View Confidence Perception-Driven Incremental Prediction for Incomplete Multi-view Multi-label Learning

Pingzhu Liu[1,†], Zunnan Xu[2,†], Chunming He[3], Zhirui Fang[2], Zitong Yu[4], Xiu Li[1], Chengliang Liu[5,*]
[1]Tsinghua University, China [2]Nanyang Technological University, Singapore
[3]Duke University, USA [4]Great Bay University, China [5]University of Macau, China
kevin_ailover@163.com, zunnanxu.work@gmail.com,
chunming.he@duke.edu, zfang009@e.ntu.edu.sg,
yuzitong@gbu.edu.cn, li.xiu@sz.tsinghua.edu.cn, liucl1996@163.com

Abstract

Incomplete Multi-view Multi-label Learning refers to a classification task where missing views and incomplete label assignments coexist. Existing methods often directly fuse heterogeneous view features, neglecting inherent imbalance where dominant views tend to overshadow non-dominant ones, resulting in unsatisfactory performance. To address this, we propose the View Confidence Perception-Driven Incremental Prediction framework, which incorporates an adaptive structural refinement strategy to balance different views via confidence-based branch expansion, thereby enabling the full exploitation of view-specific information. Specifically, we propose a View-Confidence Aware strategy, which introduces a novel metric to evaluate the predictive strength of each view-specific branch using supervision signals from the label space. By leveraging these quality assessments, we dynamically allocate optimal classifier ensembles and employ residual optimisation to improve the discriminative capability of non-dominant views. Additionally, to eliminate information redundancy and purify features, we propose a Mixture-of-Experts-based Decouple-Fusion mechanism to extract refined representations by disentangling consistent and specific features under orthogonality constraints for more precise multi-label prediction. Moreover, to simulate real-world data corruption, we introduce the Stochastic Cross-sample Fragment Swapping strategy, which interchanges feature fragments across samples to facilitate the modelling of robust global representations for enhanced generalisation. Extensive experiments across diverse benchmarks and varying missing-rate scenarios confirm that our method consistently surpasses state-of-the-art methods in multi-label classification.

1 Introduction

Multi-view Multi-label learning (MvMlL) [1–3] aims to integrate heterogeneous information from different views and achieve accurate classification across multiple labels. Learning from multiple views or modalities supports applications such as visual understanding [4, 5], video analysis [6], and text mining [7]. However, constrained by acquisition and annotation non-idealities, multi-view data frequently suffers from view and label incompleteness [8–10]. Such stochastic missingness poses a bottleneck for robust modelling, leading us to turn to the Incomplete MvMlL paradigm [11–14].

*Corresponding author. †Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Previous research [1, 15–18] has extensively explored multi-view and multimodal feature fusion. In incomplete multi-view learning, two common paradigms are latent feature synchronisation and classifier-level ensemble. (1) Latent feature synchronisation. For example, MSLPP [16] proposed an innovative model that integrates multi-source information, preserves structural information through graph embedding, and employs a masking strategy to handle missing data. Furthermore, VIQA [19] proposed a view quality module and a sample-level dual constraint, achieving dynamic weight assignment for different views to maintain feature consistency. HCLS_CGL [20] proposed a confidence graph-based consensus graph learning method to learn the cross-view consensus graph structure in multi-view learning. (2) Classifier-level ensemble. For instance, AIMNet [15] formulated a framework that imputes latent features via cross-view joint attention while capturing label correlations through a guided graph network. In addition, MTD [21] proposed a masked two-channel decoupling framework that separates shared and view-proprietary representations in order to effectively solve multi-view learning tasks. RecFormer [22] proposed a deep network framework driven by information recovery. By integrating Transformer-style cross-view autoencoders with cyclic graph constraints, it effectively captures multi-view information. While existing paradigms have achieved remarkable progress, they still face inherent limitations in multi-view learning tasks. Specifically, methods based on latent feature synchronisation often prioritise global consistency, which inadvertently struggles to deeply mine discriminative information from non-dominant views. Similarly, classifier-level ensemble methods are limited by their "one-step decision" mode, which makes it difficult to capture fine-grained semantic details and provide targeted compensation for prediction errors. In contrast, incremental prediction [23–25] offers a new path for strengthening the representation of non-dominant views and accurately repairing prediction biases through progressive residual optimisation. This mechanism enables the gradual extraction of deep semantic details from features, ultimately leading to more robust multi-label predictions. While the IMvMIL task demands refined features to mitigate the impact of missing views, existing models [11, 12, 21] often neglect to decouple consistent and specific features under orthogonality constraints. This omission hinders the elimination of redundancy and the purification of latent representations, which are essential for handling incomplete data.

To address the aforementioned challenges, we propose the View Confidence Perception-Driven Incremental Prediction (VCPIP) framework, which integrates an adaptive structural refinement scheme to balance dominant and non-dominant views through confidence-based branch expansion to fully exploit view-specific information. To identify non-dominant views, we propose a View-Confidence Aware (VCA) strategy. This strategy introduces a novel metric to evaluate the predictive strength of each view-specific branch using supervision signals from the label space, enabling the adaptive identification of non-dominant views. By leveraging auxiliary classifiers via residual optimisation, we dynamically enhance these views to rectify imbalances at the view level. To eliminate information redundancy and purify latent representations, we propose a Mixture-of-Experts-based Decouple-Fusion (MOEDF) mechanism to disentangle encoder-extracted features into consistent and specific features, enforcing an explicit orthogonality constraint. Moreover, we present a Stochastic Cross-sample Fragment Swapping (SCFS) strategy that utilises real fragments from different samples to introduce semantic interference, thereby preventing "model laziness" by forcing the network to infer context from complex cross-sample mixtures rather than simply bypassing masked positions. Our contributions are summarised as follows:

- We propose VCPIP, a robust IMvMIL framework that rectifies view-level imbalances. By introducing a novel VCA strategy to identify non-dominant views, VCPIP dynamically enhances non-dominant views through auxiliary classifiers and residual optimisation. Extensive benchmarks demonstrate that our method consistently outperforms competitive methods.
- To eliminate information redundancy and further purify the refined feature representations, we develop a novel MOEDF mechanism to disentangle view features into consistent and specific components by enforcing an explicit orthogonality constraint.
- We present the SCFS strategy to leverage semantic interference between different samples. By swapping cross-sample fragments, it forces context reconstruction from complex mixtures, significantly enhancing generalisation.

2 Problem Definition and Motivation

Problem Definition. In the IMvMlL task, we consider a dataset with n samples, m views, and c categories. We are given the v-th view's feature matrix $\mathbf{X}^{(v)} \in \mathbb{R}^{n \times d_v}$ $(v = 1, \ldots, m)$ and a

label matrix $\mathbf{Y} \in \{0,1\}^{n\times c}$, where d_v denotes the feature dimension of the v-th view. To handle data incompleteness, we introduce two binary index matrices: $\mathbf{W} \in \{0,1\}^{n\times m}$, where $\mathbf{W}_{i,j} = 1$ indicates that the j-th view of the i-th sample is available, and $\mathbf{G} \in \{0,1\}^{n\times c}$, where $\mathbf{G}_{i,j} = 1$ denotes the j-th label of the i-th sample is known. Our goal is to train a neural network for multi-label reasoning using these incomplete multi-view features and weak labels.

2.1 Motivation

We further elaborate on the core innovation of our framework, which leverages incremental prediction to address the issue of multi-view imbalance. By progressively refining predictions through a step-by-step decision-making process, this method enhances the discriminative power of non-dominant views.

