(v)}\}_{v=2}^V|\mathbf{Y}^c, \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}) + H(\mathbf{R}|\mathbf{Y}^c, \mathbf{X}^{(1)}).
\end{aligned} \tag{61}$$

Since our model achieves complete feature separation, with seamlessly integrated shared representations, no overlap among special representations, and no redundancy between shared and special features, we have the following equality:

$$I(\mathbf{X}^{(1)}; \{\mathbf{O}^{(v)}\}_{v=2}^V|\mathbf{Y}^c, \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}) = I(\mathbf{X}^{(1)}; \{\mathbf{O}^{(v)}\}_{v=2}^V) = 0. \tag{62}$$

Thus, we can further obtain

$$\begin{aligned}
H(\mathbf{R}|\mathbf{Y}^c) &\leq I(\mathbf{X}^{(1)}; \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}|\mathbf{Y}^c) + H(\mathbf{R}|\mathbf{Y}^c, \mathbf{X}^{(1)}) \\
&\leq \sum_{v=1}^V I(\mathbf{X}^{(v)}; \{\mathbf{S}^{(v)}\}_{v=1}^V, \mathbf{O}^{(v)}|\mathbf{Y}^c) + H(\mathbf{R}|\mathbf{Y}^c, \mathbf{X}^{(1)})
\end{aligned} \tag{63}$$

Putting the above back into Eq. 60, we have

$$\mathcal{P}_3 \leq 4C\mathcal{S}_{x,y} \sqrt{\sum_{c=1}^C \frac{\sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) + H(\mathbf{R} | \mathbf{Y}^c, \mathbf{X}^{(1)}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} + \log(2/\delta)}{NCq^{-1}}}. \quad (64)$$

Regarding the fourth term, as the accuracy of the view reconstruction increases, indicated by a larger value of  $\mathbf{Q}$ , the loss discrepancy  $\Delta\ell_i^c(\mathbf{Q}_{i,:}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,c})$  becomes more significant, while the corresponding generalization error bound becomes tighter. Therefore, we interpret the fourth term as the generalization error induced by the quality of view reconstruction, i.e.,

$$\begin{aligned} \mathcal{P}_4 = \overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y}) &= \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \mathbb{P}(\mathbf{Y}^c = \mathbf{y}^c) \mathbb{E}_{(\mathbf{X}, \mathbf{Y}^c) \sim \mathcal{X} \times \mathcal{Y}^*} [\Delta\ell^c(\mathbf{Q}, \mathbf{X}, \mathbf{Y}^c) | \mathbf{Y}^c = \mathbf{y}^c] \\ &\quad - \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \sum_{\mathbf{y}^c \in \mathcal{Y}^*} \Delta\ell_i^c(\mathbf{Q}_{i,:}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,c} = \mathbf{y}^c). \end{aligned} \quad (65)$$

Therefore, with probability at least  $1 - \delta$ , the generalization error bound of our model is

$$\begin{aligned} &\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{X} \times \mathcal{Y}} [\ell(f^{av}(\mathbf{X}, \mathbf{Q}), \mathbf{Y})] - \frac{1}{NC} \sum_{i=1}^N \sum_{c=1}^C \ell \left( \sum_{v=1}^V (\mathbf{Q}_{i,v} f_c(\mathbf{x}_i^{(v)})), \mathbf{Y}_{i,c} \right) \\ &\leq \frac{C\gamma}{\sqrt{NV}} \mathcal{S}_{x,y} + \mathcal{S}_{x,y}^s \frac{2\sqrt{\gamma q}}{(NV)^{1/4}} \sqrt{\frac{C \log(2/\delta)}{N}} + \overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y}) \\ &\quad + 4C\mathcal{S}_{x,y} \sqrt{q} \sqrt{\sum_{c=1}^C \frac{\sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) + H(\mathbf{R} | \mathbf{Y}^c, \mathbf{X}^{(1)}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} + \log(2/\delta)}{NC}} \\ &= \frac{\tilde{\mathcal{K}}_1}{\sqrt{NV}} + \frac{\tilde{\mathcal{K}}_2}{N^{3/4}V^{1/4}} + \tilde{\mathcal{K}}_3 \sqrt{\frac{\sum_{c=1}^C \left( \sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) + \tilde{\mathcal{K}}_4 \right)}{NC}} + \overline{\text{gen}}_{rec}(\mathbf{Q}, \mathbf{X}, \mathbf{Y}), \end{aligned} \quad (66)$$

where

$$\begin{aligned} \tilde{\mathcal{K}}_1 &= C\gamma \mathcal{S}_{x,y} \\ \tilde{\mathcal{K}}_2 &= 2\mathcal{S}_{x,y}^s \sqrt{C\gamma q \log(2/\delta)} \\ \tilde{\mathcal{K}}_3 &= 4C\mathcal{S}_{x,y} \sqrt{q} \\ \tilde{\mathcal{K}}_4 &= H(\mathbf{R} | \mathbf{Y}^c, \mathbf{X}^{(1)}) + c_\phi \sqrt{\frac{d \log(\sqrt{NV}/\gamma)}{2}} + \log(2/\delta). \end{aligned} \quad (67)$$

Since  $\mathbf{S}^{(v)}$  and  $\mathbf{O}^{(v)}$  are generated by  $\mathbf{X}^{(v)}$ , we have the following transformation:

$$\begin{aligned} &\sum_{c=1}^C \sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c) \\ &= \sum_{c=1}^C H((\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}) | \mathbf{Y}^c) \\ &= H((\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}) | \mathbf{Y}) \\ &= H((\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)})) - I((\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}); \mathbf{Y}). \end{aligned}$$

Based on Theorem 5, it can be inferred that during the model optimization process,  $I((\mathbf{S}^{(1)}, \dots, \mathbf{S}^{(V)}, \mathbf{O}^{(1)}, \dots, \mathbf{O}^{(V)}); \mathbf{Y})$  continuously approaches its maximum value, leading to a decrease in  $\sum_{c=1}^C \sum_{v=1}^V I(\mathbf{X}^{(v)}; \mathbf{S}^{(v)}, \mathbf{O}^{(v)} | \mathbf{Y}^c)$ , which indicates an improvement in generalization performance.  $\square$

## C Experiment

### C.1 Experiment Setup

Table 3: Detailed information of datasets.

