Calibratable Disambiguation Loss for Multi-Instance Partial-Label Learning

Wei Tang, Yin-Fang Yang, Weijia Zhang, and Min-Ling Zhang, *Senior Member, IEEE*

Abstract—Multi-instance partial-label learning (MIPL) is a weakly supervised framework that extends the principles of multi-instance learning (MIL) and partial-label learning (PLL) to address the challenges of inexact supervision in both instance and label spaces. However, existing MIPL approaches often suffer from poor calibration, undermining classifier reliability. In this work, we propose a plug-and-play calibratable disambiguation loss (CDL) for classification and calibration, which modulates a disambiguation objective by a top-vs-competitor prediction margin. The competitor is instantiated either as the second strongest candidate label or as the strongest non-candidate label, yielding two variants that respectively emphasize candidate-level separation and candidate-vs-non-candidate suppression. Theoretically, we analyze CDL as a margin-modulated momentum-based disambiguation loss (MDL) objective, derive a lower-bound and a pseudo-label confidence-alignment bound for calibration, and show through gradient and momentum analyses how margin shaping affects weight updates. Experimental results on benchmark and real-world MIPL datasets, together with representative PLL adaptation, confirm that our CDL significantly improves both classification accuracy and expected calibration error.

Index Terms—Multi-instance partial-label learning, partial-label learning, disambiguation loss, model calibration.

1 Introduction

Weakly supervised learning enables the construction of predictive models with limited supervision. According to [1], weak supervision can be classified into three types: inexact, inaccurate, and incomplete. Inexact supervision arises from a coarse alignment between instances and labels, a common and challenging issue in real-world applications. Two prominent frameworks that address inexact supervision from different perspectives are *multi-instance learning (MIL)* [2], [3], [4] and *partial-label learning (PLL)* [5], [6], [7]. MIL addresses inexactness in the instance space, where the positive instances within a bag are unidentified [8], [9], while PLL focuses on inexactness in the label space, where the true label remains hidden within a candidate label set [10], [11], [12].

Recent advancements have introduced *multi-instance partial-label learning (MIPL)* to jointly model these two sources of inexactness [13]. In MIPL, each training example is a multi-instance bag associated with a candidate label set. During training, both the positive instances in the instance space and the true label in the label space remain unknown. Fig. 1 shows a pathology image classification scenario. Whole-slide or high-resolution pathology images are often divided into patches for computational feasibility [14], while candidate labels can be collected from crowd-sourced annotators to reduce expert annotation cost [15], [16]. Such data naturally contain inexactness in both instance and label spaces, making MIPL a suitable formulation.

Existing MIPL approaches can be categorized into two paradigms: the instance-space paradigm and the embedded-space paradigm. The instance-space paradigm generates a predicted label for a multi-instance bag by aggregating the prediction probabilities of its constituent instances [13]. In contrast, the embedded-space paradigm classifies multi-instance bags directly by aggregating them into a single feature vector [16]. The latter paradigm often exhibits superior classification performance owing to its ability to capture global feature representations. Label disambiguation is central to MIPL, which involves identifying the true label from the candidate set and significantly influences classification performance [16], [17]. However, existing MIPL objectives are primarily designed for disambiguation.

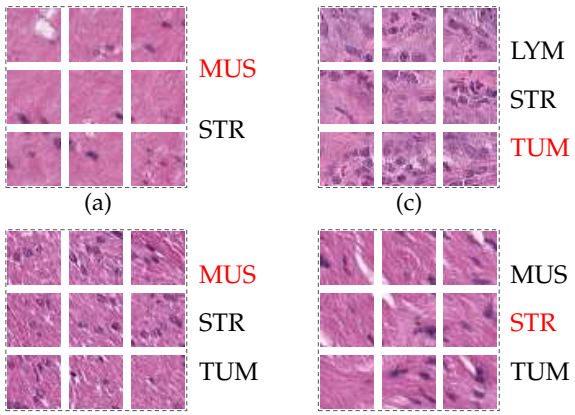


Fig. 1: Pathology image classification with crowd-sourced candidate label sets is a MIPL scenario [16], where true labels are highlighted in red and false-positive labels in black. The labels include LYM (lymphocytes), MUS (smooth muscle), STR (cancer-associated stroma), and TUM (colorectal adenocarcinoma epithelium).

- *Wei Tang, Yin-Fang Yang, and Min-Ling Zhang are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and the Key Laboratory of Computer Network and Information Integration (Southeast University), MoE, China. E-mail: {tangw, yangyf, zhangml}@seu.edu.cn.*
- *Weijia Zhang is with the School of Computer and Information Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: weijia.zhang@newcastle.edu.au.*
- *Corresponding author: Min-Ling Zhang.*

They do not explicitly ensure that the predicted confidence reflects the empirical probability of being correct, which limits their reliability in decision-sensitive applications.

This limitation is evident in Fig. 2 (a), (b), and (c), the predicted confidences of DEMIPL [16], ELIMIPL [18], and MIPLMA [19] tend to cluster around 0.25, leading to poor expected calibration error (ECE). In pathology image classification, such unreliable confidence estimates are undesirable because confidence scores may guide triage, human review, diagnosis, and treatment planning. Moreover, calibration losses designed for fully supervised data assume known true labels. Naively adapting focal-type losses to MIPL can interfere with pseudo-label evolution and lead to under-confident or over-confident predictions. Therefore, MIPL requires a training objective that improves calibration without weakening label disambiguation.

To this end, we propose a plug-and-play *Calibratable Disambiguation Loss* (CDL). CDL modulates a momentum-based disambiguation objective by a top-vs-competitor margin, where the competitor is instantiated either as the second-highest candidate-label probability or as the highest non-candidate-label probability. The two variants, namely CDL-CC and CDL-CN, respectively encourage candidate-level separation and candidate-vs-non-candidate suppression. Low-margin bags retain strong disambiguation signals, whereas well-separated bags receive less incentive for excessive probability sharpening. CDL can be seamlessly integrated into existing embedded-space MIPL frameworks. As shown in Fig. 2 (d)–(i), the resulting variants improve both classification accuracy and ECE.

Our key contributions are as follows: 1) To the best of our knowledge, this work is the first to identify model calibration issues in MIPL and show that existing disambiguation-centered methods can produce poorly calibrated confidence estimates. 2) We propose CDL, a plug-and-play top-vs-competitor margin-modulated disambiguation loss with two complementary instantiations. 3) We theoretically analyze CDL as a margin-modulated MDL objective, establish its loss-level and calibration-related properties, and explain its margin-shaping effect through gradient and momentum analyses. 4) Extensive experiments on benchmark and real-world MIPL datasets, together with representative PLL adaptation, demonstrate that CDL consistently improves both classification accuracy and calibration performance.

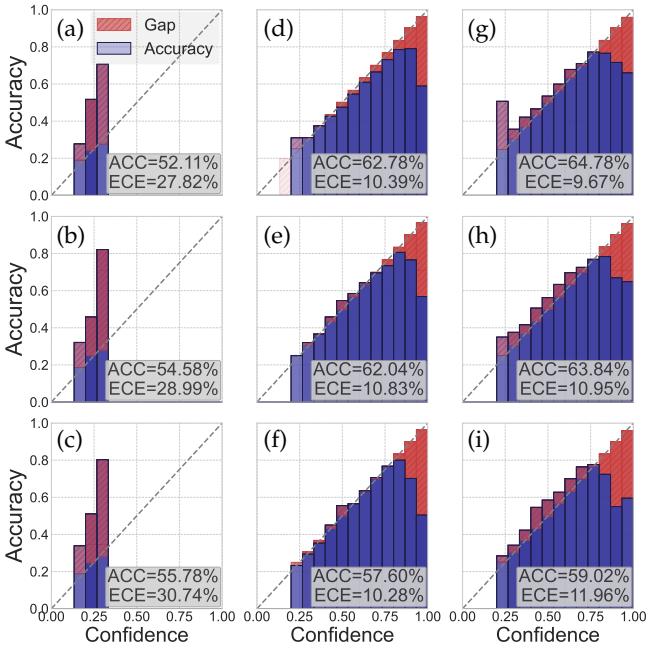


Fig. 2: Reliability diagrams of (a) DEMIPL [16], (b) ELIMIPL [18], (c) MIPLMA [19], (d) DAMCC, (e) SAMCC, (f) MAMCC, (g) DAMCN, (h) SAMCN, and (i) MAMCN on the C-KMeans test set. The diagrams display mean accuracy (ACC) and expected calibration error (ECE) from ten runs, with (d)-(i) representing our methods. The bar color intensity reflects the number of samples assigned to the corresponding confidence intervals.

2 RELATED WORK

2.1 Multi-Instance Learning

Multi-instance learning (MIL), initially developed for drug activity prediction [20], has since been applied across diverse fields, including text classification [21], [22], [23], object detection [24], and video anomaly detection [25]. Recent advancements in MIL often incorporate attention mechanisms to synthesize features from multiple instances within a bag into a cohesive representation for classification. For example, Ilse *et al.* [26] introduced both plain and gated attention mechanisms, which significantly improved the performance of binary MIL tasks. An extension of this paradigm, the loss-based attention mechanism [27] effectively addresses multi-class classification challenges. The success of attention-based MIL approaches has led to their widespread adoption in applications such as pathology image classification [28], [29] and time series classification [30]. These approaches utilize a weighted aggregation of attention scores and instances to generate a comprehensive bag-level feature.

Despite the significant advancements, existing MIL approaches remain ineffective for MIPL scenarios because they cannot directly handle ambiguous candidate label sets [13].

2.2 Partial-Label Learning

Partial-label learning (PLL) has diverse applications across numerous real-world domains, such as face naming [31], [32], [33], object classification [34], [35], bioinformatics [36], and facial age estimation [37], [38]. Recent developments have led to the emergence of several deep learning-based PLL methods. For example, Lv *et al.* [17] utilized linear classifiers and multi-layer perceptrons to generate feature representations from instances, employing progressive disambiguation techniques to identify the true labels. Building on this foundation, Feng *et al.* [39] explored the process of generating PLL data and introduced two theoretically robust algorithms for PLL. Similarly, Wen *et al.* [40] proposed a weighted loss function for disambiguation that provides a versatile approach applicable to various algorithms. Furthermore, Xu *et al.* [41] applied progressive purification of candidate labels to train classifiers within the instance-dependent PLL framework. Recently, Wang *et al.* [42] established the first PLL benchmark and introduced theoretically justified model selection criteria for PLL.

However, the inherent limitations of these PLL approaches in handling inexact supervision in the instance space impede their effectiveness in MIPL scenarios [13].

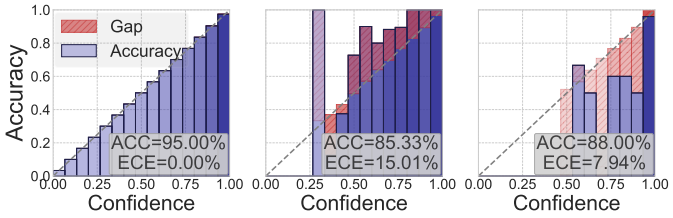
2.3 Multi-Instance Partial-Label Learning

MIPL extends MIL and PLL to handle dual inexact supervision, where both instances within a bag and labels in the candidate set are ambiguous. Tang *et al.* [13] first formalized MIPL and proposed MIPLGP, an instance-space method that assigns pseudo-negative classes, propagates bag-level candidate labels to instances, and performs Dirichlet-based disambiguation with Gaussian process regression. By transforming discrete candidate labels into continuous ones, MIPLGP enables uncertainty-aware label disambiguation. However, its bag-level prediction relies on the maximum instance-level response, limiting its ability to learn global bag representations. To address this limitation, DEMIPL [16] adopts an embedded-space paradigm, where bags are encoded by attention-based aggregation and true labels are identified through a momentum strategy. Although DEMIPL captures global bag information, it insufficiently constrains the full label space and may assign high probabilities to non-candidate labels. ELIMIPL [18] and MIPLMA [19] mitigate this issue by explicitly suppressing non-candidate labels, yielding improved performance over earlier MIPL methods. However, these approaches remain primarily discriminative and do not characterize the data-generation mechanism. PROMIPL [43] addresses this issue with a probabilistic generative model that infers latent ground-truth labels from the assumed generation process, but its reliance on a specific bag prior may restrict cross-dataset generalization. More recently, FASTMIPL [44] introduced a mixed-effects formulation to model instance-bag dependencies with improved efficiency, while its generalized linear structure may limit expressivity on complex data. Wang *et al.* [45] further explored MIPL through latent structural learning and neuro-symbolic integration, extending its application scope but relying on relatively strong assumptions.

Despite these advances, existing MIPL studies mainly emphasize label disambiguation, whereas model calibration remains insufficiently explored.

2.4 Model Calibration

Model calibration is crucial for ensuring that predicted probabilities accurately reflect true likelihoods of the respective classes [46]. Calibration methods can be broadly categorized into training-time and post-hoc approaches. Training-time calibration technologies incorporate regularization techniques during training to address overconfidence in deep neural networks. For example, label smoothing is a prominent technique that reduces overfitting and enhances calibration by mitigating excessively confident predictions [47], [48]. Implicit regularization methods, such as mixup training and focal loss, originally designed to improve generalization, have also proven effective for calibration [48], [49], [50], [51], [52]. The post-hoc calibration methods adjust model outputs after training to better align predicted probabilities with true likelihoods. Platt scaling refines binary classifier outputs using learned parameters [53], [54], whereas temperature scaling, an extension of Platt scaling, rescales logits in multi-class tasks using a learned temperature parameter [46].



(a) Perfect calibration (b) FL (c) IFL

Fig. 3: Reliability diagrams of SAM [18] with FL or IFL on the FMNIST-MIPL dataset with one false positive label ($r = 1$).

While the above calibration techniques provide substantial benefits for supervised learning tasks, using them in MIPL is challenging due to the absence of true labels.

3 PRELIMINARIES

3.1 Notations

We formalize a MIPL training dataset as $\mathcal{D} = \{(\boldsymbol{X}_i, \mathcal{S}_i) \mid 1 \leq i \leq m\}$, where $\mathcal{D}$ consists of m multi-instance bags, each associated with a corresponding candidate label set. The instance space is represented by $\mathcal{X} = \mathbb{R}^d$, and the label space is denoted as $\mathcal{Y} = \{1, 2, \ldots, k\}$, which includes k distinct class labels. Specifically, the i-th multi-instance bag $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,n_i}\}$ contains n_i instances within a d-dimensional space. The candidate label set $\mathcal{S}_i$ and the non-candidate label set $\bar{\mathcal{S}}_i$ are subsets of $\mathcal{Y}$, satisfying the constraints $\mathcal{S}_i \cup \bar{\mathcal{S}}_i = \mathcal{Y}$ and $\mathcal{S}_i \cap \bar{\mathcal{S}}_i = \varnothing$.

Notably, each bag contains at least one instance associated with the true label, referred to as a positive instance. In contrast, negative instances may correspond to background or irrelevant content, but they should not be associated with any false-positive labels in the candidate label set. For example, in pathology image classification [16], positive instances represent specific cell types, such as lymphocytes or colorectal adenocarcinoma epithelium, whereas negative instances include background regions or non-cellular areas.

3.2 Calibration

A well-calibrated model ensures that the confidence scores of the predictions reflect the true probabilities of correctness [46], [52], [55], [56], which can be formally expressed as:

$$\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p, \quad \forall p \in [0, 1], \tag{1}$$

where $\hat{y}$ and $\hat{p}$ denote the predicted label and confidence score, respectively, while y represents the true label. The confidence score p is the model's estimated probability that the prediction $\hat{y}$ is correct.