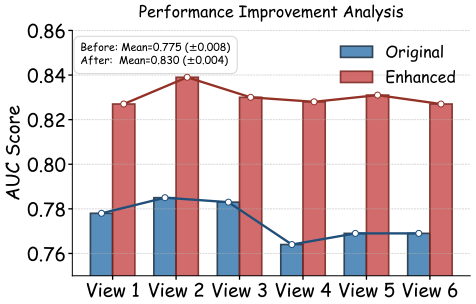


Figure 1: Performance comparison across six views before and after incremental prediction enhancement on the Corel5k dataset.

View-Dominance and Incremental Compensation. Current multi-view multi-label learning methods primarily rely on latent feature synchronization or classifier-level ensembles for view integration. However, the former methods are typically constrained by the assumption of equal view contributions. This simplistic premise frequently leads to a significant dominance effect where the model overfits high-performance views at the cost of extracting discriminative features from non-dominant ones. Furthermore, classifier-level ensembles are constrained by a single-step decision-making process. This limitation prevents them from effectively capturing fine-grained semantics or compensating for prediction biases in non-dominant views. Given our observation that the contributions of multi-view features are inherently unequal, and inspired by [10, 15, 26], we introduce the following assumption:

Assumption 1 *Given a set of multi-view representations $\mathcal{H} = \{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$ and its corresponding label space y, there exists a significant disparity in predictive strength among individual views. Formally, there exists a permutation σ of $\{1, \ldots, m\}$ such that the different view representations z can be ranked by their mutual information:*

$$I(z^{\sigma(1)}; y) \gg I(z^{\sigma(2)}; y) \gg \cdots \gg I(z^{\sigma(m)}; y).$$

Assumption 1 highlights a fundamental challenge in multi-view learning: the inherent information imbalance often leads the model to over-rely on dominant views while underutilising the latent potential of others. Given this natural imbalance, if the non-dominant views can be progressively reinforced through a streamlined mechanism, the discriminative disparities across views could be effectively mitigated. This motivates our second assumption, which posits that the discriminative capacity of non-dominant views is not static but can be systematically augmented through an iterative, residual-driven optimisation process to bridge the gaps in view-level discriminability.

Assumption 2 *The discriminative capacity of view representation $z^{(v)}$ can be progressively enhanced via sequential classifiers. Let $A_k(z^{(v)}) = \sum_{i=1}^{k} a_i(z^{(v)})$ be the ensemble mapping at stage k, where a_1 is the primary classifier and $\{a_i\}_{i=2}^{k}$ are residual-optimized auxiliary classifiers. The captured task-relevant information satisfies:*

$$I(A_k(z^{(v)}); y) \geq I(A_{k-1}(z^{(v)}); y), \quad \forall k > 1.$$

This implies that dynamic branch expansion and multi-stage residual optimization monotonically strengthen non-dominant views, narrowing view-level performance disparities.

Following Assumption 1 and 2, we propose a novel view confidence perception-driven incremental prediction framework that utilizes a custom view-evaluation metric to identify non-dominant views and trigger auxiliary classifiers with residual optimization for enhanced discrimination. As illustrated

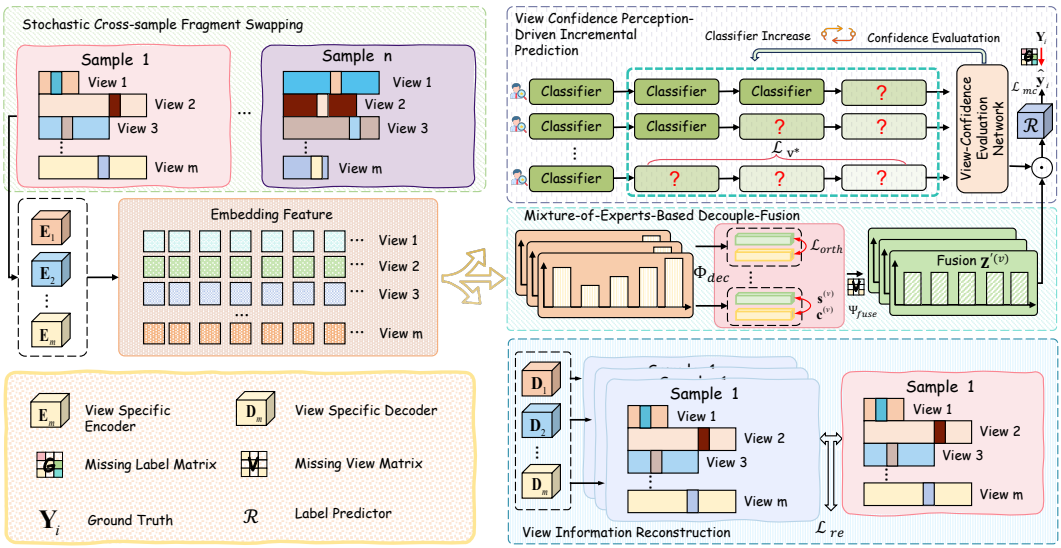


Figure 2: Overview of VCPIP. The framework consists of three core components: a Stochastic Cross-sample Fragment Swapping strategy for robust data augmentation, the view confidence perception-driven incremental prediction to enhance the discriminative power of non-dominant views, and a Mixture-of-Experts-Based Decouple-Fusion mechanism to reduce information redundancy.

in Fig. 1, we validate the effectiveness of the incremental prediction mechanism on the Corel5K dataset. The experimental results demonstrate that by introducing auxiliary classifiers for non-dominant views, this mechanism significantly enhances the discriminative power of each individual view and effectively narrows the performance gaps between them. This not only facilitates the deep exploration of complementary information across views but also enables a more balanced and efficient integration of multi-view features.

3 Method

In this section, we present the architecture of the VCPIP framework in detail. As illustrated in Fig. 2, our framework comprises three core components: the view confidence perception-driven incremental prediction architecture, the mixture-of-experts-based decouple-fusion mechanism, and the stochastic cross-sample fragment swapping strategy. In the following section, we provide a comprehensive description of the proposed component, focusing on its design philosophy and internal architecture.

3.1 View Confidence Perception-Driven Incremental Prediction

To address the view imbalance stemming from the tendency of dominant views to overshadow non-dominant ones in multi-view learning, we propose a view confidence perception-driven incremental prediction architecture. This approach identifies non-dominant views and enhances their capability by dynamically adding classifiers (we implement the classifiers using a linear layer and a ReLU operation). Specifically, a progressive residual loss is designed to guide these newly added classifiers to focus on residual information that previous classifiers failed to fit, thereby achieving targeted improvements in non-dominant-view classification.

View-Confidence Discriminability Assessment. To evaluate the quality of individual views, we propose the VCA strategy, which defines a View Confidence Score (VCS). This score enables the model to assess the confidence level of different views and effectively identify non-dominant ones. Computed on the validation set $\mathcal{D}_{\text{val}}$, the VCS assesses a view's reliability by balancing its predictive power on both positive and negative labels. For view v, let $p_{i,j}^{(v)}$ be the predicted probability for the j-th class of sample i. The VCS $\mathcal{C}_v$ is defined as:

$$\mathcal{C}^{(v)} = \frac{1}{|\Omega^+|} \sum_{(i,j)\in\Omega^+} p_{i,j}^{(v)} + \frac{1}{|\Omega^-|} \sum_{(i,j)\in\Omega^-} (1 - p_{i,j}^{(v)}), \tag{1}$$

Algorithm 1 Training process of **VCPIP**

Input: Incomplete data $\{\mathbf{X}^{(v)}\}_{v=1}^{m}$, missing-view index $\mathbf{W}$, weak labels $\mathbf{Y}$, missing-label index $\mathbf{G}$.
Output: Prediction results $\hat{\mathbf{y}}_{i,j}$.