View	Object	VOC 2007	Corel 5k	Esp Game	IAPR TC-12	MIR FLICKR
1	CH(64)	DenseHue(100)	DenseHue(100)	DenseHue(100)	DenseHue(100)	DenseHue(100)
2	CM(225)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)
3	CORR(144)	GIST(512)	GIST(512)	GIST(512)	GIST(512)	GIST(512)
4	EDH(73)	HSV(4096)	HSV(4096)	HSV(4096)	HSV(4096)	HSV(4096)
5	WT(128)	RGB(4096)	RGB(4096)	RGB(4096)	RGB(4096)	RGB(4096)
6	-	LAB(4096)	LAB(4096)	LAB(4096)	LAB(4096)	LAB(4096)
#Label	31	20	260	268	291	38
#Instance	6047	9963	4999	20770	19627	25000

**Datasets and Comparison Methods.** In our experiments, six public multi-view multi-label datasets are selected as shown in Table 3. Their specific descriptions are as follows. **Corel 5k** is composed of 4999 image samples and 260 words, where each word can be regarded as an annotation or label. **IAPRTC12** comprises 19627 high-quality natural images and each image contains 261 labels, including sports, actions, animals, cities, and so on. **ESPGAME** is a multi-view multi-label dataset containing 20770 images with 268 corresponding tags. **Pascal07** is a widely utilized dataset for visual object detection and recognition, which contains 9963 images and 20 kinds of objects. **Mirflickr** consists of 25,000 images from the Flickr platform, annotated with a total of 38 tags. **OBJECT** has 6047 instances requiring recognition, which are characterized by five distinct perspectives and annotated with 31 attributes. To validate the effectiveness of TDLSR, we compare it with nine state-of-the-art approaches, i.e., AIMNet [23], DICNet [25], DIMC [38], iMVWL [33], LMVCAT [26], MTD [24], SIP [27], LVSL [43], DM2L [28]. We also provide a comprehensive overview of their sources and functions in Table 4.

**Construction of the Application NBA Dataset.** The NBA dataset crawled from Basketball-Reference [2] includes 16,992 player samples from the regular and playoff seasons spanning 2002 to 2022. Each sample comprises six views capturing different aspects of player performance and background information: 1) *Scoring Statistics*, 20 features including shooting attempts, shooting percentages (scaled by 1000), points, and per-minute scoring efficiency, representing scoring ability and shooting efficiency. 2) *Rebounding and Physical Attributes*, 14 features such as games played, rebounds, fouls, and playing time averages, highlighting physical competitiveness and playing consistency. 3) *Technical Statistics*, 15 features including assists, steals, blocks, turnovers, and triple-doubles, reflecting player’s playmaking and defensive contributions. 4) *Advanced Efficiency Metrics*, 10 features measuring comprehensive performance like true shooting percentage, usage rate, player efficiency rating, and per-minute rates. 5) *Player Background*, 41 features encoding age and team membership via one-hot encoding. 6) *Seasonal Context*, 22 features encompassing season indicators, playoff status, also one-hot encoded. Player attributes are structured as an 18-dimensional multi-label vector per sample, which includes 10 award-related labels (e.g., All-Star selection, All-NBA teams, MVP nomination), 5 positional one-hot labels (PG, SG, SF, PF, C) and 3 career stage one-hot labels categorizing each player-season into Early Career (first 25%), Prime Career (middle 50%), and Late Career (last 25%) according to the player’s elapsed and remaining career years.

**Implementation Details.** We employ Hamming Loss (HL), Ranking Loss (RL), OneError (OE), Coverage (Cov), Average Precision (AP), and Area Under Curve (AUC) as six metrics to unify the experimental standards. Higher AP and AUC values indicate better performance, while lower HL, RL, OE, and Cov values are preferred. Their evaluation contents are described below: 1) ACC measures the proportion of correctly predicted labels across all samples. 2) RL evaluates the accuracy of the model’s ranking of predicted labels compared to true labels. 3) AP computes the area under the precision-recall curve, indicating the average precision achieved across all recall levels. 4) AUC quantifies the probability that a randomly selected positive instance is ranked higher by the model than a randomly selected negative instance across all possible threshold values. 5) OE evaluates whether the top-ranked label predicted by the model is incorrect. 6) Cov computes the number of labels the model needs to traverse to cover all true labels, reflecting the efficiency of the model’s predicted label range. The neighbor number  $k$  is fixed to 10 for all datasets. Adam optimizer with the initial learning rate of 0.0001 is used for optimization of all datasets. All methods use the same

dataset partition when conducting experiments, while the locations of view missing and label absence are recorded and kept consistent.

Table 4: Detailed information of comparison methods. ✓ represent the method is able to handle the corresponding problem.

Method	Source	Year	Multi-label	Multi-view	Missing-view	Missing-label
iMVWL	IJCAI	2018	✓	✓	✓	✓
DM2L	PR	2021	✓	✗	✗	✓
LVLS	TMM	2022	✓	✓	✗	✗
DICNet	AAAI	2023	✓	✓	✓	✓
DIMC	TNNLS	2023	✓	✓	✓	✓
LMVCAT	AAAI	2023	✓	✓	✓	✓
AIMNet	AAAI	2024	✓	✓	✓	✓
MTD	NeurIPS	2024	✓	✓	✓	✓
SIP	ICML	2024	✓	✓	✓	✓

## C.2 Experiment Result

Table 5: Experimental results of nine methods on the six datasets with 90% PER and 90% LMR. ‘Ave.R’ refers to the mean ranking of the corresponding method across all six metrics.