The expected calibration error (ECE) quantifies calibration quality from finite samples by partitioning the predicted confidences into R bins $\{B_r\}_{r=1}^{R}$ and computing:

$$\mathrm{ECE} = \sum_{r=1}^{R} \frac{|B_r|}{m} \left| A_r - P_r \right|, \tag{2}$$

where $A_r = \frac{1}{|B_r|} \sum_{i \in B_r} \mathbb{I}(\hat{y}_i = y_i)$ and $P_r = \frac{1}{|B_r|} \sum_{i \in B_r} \hat{p}_i$ represent the accuracy and average confidence in bin r,

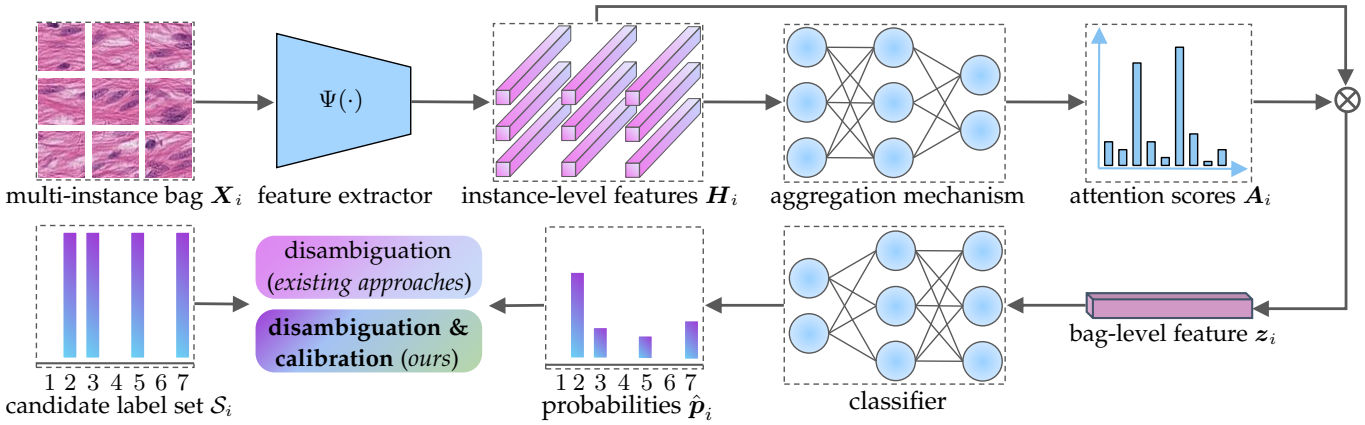


Fig. 4: Framework of MIPL approaches within the embedded-space paradigm.

respectively. This metric assesses how well-predicted confidences align with actual correctness, typically employing $R = 15$ bins. A model is well-calibrated if its predicted confidences closely reflect empirical accuracies. Fig. 3 (a) illustrates perfect calibration using generated data, where, in each bin, predicted confidences precisely match accuracies, resulting in perfect calibration, i.e., ECE $= 0$.

Focal loss (FL) [57] was introduced to mitigate class imbalance in object detection by reducing the loss weight for predicted samples with high confidence. Mukhoti *et al.* [51] demonstrated that replacing cross-entropy loss with FL enhances model calibration. Moreover, Wang *et al.* [48] proposed inverse focal loss (IFL), which improves calibration performance further. The formulations of FL and IFL are:

$$\mathcal{L}_{\text{FL}} = -(1-\hat{p}_{i,y_i})^{\gamma}\log\hat{p}_{i,y_i}, \tag{3}$$

$$\mathcal{L}_{\text{IFL}} = -(1+\hat{p}_{i,y_i})^{\gamma}\log\hat{p}_{i,y_i}, \tag{4}$$

where y_i denotes the true label of the i-th sample.

In standard supervised learning, focal losses significantly improve calibration. However, applying FL and IFL in MIPL is challenging due to the absence of true labels, as illustrated in Fig. 3 (b) and (c).

3.3 The Pitfalls of Naive Focal Losses in MIPL

One naive approach for extending the focal losses to MIPL is to select the candidate label with the highest predicted probability as a surrogate of the true label. However, this naive implementation does not create well-calibrated models and also significantly reduces the classification accuracy in MIPL. To compensate for this, we propose to employ all candidate labels with different weights. Specifically, we utilize the Scaled additive Attention Mechanism (SAM) in ELIMIPL [18] to derive the holistic features of multi-instance bags, coupled with FL or IFL as the loss function. Since FL and IFL were originally designed for fully supervised settings, we refine FL and IFL for MIPL as follows:

$$\mathcal{L}_{\text{FL}}^{\text{MIPL}} = -\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)}(1-\hat{p}_{i,c}^{(t)})^{\gamma}\log(\hat{p}_{i,c}^{(t)}), \tag{5}$$

$$\mathcal{L}_{\text{IFL}}^{\text{MIPL}} = -\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)}(1+\hat{p}_{i,c}^{(t)})^{\gamma}\log(\hat{p}_{i,c}^{(t)}), \tag{6}$$

where $\mathcal{S}_i$ is the candidate label set of the i-th multi-instance bag and $\hat{p}_{i,c}$ is the predicted probabilities of the c-th label. $w_{i,c}^{(t)}$ denotes the weights of the c-th class at the t-th epoch.

Experimental results in Fig. 3 reveal that the average classification accuracy using FL and IFL for calibration is lower compared to that of ELIMIPL (90.27%). Specifically, FL suffers from under-confidence, where classification accuracy significantly exceeds predicted confidence, while IFL exhibits over-confidence, where predicted confidence significantly surpasses classification accuracy. The results indicate that while FL and IFL can be adapted for MIPL, their effectiveness is constrained, resulting in reduced classification accuracy and notable under-confidence and over-confidence issues. Thus, despite their strong performance in standard supervised learning, applying FL and IFL in MIPL requires further refinement and optimization.

4 The Proposed Approach

The framework of the MIPL approaches based on the embedded-space paradigm is detailed in Fig. 4. First, a feature extractor $\Psi(\cdot)$ generates instance-level feature representations $\boldsymbol{H}_i$ from the i-th multi-instance bag $\boldsymbol{X}_i$. Second, an aggregation mechanism combines $\boldsymbol{H}_i$ into a unified feature vector $\boldsymbol{z}_i$. Subsequently, a classifier estimates the probability distribution $\hat{\boldsymbol{p}}_i$ for each multi-instance bag. Unlike the existing MIPL methods that focus solely on disambiguation, our approach integrates an additional emphasis on model calibration, thereby providing a more comprehensive solution.

4.1 Aggregation Mechanisms in MIPL

In MIPL, aggregation mechanisms play a crucial role in integrating information from multi-instance bags. Given a multi-instance bag $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,n_i}\}$, its instance-level feature representations $\boldsymbol{H}_i = \Psi(\boldsymbol{X}_i) = \{\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \ldots, \boldsymbol{h}_{i,n_i}\}$ are learned via a feature extractor $\Psi(\cdot)$. The significance of the j-th instance within the i-th multi-instance bag is modeled as follows:

$$\xi(h_{i,j}) = \boldsymbol{W}^{\top}(\tanh(\boldsymbol{W}_t^{\top}\boldsymbol{h}_{i,j}) \odot \text{sigm}(\boldsymbol{W}_s^{\top}\boldsymbol{h}_{i,j})), \tag{7}$$

where $\boldsymbol{W}^{\top}$, $\boldsymbol{W}_t^{\top}$, and $\boldsymbol{W}_s^{\top}$ are learnable weight matrices, including bias terms. Here, $\tanh(\cdot)$ and $\text{sigm}(\cdot)$ denote the hyperbolic tangent and sigmoid functions, respectively, and $\odot$ represents element-wise multiplication.

The central aspect of aggregation mechanisms is the computation of attention scores. Existing aggregation mechanisms in MIPL can be categorized into three types: Disambiguation Attention Mechanism (DAM) in DEMIPL [16], Scaled additive Attention Mechanism (SAM) in ELIMIPL [18], and Margin-aware Attention Mechanism (MAM) in MIPLMA [19] and PROMIPL [43]. While all three mechanisms employ attention mechanisms for aggregation, they differ in how attention scores are computed.

1) DAM determines the attention scores by applying a nonlinear transformation to $\xi(h_{i,j})$:

$$a_{i,j} = \frac{1/\{1+\exp\{-\xi(h_{i,j})\}\}}{\sum_{j'=1}^{n_i} 1/\{1+\exp\{-\xi(h_{i,j'})\}\}}. \tag{8}$$

2) SAM refines attention scores by exponentiating $\xi(h_{i,j})$ and introducing a scaling factor l:

$$a_{i,j} = \frac{\exp\{\xi(h_{i,j})/\sqrt{l}\}}{\sum_{j'=1}^{n_i} \exp\{\xi(h_{i,j'})/\sqrt{l}\}}. \tag{9}$$

3) MAM dynamically adjusts attention scores during training by varying the temperature parameter $\tau^{(t)}$:

$$\breve{a}_{i,j} = \frac{\exp\{\xi(h_{i,j})/\tau^{(t)}\}}{\sum_{j'=1}^{n_i} \exp\{\xi(h_{i,j'})/\tau^{(t)}\}}. \tag{10}$$

At the first epoch, the temperature parameter is initialized as a predefined constant and is annealed at the t-th epoch according to:

$$\tau^{(t)} = \max\{\tau_m, \tau^{(t-1)} * 0.95\}, \tag{11}$$

where τ_m denotes the minimum temperature parameter, and $\tau^{(0)}$ is the predefined initial value. To stabilize training, a normalization operation is applied to the attention scores $\breve{a}_{i,j}$:

$$a_{i,j} = \frac{\breve{a}_{i,j} - \bar{a}_i}{\sqrt{\sum_{j'=1}^{n_i}(\breve{a}_{i,j'} - \bar{a}_i)^2/(n_i-1)}}, \tag{12}$$

where $\bar{a}_i = \frac{1}{n_i}\sum_{j'=1}^{n_i} \breve{a}_{i,j'}$ represents the mean attention score for the i-th multi-instance bag.

After computing attention scores using one of the three mechanisms, the bag-level feature $\boldsymbol{z}_i$ is obtained as a weighted sum of the scores and instance-level features:

$$\boldsymbol{z}_i = \sum_{j=1}^{n_i} a_{i,j}\boldsymbol{h}_{i,j}. \tag{13}$$

In summary, all three MIPL attention mechanisms have the same workflow but differ in their computation of the attention scores.

4.2 Calibratable Disambiguation Loss

As illustrated in Fig. 3, applying FL and IFL to MIPL can lead to under-confidence and over-confidence in the predicted probabilities. To tackle these problems, we introduce a novel calibratable disambiguation loss (CDL) specifically designed to balance prediction probabilities.

Definition 1 (Calibratable Disambiguation Loss)**.** *For the i-th training bag, let $\mathcal{S}_i$ denote its candidate label set. At epoch t, we define the current top-ranked candidate label and its corresponding predicted probability as $u_i^{(t)} = \arg\max_{c\in S_i} \hat{p}_{i,c}^{(t)}$ and $q_i^{(t)} = \hat{p}_{i,u_i^{(t)}}^{(t)} = \max_{c\in S_i} \hat{p}_{i,c}^{(t)}$.*

The Calibratable Disambiguation Loss (CDL) for multi-instance partial-label learning is defined as:

$$\mathcal{L}_{CDL} = -\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)}(1 - q_i^{(t)} + \Phi(\hat{\boldsymbol{p}}_i^{(t)}))^{\gamma}\log(\hat{p}_{i,c}^{(t)}), \tag{14}$$

where $w_{i,c}^{(t)}$ denotes the pseudo-label weight for the c-th class at the t-th epoch. γ denotes the exponential factor, and $\Phi(\hat{\boldsymbol{p}}_i^{(t)})$ represents the predictive probability of a competitor.

The modulation $(1 - q_i^{(t)} + \Phi(\hat{\mathbf{p}}_i))^{\gamma}$ can be written as $(1-\beta_i)^{\gamma}$, where $\beta_i = q_i^{(t)} - \Phi(\hat{\boldsymbol{p}}_i^{(t)})$ is a top-vs-competitor margin and is computed from the current predictive distribution. Therefore, $\Phi(\hat{\mathbf{p}}_i^{(t)})$ makes CDL margin-aware: it couples the loss weight to the separability between the top candidate and its competitor, rather than to an absolute confidence value alone. This encourages the model to increase $q_i^{(t)}$ while suppressing $\Phi(\hat{\boldsymbol{p}}_i^{(t)})$, thereby reshaping the predictive distribution toward better-separated and more reliable probabilities.

For non-candidate labels $\bar{c} \in \bar{\mathcal{S}}_i$, the pseudo-label weight $w_{i,\bar{c}}^{(t)}$ is maintained at zero for all epochs $t \in \{1, 2, \ldots, T\}$, where T denotes the total number of training epochs. Conversely, for candidate labels $c \in \mathcal{S}_i$, the weight is initialized as $w_{i,c}^{(1)} = \frac{1}{|\mathcal{S}_i|}$ with $|\cdot|$ representing the set cardinality. For epochs $t \in \{2, 3, \ldots, T\}$, the weight is updated as follows:

$$w_{i,c}^{(t)} = \alpha^{(t)} w_{i,c}^{(t-1)} + (1-\alpha^{(t)})\frac{\hat{p}_{i,c}^{(t)}}{\sum_{c\in\mathcal{S}_i}\hat{p}_{i,c}^{(t)}}, \tag{15}$$

where the parameter $\alpha^{(t)} = \frac{T-t}{T}$ regulates the update rate.