Initialize model parameters Θ and hyperparameters $\{\alpha, \beta, \gamma\}$.
$t \leftarrow 0$.
while $t <$ training epochs e **do**
    Generate random feature mask matrices $\{\mathbf{M}^{(v)}\}_{v=1}^{m}$.
    Obtain augmented inputs $\{\mathbf{X}'^{(v)}\}_{v=1}^{m}$ via cross-sample fragment swapping by Eq. (8).
    Disentangle encoded features $\mathbf{Z}^{(v)}$ into consistent representations $\mathbf{c}^{(v)}$ and specific features $\mathbf{s}^{(v)}$.
    Apply vector orthogonalization to $\{\mathbf{c}^{(v)}, \mathbf{s}^{(v)}\}$.
    Generate finalised representation $\mathbf{Z}'^{(v)}$ via fusion $\{\mathbf{c}^{(v)}, \mathbf{s}^{(v)}\}$ from the fusion components.
    Calculate View Confidence Scores to identify non-dominant views.
    if performance ratio $\rho <$ threshold τ **then**
        Trigger auxiliary classifiers and execute residual optimization by Eq. (2) and Eq. (3).
    end if
    Compute view loss $\mathcal{L}_{mv}$ by Eq. (5), orthogonality restriction loss $\mathcal{L}_{orth}$ by Eq. (6) and reconstruction loss $\mathcal{L}_{re}$ by Eq. (7).
    Obtain predictions $\hat{\mathbf{y}}_{i,j}$ by Eq. (10) and compute classification loss $\mathcal{L}_{mc}$ by Eq. (11)
    Compute total loss $\mathcal{L}_{all}$ by Eq. (12).
    Update all network parameters Θ using SGD optimizer.
    $t \leftarrow t + 1$.
end while

where $\Omega^{+} = \{(i,j) \mid \mathbf{Y}_{i,j} = 1\}$ and $\Omega^{-} = \{(i,j) \mid \mathbf{Y}_{i,j} = 0\}$ denote the sets of positive and negative label indices in $\mathcal{D}_{\text{val}}$, respectively. This formulation ensures that a view is deemed proficient and assigned a higher score only when it yields high scores for positive samples and low scores for negative ones. By integrating these criteria, the VCA provides a balanced and robust strategy for assessing a view's discriminative power.

Non-Dominant View Determination And Incremental Triggering. After each training epoch, we compute the VCS set $\{\mathcal{C}^{(v)}\}_{v=1}^{m}$ to evaluate the performance of different views. To estimate the imbalance among these views, we define a relative performance ratio ρ:

$$\rho = \frac{\min_v \mathcal{C}^{(v)}}{\max_v \mathcal{C}^{(v)}}. \tag{2}$$

An incremental prediction mechanism is triggered if $\rho < \tau$ (where τ is set to 0.99), indicating a significant gap between two views. We identify the non-dominant view v^* as:

$$v^* = \arg\min_v \mathcal{C}^{(v)}. \tag{3}$$

Specifically, a new classifier is appended to view v^*, and its parameters are dynamically registered to the optimiser. This ratio-based mechanism ensures the refinement process is adaptive to the overall learning progress, eliminating the influence of varying confidence scales across training stages.

Progressive Classifier Boosting. Inspired by gradient boosting [26, 27], we design an objective to guide each incremental classifier in fitting the residual errors left by preceding stages. For the identified non-dominant view v^*, suppose it is currently at the t-th boosting stage. Let $\boldsymbol{p}_{i,k}^{(v^*)}$ denote the output probability vector of the k-th classifier for sample i. We define the residual target $\hat{\boldsymbol{y}}_{i,t}^{(v^*)}$ for the newly added classifier as:

$$\hat{\boldsymbol{y}}_{i,t}^{(v^*)} = \boldsymbol{y}_i - \sum_{k=1}^{t-1} \boldsymbol{y}_i \odot \boldsymbol{p}_{i,k}^{(v^*)}, \tag{4}$$

where $\boldsymbol{y}_i \in \{0,1\}^c$ is the ground-truth label vector for the i-th sample and $\odot$ denotes the element-wise product. This formulation forces the t-th classifier to focus on the positive signals that the existing ensemble has failed to capture. The total training objective for the view v^* at stage t is formulated as:

$$\mathcal{L}_{v^*} = (\ell(\boldsymbol{p}_{i,t}^{(v^*)}, \hat{\boldsymbol{y}}_{i,t}^{(v^*)}) + \sum_{s=1}^{t} \ell(\sum_{k=1}^{s} \boldsymbol{p}_{i,k}^{(v^*)}, \boldsymbol{y}_i)) \odot \mathbf{G}_i, \tag{5}$$

where $\ell(\cdot)$ denotes the binary cross-entropy loss [28], and $\mathbf{G}_i \in \{0,1\}^c$ is introduced to mask the loss of unknown labels for the i-th sample. Distinct from solely constraining the current and preceding states, the cumulative summation term introduces a dense multi-layer supervision. By evaluating all intermediate ensemble states from the initial stage up to t, it effectively prevents error accumulation and semantic forgetting during the progressive boosting process. For views that have not triggered the incremental process, the loss defaults to the base objective $\mathcal{L}_v = \ell(\boldsymbol{p}_{i,1}^{(v)}, \boldsymbol{y}_i) \odot \mathbf{G}_i$. The final multi-view objective is the average over all m views: $\mathcal{L}_{mv} = \frac{1}{m}\sum_{v=1}^{m}\mathcal{L}_v$. By optimizing this combined loss, the non-dominant view's discriminative ability is systematically elevated, thereby mitigating the view imbalance problem.

3.2 Mixture-of-Experts-Based Decouple-Fusion Mechanism

Most existing methods [10, 21] project multi-view information into consistent and specific features, and the absence of explicit geometric constraints beyond separate branches often triggers feature co-adaptation. This blurred granularity poses a fundamental barrier to principled feature decoupling. To address these issues, we propose a new MOEDF mechanism which leverages the diversity of experts to project encoded features $\left\{\mathbf{E}_v : \mathbf{X}'^{(v)} \to \mathbf{Z}^{(v)}\right\}_{v=1}^{m}$ into specific features $\left\{\mathbf{s}^{(v)}\right\}_{v=1}^{m} \in \mathbb{R}^{n\times d_e}$ and consistent features $\left\{\mathbf{c}^{(v)}\right\}_{v=1}^{m} \in \mathbb{R}^{n\times d_e}$ via the decoupling component $\Phi_{dec} : \mathbf{Z}^{(v)} \to \{\mathbf{s}^{(v)}, \mathbf{c}^{(v)}\}$, further enforced by orthogonal constraints. Here, d_e signifies the dimensionality of the latent embedding space where the decoupled features reside. It is noteworthy that $\mathbf{E}_v$ denotes the feature encoder for view v, which produces the latent feature $\mathbf{Z}^{(v)} \in \mathbb{R}^{n\times d_e}$ from the augmented input $\mathbf{X}'^{(v)}$ (See Section 3.3). Then, the orthogonality restriction is formulated via the weighted mean square error loss [29] as follows:

$$\mathcal{L}_{orth} = \frac{1}{n}\sum_{i=1}^{n}\sum_{v=1}^{m}\left(\frac{s_i^{(v)}}{\|s_i^{(v)}\|_2}\cdot\frac{c_i^{(v)}}{\|c_i^{(v)}\|_2}\right)^2\mathbf{W}_{i,v}. \tag{6}$$