DATA	METRIC	AIMNet	DICNet	DIMC	DM2L	iMVWL	LMVCAT	LVSL	MTD	SIP	TDLSR
COR	1-HL	0.987 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.976 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.987 <sub>0.000</sub>	0.986 <sub>0.000</sub>	0.987 <sub>0.000</sub>
	1-OE	0.277 <sub>0.007</sub>	0.242 <sub>0.006</sub>	0.239 <sub>0.009</sub>	0.171 <sub>0.010</sub>	0.181 <sub>0.005</sub>	0.229 <sub>0.008</sub>	0.247 <sub>0.006</sub>	0.274 <sub>0.012</sub>	0.289 <sub>0.014</sub>	0.374 <sub>0.010</sub>
	1-Cov	0.605 <sub>0.003</sub>	0.515 <sub>0.009</sub>	0.518 <sub>0.005</sub>	0.481 <sub>0.013</sub>	0.524 <sub>0.004</sub>	0.600 <sub>0.004</sub>	0.608 <sub>0.002</sub>	0.573 <sub>0.005</sub>	0.601 <sub>0.006</sub>	0.692 <sub>0.004</sub>
	1-RL	0.823 <sub>0.001</sub>	0.774 <sub>0.003</sub>	0.772 <sub>0.005</sub>	0.750 <sub>0.005</sub>	0.762 <sub>0.004</sub>	0.817 <sub>0.003</sub>	0.823 <sub>0.001</sub>	0.809 <sub>0.001</sub>	0.821 <sub>0.001</sub>	0.866 <sub>0.002</sub>
	AP	0.240 <sub>0.002</sub>	0.208 <sub>0.001</sub>	0.206 <sub>0.004</sub>	0.181 <sub>0.002</sub>	0.163 <sub>0.004</sub>	0.214 <sub>0.001</sub>	0.228 <sub>0.001</sub>	0.234 <sub>0.005</sub>	0.242 <sub>0.002</sub>	0.323 <sub>0.004</sub>
	AUC	0.826 <sub>0.001</sub>	0.776 <sub>0.004</sub>	0.774 <sub>0.005</sub>	0.754 <sub>0.006</sub>	0.766 <sub>0.004</sub>	0.820 <sub>0.003</sub>	0.827 <sub>0.001</sub>	0.811 <sub>0.002</sub>	0.823 <sub>0.001</sub>	0.869 <sub>0.002</sub>
ESP	Ave	3.167	6.833	7.333	8.333	9	5	3	5.667	4.167	1
	1-HL	0.982 <sub>0.000</sub>	0.982 <sub>0.000</sub>	0.982 <sub>0.000</sub>	0.983 <sub>0.000</sub>	0.969 <sub>0.000</sub>	0.982 <sub>0.000</sub>	0.983 <sub>0.000</sub>	0.982 <sub>0.000</sub>	0.982 <sub>0.000</sub>	0.982 <sub>0.000</sub>
	1-OE	0.310 <sub>0.007</sub>	0.289 <sub>0.012</sub>	0.283 <sub>0.006</sub>	0.210 <sub>0.001</sub>	0.204 <sub>0.009</sub>	0.266 <sub>0.023</sub>	0.265 <sub>0.004</sub>	0.302 <sub>0.008</sub>	0.327 <sub>0.008</sub>	0.385 <sub>0.010</sub>
	1-Cov	0.508 <sub>0.003</sub>	0.464 <sub>0.000</sub>	0.456 <sub>0.004</sub>	0.447 <sub>0.003</sub>	0.421 <sub>0.007</sub>	0.468 <sub>0.003</sub>	0.489 <sub>0.001</sub>	0.492 <sub>0.003</sub>	0.500 <sub>0.004</sub>	0.576 <sub>0.002</sub>
	1-RL	0.792 <sub>0.001</sub>	0.773 <sub>0.001</sub>	0.769 <sub>0.002</sub>	0.758 <sub>0.002</sub>	0.729 <sub>0.004</sub>	0.771 <sub>0.001</sub>	0.783 <sub>0.001</sub>	0.786 <sub>0.000</sub>	0.785 <sub>0.001</sub>	0.825 <sub>0.001</sub>
	AP	0.222 <sub>0.003</sub>	0.210 <sub>0.002</sub>	0.207 <sub>0.003</sub>	0.172 <sub>0.002</sub>	0.155 <sub>0.004</sub>	0.201 <sub>0.006</sub>	0.204 <sub>0.001</sub>	0.219 <sub>0.002</sub>	0.225 <sub>0.002</sub>	0.271 <sub>0.003</sub>
IAP	AUC	0.979 <sub>0.001</sub>	0.777 <sub>0.000</sub>	0.772 <sub>0.002</sub>	0.762 <sub>0.002</sub>	0.733 <sub>0.005</sub>	0.775 <sub>0.002</sub>	0.787 <sub>0.000</sub>	0.790 <sub>0.000</sub>	0.790 <sub>0.002</sub>	0.830 <sub>0.001</sub>
	Ave	3.333	5.333	6.5	7.667	10	6.333	5.167	4	4.167	1.333
	1-HL	0.980 <sub>0.000</sub>	0.980 <sub>0.000</sub>	0.980 <sub>0.000</sub>	0.983 <sub>0.000</sub>	0.966 <sub>0.000</sub>	0.980 <sub>0.000</sub>				
	1-OE	0.342 <sub>0.004</sub>	0.333 <sub>0.009</sub>	0.318 <sub>0.002</sub>	0.247 <sub>0.004</sub>	0.245 <sub>0.011</sub>	0.290 <sub>0.007</sub>	0.294 <sub>0.004</sub>	0.344 <sub>0.007</sub>	0.355 <sub>0.005</sub>	0.397 <sub>0.009</sub>
	1-Cov	0.521 <sub>0.002</sub>	0.472 <sub>0.004</sub>	0.468 <sub>0.002</sub>	0.443 <sub>0.005</sub>	0.438 <sub>0.009</sub>	0.471 <sub>0.005</sub>	0.496 <sub>0.002</sub>	0.510 <sub>0.006</sub>	0.519 <sub>0.003</sub>	0.616 <sub>0.001</sub>
	1-RL	0.818 <sub>0.003</sub>	0.799 <sub>0.004</sub>	0.795 <sub>0.002</sub>	0.780 <sub>0.003</sub>	0.761 <sub>0.005</sub>	0.793 <sub>0.000</sub>	0.807 <sub>0.001</sub>	0.816 <sub>0.003</sub>	0.817 <sub>0.003</sub>	0.860 <sub>0.001</sub>
MIR	AP	0.229 <sub>0.002</sub>	0.222 <sub>0.004</sub>	0.215 <sub>0.002</sub>	0.186 <sub>0.003</sub>	0.167 <sub>0.004</sub>	0.202 <sub>0.003</sub>	0.208 <sub>0.001</sub>	0.232 <sub>0.002</sub>	0.235 <sub>0.002</sub>	0.278 <sub>0.003</sub>
	AUC	0.822 <sub>0.002</sub>	0.801 <sub>0.003</sub>	0.798 <sub>0.002</sub>	0.783 <sub>0.003</sub>	0.766 <sub>0.005</sub>	0.797 <sub>0.001</sub>	0.811 <sub>0.001</sub>	0.819 <sub>0.002</sub>	0.820 <sub>0.002</sub>	0.861 <sub>0.001</sub>
	Ave	2.5	4.833	5.833	7.667	10	6.667	5	3.167	2.333	1