CDL can be instantiated in various ways by adjusting the transformation function $\Phi(\hat{\boldsymbol{p}}_i^{(t)})$. Specifically, we explore two distinct implementations: one where $\Phi(\hat{\boldsymbol{p}}_i^{(t)})$ is the second highest predicted probability among the candidate labels, i.e., $\Phi(\hat{\boldsymbol{p}}_i^{(t)}) = \max_{c'\in\mathcal{S}_i\setminus\{u_i\}} \hat{p}_{i,c'}^{(t)}$, and the corresponding loss function is defined as:

$$\mathcal{L}_{\text{CDL-CC}} = -\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)}(1 - q_i^{(t)} + \max_{c'\in\mathcal{S}_i\setminus\{u_i\}} \hat{p}_{i,c'}^{(t)})^{\gamma}\log(\hat{p}_{i,c}^{(t)}), \tag{16}$$

and another in which $\Phi(\hat{\boldsymbol{p}}_i^{(t)})$ is the maximum predicted probability among non-candidate labels, i.e., $\Phi(\hat{\boldsymbol{p}}_i^{(t)}) = \max_{\bar{c}\in\bar{\mathcal{S}}_i} \hat{p}_{i,\bar{c}}^{(t)}$ is formulated as follows:

$$\mathcal{L}_{\text{CDL-CN}} = -\sum_{c\in\mathcal{S}_i} w_{i,c}^{(t)}(1 - q_i^{(t)} + \max_{\bar{c}\in\bar{\mathcal{S}}_i} \hat{p}_{i,\bar{c}}^{(t)})^{\gamma}\log(\hat{p}_{i,c}^{(t)}). \tag{17}$$

The two CDL instantiations adjust the learning dynamics through the top-vs-competitor margin β_i. When the prediction is ambiguous and the top candidate is not well separated from its competitor, β_i is small and the modulation $(1-\beta_i)^{\gamma}$ stays close to one. In this case, the loss is not down-weighted and the model keeps a strong disambiguation signal, which helps alleviate under-confidence by continuing to sharpen the distribution only after sufficient evidence is accumulated. When the prediction becomes well separated, β_i is large and $(1-\beta_i)^{\gamma}$ decreases, which

Algorithm 1 Pseudo-Code of CDL

Inputs:
$\mathcal{D}$: MIPL training set $\{(\boldsymbol{X}_i, \mathcal{S}_i) \mid 1 \leq i \leq m\}$, where $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,n_i}\}$, $\boldsymbol{x}_{i,j} \in \mathcal{X}$, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{S}_i \subset \mathcal{Y}$, and $\mathcal{Y} = \{1, 2, \ldots, k\}$
T: Total number of training epochs
$\boldsymbol{X}_*$: Unseen multi-instance bag with n_* instances
Outputs:
$\boldsymbol{Y}_*$: Predicted label for $\boldsymbol{X}_* = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,n_*}\}$
Process:

1: Initialize the weights $w_{i,c}^{(1)} = \frac{1}{|\mathcal{S}_i|}$ for the candidate labels
2: **for** $t = 1$ to T **do**
3: Shuffle the training set $\mathcal{D}$ into B mini-batches
4: **for** $b = 1$ to B **do**
5: Learn instance-level features using $\Psi(\cdot)$
6: Compute attention scores based on Eq. (8) for DAM, Eq. (9) for SAM, or Eqs. (10, 11, 12) for MAM
7: Aggregate instance-level features into bag-level features according to Eq. (13)
8: Classify the multi-instance bag and obtain predicted probabilities
9: Update the weights $w_{i,c}^{(t)}$ using Eq. (15)
10: Compute the calibratable disambiguation loss according to Eq. (16) or Eq. (17)
11: Update model parameters with the SGD optimizer
12: **end for**
13: **end for**
14: Learn instance-level features of the unknown multi-instance bag $\boldsymbol{X}_*$ in the test set
15: Compute attention scores and aggregate instance-level features into a single vector $\boldsymbol{z}_*$
16: Return $Y_* = \arg\max_{c \in \mathcal{Y}} \hat{p}_{*,c}$

reduces the incentive to further increase already confident probabilities and limits excessive sharpening, thereby mitigating over-confidence. Moreover, CDL is plug-and-play, enabling seamless integration with existing MIPL approaches. We propose a systematic integration of the two CDL instantiations with three established attention mechanisms: Disambiguation Attention Mechanism (DAM) [16], Scaled additive Attention Mechanism (SAM) [18], and Margin-aware Attention Mechanism (MAM) [19]. This integration yields six variants, i.e., DAMCC, DAMCN, SAMCC, SAMCN, MAMCC, and MAMCN.

4.3 Pseudo-Code of Calibratable Disambiguation Loss

Based on the provided dataset and parameters, Algorithm 1 presents the pseudo-code of our CDL. Initially, the algorithm uniformly initializes the weights of the candidate label set (Step 1). During each epoch, the training data is partitioned into multiple mini-batches (Step 3). The models then extract instance-level features and compute attention scores (Steps 5-6). These features are aggregated into bag-level representations for each mini-batch (Step 7). Subsequently, the model classifies each multi-instance bag and updates the weights of the candidate label set (Steps 8-9). Finally, the calibratable disambiguation loss is computed, and the model parameters are updated (Steps 10-11).

For unseen multi-instance bags, the process starts with extracting instance-level features using $\Psi(\cdot)$ (Step 14). Next, these features are aggregated into a bag-level representation using the attention mechanism DAM, SAM, or MAM (Step 15). Finally, the predicted label is obtained by selecting the category with the highest prediction probability (Step 16).

5 Theoretical Analysis

This section theoretically analyzes the proposed calibratable disambiguation loss (CDL). The analysis is organized around three complementary questions. First, from the loss perspective, we rewrite CDL as a margin-modulated version of the momentum-based disambiguation loss (MDL) and show that CDL preserves an MDL-like fitting signal on ambiguous low-margin bags. Second, from the calibration perspective, we relate population top-label calibration to pseudo-label confidence alignment and then show that the CDL risk controls this alignment under a non-degenerate modulation condition. Third, from the optimization perspective, we show that CDL is not merely a fixed reweighting of MDL: its gradient contains an additional margin-shaping term, whose logit-level effect is to increase the active top-vs-competitor margin, and whose influence is carried into later pseudo-label weights through the momentum update. For notational simplicity, we omit the training-epoch superscript and the explicit dependence on model parameters unless needed.

5.1 Notation and Margin-Modulated Form of CDL

Let $(\boldsymbol{X}_i, \mathcal{S}_i)$ denote the i-th MIPL example, where $\boldsymbol{X}_i$ is a multi-instance bag and $\mathcal{S}_i \subseteq \mathcal{Y} = \{1, 2, \ldots, k\}$ is its candidate label set. Let $Y_i \in \mathcal{Y}$ be the latent ground-truth label. The bag-level predictive distribution, latent posterior, and pseudo-label weight vector are denoted by

$$\begin{aligned} \hat{\boldsymbol{p}}_i &= \big(\hat{p}_{i,1}, \hat{p}_{i,2}, \ldots, \hat{p}_{i,k}\big) \in \Delta^{k-1}, \\ \boldsymbol{\eta}_i &= \big(\eta_{i,1}, \eta_{i,2}, \ldots, \eta_{i,k}\big) \in \Delta^{k-1}, \\ \boldsymbol{w}_i &= \big(w_{i,1}, w_{i,2}, \ldots, w_{i,k}\big) \in \Delta^{k-1}, \end{aligned} \tag{18}$$

where

$$\eta_{i,c} = \mathbb{P}(Y_i = c \mid \boldsymbol{X}_i, \mathcal{S}_i), \qquad \eta_{i,c} = w_{i,c} = 0 \text{ for all } c \notin \mathcal{S}_i. \tag{19}$$

For each bag, define the top candidate label and its probability as follows:

$$u_i = \arg\max_{c \in \mathcal{S}_i} \hat{p}_{i,c}, \qquad q_i = \hat{p}_{i,u_i} = \max_{c \in \mathcal{S}_i} \hat{p}_{i,c}. \tag{20}$$

For CDL-CC, the competitor is the second strongest candidate label, i.e., $\phi_{i,\mathrm{CC}} = \max_{c' \in \mathcal{S}_i \setminus \{u_i\}} \hat{p}_{i,c'}$. For CDL-CN, the competitor is the strongest non-candidate label, i.e., $\phi_{i,\mathrm{CN}} = \max_{\bar{c} \in \bar{\mathcal{S}}_i} \hat{p}_{i,\bar{c}}$. For either CDL instantiation, write $\phi_i = \Phi(\hat{\boldsymbol{p}}_i)$, and define the top-vs-competitor margin and the CDL modulation factor by

$$\beta_i = q_i - \phi_i, \qquad \lambda_i = \big(1 - \beta_i\big)^{\gamma}. \tag{21}$$

The per-bag MDL objective induced by the current pseudo-label weights is

$$\ell_i^{\mathrm{MDL}} = -\sum_{c \in \mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c}. \tag{22}$$

With the notation in Eq. (21), the per-bag CDL objective can be written as a margin-modulated MDL objective:

$$\ell_i^{\mathrm{CDL}} = \lambda_i \ell_i^{\mathrm{MDL}} = \left(1-\beta_i\right)^{\gamma} \left(-\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c} \right). \quad (23)$$

Because $\boldsymbol{w}_i$ is supported on $\mathcal{S}_i$, the MDL objective admits the following KL-entropy decomposition:

$$\begin{aligned} \ell_i^{\mathrm{MDL}} &= \mathrm{KL}(\boldsymbol{w}_i \| \hat{\boldsymbol{p}}_i) + \mathbb{H}(\boldsymbol{w}_i)\,, \\ \mathbb{H}(\boldsymbol{w}_i) &= -\sum_{c\in\mathcal{S}_i} w_{i,c} \log w_{i,c}. \end{aligned} \quad (24)$$

For a training set of m bags, define the empirical MDL and CDL objectives by

$$\mathcal{L}_{\mathrm{MDL}} = \frac{1}{m}\sum_{i=1}^{m} \ell_i^{\mathrm{MDL}}, \qquad \mathcal{L}_{\mathrm{CDL}} = \frac{1}{m}\sum_{i=1}^{m} \ell_i^{\mathrm{CDL}}. \quad (25)$$

5.2 Linear Lower Bound of CDL

The first result formalizes the loss-level connection between CDL and MDL. It shows that the modulation does not remove the MDL fitting signal on low-margin bags, which is important for preserving disambiguation while allowing the modulation to down-weight already separated predictions.

Theorem 1 (Linear lower bound of CDL)**.** *Let $\gamma \geq 1$. For the i-th training bag, assume that $\hat{\boldsymbol{p}}_i \in \Delta^{k-1}$, $\hat{p}_{i,c} > 0$ for all $c \in \mathcal{Y}$, and $\boldsymbol{w}_i \in \Delta^{k-1}$ with $w_{i,c} = 0$ for all $c \notin \mathcal{S}_i$. Let $u_i = \arg\max_{c\in\mathcal{S}_i} \hat{p}_{i,c}$, $q_i = \hat{p}_{i,u_i} = \max_{c\in\mathcal{S}_i} \hat{p}_{i,c}$. Let*

$$\ell_i^{\mathrm{MDL}} = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c}, \qquad \ell_i^{\mathrm{CDL}} = \lambda_i \ell_i^{\mathrm{MDL}}. \quad (26)$$

Then, for each training bag,

$$\ell_i^{\mathrm{CDL}} \geq (1-\gamma\beta_i)\ell_i^{\mathrm{MDL}} = (1-\gamma\beta_i)\left[\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) + \mathbb{H}(\boldsymbol{w}_i)\right]. \quad (27)$$

Consequently,

$$\mathcal{L}_{\mathrm{CDL}} \geq \frac{1}{m}\sum_{i=1}^{m} (1-\gamma\beta_i)\ell_i^{\mathrm{MDL}}. \quad (28)$$

Moreover, if $\beta_{\max} = \max_{1\leq i\leq m} \beta_i$, then

$$\mathcal{L}_{\mathrm{CDL}} \geq (1-\gamma\beta_{\max})\mathcal{L}_{\mathrm{MDL}}. \quad (29)$$

Theorem 1 shows that CDL retains an MDL-like disambiguation when the margin β_i is small. When $\beta_i > 0$ is large, the modulation $\lambda_i = (1-\beta_i)^{\gamma}$ reduces the loss contribution of already well-separated bags. For CDL-CN, the case $\phi_{i,\mathrm{CN}} > q_i$ gives $\beta_i < 0$, and CDL intentionally amplifies the loss $\phi_{i,\mathrm{CN}} > q_i \Longrightarrow \lambda_i = \left(1+\phi_{i,\mathrm{CN}}-q_i\right)^{\gamma} > 1$. Therefore, CDL-CC mainly suppresses separated candidate-level predictions, while CDL-CN can additionally penalize bags whose non-candidate probabilities dominate the top candidate probability. Having established that CDL remains tied to the MDL fitting objective, we next connect this pseudo-label fitting view to the population top-label calibration.

5.3 Population Top-Label Calibration and Pseudo-Label Approximation

We use a generic MIPL draw $(\boldsymbol{X}_i, \mathcal{S}_i, Y_i)$ to keep the population analysis consistent with the empirical notation. The top-label prediction and its confidence score are defined over the full label space by

$$\hat{y}_i = \arg\max_{c\in\mathcal{Y}} \hat{p}_{i,c}, \qquad C_i = \hat{p}_{i,\hat{y}_i} = \max_{c\in\mathcal{Y}} \hat{p}_{i,c}. \quad (30)$$

Here $\hat{y}_i$ is used for calibration and may differ from the top candidate label used by CDL. The population top-label calibration error is

$$E_{\mathrm{cal}} = \mathbb{E}\left[\left|\mathbb{E}[\mathbb{I}\{Y_i = \hat{y}_i\} \mid C_i] - C_i\right|\right]. \quad (31)$$

The empirical ECE in Eq. (2) is a binned finite-sample approximation of Eq. (31). To connect calibration with pseudo-label learning, define

$$\begin{aligned} E_{\mathrm{conf}} &= \mathbb{E}\left[\left|\eta_{i,\hat{y}_i} - C_i\right|\right], \\ E_{\mathrm{pconf}} &= \mathbb{E}\left[\left|w_{i,\hat{y}_i} - C_i\right|\right], \\ \delta_w^{\mathrm{TV}} &= \mathbb{E}\left[d_{\mathrm{TV}}(\boldsymbol{\eta}_i, \boldsymbol{w}_i)\right]. \end{aligned} \quad (32)$$

The total variation distance is

$$d_{\mathrm{TV}}(\boldsymbol{\eta}_i, \boldsymbol{w}_i) = \frac{1}{2}\|\boldsymbol{\eta}_i - \boldsymbol{w}_i\|_1. \quad (33)$$

Proposition 1 (Calibration is controlled by pseudo-label confidence error)**.** *Fix the current training state. Assume that $\hat{\boldsymbol{p}}_i$ and $\boldsymbol{w}_i$ are measurable with respect to $\mathcal{G}_i = \sigma(\boldsymbol{X}_i, \mathcal{S}_i)$. Then*

$$E_{\mathrm{cal}} \leq E_{\mathrm{conf}} \leq E_{\mathrm{pconf}} + \delta_w^{\mathrm{TV}}. \quad (34)$$

Proposition 1 provides an upper-bound decomposition of the calibration error. Specifically, E_{cal} is controlled by the pseudo-label confidence error E_{pconf}, which reflects the discrepancy between the model confidence and the assigned pseudo-label weight, and the posterior approximation error δ_w^{TV}, which measures the quality of the pseudo-label weights as approximations to the true posterior distribution. Thus, this result identifies the optimizable component that CDL can influence and separates it from the quality of the pseudo-label approximation. The following theorem then links this optimizable component directly to the CDL risk.

5.4 Confidence Alignment Bound for CDL

We next show that CDL controls E_{pconf} under a non-degenerate modulation condition. For the same generic bag $\boldsymbol{X}_i$, the corresponding MDL and CDL losses are shown as Eqs. (22) and (23). Define the population risks and the expected pseudo-label entropy as

$$R_{\mathrm{MDL}}^{w} = \mathbb{E}\left[\ell_i^{\mathrm{MDL}}\right], \quad R_{\mathrm{CDL}}^{w} = \mathbb{E}\left[\ell_i^{\mathrm{CDL}}\right], \quad \mathcal{H}_w = \mathbb{E}\left[\mathbb{H}(\boldsymbol{w}_i)\right], \quad (35)$$

where

$$\mathbb{H}(\boldsymbol{w}_i) = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log w_{i,c}. \quad (36)$$

Theorem 2 (Pseudo-label confidence alignment bound)**.** *Assume that $\hat{\boldsymbol{p}}_i$ is induced by finite logits, so that $\hat{p}_{i,c} > 0$ for all $c \in \mathcal{Y}$. Assume further that the relevant competitor set in CDL is nonempty and that the CDL modulation satisfies $\lambda_i \geq \lambda_0 > 0$. Then*

$$E_{\mathrm{pconf}} \leq \sqrt{2\left(\lambda_0^{-1} R_{\mathrm{CDL}}^{w} - \mathcal{H}_w\right)}. \quad (37)$$

Consequently,

$$E_{\mathrm{cal}} \le E_{\mathrm{conf}} \le \sqrt{2\left(\lambda_0^{-1} R^{w}_{\mathrm{CDL}} - \mathcal{H}_w\right)} + \delta_w^{\mathrm{TV}}. \tag{38}$$

Theorem 2 is conditional on the current pseudo-label weights. If $\boldsymbol{w}_i$ is close to the latent posterior $\boldsymbol{\eta}_i$, then δ_w^{TV} is small and controlling the CDL risk improves confidence alignment. If the pseudo-label weights are inaccurate, the bound remains valid but can be loose. The assumption $\lambda_i \ge \lambda_0 > 0$ is a non-degeneracy condition preventing the CDL modulation from vanishing. The result explains why reducing CDL can improve the optimizable calibration component, while Proposition 1 clarifies that the final calibration bound also depends on the posterior quality of the pseudo-label weights. We next move from this risk-level view to the local optimization dynamics induced by the margin-dependent modulation.