Subsequently, the fusion component Ψ_{fuse} adaptively aggregates the specific $\mathbf{s}^{(v)}$ and consistent features $\mathbf{c}^{(v)}$ to generate the integrated representation $\mathbf{Z}'^{(v)} = \Psi_{fuse}(\mathbf{s}^{(v)}, \mathbf{c}^{(v)})$. The resulting integrated features $\mathbf{Z}'^{(v)}$ are then fed into the decoder $\left\{\mathbf{D}_v : \mathbf{Z}'^{(v)} \to \bar{\mathbf{X}}^{(v)}\right\}_{v=1}^{m}$ for reconstruction, aiming to reinforce the representational capacity of the features, where $\mathbf{D}_v$ is the decoder corresponding to v-th view. $\bar{\mathbf{X}}^{(v)}$ is the reconstructed feature. Finally, we measure the reconstruction quality using the same weighted mean square error loss:

$$\mathcal{L}_{re} = \frac{1}{n}\sum_{v=1}^{m}\sum_{i=1}^{n}\frac{1}{d_v}\|\bar{\mathbf{X}}_{i,:}^{(v)} - \mathbf{X}_{i,:}^{(v)}\|_2^2\mathbf{W}_{i,v}. \tag{7}$$

3.3 Stochastic Cross-sample Fragment Swapping Strategy

To further enhance the generalization and robustness of the model, and to mitigate the interference caused by redundant information inherent in the raw data, we propose a SCFS strategy. Inspired by CutMix [30], our SCFS strategy encourages the model to extract consistent semantic features amidst dynamic cross-sample interference by substituting local feature fragments with heterogeneous data. For any view v, given the input feature $\mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v}$, where n is the batch size and d_v is the feature dimension, we first define a swapping proportion $\eta \in (0,1)$ to determine the fragment length $l = \lfloor \eta \cdot d_v \rfloor$. To execute the swapping, we generate a set of random starting points $\{b_1, b_2, \ldots, b_n\}$ sampled from the range $[1, d_v - l]$. For each instance i in the batch, a donor index j is randomly assigned such that $j \neq i$ to ensure the swapped information originates from a heterogeneous sample. Then, we construct a binary swapping mask $\mathbf{M}^{(v)} \in \{0,1\}^{n\times d_v}$, where elements in the feature index range $[b_i, b_i + l]$ for the i-th sample are set to 1. The augmented feature $\mathbf{X}'^{(v)}$ is formulated by substituting the local segments of the original data with the corresponding fragments from the other samples:

$$\mathbf{X}'^{(v)} = (\mathbf{1} - \mathbf{M}^{(v)}) \odot \mathbf{X}^{(v)} + \mathbf{M}^{(v)} \odot \mathbf{X}_{other}^{(v)}, \tag{8}$$

where $\mathbf{X}_{other}^{(v)}$ represents the features from other samples within the same mini-batch, and $\odot$ denotes element-wise multiplication. The SCFS strategy serves two primary purposes by not only forcing the encoders to maintain semantic consistency when local feature structures are corrupted by foreign data

Table 1: Experimental results of ten methods on the five databases with 50% missing-view rate, 50% missing-label rate, and 70% training samples.

Data	Metric	DM2L	LVSL	iMVWL	NAIM3L	MTD	MSLPP	AIMNet	VIQA	VCPIP
Corel5k	AP	$0.262_{0.005}$	$0.342_{0.004}$	$0.283_{0.008}$	$0.309_{0.004}$	$0.415_{0.008}$	$0.411_{0.007}$	$0.400_{0.01}$	$0.411_{0.008}$	$0.425_{0.01}$
	1-HL	$0.987_{0.000}$	$0.987_{0.000}$	$0.978_{0.000}$	$0.987_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$
	1-RL	$0.843_{0.002}$	$0.881_{0.003}$	$0.865_{0.005}$	$0.878_{0.002}$	$0.893_{0.004}$	$0.899_{0.004}$	$0.902_{0.002}$	$0.900_{0.004}$	$0.907_{0.004}$
	AUC	$0.845_{0.002}$	$0.884_{0.003}$	$0.868_{0.005}$	$0.881_{0.002}$	$0.896_{0.004}$	$0.902_{0.004}$	$0.902_{0.004}$	$0.902_{0.004}$	$0.909_{0.004}$
	1-OE	$0.295_{0.014}$	$0.391_{0.009}$	$0.311_{0.015}$	$0.350_{0.009}$	$0.491_{0.012}$	$0.485_{0.01}$	$0.475_{0.018}$	$0.492_{0.013}$	$0.504_{0.018}$
	1-Cov	$0.647_{0.005}$	$0.718_{0.006}$	$0.702_{0.008}$	$0.725_{0.005}$	$0.749_{0.009}$	$0.766_{0.009}$	$0.771_{0.018}$	$0.761_{0.009}$	$0.784_{0.009}$
Pascal07	AP	$0.471_{0.008}$	$0.504_{0.005}$	$0.437_{0.018}$	$0.488_{0.003}$	$0.551_{0.004}$	$0.545_{0.009}$	$0.548_{0.008}$	$0.556_{0.007}$	$0.559_{0.008}$
	1-HL	$0.928_{0.001}$	$0.930_{0.000}$	$0.882_{0.004}$	$0.928_{0.001}$	$0.932_{0.001}$	$0.933_{0.001}$	$0.931_{0.001}$	$0.931_{0.001}$	$0.933_{0.001}$
	1-RL	$0.761_{0.005}$	$0.806_{0.003}$	$0.736_{0.015}$	$0.783_{0.001}$	$0.831_{0.003}$	$0.820_{0.003}$	$0.831_{0.004}$	$0.828_{0.004}$	$0.837_{0.004}$
	AUC	$0.779_{0.004}$	$0.832_{0.002}$	$0.767_{0.015}$	$0.811_{0.001}$	$0.851_{0.003}$	0.842_{0003}	0.851_{0004}	$0.847_{0.004}$	$0.856_{0.004}$
	1-OE	$0.420_{0.011}$	$0.419_{0.008}$	$0.362_{0.023}$	$0.421_{0.006}$	$0.459_{0.007}$	$0.466_{0.014}$	$0.461_{0.013}$	$0.471_{0.011}$	$0.467_{0.010}$
	1-Cov	$0.692_{0.004}$	$0.751_{0.003}$	$0.677_{0.015}$	$0.727_{0.002}$	$0.784_{0.003}$	$0.771_{0.003}$	$0.783_{0.004}$	$0.780_{0.005}$	$0.791_{0.005}$
ESPGame	AP	$0.212_{0.002}$	$0.285_{0.003}$	$0.244_{0.005}$	$0.246_{0.002}$	$0.306_{0.003}$	$0.312_{0.002}$	$0.305_{0.004}$	$0.307_{0.004}$	$0.319_{0.003}$
	1-HL	$0.982_{0.000}$	$0.983_{0.000}$	$0.972_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$
	1-RL	$0.781_{0.001}$	$0.829_{0.001}$	$0.808_{0.002}$	$0.818_{0.002}$	$0.837_{0.002}$	$0.846_{0.002}$	$0.846_{0.002}$	$0.842_{0.001}$	$0.855_{0.002}$
	AUC	$0.785_{0.001}$	$0.833_{0.002}$	$0.813_{0.002}$	$0.824_{0.002}$	$0.842_{0.002}$	$0.850_{0.002}$	$0.850_{0.002}$	$0.846_{0.001}$	$0.859_{0.002}$
	1-OE	$0.294_{0.006}$	$0.389_{0.004}$	$0.343_{0.013}$	$0.339_{0.003}$	$0.447_{0.009}$	$0.457_{0.012}$	$0.442_{0.011}$	$0.448_{0.009}$	$0.463_{0.007}$
	1-Cov	$0.488_{0.003}$	$0.567_{0.005}$	$0.548_{0.004}$	$0.571_{0.003}$	$0.602_{0.004}$	$0.622_{0.005}$	$0.624_{0.005}$	$0.651_{0.004}$	$0.646_{0.005}$
IAPRTC12	AP	$0.234_{0.003}$	$0.304_{0.004}$	$0.237_{0.003}$	$0.261_{0.001}$	$0.332_{0.003}$	$0.336_{0.005}$	$0.329_{0.005}$	$0.338_{0.004}$	$0.353_{0.005}$
	1-HL	$0.980_{0.000}$	$0.981_{0.000}$	$0.969_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$
	1-RL	$0.823_{0.002}$	$0.861_{0.002}$	$0.833_{0.002}$	$0.848_{0.001}$	$0.875_{0.002}$	$0.880_{0.002}$	$0.883_{0.003}$	$0.882_{0.002}$	$0.893_{0.002}$
	AUC	$0.825_{0.001}$	$0.863_{0.001}$	$0.835_{0.001}$	$0.850_{0.001}$	$0.876_{0.001}$	$0.881_{0.002}$	$0.885_{0.003}$	$0.883_{0.002}$	$0.893_{0.002}$
	1-OE	$0.340_{0.006}$	$0.429_{0.009}$	$0.352_{0.008}$	$0.390_{0.005}$	$0.467_{0.005}$	$0.474_{0.008}$	$0.459_{0.008}$	$0.471_{0.011}$	$0.484_{0.006}$
	1-Cov	$0.529_{0.004}$	$0.597_{0.004}$	$0.564_{0.005}$	$0.592_{0.004}$	$0.649_{0.004}$	$0.672_{0.006}$	$0.673_{0.006}$	$0.668_{0.005}$	$0.697_{0.005}$
MIRFLICKR	AP	$0.514_{0.006}$	$0.553_{0.002}$	$0.490_{0.012}$	$0.551_{0.002}$	$0.607_{0.004}$	$0.615_{0.003}$	$0.602_{0.004}$	$0.616_{0.004}$	$0.622_{0.005}$
	1-HL	$0.878_{0.001}$	$0.885_{0.001}$	$0.839_{0.002}$	$0.882_{0.001}$	$0.891_{0.001}$	$0.892_{0.001}$	$0.890_{0.001}$	$0.892_{0.001}$	$0.893_{0.001}$
	1-RL	$0.831_{0.003}$	$0.856_{0.001}$	$0.803_{0.008}$	$0.844_{0.001}$	$0.875_{0.002}$	$0.878_{0.001}$	$0.873_{0.002}$	$0.880_{0.003}$	$0.881_{0.003}$
	AUC	$0.828_{0.003}$	$0.844_{0.001}$	$0.787_{0.012}$	$0.837_{0.001}$	$0.862_{0.002}$	$0.864_{0.001}$	$0.861_{0.001}$	$0.866_{0.002}$	$0.866_{0.003}$
	1-OE	$0.510_{0.008}$	$0.607_{0.004}$	$0.511_{0.022}$	$0.585_{0.003}$	$0.655_{0.004}$	$0.667_{0.001}$	$0.652_{0.006}$	$0.662_{0.008}$	$0.671_{0.004}$
	1-Cov	$0.604_{0.005}$	$0.636_{0.001}$	$0.572_{0.013}$	$0.631_{0.002}$	$0.676_{0.004}$	$0.679_{0.003}$	$0.671_{0.004}$	$0.683_{0.003}$	$0.686_{0.004}$