	1-HL	0.875 <sub>0.001</sub>	0.879 <sub>0.001</sub>	0.877 <sub>0.002</sub>	0.876 <sub>0.000</sub>	0.827 <sub>0.004</sub>	0.865 <sub>0.004</sub>	0.874 <sub>0.000</sub>	0.880 <sub>0.001</sub>	0.875 <sub>0.002</sub>	0.885 <sub>0.002</sub>
	1-OE	0.506 <sub>0.023</sub>	0.533 <sub>0.005</sub>	0.511 <sub>0.005</sub>	0.439 <sub>0.005</sub>	0.406 <sub>0.023</sub>	0.470 <sub>0.020</sub>	0.485 <sub>0.006</sub>	0.535 <sub>0.008</sub>	0.540 <sub>0.009</sub>	0.612 <sub>0.006</sub>
	1-Cov	0.598 <sub>0.006</sub>	0.549 <sub>0.003</sub>	0.589 <sub>0.001</sub>	0.572 <sub>0.003</sub>	0.530 <sub>0.012</sub>	0.581 <sub>0.002</sub>	0.584 <sub>0.001</sub>	0.606 <sub>0.006</sub>	0.604 <sub>0.006</sub>	0.645 <sub>0.005</sub>
OBJ	1-RL	0.827 <sub>0.005</sub>	0.828 <sub>0.001</sub>	0.823 <sub>0.001</sub>	0.809 <sub>0.002</sub>	0.765 <sub>0.011</sub>	0.817 <sub>0.003</sub>	0.819 <sub>0.001</sub>	0.834 <sub>0.003</sub>	0.830 <sub>0.002</sub>	0.861 <sub>0.003</sub>
	AP	0.494 <sub>0.017</sub>	0.512 <sub>0.001</sub>	0.501 <sub>0.002</sub>	0.467 <sub>0.004</sub>	0.415 <sub>0.009</sub>	0.485 <sub>0.010</sub>	0.482 <sub>0.002</sub>	0.519 <sub>0.006</sub>	0.519 <sub>0.002</sub>	0.575 <sub>0.002</sub>
	AUC	0.820 <sub>0.003</sub>	0.823 <sub>0.001</sub>	0.818 <sub>0.001</sub>	0.805 <sub>0.002</sub>	0.769 <sub>0.007</sub>	0.808 <sub>0.004</sub>	0.816 <sub>0.001</sub>	0.827 <sub>0.002</sub>	0.823 <sub>0.001</sub>	0.852 <sub>0.000</sub>
	Ave	5.333	4	5.333	8.333	10	8	7.333	2.167	3.333	1
	1-HL	0.937 <sub>0.001</sub>	0.938 <sub>0.001</sub>	0.938 <sub>0.000</sub>	0.935 <sub>0.000</sub>	0.882 <sub>0.005</sub>	0.927 <sub>0.002</sub>	0.934 <sub>0.000</sub>	0.938 <sub>0.000</sub>	0.937 <sub>0.001</sub>	0.946 <sub>0.001</sub>
	1-OE	0.468 <sub>0.005</sub>	0.453 <sub>0.009</sub>	0.439 <sub>0.005</sub>	0.415 <sub>0.009</sub>	0.335 <sub>0.036</sub>	0.405 <sub>0.019</sub>	0.364 <sub>0.004</sub>	0.474 <sub>0.004</sub>	0.485 <sub>0.013</sub>	0.603 <sub>0.020</sub>
PAS	1-Cov	0.727 <sub>0.009</sub>	0.720 <sub>0.000</sub>	0.709 <sub>0.010</sub>	0.682 <sub>0.002</sub>	0.657 <sub>0.023</sub>	0.705 <sub>0.017</sub>	0.712 <sub>0.002</sub>	0.740 <sub>0.007</sub>	0.727 <sub>0.006</sub>	0.795 <sub>0.003</sub>
	1-RL	0.829 <sub>0.003</sub>	0.823 <sub>0.006</sub>	0.814 <sub>0.008</sub>	0.800 <sub>0.000</sub>	0.768 <sub>0.012</sub>	0.806 <sub>0.011</sub>	0.811 <sub>0.001</sub>	0.835 <sub>0.003</sub>	0.828 <sub>0.004</sub>	0.881 <sub>0.002</sub>
	AP	0.506 <sub>0.010</sub>	0.502 <sub>0.006</sub>	0.489 <sub>0.010</sub>	0.470 <sub>0.002</sub>	0.394 <sub>0.026</sub>	0.476 <sub>0.009</sub>	0.446 <sub>0.001</sub>	0.519 <sub>0.008</sub>	0.522 <sub>0.009</sub>	0.624 <sub>0.011</sub>
	AUC	0.842 <sub>0.003</sub>	0.836 <sub>0.005</sub>	0.828 <sub>0.008</sub>	0.814 <sub>0.000</sub>	0.784 <sub>0.012</sub>	0.821 <sub>0.011</sub>	0.827 <sub>0.011</sub>	0.848 <sub>0.003</sub>	0.840 <sub>0.004</sub>	0.891 <sub>0.001</sub>
	Ave	3.667	4.667	5.833	8.167	10	8	7.667	2.333	3.667	1
	1-HL	0.923 <sub>0.003</sub>	0.927 <sub>0.001</sub>	0.927 <sub>0.001</sub>	0.926 <sub>0.000</sub>	0.871 <sub>0.003</sub>	0.921 <sub>0.003</sub>	0.926 <sub>0.000</sub>	0.926 <sub>0.001</sub>	0.923 <sub>0.003</sub>	0.927 <sub>0.000</sub>
LMVCAT	1-OE	0.382 <sub>0.023</sub>	0.402 <sub>0.002</sub>	0.403 <sub>0.002</sub>	0.380 <sub>0.007</sub>	0.306 <sub>0.037</sub>	0.376 <sub>0.021</sub>	0.415 <sub>0.001</sub>	0.395 <sub>0.011</sub>	0.389 <sub>0.011</sub>	0.415 <sub>0.016</sub>
	1-Cov	0.658 <sub>0.002</sub>	0.636 <sub>0.013</sub>	0.626 <sub>0.013</sub>	0.636 <sub>0.009</sub>	0.589 <sub>0.014</sub>	0.630 <sub>0.018</sub>	0.654 <sub>0.005</sub>	0.674 <sub>0.005</sub>	0.668 <sub>0.018</sub>	0.753 <sub>0.006</sub>
	1-RL	0.727 <sub>0.006</sub>	0.710 <sub>0.011</sub>	0.703 <sub>0.010</sub>	0.698 <sub>0.008</sub>	0.658 <sub>0.014</sub>	0.693 <sub>0.016</sub>	0.726 <sub>0.004</sub>	0.740 <sub>0.003</sub>	0.729 <sub>0.013</sub>	0.808 <sub>0.007</sub>
	AP	0.440 <sub>0.009</sub>	0.440 <sub>0.005</sub>	0.434 <sub>0.004</sub>	0.421 <sub>0.006</sub>	0.368 <sub>0.021</sub>	0.430 <sub>0.005</sub>	0.444 <sub>0.002</sub>	0.454 <sub>0.007</sub>	0.449 <sub>0.006</sub>	0.504 <sub>0.009</sub>
	AUC	0.754 <sub>0.003</sub>	0.737 <sub>0.011</sub>	0.727 <sub>0.012</sub>	0.733 <sub>0.007</sub>	0.690 <sub>0.013</sub>	0.727 <sub>0.010</sub>	0.754 <sub>0.003</sub>	0.766 <sub>0.006</sub>	0.761 <sub>0.013</sub>	0.831 <sub>0.006</sub>
	Ave	5.333	5.167	6	7.167	10	8.5	4.167	3.167	4.167	1