5.5 Optimization-Level Margin Shaping and Momentum Dynamics

Although Eq. (23) writes CDL as a margin-modulated MDL objective, the modulation factor depends on the current prediction and hence on the model parameters. Therefore, CDL is not merely a fixed sample reweighting of MDL. This subsection makes this distinction explicit at the optimization level. We first decompose the CDL gradient into an MDL fitting term and an additional margin-shaping term, then interpret the latter in the softmax-logit space, and finally show how prediction-level margins are recursively transferred to future pseudo-label margins through the momentum update. Throughout this subsection, ∇ denotes the gradient with respect to the model parameters.

Proposition 2 (Gradient decomposition of CDL on differentiable regions)**.** *Fix a training bag $(\boldsymbol{X}_i, \mathcal{S}_i)$, and let θ denote the model parameters. Suppose that there is an open parameter region $\mathcal{U}$ on which each $\hat{p}_{i,c}(\theta)$ is differentiable and strictly positive. Assume that the top candidate label $u_i = \arg\max_{c\in\mathcal{S}_i} \hat{p}_{i,c}(\theta)$ is uniquely attained and remains unchanged on $\mathcal{U}$. For CDL-CC, let $\mathcal{C}_i = \mathcal{S}_i \setminus \{u_i\}$; for CDL-CN, let $\mathcal{C}_i = \bar{\mathcal{S}}_i$. Assume that $\mathcal{C}_i \neq \varnothing$ and that the competitor $v_i = \arg\max_{c\in\mathcal{C}_i} \hat{p}_{i,c}(\theta)$ is also uniquely attained and remains unchanged on $\mathcal{U}$. Define*

$$q_i(\theta) = \hat{p}_{i,u_i}(\theta), \qquad \phi_i(\theta) = \hat{p}_{i,v_i}(\theta), \tag{39}$$

$$\beta_i(\theta) = q_i(\theta) - \phi_i(\theta), \qquad \lambda_i(\theta) = \big(1-\beta_i(\theta)\big)^{\gamma}. \tag{40}$$

During the current gradient computation, regard the pseudo-label weights $\boldsymbol{w}_i$ as fixed, and write

$$\ell_i^{\mathrm{MDL}}(\theta) = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c}(\theta), \quad \ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\ell_i^{\mathrm{MDL}}(\theta). \tag{41}$$

Then, for every $\theta \in \mathcal{U}$,

$$\begin{aligned}\nabla_\theta \ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\nabla_\theta \ell_i^{\mathrm{MDL}}(\theta) \\ - \gamma\big(1-\beta_i(\theta)\big)^{\gamma-1}\ell_i^{\mathrm{MDL}}(\theta)\nabla_\theta\beta_i(\theta).\end{aligned} \tag{42}$$

Proposition 2 is the basic differential identity behind the optimization-level effect of CDL. The first term in (42) is the MDL fitting term scaled by the current modulation factor, whereas the second term is absent from MDL. In a gradient-descent step, this second term contributes the update component

$$\delta_\beta = \eta\gamma\big(1-\beta_i(\theta)\big)^{\gamma-1}\ell_i^{\mathrm{MDL}}(\theta)\nabla_\theta\beta_i(\theta), \tag{43}$$

where $\eta > 0$ is the step size. Therefore, whenever $\nabla_\theta\beta_i(\theta) \neq 0$, a first-order Taylor expansion gives

$$\beta_i(\theta+\delta_\beta) - \beta_i(\theta) = \eta\gamma\big(1-\beta_i(\theta)\big)^{\gamma-1}\ell_i^{\mathrm{MDL}}(\theta)\|\nabla_\theta\beta_i(\theta)\|^2 + o(\eta). \tag{44}$$

Therefore, the additional CDL component explicitly promotes an increase of the active top-vs-competitor margin. At exact ties, the max-based margin is not differentiable; the differentiable statement applies on each region with a fixed active top candidate and competitor. The following corollary translates this abstract parameter-space margin term into a concrete logit-space effect.

Corollary 1 (Logit-space effect of margin shaping)**.** *Fix a training bag i. Let $s_{i,c} \in \mathbb{R}$ be the logit of class c, and let*

$$\hat{p}_{i,c} = \frac{\exp(s_{i,c})}{\sum_{a\in\mathcal{Y}}\exp(s_{i,a})}, \qquad c \in \mathcal{Y}. \tag{45}$$

Assume that, in a neighborhood of the current logits, the top candidate u and the competitor v are unique, distinct, and fixed. Then $\beta_i = \hat{p}_{i,u} - \hat{p}_{i,v}$ is differentiable in this neighborhood, and

$$\frac{\partial\beta_i}{\partial s_{i,u}} = \hat{p}_{i,u}(1-\hat{p}_{i,u}+\hat{p}_{i,v}) > 0, \tag{46}$$

$$\frac{\partial\beta_i}{\partial s_{i,v}} = -\hat{p}_{i,v}(1+\hat{p}_{i,u}-\hat{p}_{i,v}) < 0, \tag{47}$$

$$\frac{\partial\beta_i}{\partial s_{i,c}} = \hat{p}_{i,c}(\hat{p}_{i,v}-\hat{p}_{i,u}), \quad c \notin \{u,v\}. \tag{48}$$

Thus a positive step along $\nabla_{s_i}\beta_i$ locally increases the top-candidate logit and decreases the active-competitor logit.

Corollary 1 provides the logit-space interpretation of the margin-shaping term in Proposition 2. It shows that increasing the active margin locally raises the top-candidate logit and suppresses the active-competitor logit. The preceding two results describe the current optimization step. However, the pseudo-label weights are updated across epochs by a momentum rule in MIPL. The following lemma records how candidate-level prediction margins are transferred into future pseudo-label margins. It is stated for two candidate labels $u, v \in \mathcal{S}_i$, because the pseudo-label weight is supported on the candidate set. For CDL-CN, the non-candidate competitor affects this recursion indirectly through the candidate-normalized predictions.

Lemma 1 (Momentum recursion for candidate pseudo-label margins)**.** *Fix a training bag $(\boldsymbol{X}_i, \mathcal{S}_i)$ and two candidate labels $u, v \in \mathcal{S}_i$. For $t = 2, \dots, T$, define*

$$\tilde{p}_{i,c}^{(t)} = \frac{\hat{p}_{i,c}^{(t)}}{\sum_{a\in\mathcal{S}_i}\hat{p}_{i,a}^{(t)}}, \quad M_{i,uv}^{(t)} = w_{i,u}^{(t)} - w_{i,v}^{(t)}, \quad \Delta_{i,uv}^{(t)} = \tilde{p}_{i,u}^{(t)} - \tilde{p}_{i,v}^{(t)}. \tag{49}$$

Assume that the pseudo-label weights are updated by

$$w_{i,c}^{(t)} = \alpha^{(t)} w_{i,c}^{(t-1)} + \big(1-\alpha^{(t)}\big)\tilde{p}_{i,c}^{(t)}, \quad c \in \mathcal{S}_i. \tag{50}$$

Then, for every $t = 2, \dots, T$,

$$M_{i,uv}^{(t)} = \alpha^{(t)} M_{i,uv}^{(t-1)} + \big(1-\alpha^{(t)}\big)\Delta_{i,uv}^{(t)}. \tag{51}$$

Equivalently, with $A_{a:b} = \prod_{\tau=a}^{b} \alpha^{(\tau)}$ *and the empty product defined as one,*

$$M_{i,uv}^{(t)} = A_{2:t} M_{i,uv}^{(1)} + \sum_{s=2}^{t} \left(1 - \alpha^{(s)}\right) A_{s+1:t} \Delta_{i,uv}^{(s)}. \tag{52}$$

Moreover,

$$\Delta_{i,uv}^{(s)} = \frac{\hat{p}_{i,u}^{(s)} - \hat{p}_{i,v}^{(s)}}{\sum_{a \in \mathcal{S}_i} \hat{p}_{i,a}^{(s)}}. \tag{53}$$

Thus past candidate prediction margins enter the current pseudo-label margin through the momentum coefficients. If $\alpha^{(s)} \in [0, 1]$ *for all* s*, then* $M_{i,uv}^{(t)}$ *is a convex combination of* $M_{i,uv}^{(1)}$ *and* $\Delta_{i,uv}^{(2)}, \ldots, \Delta_{i,uv}^{(t)}$.

Lemma 1 complements the local gradient analysis by showing that the margins shaped at the prediction level are not transient. They enter the subsequent pseudo-label weights through the momentum rule. This completes the optimization-level explanation of CDL.

Taken together, the above results provide a coherent account of CDL from three complementary perspectives. At the loss level, the margin-modulated form in Eq. (23) and Theorem 1 relate CDL to MDL and clarify why low-margin bags still receive an MDL-like disambiguation signal. At the calibration level, Proposition 1 separates the calibration error into a pseudo-label confidence-alignment term and a pseudo-label approximation term, while Theorem 2 connects the former to the CDL risk. At the optimization level, Proposition 2, Corollary 1, and Lemma 1 explain how the margin-dependent term reshapes the active top-vs-competitor margin and how this effect is propagated to subsequent pseudo-label updates.

These theoretical implications motivate the empirical evaluation in the next section. In particular, the analysis suggests that CDL should preserve the disambiguation ability of MDL, mitigate excessive confidence growth through margin modulation, and transmit margin information across epochs through the momentum update. Therefore, we evaluate CDL on benchmark and real-world MIPL datasets in terms of both classification accuracy and calibration quality.

6 Experiments

6.1 Experimental Configurations

6.1.1 Datasets

We adopt the experimental protocol established by the prior research [16], which includes a diverse range of datasets, including four benchmark datasets and seven real-world datasets. Specifically, the benchmark datasets are MNIST-MIPL, FMNIST-MIPL, Birdsong-MIPL, and SIVAL-MIPL, which span various domains from image analysis to biological data [36], [58], [59], [60]. Additionally, we include CRC-MIPL, a real-world dataset for colorectal cancer classification. The candidate label sets are curated by professionally trained crowdsourcing workers. In the prior research [16], [18], four types of multi-instance features have been employed across different sub-datasets: CRC-MIPL-Row (C-Row), CRC-MIPL-SBN (C-SBN), CRC-MIPL-KMeansSeg (C-KMeans), and CRC-MIPL-SIFT (C-SIFT). These features are generated by four image bag generators [61], including Row, single blob with neighbors (SBN), k-means segmentation (KMeansSeg), and scale-invariant feature transform (SIFT).

Furthermore, a novel feature extraction strategy is proposed for the CRC-MIPL dataset by leveraging ResNet-34. Each image is partitioned into N non-overlapping patches, and ResNet-34 is applied to obtain feature representations for each patch. Three configurations are investigated, with N set to 9, 16, and 25, resulting in the new datasets C-R34-9, C-R34-16, and C-R34-25, respectively. Notably, the C-R34-9 dataset is first introduced in this work. Table 1 provides further details on these datasets. More information on the benchmark datasets and the CRC-MIPL dataset can be found in the literature [13] and [16], respectively.

6.1.2 Comparative Methods

To our knowledge, there are six available methods relevant to our MIPL settings, namely MIPLGP [13], DEMIPL [16], ELIMIPL [18], MIPLMA [19], PROMIPL [43], and FASTMIPL [44]. All corresponding code implementations for these methods are publicly available. We systematically compare our six methods against the six MIPL methods to evaluate their classification accuracy and calibration performance. The parameters for all compared methods were set following the recommendations from the original literature or were further optimized to enhance performance.

6.1.3 Implementation Details

Our algorithm is implemented using PyTorch and executed on an NVIDIA Tesla V100 GPU. We utilize stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001 for optimization. We adopt the same feature extractor $\Psi(\cdot)$ as used in previous MIPL methods [16], [18], [19], [43], [44]. For MNIST-MIPL and FMNIST-MIPL, we extract instance-level features using a two-layer convolutional neural network followed by a fully connected network. For preprocessed datasets, i.e., Birdsong-MIPL and SIVAL-MIPL, we use a fully connected network for feature extraction. For CRC-MIPL, we explore four image bag generators and use ResNet-34 to extract instance-level features. We select the learning rate from $\{0.01, 0.05\}$ and adjust it using cosine annealing. DAMCC and DAMCN are trained for 200 epochs on CRC-MIPL and 100 epochs on benchmark datasets. Meanwhile, SAMCC, SAMCN, MAMCC, and MAMCN are trained for 100 epochs on all datasets. We set the CDL parameter γ to 1 in all experiments. We report the mean and standard deviation (std) of classification accuracy and expected calibration error over ten random train/test splits with a ratio of 7:3. In our experiments, the expected calibration error is evaluated with 15 bins. For each dataset, the highest classification accuracy and the lowest expected calibration error are highlighted in bold to facilitate comparison. In the following tables, $\uparrow$ and $\downarrow$ denote improvements and reductions over DEMIPL [16], ELIMIPL [18], or MIPLMA [19], respectively. The code of this work can be found at https://github.com/tangw-seu/MIPLCDL.

6.2 Classification Performance

6.2.1 Accuracy on the Benchmark Datasets

Table 2 demonstrates that our methods outperform the comparative methods in 67 of the 72 cases. In 12 of the

TABLE 1: Characteristics of the benchmark and real-world MIPL datasets. The C-R34-9 is introduced in this work as a new MIPL dataset. "n/a" is abbreviated for not applicable as the instance-level labels of the CRC-MIPL dataset are unknown.

Dataset	#bag	#ins	max. #ins	min. #ins	avg. #ins	max. #pos	min. #pos	avg. #pos	#dim	#C	avg. #CLs
MNIST-MIPL	500	20664	48	35	41.33	9.1%	7.0%	8.0%	784	5	2, 3, 4
FMNIST-MIPL	500	20810	48	36	41.62	9.1%	7.0%	8.0%	784	5	2, 3, 4
Birdsong-MIPL	1300	48425	76	25	37.25	10.0%	6.9%	8.3%	38	13	2, 3, 4
SIVAL-MIPL	1500	47414	32	31	31.61	90.62%	3.12%	25.6%	30	25	2, 3, 4
C-Row	7000	56000	8	8	8	n/a	n/a	n/a	9	7	2.08
C-SBN	7000	63000	9	9	9	n/a	n/a	n/a	15	7	2.08
C-KMeans	7000	30178	6	3	4.311	n/a	n/a	n/a	6	7	2.08
C-SIFT	7000	175000	25	25	25	n/a	n/a	n/a	128	7	2.08
C-R34-9	7000	63000	9	9	9	n/a	n/a	n/a	1000	7	2.08
C-R34-16	7000	112000	16	16	16	n/a	n/a	n/a	1000	7	2.08
C-R34-25	7000	175000	25	25	25	n/a	n/a	n/a	1000	7	2.08

TABLE 2: Classification accuracy (mean±std%) on the benchmark datasets with varying numbers of false-positive labels.