but also preventing the model from converging to trivial solutions through increased reconstruction complexity. In our implementation, we empirically set the swapping rate η to 0.2 to strike a balance between data diversity and semantic stability.

3.4 Weighted multi-label classification and overall loss function

The final global prediction $\hat{\boldsymbol{y}}_i$ is synergistically generated by integrating the classification results from the boosting stages and the structural features from the MOEDF mechanism. Specifically, let $\mathbf{P}_i^{(v)} = \sum_{k=1}^{t} \boldsymbol{p}_{i,k}^{(v)}$ be the aggregated semantic output from the incremental classifiers of view v (as defined in Eq. 4). Concurrently, let $\mathbf{Z}'^{(v)}$ denote the refined feature provided by the MOEDF mechanism. To adaptively assess the reliability of each view, we employ a view-confidence evaluation network $\mathscr{D}(\cdot)$, a multi-layer perceptron (MLP), to generate the adaptive weights $\alpha_i^{(v)}$:

$$\boldsymbol{\alpha}_i = \delta\left(\mathscr{D}\left(\{\mathbf{P}_i^{(v)}\}_{v=1}^{m}\right)\right), \tag{9}$$

where $\delta(\cdot)$ denotes the softmax function. Finally, we obtain the final prediction results:

$$\hat{\boldsymbol{y}}_i = \mathscr{R}(\text{RELU}(\sum_{v=1}^{m} \alpha_i^{(v)} \mathbf{P}_i^{(v)} \odot \delta(\sum_{v=1}^{m} \alpha_i^{(v)} \mathbf{Z}'^{(v)}_i))) \tag{10}$$

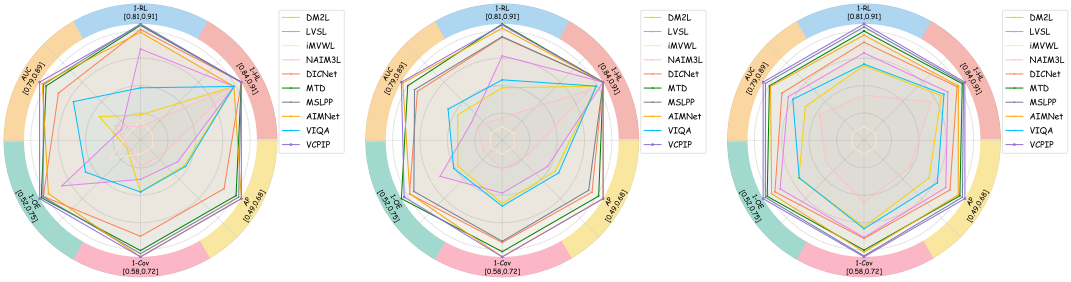
where $\mathscr{R}(\cdot)$ denotes the final regression layer and the Sigmoid activation function. Subsequently, the following multi-label cross-entropy loss is employed as the primary classification objective:

$$\mathscr{L}_{mc} = -\frac{1}{nc} \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mathbf{Y}_{i,j} \log(\hat{\boldsymbol{y}}_{i,j}) + (1 - \mathbf{Y}_{i,j}) \log(1 - \hat{\boldsymbol{y}}_{i,j})\right) \mathbf{G}_{i,j}. \tag{11}$$

Combined with different views loss 5, feature orthogonal loss 6, reconstructed loss 7, and weighted multi-label classification loss 11, our total loss function can be expressed as:

$$\mathscr{L}_{all} = \mathscr{L}_{mc} + \mathscr{L}_{mv} + \alpha \mathscr{L}_{orth} + \beta \mathscr{L}_{re}, \tag{12}$$

where α and β are corresponding penalty parameters. Algorithm 1 shows the training process.



(a) ESPGame (b) IAPRTC12 (c) MIRFLICKR

Figure 3: Experimental results of ten methods on the three full databases without any missing views or labels. The center of the radar map shows the worst results and the vertexes mean the best results on the six metrics.

4 Experiments

4.1 Datasets and metrics

Following the experimental protocols of established IMvMlL benchmarks [9, 15, 16, 31], we evaluate our framework on five widely-recognized multi-view multi-label datasets: Corel5k [32], Pascal07 [33], ESPGame [34], IAPRTC12 [35], and MIRFLICKR [36]. To represent different views, we utilize six distinct feature types, which include GIST, HSV, Hue, Sift, RGB, and LAB. To ensure a strictly equitable comparison, the data partitioning and feature configurations in our experiments conform to the standards set by prior literature.