**Comparison Experiment.** To validate that our method can adapt to varying degrees of data missing, we conduct comparison experiments with PER and LMR ranging from {30%, 50%, 70%, 90%}. Table 5 presents the comparison results and algorithm rankings when PER and LMR are fixed at 90%. Fig. 8, 9 and 10 illustrate the distributional trends of all metrics as PER increases from 30% to 70%, with LER fixed at 30%, 50%, and 70%, respectively. It can be observed that our method

outperforms the other nine methods in almost all cases, which demonstrates the robustness of our TDLSR in handling incomplete multi-view multi-label problems. Moreover, the effectiveness of our method is especially evident when dealing with high levels of data unavailability. For instance, as shown in Fig. 8, on Corel 5k, the performance margin of our method over the second-best gradually widens. Additionally, as reported in Table 5, our method achieves more than a 10% improvement in the most representative metric AP on both OBJECT and Pascal07.

**Parameter Determination and Convergence.** The parameters  $\lambda_1$  and  $\lambda_2$  are used to balance the effects of  $\mathcal{L}_{IB}$  and  $\mathcal{L}_{le}$ . Two parameters are selected from the range of  $\{0.01, 0.1, 1, 10, 100, 1000\}$  and the joint influence are presented in the heatmap as shown in Fig. 6. From the result, the performance of TDLSR exhibits variability under different parameter setting. Besides, the optimal results are typically achieved when  $\lambda_1$  falls within the range of  $(0.01, 1)$ . Overall, the parameter sensitivity is relatively low on certain datasets, such as on ESPgame and OBJECT. We also present the simultaneous evolution of training loss and performance metric on the validation set throughout the training process in Fig. 7. The results show that our TDLSR demonstrates great convergence and gradually approaches the optimal network parameters.