	MNIST-MIPL	FMNIST-MIPL	Birdsong-MIPL	SIVAL-MIPL
		$r = 1$		
MIPLGP	94.87±1.55	84.67±2.98	71.62±2.59	66.87±1.95
PROMIPL	99.92±0.29	92.20±2.39	77.63±1.50	68.20±3.22
FASTMIPL	99.91±0.23	90.99±2.19	79.60±2.35	**77.65±3.02**
DEMIPL	97.60±0.80	88.01±2.08	74.36±1.57	63.53±4.10
DAMCC (ours)	99.33±0.52 (1.73 ↑)	92.07±2.18 (4.06 ↑)	79.03±1.49 (4.67 ↑)	73.87±2.25 (10.34 ↑)
DAMCN (ours)	99.40±0.55 (1.80 ↑)	91.87±1.90 (3.86 ↑)	**80.38±2.11** (6.02 ↑)	75.49±1.99 (11.96 ↑)
ELIMIPL	99.20±0.65	90.27±1.84	77.13±1.80	67.49±2.18
SAMCC (ours)	99.80±0.43 (0.60 ↑)	90.27±1.98 (0.00 ↑)	79.21±2.26 (2.08 ↑)	71.96±2.34 (4.47 ↑)
SAMCN (ours)	99.73±0.44 (0.53 ↑)	90.33±2.09 (0.06 ↑)	79.54±1.97 (2.41 ↑)	73.64±1.75 (6.15 ↑)
MIPLMA	98.47±1.03	91.53±1.61	77.56±2.05	70.33±2.63
MAMCC (ours)	**99.93±0.20** (1.46 ↑)	92.73±1.21 (1.20 ↑)	79.54±1.59 (1.98 ↑)	72.40±1.86 (2.07 ↑)
MAMCN (ours)	**99.93±0.20** (1.46 ↑)	**93.20±1.33** (1.67 ↑)	79.95±1.23 (2.39 ↑)	74.40±2.21 (4.07 ↑)
		$r = 2$		
MIPLGP	81.73±3.01	79.07±2.74	67.18±1.50	61.31±2.61
PROMIPL	99.86±0.24	88.83±2.26	71.84±1.96	63.35±2.32
FASTMIPL	99.77±0.41	**90.15±2.53**	**78.91±2.29**	70.72±2.67
DEMIPL	94.27±2.72	82.33±2.85	70.13±2.40	55.40±5.09
DAMCC (ours)	99.40±0.81 (5.13 ↑)	89.33±2.88 (7.00 ↑)	77.03±1.64 (6.90 ↑)	70.24±2.39 (14.84 ↑)
DAMCN (ours)	99.53±0.60 (5.26 ↑)	89.40±2.69 (7.07 ↑)	78.00±1.54 (7.87 ↑)	**71.93±1.82** (16.53 ↑)
ELIMIPL	98.67±0.99	84.53±2.56	74.46±1.53	61.58±2.54
SAMCC (ours)	98.80±0.83 (0.13 ↑)	85.33±3.24 (0.80 ↑)	77.21±2.03 (2.75 ↑)	66.02±2.79 (4.44 ↑)
SAMCN (ours)	98.93±0.85 (0.26 ↑)	85.40±2.88 (0.87 ↑)	77.54±2.20 (3.08 ↑)	66.87±2.32 (5.29 ↑)
MIPLMA	97.87±1.39	86.67±2.76	76.15±1.55	66.82±3.10
MAMCC (ours)	99.80±0.31 (1.93 ↑)	89.47±1.86 (2.80 ↑)	77.36±2.13 (1.21 ↑)	69.91±2.34 (3.09 ↑)
MAMCN (ours)	**99.87±0.27** (2.00 ↑)	90.07±1.50 (3.40 ↑)	78.46±1.92 (2.31 ↑)	71.42±1.75 (4.60 ↑)
		$r = 3$		
MIPLGP	62.13±6.39	67.00±5.24	62.51±1.47	56.93±3.21
PROMIPL	78.32±11.64	65.89±4.14	69.33±2.11	53.88±2.41
FASTMIPL	97.43±7.37	**81.60±7.01**	77.30±2.24	61.49±3.48
DEMIPL	70.93±8.83	65.73±2.48	69.62±2.39	50.31±1.84
DAMCC (ours)	85.27±9.06 (14.37 ↑)	72.87±7.20 (7.14 ↑)	76.90±1.73 (7.28 ↑)	66.76±2.40 (16.45 ↑)
DAMCN (ours)	78.73±15.67 (7.80 ↑)	74.60±4.63 (8.87 ↑)	77.69±1.75 (8.07 ↑)	68.89±1.94 (18.58 ↑)
ELIMIPL	74.80±14.41	70.20±5.54	71.67±1.67	60.02±2.89
SAMCC (ours)	88.40±9.09 (13.60 ↑)	72.53±7.86 (2.33 ↑)	75.13±2.17 (3.46 ↑)	61.62±1.93 (1.60 ↑)
SAMCN (ours)	86.13±9.48 (11.33 ↑)	74.27±4.09 (4.07 ↑)	77.05±1.46 (5.38 ↑)	63.16±1.94 (3.14 ↑)
MIPLMA	74.93±10.32	65.40±5.53	74.56±1.29	62.73±2.38
MAMCC (ours)	95.07±7.26 (20.14 ↑)	74.00±7.57 (8.60 ↑)	77.18±1.57 (2.62 ↑)	67.20±2.73 (4.47 ↑)
MAMCN (ours)	**98.00±1.49** (23.07 ↑)	77.73±4.95 (12.33 ↑)	**77.82±1.32** (3.26 ↑)	**69.13±1.47** (6.40 ↑)

72 cases, the improvement exceeds 10%. FASTMIPL outperforms our models in 4 of the 72 cases. We speculate that FASTMIPL benefits from training all samples in a single batch, leveraging greater computational resources.

For MNIST-MIPL, the potential for accuracy improvement is limited for $r = 1$ and 2. However, for $r = 3$, DAMCC and SAMCC achieve significant accuracy improvements of 14.37% and 13.60%, respectively. Notably, MAMCC and MAMCN surpass MIPLMA by 20.14% and 23.07% on MNIST-MIPL with $r = 3$. On the FMNIST-MIPL and Birdsong-MIPL datasets, DAMCN achieves improvements exceeding 8%. Furthermore, on SIVAL-MIPL, both DAMCC and DAMCN achieve gains exceeding 10% for $r \in \{1, 2, 3\}$, with a peak improvement of 18.58%. As r increases from 1 to 3, the mean accuracy improvement across benchmark datasets is 3.27%, 4.57%, and 8.93%. These results indicate that CDL effectively handles more challenging scenarios.

6.2.2 Accuracy on the Real-World Datasets

The experimental results in Table 3 demonstrate that our methods consistently achieve superior classification accuracy compared to all baseline methods. Some MIPLGP results are marked as "–" due to computational limitations, which hinder evaluation on certain datasets. This limitation

TABLE 3: Classification accuracy (mean±std%) on the real-world datasets. – indicates computational constraints.

	C-Row	C-SBN	C-KMeans	C-SIFT	C-R34-9	C-R34-16	C-R34-25
MIPLGP	43.13±0.57	33.49±0.63	32.90±1.25	–	–	–	–
PROMIPL	43.51±0.93	51.56±1.23	56.52±1.30	56.25±1.11	60.62±1.02	64.63±0.75	67.24±0.80
FASTMIPL	48.68±3.81	57.21±3.08	57.30±1.28	52.55±2.95	56.78±1.61	61.85±1.48	64.24±1.64
DEMIPL	40.78±1.01	48.58±1.40	52.11±1.20	53.17±1.27	58.84±1.40	62.34±0.82	65.19±0.85
DAMCC (ours)	48.10±0.60 (7.32 ↑)	56.54±0.84 (7.96 ↑)	62.78±1.30 (10.67 ↑)	53.59±1.05 (0.42 ↑)	62.13±1.21 (3.29 ↑)	63.99±0.87 (1.65 ↑)	65.34±0.74 (0.15 ↑)
DAMCN (ours)	**49.06±0.71** (8.28 ↑)	**57.91±0.60** (9.33 ↑)	**64.78±1.02** (12.67 ↑)	55.34±1.05 (2.17 ↑)	63.74±0.94 (4.90 ↑)	64.97±1.00 (2.63 ↑)	66.64±0.85 (1.45 ↑)
ELIMIPL	43.26±0.82	50.90±0.79	54.58±1.18	54.05±1.02	60.97±1.22	63.08±0.66	66.50±0.67
SAMCC (ours)	46.58±0.57 (3.32 ↑)	55.91±0.95 (5.01 ↑)	62.04±1.78 (7.46 ↑)	**57.20±0.94** (3.15 ↑)	63.95±0.98 (2.98 ↑)	67.41±0.70 (4.33 ↑)	69.43±0.94 (2.93 ↑)
SAMCN (ours)	47.50±0.54 (4.24 ↑)	57.24±0.90 (6.34 ↑)	63.84±1.01 (9.26 ↑)	57.10±0.76 (3.05 ↑)	64.58±1.07 (3.61 ↑)	67.81±0.72 (4.73 ↑)	69.75±0.83 (3.25 ↑)
MIPLMA	44.37±0.99	52.46±0.70	55.73±1.01	55.29±0.94	59.38±1.24	63.12±0.77	68.70±1.08
MAMCC (ours)	47.66±0.57 (3.29 ↑)	56.83±1.06 (4.37 ↑)	57.60±1.28 (1.87 ↑)	56.97±0.97 (1.68 ↑)	**64.74±0.98** (5.36 ↑)	**67.88±0.85** (4.76 ↑)	**69.83±0.96** (1.13 ↑)
MAMCN (ours)	47.99±0.50 (3.62 ↑)	57.40±0.94 (4.94 ↑)	59.02±0.96 (3.29 ↑)	57.07±0.86 (1.78 ↑)	64.73±1.02 (5.35 ↑)	67.72±0.78 (4.60 ↑)	69.81±1.00 (1.11 ↑)

TABLE 4: Expected calibration error (mean±std%) on the benchmark datasets with varying numbers of false-positive labels.

	MNIST-MIPL	FMNIST-MIPL	Birdsong-MIPL	SIVAL-MIPL
		$r = 1$		
PROMIPL	59.62±0.16	53.56±2.20	60.77±1.52	59.46±3.12
FASTMIPL	3.47±0.54	5.53±1.35	7.15±1.72	**10.90±2.11**
DEMIPL	59.40±0.77	51.50±1.54	57.86±1.16	56.89±3.01
DAMCC (ours)	1.08±0.41 (58.32 ↓)	5.09±1.05 (46.41 ↓)	4.68±1.10 (53.18 ↓)	12.74±1.63 (44.15 ↓)
DAMCN (ours)	1.11±0.38 (58.29 ↓)	5.44±1.49 (46.06 ↓)	7.53±1.63 (50.33 ↓)	16.25±1.71 (40.64 ↓)
ELIMIPL	59.26±0.38	51.55±1.44	60.16±1.83	58.83±2.15
SAMCC (ours)	1.45±0.38 (57.81 ↓)	5.60±1.64 (45.95 ↓)	5.85±1.43 (54.31 ↓)	12.85±1.52 (45.98 ↓)
SAMCN (ours)	1.34±0.35 (57.92 ↓)	5.75±1.04 (45.80 ↓)	6.32±2.26 (53.84 ↓)	16.75±1.49 (42.08 ↓)
MIPLMA	58.65±0.85	52.38±1.60	60.89±2.04	61.67±2.58
MAMCC (ours)	1.00±0.15 (57.65 ↓)	4.84±1.43 (47.54 ↓)	**4.53±0.68** (56.36 ↓)	12.15±1.02 (49.52 ↓)
MAMCN (ours)	**0.94±0.17** (57.71 ↓)	**4.18±1.57** (48.20 ↓)	6.87±1.75 (54.02 ↓)	15.91±1.38 (45.76 ↓)
		$r = 2$		
PROMIPL	59.64±0.20	50.57±1.96	55.76±1.76	54.83±2.21
FASTMIPL	11.05±1.31	10.02±1.98	5.34±1.19	11.395±2.07
DEMIPL	59.62±1.09	48.84±2.18	51.65±1.63	51.55±3.68
DAMCC (ours)	1.20±0.58 (58.42 ↓)	6.62±2.04 (42.22 ↓)	5.23±0.74 (46.42 ↓)	**10.32±1.20** (41.23 ↓)
DAMCN (ours)	**0.86±0.42** (58.76 ↓)	6.33±2.41 (42.51 ↓)	5.54±1.38 (46.11 ↓)	14.45±1.63 (37.10 ↓)
ELIMIPL	60.02±1.04	48.89±2.55	57.77±1.56	53.18±2.49
SAMCC (ours)	3.13±0.51 (56.89 ↓)	7.53±1.35 (41.36 ↓)	**4.98±0.83** (52.79 ↓)	10.66±1.81 (42.52 ↓)
SAMCN (ours)	2.72±0.32 (57.30 ↓)	6.93±1.74 (41.96 ↓)	6.70±1.21 (51.07 ↓)	16.53±1.97 (36.65 ↓)
MIPLMA	59.31±1.11	48.54±3.00	59.50±1.61	58.31±3.07
MAMCC (ours)	1.20±0.26 (58.11 ↓)	**6.26±1.30** (42.28 ↓)	5.01±0.73 (54.49 ↓)	10.60±1.54 (47.71 ↓)
MAMCN (ours)	1.16±0.26 (58.15 ↓)	6.46±1.71 (42.08 ↓)	5.71±0.97 (53.79 ↓)	14.38±1.59 (43.93 ↓)
		$r = 3$		
PROMIPL	46.41±11.20	35.26±4.25	54.40±2.22	45.89±2.26
FASTMIPL	21.46±3.29	**16.92±4.35**	6.77±1.24	13.91±2.48
DEMIPL	44.55±12.76	40.04±3.27	53.42±1.19	46.26±2.70
DAMCC (ours)	12.99±7.43 (31.56 ↓)	23.28±6.11 (16.76 ↓)	5.63±0.88 (47.79 ↓)	9.48±1.60 (36.78 ↓)
DAMCN (ours)	17.99±13.08 (26.56 ↓)	22.16±4.02 (17.88 ↓)	**4.52±0.87** (48.90 ↓)	15.09±2.06 (31.17 ↓)
ELIMIPL	43.59±15.00	40.65±5.38	55.31±1.64	51.63±2.85
SAMCC (ours)	9.92±7.18 (33.67 ↓)	22.70±7.34 (17.95 ↓)	5.03±0.89 (50.28 ↓)	9.05±1.45 (42.58 ↓)
SAMCN (ours)	11.24±7.52 (32.35 ↓)	20.89±4.20 (19.76 ↓)	4.87±1.04 (50.44 ↓)	15.35±2.51 (36.28 ↓)
MIPLMA	44.90±10.56	38.04±5.32	58.02±1.22	54.35±2.33
MAMCC (ours)	4.55±5.62 (40.35 ↓)	21.33±7.25 (16.71 ↓)	5.55±1.12 (52.47 ↓)	**8.88±1.45** (45.47 ↓)
MAMCN (ours)	**1.72±0.92** (43.18 ↓)	18.90±4.59 (19.14 ↓)	5.34±1.32 (52.68 ↓)	14.02±1.42 (40.33 ↓)

indicates that MIPLGP may struggle to integrate with deep learning-based feature representations.

The integration of CDL with DAM, SAM, or MAM consistently improves classification performance over DEMIPL, ELIMIPL, and MIPLMA. On the C-KMeans dataset, DAMCC and DAMCN achieve significant accuracy improvements of 10.67% and 12.67%, respectively. For image bag generators such as Row, SBN, and KMeans, DAM outperforms SAM and MAM in classification accuracy. In contrast, for instance-level feature learning with SIFT or ResNet-34, SAM and MAM outperform DAM. Furthermore, in the three CRC-MIPL datasets utilizing ResNet-34, all MIPL approaches exhibit improved performance as the number of partitioned patches increases. This trend indicates that in CRC-MIPL datasets, leveraging deep learning-based features contributes to enhancing classification performance.