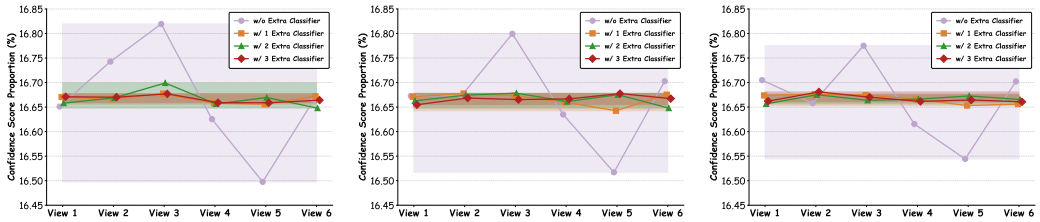
To ensure a thorough evaluation, we adopt six standard performance indicators following the precedents set by [9, 15, 16]: Ranking Loss (**RL**), Average Precision (**AP**), Hamming Loss (**HL**), adapted **AUC**, One-Error (**OE**), and Coverage (**Cov**). For ease of interpretation, we transform **HL**, **RL**, **OE**, and **Cov** into their complementary forms (e.g., **1-HL**); consequently, a higher value across all reported metrics signifies enhanced model performance.

4.2 Comparison Methods

To evaluate the performance of our proposed method, we compare it against eight competitive approaches, including **DM2L** [37], **LVSL** [1], **NAIM3L** [17], **iMVWL** [38], **MTD** [21], **MSLPP** [16], **AIMNet** [15], and **VIQA** [19] to compare with our VCPIP on the five incomplete multi-view weak multi-label datasets. It is worth noting that not all methods are inherently equipped to handle the dual-incompleteness of views and labels. Specifically, since DM2L is a single-view multi-label classification approach tailored for incomplete labels, we execute it on each individual view separately and report the peak performance. Conversely, LVSL is a conventional MvMLC paradigm that lacks the capacity to manage missing data. To ensure its applicability, we compensate for missing views via mean imputation and initialize unobserved label entries with zeros.

4.3 Experimental Results and Analysis

To comprehensively evaluate the performance of our proposed method in scenarios with missing views and labels, we compare it with eight state-of-the-art algorithms across five datasets. We use fixed data splits across compared configurations and repeat each experiment with controlled random seeds; Table 1 reports the mean and standard deviation across repeated runs (with the standard deviation shown as the subscript). As shown in Table 1, VCPIP achieves state-of-the-art performance across multiple metrics on five datasets. For instance, compared with the state-of-the-art methods VIQA, VCPIP achieves AP and AUC gains of 1.4% and 0.7% on the Pascal07 dataset, and 1.2% and 1.3% on the ESPGame dataset, respectively. Similarly, VCPIP outperforms AIMNet, yielding AP and AUC gains of 2.5% and 0.7% on the Corel5k dataset and 1.1% and 0.5% on the Pascal07 dataset. These experimental results demonstrate the effectiveness of VCPIP in augmenting overall performance via the enhancement of non-dominant view representations. Meanwhile, to further



(a) Epoch 90 (b) Epoch 140 (c) Epoch 190

Figure 4: Comparison of view weight distributions across varying numbers of classifiers. By normalizing raw confidence scores into weights, this figure intuitively illustrates how additional classifiers balance the relative contributions of individual views.

explore the performance ceiling of the models, we evaluate all methods in a complete data scenario where both view and label missing rates are set to 0. The radar charts in Fig. 3 illustrate the results across three representative datasets, where a curve closer to the centre indicates inferior performance. Notably, while data incompleteness inevitably compromises the performance of all models, our proposed method demonstrates exceptional resilience under missing-view scenarios and consistently surpasses some baselines in the complete data setting, thereby maintaining a clear competitive advantage.

To further investigate the impact of extra classifiers on the evolution of view weights, we conduct experiments on the Corel5k dataset under a challenging setting of 50% missing views and missing labels. We evaluated the model at different training stages (epochs 90, 140, and 190), comparing configurations without extra classifiers against those with 1, 2, and 3 additional classifiers. As illustrated in Fig. 4, the inclusion of extra classifiers effectively bolsters the weights of non-dominant views, promoting a more uniform weight distribution that converges toward the mean.

Table 2: Ablation results on Corel5k and Pascal07 datasets with 50% missing views and 50% missing labels. 'w/o' means 'without'. 'w/' means 'with'

Method	Corel5k						Pascal07					
	AP	1-HL	1-RL	AUC	1-OE	1-Cov	AP	1-HL	1-RL	AUC	1-OE	1-Cov
w/ static prediction	0.394	0.988	0.897	0.900	0.478	0.766	0.549	0.932	0.815	0.836	0.475	0.762
w/o MOEDF	0.394	0.988	0.896	0.898	0.486	0.764	0.546	0.931	0.826	0.845	0.461	0.775
w/o SCFS	0.398	0.988	0.900	0.902	0.497	0.770	0.543	0.932	0.819	0.839	0.456	0.769
VCPIP	**0.425**	**0.988**	**0.907**	**0.909**	**0.504**	**0.784**	**0.559**	**0.933**	**0.837**	**0.856**	**0.467**	**0.791**
w/o $\mathcal{L}_{orth}$	0.415	0.988	0.906	0.909	0.497	0.784	0.555	0.933	0.834	0.853	0.463	0.786
w/o $\mathcal{L}_{mv}$	0.408	0.988	0.905	0.907	0.495	0.781	0.544	0.933	0.826	0.848	0.461	0.775
w/o $\mathcal{L}_{re}$	0.405	0.988	0.904	0.906	0.497	0.776	0.554	0.933	0.834	0.854	0.467	0.786

4.4 Ablation Study

Our framework incorporates several innovative components designed to optimize multi-view learning. To rigorously evaluate the individual efficacy of these components, we conduct extensive ablation studies on the Corel5k and Pascal07 datasets under a challenging configuration of 50% missing labels and views, with the quantitative results summarized in Table 2. Notably, the inclusion of the static prediction resulted in AP drops of 0.031 and 0.016 on Corel5k and Pascal07, respectively. Furthermore, our ablation analysis of the corresponding loss functions confirms that each objective term contributes substantially to the overall performance. These empirical findings validate that our architectural design effectively harnesses complementary multi-view information, thereby bolstering the representation learning capability of individual views.

5 Conclusion

In this paper, we present VCPIP, a novel view-confidence perception-driven incremental prediction framework designed to tackle the formidable challenge of dual-incompleteness in multi-view multi-

label classification. Additionally, we introduce the VCA strategy that achieves dynamic identification of non-dominant views and refines predictions through residual optimisation. Moreover, the proposed MOEDF mechanism further enhances feature representation by disentangling consistent and specific features under rigorous orthogonality constraints. Furthermore, we introduce the SCFS strategy, which interchanges sample-wise feature fragments to constrain the model to learn robust global features for superior generalization. Extensive experimental results across diverse benchmarks validate that VCPIP not only exhibits superior robustness in the dual incompleteness of view and label but also maintains excellent performance in dual completeness of views and labels.

Acknowledgments and Disclosure of Funding

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2026-08-077).

References

[1] Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 25:7235–7247, 2022.

[2] Buyue Qian, Xiang Wang, Jieping Ye, and Ian Davidson. A reconstruction error based framework for multi-label and multi-view learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):594–607, 2014.

[3] Yongshan Zhang, Jia Wu, Zhihua Cai, and Philip S Yu. Multi-view multi-label learning with sparse feature selection for image annotation. *IEEE Transactions on Multimedia*, 22(11):2844–2857, 2020.

[4] Pingzhu Liu, Wenbin Qian, Jintao Huang, Yanqiang Tu, and Yiu-Ming Cheung. Transformer-driven feature fusion network and visual feature coding for multi-label image classification. *Pattern Recognition*, 164:111584, 2025.

[5] Zunnan Xu, Jiaqi Huang, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Enhancing fine-grained multi-modal alignment via adapters: A parameter-efficient training framework for referring image segmentation. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024.