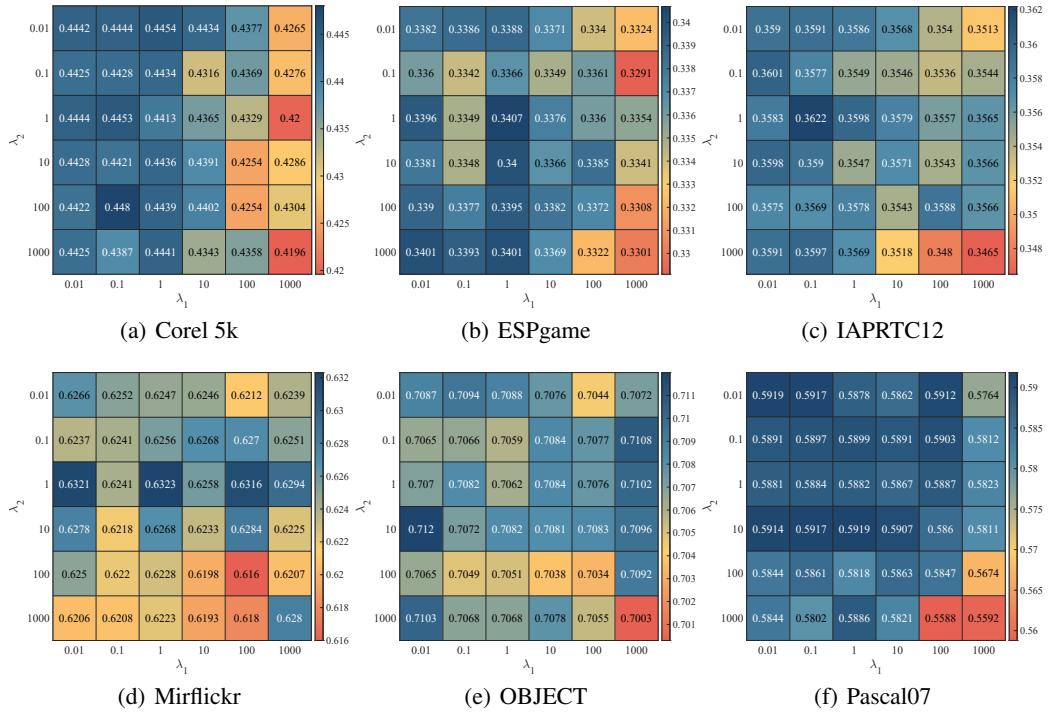


Figure 6: Parameter analysis of the trade-off parameters  $\lambda_1$  and  $\lambda_2$ .

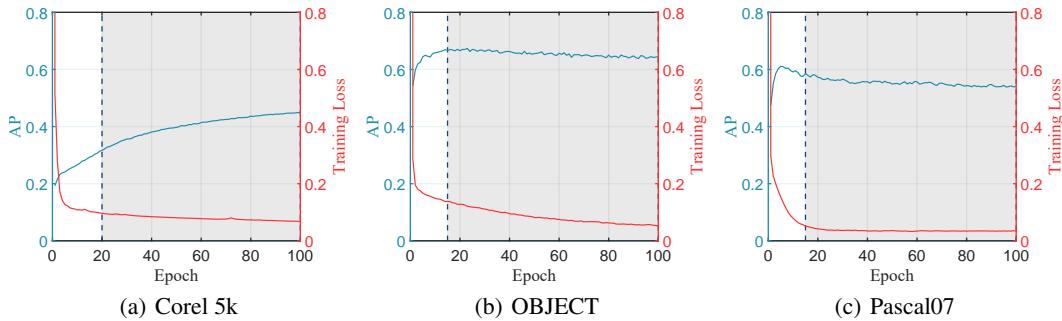


Figure 7: The convergence behavior of our TDLSR during training.