Additionally, DAM performs well on simple feature representations in the CRC-MIPL dataset but struggles with

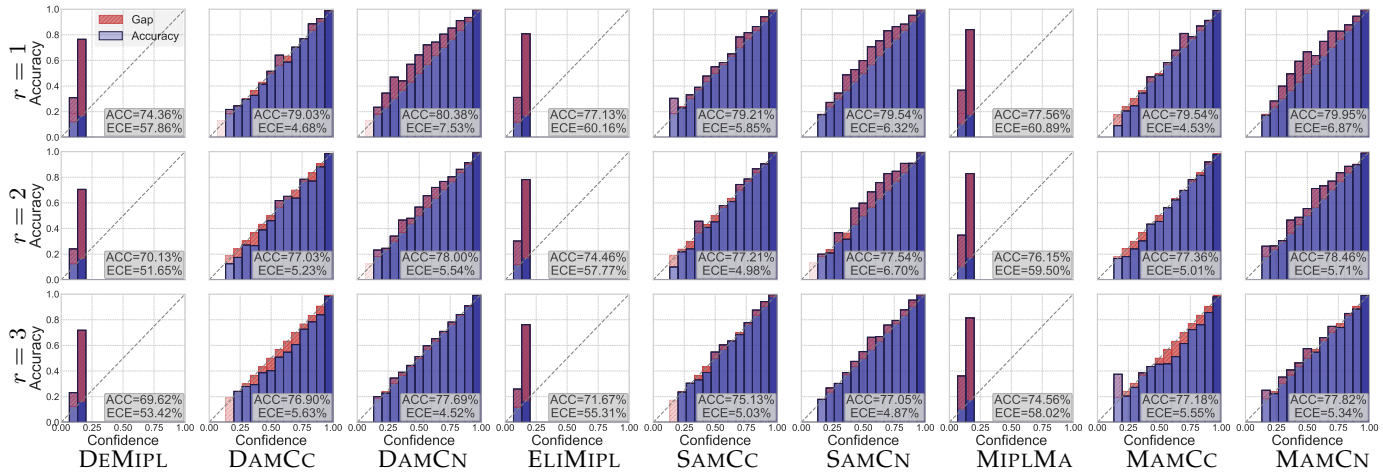


Fig. 5: Reliability diagrams of DEMIPL [16], ELIMIPL [18], and MIPLMA [19], and our methods on the Birdsong-MIPL dataset with varying numbers of false-positive labels. The bar color intensity indicates that more samples are assigned to the corresponding confidence intervals.

complex ones. In contrast, SAM and MAM excel at complex feature representations in the CRC-MIPL dataset but are not the best with simpler ones. Therefore, the choice of attention mechanism should align with the dataset's feature complexity. For simple feature representations, DAM is optimal, whereas complex representations, such as those derived from deep learning-based feature extractors, benefit from the advanced capabilities of SAM and MAM.

6.3 Calibration Performance

6.3.1 ECE on the Benchmark Datasets

Table 4 displays the expected calibration error for our methods with five comparative MIPL methods on the benchmark datasets. Notably, MIPLGP follows the instance-space paradigm, where bag-level predicted labels are aggregated from the instance-level predictions. This method fits the predicted probability and label to the most prominent instance rather than the entire multi-instance bag. Consequently, MIPLGP generally underperforms compared to MIPL approaches based on the embedded-space paradigm. Therefore, we exclude the calibration results for MIPLGP.

The calibration results demonstrate that our CDL significantly reduces the ECE on the MIPL datasets. On the benchmark datasets, CDL reduces the ECE by over 50% in 26 out of 60 cases, with the minimum reduction being 16.76%, the maximum reduction being 58.76%, and the mean reduction being 44.76%. Additionally, MAM outperforms DAM and SAM in most cases, which is consistent with the accuracy observed on the benchmark datasets.

To evaluate the calibration performance of our methods relative to baseline methods, we present reliability diagrams on the test set of the Birdsong-MIPL dataset. Fig. 5 presents the reliability diagrams on the Birdsong-MIPL dataset. The diagrams display the classification accuracies and ECE based on ten random runs. The baseline methods, DEMIPL, ELIMIPL, and MIPLMA, consistently produce predicted probabilities below 0.25. Although these methods achieve over 70% classification accuracy, their calibration performance is notably subpar. In contrast, our methods exhibit significant improvements in both calibration performance and classification performance.

6.3.2 ECE on the Real-World Datasets

Table 5 presents ECE across the real-world datasets, comparing our proposed methods against state-of-the-art methods in MIPL. Our proposed methods consistently achieve lower calibration error than all the comparative methods, demonstrating their effectiveness in improving model calibration. Notably, MAM yields the best overall performance, with MAMCC attaining the lowest ECE across most datasets. The most significant improvement occurs on C-R34-25, where MAMCC reduces the ECE by 33.32% compared to MIPLMA. Our methods achieve over a 10% reduction in ECE in 33 out of 42 cases. These results indicate that our CDL effectively aligns predicted confidences with classification accuracies, thereby enhancing model calibration.

Our methods exhibit particularly strong performance gains on the C-R34-9, C-R34-16, and C-R34-25 datasets compared to previous MIPL methods such as DEMIPL, ELIMIPL, and MIPLMA. Specifically, when combined with DAM, SAM, or MAM, our CDL achieves up to 19%, 24%, and 33% lower ECE on the C-R34-25 dataset. These findings highlight our framework's ability to leverage fine-grained image representations for improved calibration, demonstrating its effectiveness on challenging datasets.

6.4 Discussion

Appendices B and C provide supplementary empirical evidence for CDL. Appendix B analyzes its feature aggregation and label disambiguation mechanisms, examines sensitivity to γ, compares CDL with FL/IFL and PLL baselines, and further evaluates robustness, scalability, computational cost, and statistical significance. Appendix C presents a CRC-MIPL engineering study on patch granularity and confidence-based triage for practical deployment.

Our evaluation confirms that CDL achieves state-of-the-art performance across the benchmark and real-world MIPL datasets. The experimental results highlight three key insights: (1) CDL significantly improves classification

TABLE 5: Expected calibration error (mean±std%) on the real-world datasets.

	C-Row	C-SBN	C-KMeans	C-SIFT	C-R34-9	C-R34-16	C-R34-25
PROMIPL	17.99±1.64	25.48±0.65	31.52±1.11	31.42±1.12	34.30±0.99	37.72±0.72	41.31±0.85
FASTMIPL	50.54±3.78	42.17±3.07	41.62±1.27	43.48±2.75	41.01±1.60	36.19±1.46	33.97±1.62
DEMIPL	16.40±1.06	22.59±1.10	27.82±1.09	28.17±0.99	32.89±1.37	36.37±0.80	39.23±0.85
DAMCC (ours)	16.10±0.57 (0.30 ↓)	15.13±0.61 (7.46 ↓)	10.39±0.71 (17.43 ↓)	23.42±1.21 (4.75 ↓)	20.90±1.37 (11.99 ↓)	20.42±0.84 (15.95 ↓)	19.83±0.98 (19.40 ↓)
DAMCN (ours)	10.54±0.86 (5.86 ↓)	12.12±0.57 (10.47 ↓)	**9.67±0.44** (18.15 ↓)	21.38±1.08 (6.79 ↓)	19.96±0.78 (12.93 ↓)	20.12±0.77 (16.25 ↓)	19.59±1.01 (19.64 ↓)
ELIMIPL	18.21±0.83	24.91±0.78	28.99±1.18	28.76±1.04	34.47±1.17	36.70±0.67	40.24±0.66
SAMCC (ours)	15.52±0.51 (2.69 ↓)	14.58±0.93 (10.33 ↓)	10.83±0.69 (18.16 ↓)	17.73±0.86 (11.03 ↓)	18.38±1.03 (16.09 ↓)	16.88±0.60 (19.82 ↓)	16.04±0.87 (24.2 ↓)
SAMCN (ours)	9.82±0.86 (8.39 ↓)	10.66±0.60 (14.25 ↓)	10.95±1.27 (18.04 ↓)	18.91±0.68 (9.85 ↓)	20.45±0.93 (14.02 ↓)	18.53±0.71 (18.17 ↓)	16.11±0.84 (24.13 ↓)
MIPLMA	18.40±0.99	25.24±0.68	30.74±0.90	29.48±0.92	33.30±1.25	37.05±0.73	41.58±1.05
MAMCC (ours)	**6.01±0.85** (12.39 ↓)	**6.23±1.14** (19.01 ↓)	10.28±0.74 (20.46 ↓)	6.80±0.43 (22.68 ↓)	**9.46±0.81** (23.84 ↓)	**8.77±0.68** (28.28 ↓)	**8.26±0.59** (33.32 ↓)
MAMCN (ours)	10.76±0.65 (7.64 ↓)	8.97±1.16 (16.27 ↓)	11.96±1.03 (18.78 ↓)	**6.68±0.34** (22.8 ↓)	10.25±0.73 (23.05 ↓)	9.25±0.68 (27.8 ↓)	8.50±0.76 (33.08 ↓)

accuracy, particularly in high-ambiguity scenarios ($r = 3$), achieving a mean improvement of 8.93% on benchmark datasets. (2) CDL simultaneously enhances calibration, reducing ECE by an average of 44.76% on benchmark datasets due to its dynamic sample weighting mechanism. (3) Visualization analyses and probability distribution studies confirm that these improvements result from CDL's ability to encourage more separable feature distributions while adaptively prioritizing ambiguous labels during training.

The two instantiations of CDL consistently enhance both classification and calibration compared to the baseline models. In terms of classification accuracy, the second instantiation $\mathcal{L}_{\text{CDL-CN}}$ outperforms $\mathcal{L}_{\text{CDL-CC}}$ in most cases. Conversely, $\mathcal{L}_{\text{CDL-CC}}$ generally achieves lower expected calibration error than $\mathcal{L}_{\text{CDL-CN}}$. These observations lead to several insights: First, incorporating non-candidate label confidences can facilitate improved learning in MIPL models. Second, $\mathcal{L}_{\text{CDL-CC}}$ appears more effective when the candidate label set contains semantically similar labels. Third, $\mathcal{L}_{\text{CDL-CN}}$ tends to perform better in scenarios with a high degree of label randomness, as commonly seen in benchmark datasets. Moreover, performance differences among the three attention mechanisms on real-world datasets suggest that architectural complexity should be matched to the characteristics of the feature space. For example, DAM is more efficient for relatively simple features, while SAM and MAM are better suited for complex representations. No single CDL instantiation or attention mechanism is optimal for both classification and calibration, highlighting the importance of selecting the appropriate combination based on specific task requirements.

7 CONCLUSION

This paper investigates the calibration performance of multi-instance partial-label learning. We propose a calibratable disambiguation loss (CDL), a plug-and-play top-vs-competitor margin-modulated disambiguation loss. Theoretically, we show that CDL can be viewed as a margin-modulated MDL objective with an adaptive regularization effect. We further relate top-label calibration to weight alignment and analyze how the additional margin-dependent gradient term and momentum-based weight updates propagate margin information. Extensive experiments demonstrate CDL's effectiveness, achieving superior classification performance in 105 out of 110 cases and superior calibration performance in 93 out of 95 cases compared to state-of-the-art MIPL approaches. These results highlight CDL's robustness in handling high-ambiguity scenarios and improving the alignment between model confidence and accuracy.

While our approach demonstrates superior classification and calibration performance on challenging datasets like the C-R34-25 dataset, it does not achieve perfect calibration, where the expected calibration error reaches zero. Moreover, CDL introduces a focusing factor γ to control the strength of adaptive down-weighting. We use $\gamma = 1$ as a robust default in all main experiments, while tuning within the valid range can bring additional calibration gains. Finally, we follow the standard MIPL assumption that the candidate label set contains the true label. When this assumption is violated (missing-true-label cases), learning becomes substantially more difficult, as reflected by our stress test (Appendix B.6). Future work will investigate adaptive strategies for selecting γ and extend calibration as well as disambiguation to more challenging MIPL settings with missing true labels and improved uncertainty-aware learning frameworks [62].

REFERENCES

[1] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[2] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[3] W. Zhang, X. Zhang, H.-W. Deng, and M.-L. Zhang, "Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization," in *Advances in Neural Information Processing Systems 35, New Orleans, LA, USA*, 2022, pp. 34 940–34 953.

[4] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA*, 2022, pp. 18 802–18 812.

[5] F. Zhang, L. Feng, B. Han, T. Liu, G. Niu, T. Qin, and M. Sugiyama, "Exploiting class activation value for partial-label learning," in *Proceedings of the 10th International Conference on Learning Representations, Virtual Event*, 2022, pp. 1–17.

[6] S. He, L. Feng, F. Lv, W. Li, and G. Yang, "Partial label learning with semantic label representations," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA*, 2022, pp. 545–553.

[7] X. Gong, D. Yuan, W. Bao, and F. Luo, "A unifying probabilistic framework for partially labeled data learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8036–8048, 2023.

[8] G. S. A. Hajj, A. Hubin, C. Kanduri, M. Pavlovic, K. D. Rand, M. Widrich, A. S. Solberg, V. Greiff, J. Pensar, G. Klambauer, and G. K. Sandve, "Incorporating probabilistic domain knowledge into deep multiple instance learning," in *Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria*, 2024, pp. 17 279–17 297.

[9] Y. Zhang, Z. Zhou, X. He, A. R. Adhikary, and B. Dutta, "Data-driven knowledge fusion for deep multi-instance learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024.

[10] X. Li, Y. Jiang, C. Li, Y. Wang, and J. Ouyang, "Learning with partial labels from semi-supervised perspective," in *Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA*, 2023, pp. 8666–8674.

[11] D.-D. Wu, D.-B. Wang, and M.-L. Zhang, "Distilling reliable knowledge for instance-dependent partial label learning," in *Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, Canada*, 2024, pp. 15 888–15 896.

[12] Y.-Z. Wang, W. Zhang, and M.-L. Zhang, "Partial label causal representation learning for instance-dependent supervision and domain generalization," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA*, 2025, pp. 1–9.

[13] W. Tang, W. Zhang, and M.-L. Zhang, "Multi-instance partial-label learning: Towards exploiting dual inexact supervision," *Science China Information Sciences*, vol. 67, no. 3, pp. 132 103:1–132 103:14, 2024.

[14] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[15] A. Grote, N. S. Schaadt, G. Forestier, C. Wemmert, and F. Feuerhake, "Crowdsourcing of histological image labeling and object delineation by medical students," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1284–1294, 2019.

[16] W. Tang, W. Zhang, and M.-L. Zhang, "Disambiguated attention embedding for multi-instance partial-label learning," in *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, 2023, pp. 56 756–56 771.

[17] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama, "Progressive identification of true labels for partial-label learning," in *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*, 2020, pp. 6500–6510.

[18] W. Tang, W. Zhang, and M.-L. Zhang, "Exploiting conjugate label information for multi-instance partial-label learning," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, South Korea*, 2024, pp. 4973–4981.

[19] W. Tang, Y.-F. Yang, Z. Wang, W. Zhang, and M.-L. Zhang, "Multi-instance partial-label learning with margin adjustment," in *Advances in Neural Information Processing Systems 37, Vancouver, Canada*, 2024, pp. 26 331–26 354.

[20] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[21] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *Proceedings of the 26th International Conference on Machine Learning, Montreal, Quebec, Canada*, 2009, pp. 1249–1256.

[22] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[23] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[24] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, "Multiple instance active learning for object detection," in *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event*, 2021, pp. 5330–5339.

[25] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada*, 2023, pp. 8022–8031.

[26] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden*, 2018, pp. 2132–2141.