[6] Weifeng Hu, Yu Zhang, Yujun Li, Jia Zhao, Xifeng Hu, Yan Cui, and Xuejing Wang. Query-based video summarization with multi-label classification network. *Multimedia Tools and Applications*, 82(24):37529–37549, 2023.

[7] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124, 2017.

[8] Jie Wen, Yadong Liu, Zhanyan Tang, Yuting He, Yulong Chen, Mu Li, and Chengliang Liu. Learning compact semantic information for incomplete multi-view missing multi-label classification. In *Forty-second International Conference on Machine Learning*, 2025.

[9] Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *Forty-first international conference on machine learning*, 2024.

[10] Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[11] Chengliang Liu, Bo Li, Bob Zhang, Xiaoling Luo, Yabo Liu, and Jie Wen. Permutation-consistent variational encoding for incomplete multi-view multi-label classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Quanjiang Li, Tianxiang Xu, Tingjin Luo, Yan Zhong, Yang Li, Yiyun Zhou, and Chenping Hou. Theory-driven label-specific representation for incomplete multi-view multi-label learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[13] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. 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, 2023.

[14] Jie Wen, Lian Zhao, Xiaohuan Lu, Chengliang Liu, Li Shen, Chao Huang, and Yong Xu. Disentangling consistent and specific information for double incomplete multi-view multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[15] Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13864–13872, 2024.

[16] Jiang Long, Qi Zhang, Xiaohuan Lu, Jie Wen, Lian Zhao, and Wulin Xie. Multi-scale locality preserving projection for partial multi-view incomplete multi-label learning. *Neural Networks*, 180:106748, 2024.

[17] Xiang Li and Songcan Chen. 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, 2022.

[18] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023.

[19] Xiaohuan Lu, Jiang Long, Haitao Zhang, Wulin Xie, Lian Zhao, Yinghao Ye, and Jie Wen. Partial multi-view incomplete multi-label learning network with quality-aware representation fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[20] Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15712–15721, 2023.

[21] Chengliang Liu, Jie Wen, Yabo Liu, Chao Huang, Zhihao Wu, Xiaoling Luo, and Yong Xu. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. *Advances in neural information processing systems*, 36:32387–32400, 2023.

[22] Chengliang Liu, Jie Wen, Zhihao Wu, Xiaoling Luo, Chao Huang, and Yong Xu. Information recovery-driven deep incomplete multiview clustering network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15442–15452, 2023.

[23] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[24] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13926–13935, 2020.

[25] Zijun Que, Xiaohang Jin, Zhengguo Xu, and Chang Hu. Remaining useful life prediction based on incremental learning. *IEEE Transactions on Reliability*, 73(2):876–884, 2023.

[26] QingYuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. *arXiv preprint arXiv:2502.20120*, 2025.

[27] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[28] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. pmlr, 2023.

[29] Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *arXiv preprint arXiv:2105.08919*, 2021.

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[31] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8816–8824, 2023.

[32] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, pages 97–112. Springer, 2002.

[33] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009.

[34] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

[35] M Henning, D Thomas, et al. The iapr benchmark: Anew evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*, pages 1–11, 2006.

[36] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[37] Zhongchen Ma and Songcan Chen. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. *Pattern Recognition*, 111:107675, 2021.

[38] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *International Joint Conference on Artificial Intelligence*, pages 2703–2709, 2018.

A Datasets

To provide a multi-perspective representation, each of the five image benchmarks is characterized by a comprehensive set of six distinct feature descriptors, namely GIST, HSV, Hue, SIFT, RGB, and LAB. The intrinsic properties and quantitative profiles of these datasets are summarized in Table 3.

Table 3: Detailed statistics about five multi-view multi-label databases.

Database	# Sample	# Category	# View	# Category/#Sample
Corel5k	4999	260	6	3.40
Pascal07	9963	20	6	1.47
ESPGame	20770	268	6	4.69
IAPRTC12	19627	291	6	5.72
MIRFLICKR	25000	38	6	4.72

B Time cost

To substantiate the computational efficiency of VCPIP, Table 4 presents a comparative analysis of execution times across ten benchmarks using a 70% training split of the Corel5k dataset. Recognizing that training duration is contingent upon specific convergence criteria, we evaluate all models under their native termination settings to ensure a fair comparison. For single-view paradigms, we report the aggregate training latency across all modalities alongside the inference time for a representative view. Notably, for MTD, MSLPP, AIMNet, VIQA, and VCPIP, the training process is standardized to 100 epochs.

Table 4: Time efficiency of training and testing phases of the nine methods on the Corel5k dataset with 70% training samples. (Unit: second)

Phase \ Method	DM2L	LVSL	iMVWL	NAIM3L	MTD	MSLPP	AIMNet	VIQA	VCPIP
Training	713.37	63.73	165.82	143.63	687.52	4889.84	244.12	184.55	484.53
Test	0.04	0.64	0.02	0.01	0.15	0.05	0.03	0.09	0.02

C Incomplete Multi-View Weak Multi-Label Dataset Setting

To emulate dual-incomplete scenarios, a stochastic masking mechanism is applied to both views and label. Specifically, we discard individual views with a constant probability, maintaining a minimum of one view per instance to preserve data integrity. Concurrently, label components are obscured at an equivalent ratio. In our standard configuration, we initialize both corruption rates at 0.5. The final data pool is divided into training, validation, and testing subsets following a 7:1.5:1.5 distribution.

D Implementation Details

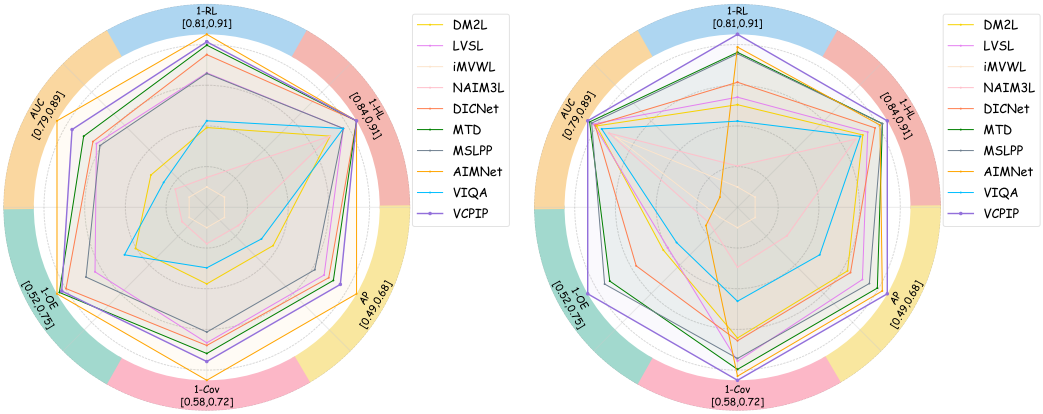
The proposed VCPIP framework is developed using the PyTorch library on an Ubuntu platform, powered by an Intel i9-14900 CPU and a single RTX 4060 GPU. During the training phase, we employ the SGD optimizer with an initial learning rate of 0.1. Across all five benchmarks, the batch size and momentum are consistently maintained at 128 and 0.9, respectively. Additional hyperparameter configurations are detailed in the supplementary material. The source code for our full implementation can be found at `https://github.com/LPZliu/VCPIP`.

E Comparison Methods

In this section, we give details of the eight comparison methods in Table 5.

Table 5: Comparison of multi-label classification approaches: 'Multi-view' identifies methods tailored for multi-view scenarios; 'Missing-view' and 'Missing-label' indicate robustness against deficient views and sparse labeling, respectively.