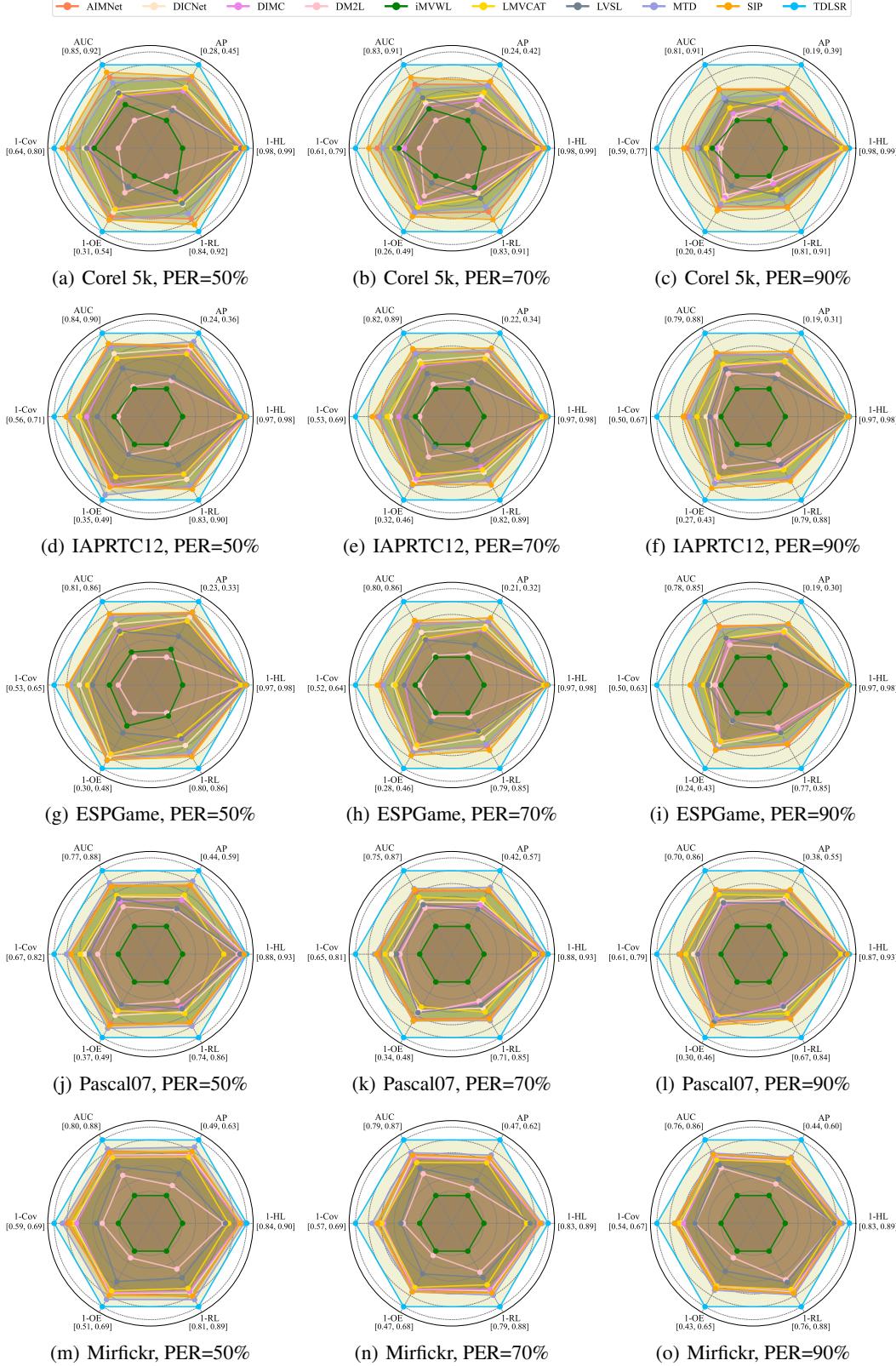


Figure 8: Experimental results of ten methods on five datasets with PER varying from 50% to 90% while LMR= 50%.

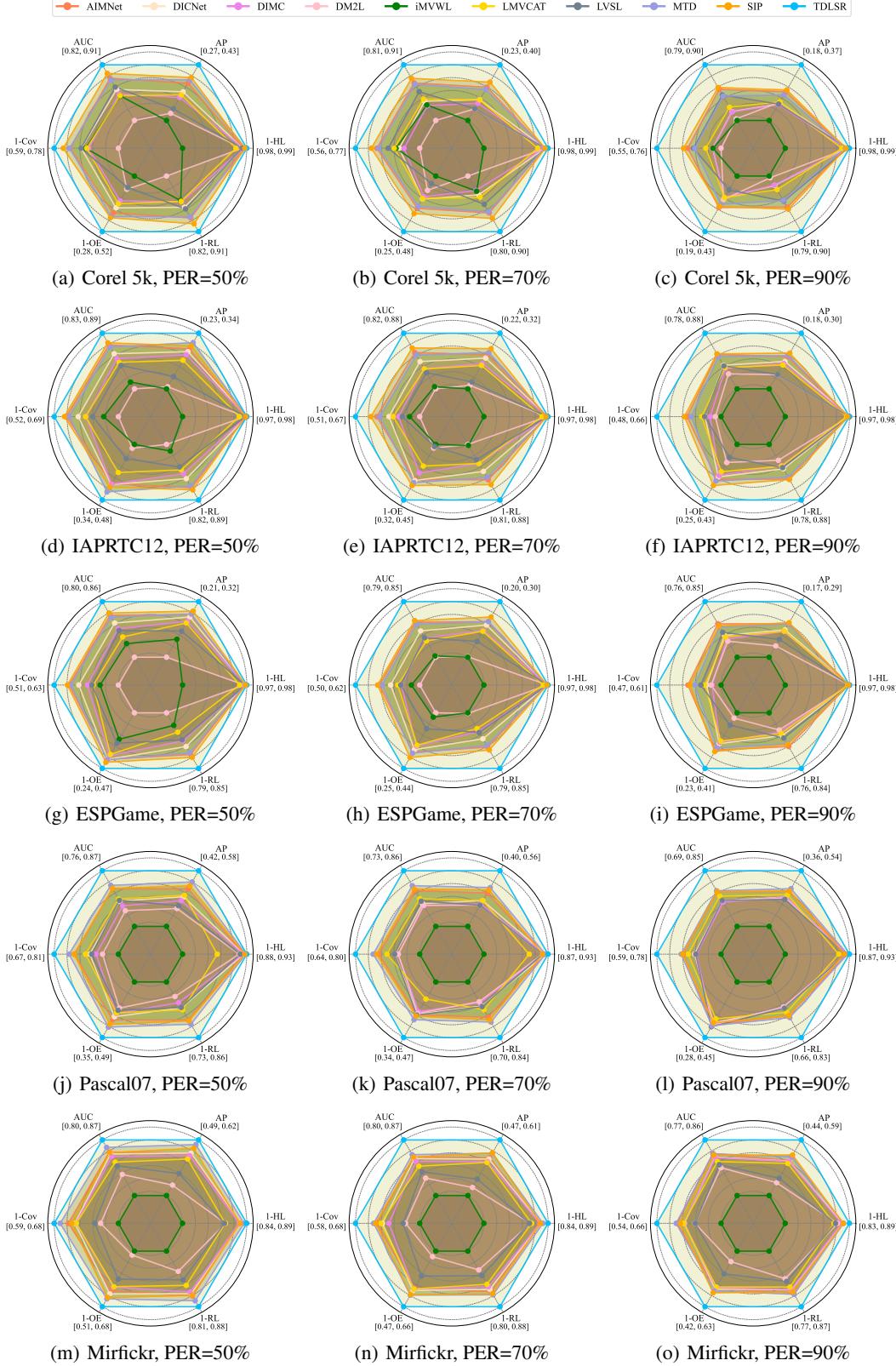


Figure 9: Experimental results of ten methods on five datasets with PER varying from 50% to 90% while LMR= 70%.