[27] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang, "Loss-based attention for deep multiple instance learning," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA*, 2020, pp. 5742–5749.

[28] Y. Cui, Z. Liu, X. Liu, X. Liu, C. Wang, T.-W. Kuo, C. J. Xue, and A. B. Chan, "Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images," in *Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda*, 2023, pp. 1–17.

[29] O. Fourkioti, M. D. Vries, and C. Bakal, "CAMIL: context-aware multiple instance learning for cancer detection and subtyping in whole slide images," in *Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria*, 2024, pp. 1–16.

[30] J. Early, G. K. C. Cheung, K. Cutajar, H. Xie, J. Kandola, and N. Twomey, "Inherently interpretable time series classification via multiple instance learning," in *Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria*, 2024, paper 1–29.

[31] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *The Journal of Machine Learning Research*, vol. 12, pp. 1501–1536, 2011.

[32] G. Lyu, S. Feng, T. Wang, C. Lang, and Y. Li, "GM-PLL: Graph matching based partial label learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 2, pp. 521–535, 2019.

[33] Y. Yao, J. Deng, X. Chen, C. Gong, J. Wu, and J. Yang, "Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA*, 2020, pp. 12 669–12 676.

[34] L. Liu and T. G. Dietterich, "A conditional multinomial mixture model for superset label learning," in *Advances in Neural Information Processing Systems 25, Cambridge, MA, USA*, 2012, pp. 548–556.

[35] K. He, W. Tang, T. Wei, and M. Zhang, "Tuning the right foundation models is what you need for partial label learning," *CoRR*, vol. abs/2506.05027, 2025.

[36] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, 2012, pp. 534–542.

[37] W. Wang and M. Zhang, "Semi-supervised partial label learning via confidence-rated margin maximization," in *Advances in Neural Information Processing Systems 33, Virtual Event*, 2020, pp. 6982–6993.

[38] D.-B. Wang, M.-L. Zhang, and L. Li, "Adaptive graph guided disambiguation for partial label learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8796–8811, 2022.

[39] L. Feng, J. Lv, B. Han, M. Xu, G. Niu, X. Geng, B. An, and M. Sugiyama, "Provably consistent partial-label learning," in *Advances in Neural Information Processing Systems 33, Virtual Event*, 2020, pp. 10 948–10 960.

[40] H. Wen, J. Cui, H. Hang, J. Liu, Y. Wang, and Z. Lin, "Leveraged weighted loss for partial label learning," in *Proceedings of the 38th International Conference on Machine Learning, Virtual Event*, 2021, pp. 11 091–11 100.

[41] N. Xu, B. Liu, J. Lv, C. Qiao, and X. Geng, "Progressive purification for instance-dependent partial label learning," in *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 38 551–38 565.

[42] W. Wang, D.-D. Wu, J. Wang, G. Niu, M.-L. Zhang, and M. Sugiyama, "Realistic evaluation of deep partial-label learning algorithms," in *Proceedings of the 13th International Conference on Learning Representations, Singapore*, 2025, pp. 1–25.

[43] Y.-F. Yang, W. Tang, and M.-L. Zhang, "ProMIPL: A probabilistic generative model for multi-instance partial-label learning," in *Proceedings of the 24th IEEE International Conference on Data Mining, Abu Dhabi, UAE*, 2024, pp. 560–569.

[44] Y.-F. Yang, W. Tang, and M.-L. Zhang, "Fast multi-instance partial-label learning," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA*, 2025, pp. 1–9.

[45] K. Wang, E. Tsamoura, and D. Roth, "On learning latent models with multi-instance weak supervision," in *Advances in Neural*

Information Processing Systems 36, New Orleans, LA, USA, 2023, pp. 9661–9694.
[46] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia*, 2017, pp. 1321–1330.
[47] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada*, 2019, pp. 4696–4705.
[48] D.-B. Wang, L. Feng, and M.-L. Zhang, "Rethinking calibration of deep neural networks: Do not be afraid of overconfidence," in *Advances in Neural Information Processing Systems 34, Virtual Event*, 2021, pp. 11 809–11 820.
[49] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada*, 2019, pp. 13 888–13 899.
[50] L. Zhang, Z. Deng, K. Kawaguchi, and J. Zou, "When and how mixup improves calibration," in *Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA*, vol. 162, 2022, pp. 26 135–26 160.
[51] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," in *Advances in Neural Information Processing Systems 33, Virtual Event*, 2020, pp. 15 288–15 299.
[52] L. Tao, M. Dong, and C. Xu, "Dual focal loss for calibration," in *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA*, 2023, pp. 33 833–33 849.
[53] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
[54] C. Tomani, S. Gruber, M. E. Erdem, D. Cremers, and F. Buettner, "Post-hoc uncertainty calibration for domain drift scenarios," in *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, June 19-25, 2021*, 2021, pp. 10 124–10 132.
[55] D.-B. Wang, L. Li, P. Zhao, P.-A. Heng, and M.-L. Zhang, "On the pitfall of mixup for uncertainty calibration," in *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada*, 2023, pp. 7609–7618.
[56] D.-B. Wang and M.-L. Zhang, "Calibration bottleneck: Over-compressed representations are less calibratable," in *Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria*, 2024, pp. 52 156–52 170.
[57] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
[58] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[59] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017. [Online]. Available: http://arxiv.org/abs/1708.07747
[60] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems 20, Vancouver, British Columbia, Canada*, 2007, pp. 1289–1296.
[61] X.-S. Wei and Z.-H. Zhou, "An empirical study on image bag generators for multi-instance learning," *Machine Learning*, vol. 105, pp. 155–198, 2016.
[62] V. H. Moghaddam and J. Hamidzadeh, "New hermite orthogonal polynomial kernel and combined kernels in support vector machine classifier," *Pattern Recognition*, vol. 60, pp. 921–935, 2016.

Appendix for Calibratable Disambiguation Loss for Multi-Instance Partial-Label Learning

Wei Tang, Yin-Fang Yang, Weijia Zhang, and Min-Ling Zhang, *Senior Member, IEEE*

APPENDIX A
PROOFS FOR THEORETICAL ANALYSIS

This section provides the complete proofs for the theoretical results presented in the main text, including the lower-bound, calibration, gradient, logit-space effect, and pseudo-label momentum analyses of CDL. Unless otherwise stated, we use the notation of the main text.

A.1 Proof of Theorem 1

Theorem 1 (Linear lower bound of CDL)**.** *Let $\gamma \geq 1$. For the i-th training bag, assume that $\hat{\boldsymbol{p}}_i \in \Delta^{k-1}$, $\hat{p}_{i,c} > 0$ for all $c \in \mathcal{Y}$, and $\boldsymbol{w}_i \in \Delta^{k-1}$ with $w_{i,c} = 0$ for all $c \notin \mathcal{S}_i$. Let $u_i = \arg\max_{c\in\mathcal{S}_i} \hat{p}_{i,c}$, $q_i = \hat{p}_{i,u_i} = \max_{c\in\mathcal{S}_i} \hat{p}_{i,c}$. Let*

$$\ell_i^{\mathrm{MDL}} = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c}, \qquad \ell_i^{\mathrm{CDL}} = \lambda_i \ell_i^{\mathrm{MDL}}. \tag{A1}$$

Then, for each training bag,

$$\ell_i^{\mathrm{CDL}} \geq (1-\gamma\beta_i)\ell_i^{\mathrm{MDL}} = (1-\gamma\beta_i)\left[\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) + \mathbb{H}(\boldsymbol{w}_i)\right]. \tag{A2}$$

Consequently,

$$\mathcal{L}_{\mathrm{CDL}} \geq \frac{1}{m}\sum_{i=1}^{m}(1-\gamma\beta_i)\ell_i^{\mathrm{MDL}}. \tag{A3}$$

Moreover, if $\beta_{\max} = \max_{1\leq i\leq m} \beta_i$, then

$$\mathcal{L}_{\mathrm{CDL}} \geq (1-\gamma\beta_{\max})\mathcal{L}_{\mathrm{MDL}}. \tag{A4}$$

Proof. Fix an arbitrary training bag indexed by i. Since $\hat{\boldsymbol{p}}_i \in \Delta^{k-1}$, we have $0 \leq \hat{p}_{i,c} \leq 1$ for every $c \in \mathcal{Y}$, and therefore $q_i \in [0,1]$. By assumption, the competitor probability ϕ_i is well-defined and satisfies $\phi_i \in [0,1]$. Hence

$$\beta_i = q_i - \phi_i \leq 1. \tag{A5}$$

It follows from Eq. (A5) that $1-\beta_i \geq 0$, so the quantity $(1-\beta_i)^\gamma$ is well-defined for every real exponent $\gamma \geq 1$. Consider the function $f(t) = t^\gamma$ on $[0,\infty)$. Since $\gamma \geq 1$, f is convex on $[0,\infty)$, and the first-order supporting hyperplane inequality at $t = 1$ gives

$$t^\gamma \geq f(1) + f'(1)(t-1) = 1 + \gamma(t-1), \qquad t \geq 0. \tag{A6}$$

Taking $t = 1-\beta_i$ in Eq. (A6) yields the Bernoulli-type bound

$$(1-\beta_i)^\gamma \geq 1 - \gamma\beta_i. \tag{A7}$$

Next, since $\hat{p}_{i,c} > 0$ and $\hat{p}_{i,c} \leq 1$ for all $c \in \mathcal{Y}$, we have $\log \hat{p}_{i,c} \leq 0$, and since $w_{i,c} \geq 0$, every summand in the MDL loss is nonnegative. Therefore

$$\ell_i^{\mathrm{MDL}} = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c} \geq 0. \tag{A8}$$

Multiplying both sides of Eq. (A7) by the nonnegative scalar and using the definition $\ell_i^{\mathrm{CDL}} = (1-\beta_i)^\gamma \ell_i^{\mathrm{MDL}}$, we obtain

$$\ell_i^{\mathrm{CDL}} = (1-\beta_i)^\gamma \ell_i^{\mathrm{MDL}} \geq (1-\gamma\beta_i)\ell_i^{\mathrm{MDL}}. \tag{A9}$$

We now verify the KL-entropy decomposition of ℓ_i^{MDL}. Since $\boldsymbol{w}_i \in \Delta^{k-1}$, $w_{i,c} = 0$ for $c \notin \mathcal{S}_i$, and $\hat{p}_{i,c} > 0$ for all $c \in \mathcal{Y}$, the KL divergence is finite and can be written as

$$\begin{aligned}\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) &= \sum_{c\in\mathcal{Y}} w_{i,c} \log \frac{w_{i,c}}{\hat{p}_{i,c}} \\ &= \sum_{c\in\mathcal{Y}} w_{i,c} \log w_{i,c} - \sum_{c\in\mathcal{Y}} w_{i,c} \log \hat{p}_{i,c},\end{aligned} \tag{A10}$$

where the convention $0 \log 0 = 0$ is used. By the definition of entropy,

$$\mathbb{H}(\boldsymbol{w}_i) = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log w_{i,c}. \tag{A11}$$

Combining Eqs. (A10) and (A11) gives

$$\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) + \mathbb{H}(\boldsymbol{w}_i) = -\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c}. \tag{A12}$$

Because $w_{i,c} = 0$ for every $c \notin \mathcal{S}_i$, the right-hand side of Eq. (A12) reduces to the MDL loss:

$$-\sum_{c\in\mathcal{S}_i} w_{i,c} \log \hat{p}_{i,c} = \ell_i^{\mathrm{MDL}}. \tag{A13}$$

Hence

$$\ell_i^{\mathrm{MDL}} = \mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) + \mathbb{H}(\boldsymbol{w}_i). \tag{A14}$$

Substituting Eq. (A14) into Eq. (A9) proves the per-bag lower bound Eq. (A2). Averaging Eq. (A9) over all m training bags gives

$$\mathcal{L}_{\mathrm{CDL}} = \frac{1}{m}\sum_{i=1}^{m}\ell_i^{\mathrm{CDL}} \geq \frac{1}{m}\sum_{i=1}^{m}(1-\gamma\beta_i)\ell_i^{\mathrm{MDL}}, \tag{A15}$$

- *Wei Tang, Yin-Fang Yang, and Min-Ling Zhang are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and the Key Laboratory of Computer Network and Information Integration (Southeast University), MoE, China. E-mail: {tangw, yangyf, zhangml}@seu.edu.cn.*
- *Weijia Zhang is with the School of Computer and Information Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: weijia.zhang@newcastle.edu.au.*
- *Corresponding author: Min-Ling Zhang.*

which proves Eq. (A3). Finally, by the definition $\beta_{\max} = \max_{1\le i\le m}\beta_i$, we have $\beta_i \le \beta_{\max}$ for all i, and since $\gamma > 0$,

$$1-\gamma\beta_i \ge 1-\gamma\beta_{\max}. \tag{A16}$$

Using $\ell_i^{\mathrm{MDL}} \ge 0$ again, Eq. (A16) implies

$$\frac{1}{m}\sum_{i=1}^{m}(1-\gamma\beta_i)\ell_i^{\mathrm{MDL}} \ge (1-\gamma\beta_{\max})\frac{1}{m}\sum_{i=1}^{m}\ell_i^{\mathrm{MDL}}. \tag{A17}$$

Since $\mathcal{L}_{\mathrm{MDL}} = \frac{1}{m}\sum_{i=1}^{m}\ell_i^{\mathrm{MDL}}$, combining Eqs. (A15) and (A17) gives

$$\mathcal{L}_{\mathrm{CDL}} \ge (1-\gamma\beta_{\max})\mathcal{L}_{\mathrm{MDL}}. \tag{A18}$$

The factor $1-\gamma\beta_{\max}$ is positive exactly when $1-\gamma\beta_{\max} > 0$, equivalently $\gamma\beta_{\max} < 1$. This completes the proof. □

A.2 Proof of Proposition 1

Proposition 1 (Calibration is controlled by pseudo-label confidence error)**.** *Fix the current training state. Assume that $\hat{\boldsymbol{p}}_i$ and $\boldsymbol{w}_i$ are measurable with respect to $\mathcal{G}_i = \sigma(\boldsymbol{X}_i, \mathcal{S}_i)$. Then*

$$E_{\mathrm{cal}} \le E_{\mathrm{conf}} \le E_{\mathrm{pconf}} + \delta_w^{\mathrm{TV}}. \tag{A19}$$

Proof. Since $\hat{\boldsymbol{p}}_i$ is $\mathcal{G}_i$-measurable and the tie-breaking rule is fixed, both $\hat{y}_i$ and C_i are $\mathcal{G}_i$-measurable. Moreover,

$$\mathbb{E}\left[\mathbb{I}\{Y_i = \hat{y}_i\} \mid \mathcal{G}_i\right] = \eta_{i,\hat{y}_i}. \tag{A20}$$

Using the tower property and the fact that C_i is $\mathcal{G}_i$-measurable gives

$$\mathbb{E}\left[\mathbb{I}\{Y_i = \hat{y}_i\} \mid C_i\right] = \mathbb{E}\left[\eta_{i,\hat{y}_i} \mid C_i\right]. \tag{A21}$$

Therefore, Jensen's inequality yields

$$E_{\mathrm{cal}} = \mathbb{E}\left[\left|\mathbb{E}\left[\eta_{i,\hat{y}_i} - C_i \mid C_i\right]\right|\right] \le \mathbb{E}\left[\left|\eta_{i,\hat{y}_i} - C_i\right|\right] = E_{\mathrm{conf}}. \tag{A22}$$

For the second inequality, the triangle inequality gives

$$\left|\eta_{i,\hat{y}_i} - C_i\right| \le \left|w_{i,\hat{y}_i} - C_i\right| + \left|\eta_{i,\hat{y}_i} - w_{i,\hat{y}_i}\right|. \tag{A23}$$