Method	Sources	Multi-view	Missing-view	Missing-label
DM2L[37]	PR'21	✗	✗	✓
LVSL[1]	TMM'22	✓	✗	✗
iMVWL[38]	IJCAI'18	✓	✓	✓
NAIM3L[17]	TPAMI'22	✓	✓	✓
MTD[21]	NIPS'23	✓	✓	✓
MSLPP[16]	NN'24	✓	✓	✓
AIMNet[15]	AAAI'24	✓	✓	✓
VIQA[19]	TCSVT'25	✓	✓	✓



(a) Corel5k (b) Pascal07

Figure 5: Experimental results of ten methods on the two full databases without any missing views or labels. The center of the radar map shows the worst results and the vertexes mean the best results on the six metrics.

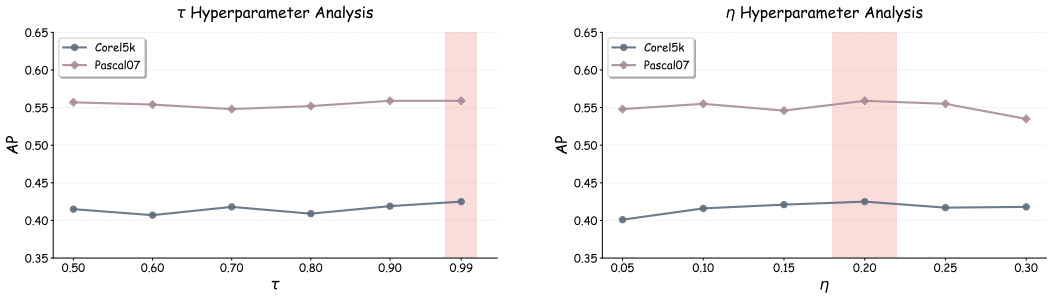
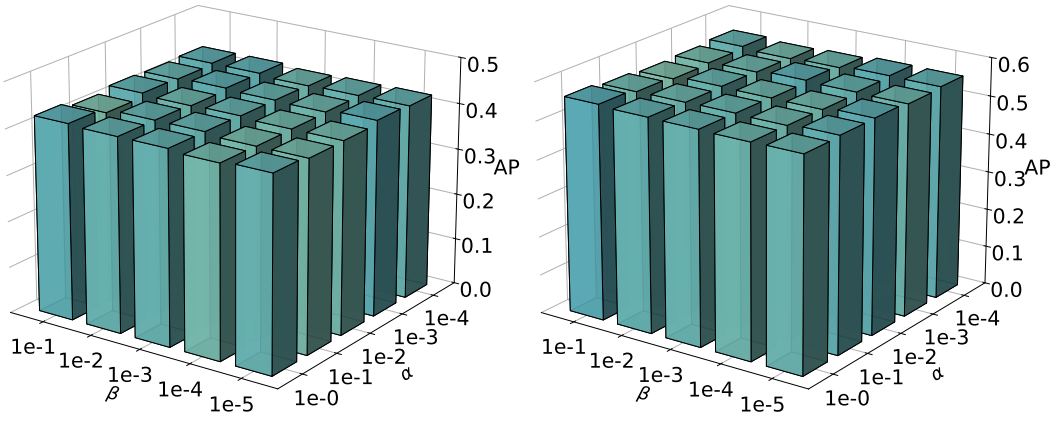


Figure 6: Hyperparameter sensitivity analysis. The left plot illustrates the impact of τ on model performance, while the right plot shows the analysis for η.

F Comparison with Fully-Supervised Methods on Two Other Datasets

In this section, we show the results of ten methods on two datasets without any missing views and labels in Fig. 5.



(a) Corel5k (b) Pascal07

Figure 7: Sensitivity analysis of hyperparameters α and β conducted on the Corel5k and Pascal07 datasets.

G Hyperparameters Study

To further validate the influence of hyperparameters, we conduct extensive experiments on the Corel5k and Pascal07 datasets. Following the standard protocol for incomplete multi-view multi-label learning, both datasets are configured with a 50% missing-label rate, a 50% missing-view rate, and 70% training samples. As illustrated in Fig. 6 and Fig. 7, our method exhibits consistent performance across a broad spectrum of hyperparameter settings. Specifically, we observe that the optimal range for α spans from $1e-4$ to $1e-0$, while β performs effectively within the range of $[1e-5, 1e-1]$ for both the Corel5k and Pascal07 datasets. For implementation consistency, we empirically set $\alpha = 1e-0$ and $\beta = 1e-1$ as the default configurations in our experiments. Furthermore, the sensitivity analysis indicates that the framework is highly robust to variations in τ and η. Consequently, we uniformly fix $\tau = 0.99$ and $\eta = 0.2$ across all five datasets to maintain a balance between stability and generalizability. And it is worth noting that the MOEDF module is structured with four experts.

H Limitations and Future Work

While VCPIP demonstrates robust performance across diverse benchmarks by rectifying view-level imbalances, certain challenges remain for future exploration. A potential performance bottleneck in the IMvMIL task involves the further integration of higher-order semantic correlations across heterogeneous view information. Although our MOEDF mechanism effectively disentangles consistent and specific features, capturing the more intricate, non-linear dependencies between views could further enhance discriminative power. Furthermore, current hyperparameters, such as the penalty parameters $\alpha, \beta, \gamma, \tau$ and the swapping rate η, are determined through extensive empirical validation. While these settings yield state-of-the-art results, they may not be optimal for all possible data distributions. In future research, we aim to integrate automated machine learning (AutoML) or Bayesian optimization frameworks to enable adaptive hyperparameter selection. This transition from manual tuning to self-evolving configurations will significantly bolster the model's generalisation and deployment efficiency in complex, real-world multi-view scenarios.

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 outline the Incomplete Multi-view Multi-label Learning (IMvMIL) problem and accurately state the proposed VCPIP framework's contributions, which align with the subsequent empirical validation.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: A dedicated "Limitations and Future Work" section (Appendix H) discusses the challenge of integrating higher-order semantic correlations and the need for adaptive hyperparameter selection in future research.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The paper proposes an empirical deep learning framework and does not focus on proving novel theoretical results or theorems.

Guidelines:

- The answer [N/A] 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 paper describes the architecture, evaluation metrics, and implementation details (e.g., PyTorch framework, SGD optimizer, learning rate, batch size) in Section 4 and Appendix D.

Guidelines:

- The answer [N/A] 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: While implementation details are provided, the paper currently does not include a direct URL to an open-source code repository or data package.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental settings, including dataset partitioning (7:1.5:1.5), hyperparameters (a, β, τ, η), optimizer, and learning rate, are specified in the main text and Appendices C, D, and G.

   Guidelines:

   - The answer [N/A] 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: The main experimental results in Table 1 report both the mean and standard deviation for the evaluated metrics across the datasets.

   Guidelines:

   - The answer [N/A] 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: Appendix D explicitly states that experiments were conducted on an Ubuntu platform with an Intel i9-14900 CPU and a single RTX 4060 GPU, and Appendix B lists the specific training and testing time costs.

   Guidelines:

   - The answer [N/A] 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 methodology utilizes standard, publicly available image benchmarks and strictly conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The paper focuses on foundational machine learning methodology for multi-view multi-label classification and does not explicitly explore direct societal impacts or specific deployment applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The proposed framework and the standard public datasets used (e.g., Corel5k, Pascal07) do not present a high risk for misuse.

Guidelines:

- The answer [N/A] 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: [Yes]

Justification: The paper appropriately cites the original sources and creators for the widely recognized benchmarks (Corel5k, Pascal07, ESPGame, IAPRTC12, MIRFLICKR) employed in the experiments.

Guidelines:

- The answer [N/A] 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` 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: [N/A]

    Justification: The research evaluates its methodology on existing, standard public datasets and does not introduce or release any new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The experiments are based on existing datasets and do not involve crowdsourcing or any research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: The research solely relies on publicly available datasets and does not involve human subjects requiring IRB approval.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Large Language Models (LLMs) are not utilized as an important, original, or non-standard component of the core methods developed in this research.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.