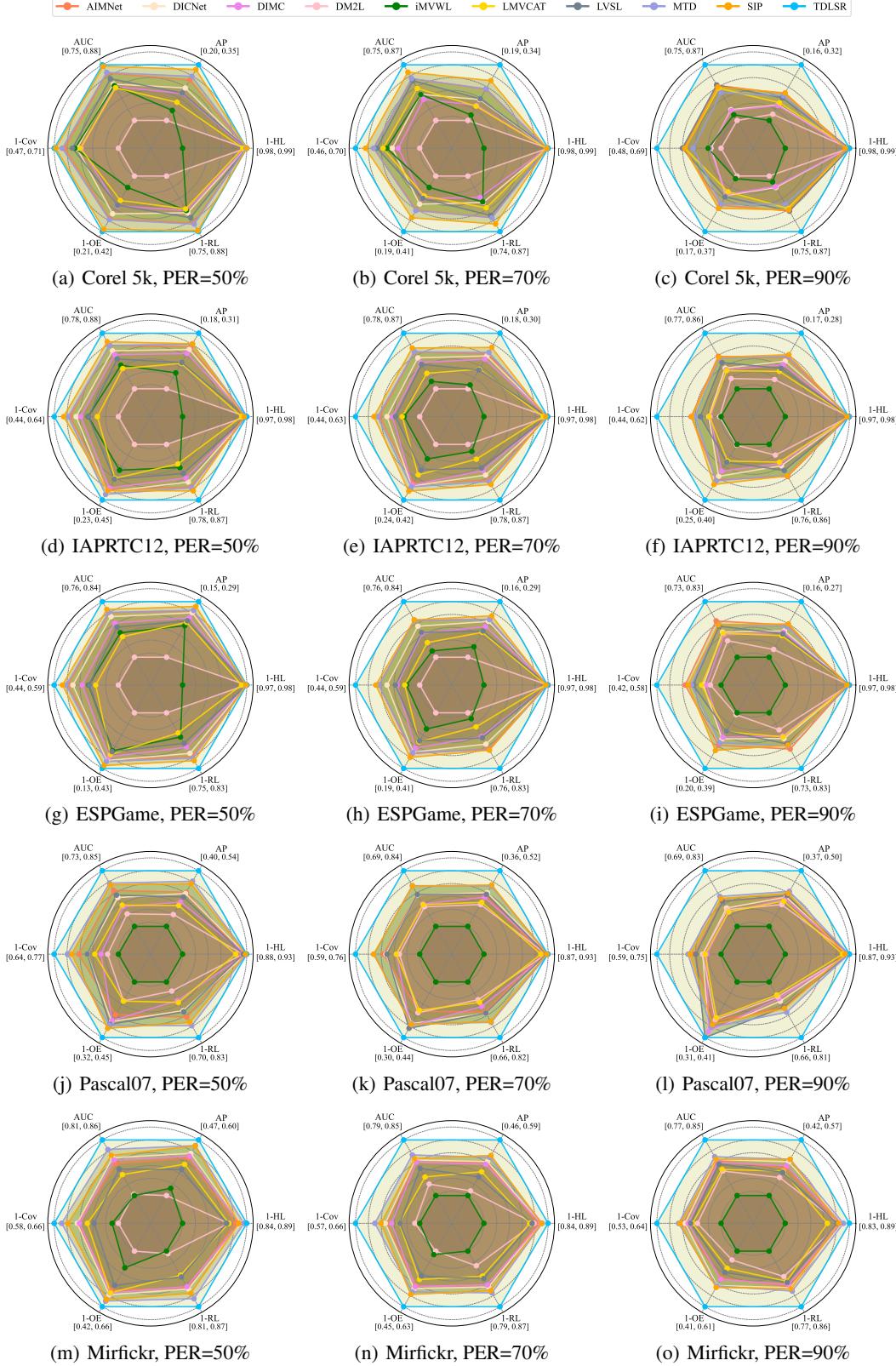


Figure 10: Experimental results of ten methods on five datasets with PER varying from 50% to 90% while LMR= 90%.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly describe the algorithm implementation, theoretical contributions, and comprehensive experiments presented in this paper, consistent with the subsequent Sections 2 and 3.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Sections 2 and 5 mention that the model needs to fully separate feature semantics in order to achieve optimal generalization performance, and the prior semantics among multiple labels also need to be leveraged, respectively, which are key challenges in model optimization and design.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provide the full set of assumptions and a complete proof needed for each theoretical result in Section 2 and Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The network architecture, datasets, model parameters, and training approaches involved in the experimental implementation are described in detail in both the main text and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: **[No]**

Justification: We commit to sharing our code upon the acceptance of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: **[Yes]**

Justification: Experimental settings, including data splits, hyper parameters, optimizer and learning rate are discussed in both the main text and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: **[Yes]**

Justification: Fig 3 and Table 1 present the results as the mean and standard deviation over 10 random trials, reflecting the statistical variability.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Section 3 mentiones that our model is implemented by PyTorch on one NVIDIA GeForce RTX 4090 GPU of 24GB memory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: The research focuses on algorithm design and theoretical analysis. It does not involve human subjects or ethically sensitive applications. We believe it complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper properly cites the sources for existing assets, such as baseline methods and data sources. However, the specific licenses and terms of use for these assets are not explicitly stated in the main text or Appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.