Since total variation distance dominates the discrepancy of every singleton event,

$$\left|\eta_{i,\hat{y}_i} - w_{i,\hat{y}_i}\right| \le d_{\mathrm{TV}}(\boldsymbol{\eta}_i, \boldsymbol{w}_i). \tag{A24}$$

Combining Eqs. (A23) and (A24), and then taking expectations, gives

$$E_{\mathrm{conf}} \le E_{\mathrm{pconf}} + \delta_w^{\mathrm{TV}}. \tag{A25}$$

The proof follows by combining Eqs. (A22) and (A25). □

A.3 Proof of Theorem 2

Theorem 2 (Pseudo-label confidence alignment bound)**.** *Assume that $\hat{\boldsymbol{p}}_i$ is induced by finite logits, so that $\hat{p}_{i,c} > 0$ for all $c \in \mathcal{Y}$. Assume further that the relevant competitor set in CDL is nonempty and that the CDL modulation satisfies $\lambda_i \ge \lambda_0 > 0$. Then*

$$E_{\mathrm{pconf}} \le \sqrt{2\left(\lambda_0^{-1}R_{\mathrm{CDL}}^{w} - \mathcal{H}_w\right)}. \tag{A26}$$

Consequently,

$$E_{\mathrm{cal}} \le E_{\mathrm{conf}} \le \sqrt{2\left(\lambda_0^{-1}R_{\mathrm{CDL}}^{w} - \mathcal{H}_w\right)} + \delta_w^{\mathrm{TV}}. \tag{A27}$$

Proof. Since $C_i = \hat{p}_{i,\hat{y}_i}$, we have

$$\left|w_{i,\hat{y}_i} - C_i\right| = \left|w_{i,\hat{y}_i} - \hat{p}_{i,\hat{y}_i}\right|. \tag{A28}$$

A single-coordinate discrepancy is bounded by the full ℓ_1 distance:

$$\left|w_{i,\hat{y}_i} - \hat{p}_{i,\hat{y}_i}\right| \le \sum_{c\in\mathcal{Y}}\left|w_{i,c} - \hat{p}_{i,c}\right| = \left\|\boldsymbol{w}_i - \hat{\boldsymbol{p}}_i\right\|_1. \tag{A29}$$

By Pinsker's inequality,

$$\left\|\boldsymbol{w}_i - \hat{\boldsymbol{p}}_i\right\|_1 \le \sqrt{2\,\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i)}. \tag{A30}$$

Combining Eqs. (A28), (A29), and (A30) gives

$$\left|w_{i,\hat{y}_i} - C_i\right| \le \sqrt{2\,\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i)}. \tag{A31}$$

Taking expectations and applying Jensen's inequality to the concave square-root function yield

$$E_{\mathrm{pconf}} \le \sqrt{2\,\mathbb{E}\left[\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i)\right]}. \tag{A32}$$

It remains to upper bound the expected KL divergence by the CDL risk. Since $w_{i,c} = 0$ for $c \notin \mathcal{S}_i$ and $\hat{p}_{i,c} > 0$, the KL-entropy decomposition gives

$$\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i) = \ell_i^{\mathrm{MDL}} - \mathbb{H}(\boldsymbol{w}_i). \tag{A33}$$

Taking expectations gives

$$\mathbb{E}\left[\mathrm{KL}(\boldsymbol{w}_i\|\hat{\boldsymbol{p}}_i)\right] = R_{\mathrm{MDL}}^{w} - \mathcal{H}_w. \tag{A34}$$

Because $\ell_i^{\mathrm{MDL}} \ge 0$ and $\lambda_i \ge \lambda_0 > 0$ almost surely,

$$R_{\mathrm{MDL}}^{w} = \mathbb{E}\left[\ell_i^{\mathrm{MDL}}\right] \le \lambda_0^{-1}\mathbb{E}\left[\lambda_i\ell_i^{\mathrm{MDL}}\right] = \lambda_0^{-1}R_{\mathrm{CDL}}^{w}. \tag{A35}$$

Combining Eqs. (A32), (A34), and (A35) proves (A26). Finally, Eq. (A27) follows directly from Proposition 1. The proof is complete. □

A.4 Proof of Proposition 2

Proposition 2 (Gradient decomposition of CDL on differentiable regions)**.** *Fix a training bag $(\boldsymbol{X}_i, \mathcal{S}_i)$, and let θ denote the model parameters. Suppose that there is an open parameter region $\mathcal{U}$ on which each $\hat{p}_{i,c}(\theta)$ is differentiable and strictly positive. Assume that the top candidate label $u_i = \arg\max_{c\in\mathcal{S}_i}\hat{p}_{i,c}(\theta)$ is uniquely attained and remains unchanged on $\mathcal{U}$. For CDL-CC, let $\mathcal{C}_i = \mathcal{S}_i \setminus \{u_i\}$; for CDL-CN, let $\mathcal{C}_i = \bar{\mathcal{S}}_i$. Assume that $\mathcal{C}_i \ne \varnothing$ and that the competitor $v_i = \arg\max_{c\in\mathcal{C}_i}\hat{p}_{i,c}(\theta)$ is also uniquely attained and remains unchanged on $\mathcal{U}$. Define*

$$q_i(\theta) = \hat{p}_{i,u_i}(\theta), \qquad \phi_i(\theta) = \hat{p}_{i,v_i}(\theta), \tag{A36}$$

$$\beta_i(\theta) = q_i(\theta) - \phi_i(\theta), \qquad \lambda_i(\theta) = \big(1-\beta_i(\theta)\big)^{\gamma}. \tag{A37}$$

During the current gradient computation, regard the pseudo-label weights $\boldsymbol{w}_i$ as fixed, and write

$$\ell_i^{\mathrm{MDL}}(\theta) = -\sum_{c\in\mathcal{S}_i} w_{i,c}\log\hat{p}_{i,c}(\theta), \quad \ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\ell_i^{\mathrm{MDL}}(\theta). \tag{A38}$$

Then, for every $\theta \in \mathcal{U}$,

$$\begin{aligned}\nabla_\theta\ell_i^{\mathrm{CDL}}(\theta) = {}& \lambda_i(\theta)\nabla_\theta\ell_i^{\mathrm{MDL}}(\theta) \\ & - \gamma\big(1-\beta_i(\theta)\big)^{\gamma-1}\ell_i^{\mathrm{MDL}}(\theta)\nabla_\theta\beta_i(\theta).\end{aligned} \tag{A39}$$

Proof. Fix an arbitrary $\theta \in \mathcal{U}$. Since the active top candidate and the active competitor are fixed on $\mathcal{U}$, the two max operations reduce locally to ordinary coordinate projections:

$$\max_{c\in\mathcal{S}_i}\hat{p}_{i,c}(\theta) = \hat{p}_{i,u_i}(\theta), \qquad \max_{c\in\mathcal{C}_i}\hat{p}_{i,c}(\theta) = \hat{p}_{i,v_i}(\theta). \tag{A40}$$

Consequently, the local margin is an ordinary differentiable scalar function:

$$\beta_i(\theta) = \hat{p}_{i,u_i}(\theta) - \hat{p}_{i,v_i}(\theta). \tag{A41}$$

Taking the gradient of Eq. (A41) gives

$$\nabla_\theta \beta_i(\theta) = \nabla_\theta \hat{p}_{i,u_i}(\theta) - \nabla_\theta \hat{p}_{i,v_i}(\theta). \tag{A42}$$

The positivity of the predictive probabilities ensures that the logarithms in ℓ_i^{MDL} are well defined. Moreover, since u_i and v_i are distinct labels and the predictive vector is a strictly positive probability distribution, we have

$$1 - \beta_i(\theta) = 1 - \hat{p}_{i,u_i}(\theta) + \hat{p}_{i,v_i}(\theta) > 0. \tag{A43}$$

Hence $\lambda_i(\theta) = \big(1 - \beta_i(\theta)\big)^\gamma$ is differentiable on $\mathcal{U}$. By the definition of CDL in the margin-modulated form,

$$\ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\ell_i^{\mathrm{MDL}}(\theta). \tag{A44}$$

Applying the product rule to Eq. (A44) yields

$$\nabla_\theta \ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\nabla_\theta \ell_i^{\mathrm{MDL}}(\theta) + \ell_i^{\mathrm{MDL}}(\theta)\nabla_\theta \lambda_i(\theta). \tag{A45}$$

It remains to compute $\nabla_\theta \lambda_i(\theta)$. By the chain rule,

$$\nabla_\theta \lambda_i(\theta) = \nabla_\theta \big(1 - \beta_i(\theta)\big)^\gamma = -\gamma\big(1 - \beta_i(\theta)\big)^{\gamma-1}\nabla_\theta \beta_i(\theta). \tag{A46}$$

Substituting Eq. (A46) into Eq. (A45) gives

$$\begin{aligned}\nabla_\theta \ell_i^{\mathrm{CDL}}(\theta) = \lambda_i(\theta)\nabla_\theta \ell_i^{\mathrm{MDL}}(\theta) \\ - \gamma\big(1 - \beta_i(\theta)\big)^{\gamma-1}\ell_i^{\mathrm{MDL}}(\theta)\nabla_\theta \beta_i(\theta).\end{aligned} \tag{A47}$$

This is Eq. (A39). When $\boldsymbol{w}_i$ is detached during the current optimization step, the MDL gradient appearing above is

$$\begin{aligned}\nabla_\theta \ell_i^{\mathrm{MDL}}(\theta) &= -\sum_{c\in\mathcal{S}_i} w_{i,c}\nabla_\theta \log \hat{p}_{i,c}(\theta) \\ &= -\sum_{c\in\mathcal{S}_i} \frac{w_{i,c}}{\hat{p}_{i,c}(\theta)}\nabla_\theta \hat{p}_{i,c}(\theta).\end{aligned} \tag{A48}$$

If one differentiates through the pseudo-label update itself, the same product-rule identity still holds, but $\nabla_\theta \ell_i^{\mathrm{MDL}}(\theta)$ must then be interpreted as the full derivative, including the derivatives of $w_{i,c}(\theta)$. The ordinary-gradient statement above is local: at exact ties of the top candidate or the competitor, the max-based margin is generally not differentiable and must instead be handled with subdifferentials. □

A.5 Proof of Corollary 1

Corollary 1 (Logit-space effect of margin shaping)**.** *Fix a training bag i. Let $s_{i,c} \in \mathbb{R}$ be the logit of class c, and let*

$$\hat{p}_{i,c} = \frac{\exp(s_{i,c})}{\sum_{a\in\mathcal{Y}} \exp(s_{i,a})}, \qquad c \in \mathcal{Y}. \tag{A49}$$

Assume that, in a neighborhood of the current logits, the top candidate u and the competitor v are unique, distinct, and fixed. Then $\beta_i = \hat{p}_{i,u} - \hat{p}_{i,v}$ is differentiable in this neighborhood, and

$$\frac{\partial \beta_i}{\partial s_{i,u}} = \hat{p}_{i,u}(1 - \hat{p}_{i,u} + \hat{p}_{i,v}) > 0, \tag{A50}$$

$$\frac{\partial \beta_i}{\partial s_{i,v}} = -\hat{p}_{i,v}(1 + \hat{p}_{i,u} - \hat{p}_{i,v}) < 0, \tag{A51}$$

$$\frac{\partial \beta_i}{\partial s_{i,c}} = \hat{p}_{i,c}(\hat{p}_{i,v} - \hat{p}_{i,u}), \quad c \notin \{u, v\}. \tag{A52}$$

Thus a positive step along $\nabla_{s_i}\beta_i$ locally increases the top-candidate logit and decreases the active-competitor logit.

Proof. The local uniqueness assumption fixes the active indices u and v, so no derivative of the max operator is involved. For the softmax map,

$$\frac{\partial \hat{p}_{i,a}}{\partial s_{i,b}} = \hat{p}_{i,a}\big(\mathbb{I}\{a = b\} - \hat{p}_{i,b}\big), \quad a, b \in \mathcal{Y}. \tag{A53}$$

Using $\beta_i = \hat{p}_{i,u} - \hat{p}_{i,v}$ and $u \neq v$, we obtain

$$\frac{\partial \beta_i}{\partial s_{i,u}} = \hat{p}_{i,u}(1 - \hat{p}_{i,u}) + \hat{p}_{i,u}\hat{p}_{i,v} = \hat{p}_{i,u}(1 - \hat{p}_{i,u} + \hat{p}_{i,v}), \tag{A54}$$

$$\frac{\partial \beta_i}{\partial s_{i,v}} = -\hat{p}_{i,u}\hat{p}_{i,v} - \hat{p}_{i,v}(1 - \hat{p}_{i,v}) = -\hat{p}_{i,v}(1 + \hat{p}_{i,u} - \hat{p}_{i,v}), \tag{A55}$$

$$\frac{\partial \beta_i}{\partial s_{i,c}} = -\hat{p}_{i,u}\hat{p}_{i,c} + \hat{p}_{i,v}\hat{p}_{i,c} = \hat{p}_{i,c}(\hat{p}_{i,v} - \hat{p}_{i,u}), \quad c \notin \{u, v\}. \tag{A56}$$

Since the logits are finite, all softmax probabilities are strictly positive. Moreover,

$$1 - \hat{p}_{i,u} + \hat{p}_{i,v} \geq 2\hat{p}_{i,v} > 0, \quad 1 + \hat{p}_{i,u} - \hat{p}_{i,v} \geq 2\hat{p}_{i,u} > 0. \tag{A57}$$

The strict signs in Eqs. (A50) and (A51) follow immediately. □

A.6 Proof of Lemma 1

Lemma 1 (Momentum recursion for candidate pseudo-label margins)**.** *Fix a training bag $(\boldsymbol{X}_i, \mathcal{S}_i)$ and two candidate labels $u, v \in \mathcal{S}_i$. For $t = 2, \ldots, T$, define*

$$\tilde{p}_{i,c}^{(t)} = \frac{\hat{p}_{i,c}^{(t)}}{\sum_{a\in\mathcal{S}_i} \hat{p}_{i,a}^{(t)}}, \quad M_{i,uv}^{(t)} = w_{i,u}^{(t)} - w_{i,v}^{(t)}, \quad \Delta_{i,uv}^{(t)} = \tilde{p}_{i,u}^{(t)} - \tilde{p}_{i,v}^{(t)}. \tag{A58}$$

Assume that the pseudo-label weights are updated by

$$w_{i,c}^{(t)} = \alpha^{(t)} w_{i,c}^{(t-1)} + \big(1 - \alpha^{(t)}\big)\tilde{p}_{i,c}^{(t)}, \quad c \in \mathcal{S}_i. \tag{A59}$$

Then, for every $t = 2, \ldots, T$,

$$M_{i,uv}^{(t)} = \alpha^{(t)} M_{i,uv}^{(t-1)} + \big(1 - \alpha^{(t)}\big)\Delta_{i,uv}^{(t)}. \tag{A60}$$

Equivalently, with $A_{a:b} = \prod_{\tau=a}^{b} \alpha^{(\tau)}$ and the empty product defined as one,

$$M_{i,uv}^{(t)} = A_{2:t} M_{i,uv}^{(1)} + \sum_{s=2}^{t} \big(1 - \alpha^{(s)}\big) A_{s+1:t} \Delta_{i,uv}^{(s)}. \tag{A61}$$

Moreover,

$$\Delta_{i,uv}^{(s)} = \frac{\hat{p}_{i,u}^{(s)} - \hat{p}_{i,v}^{(s)}}{\sum_{a\in\mathcal{S}_i} \hat{p}_{i,a}^{(s)}}. \tag{A62}$$

Thus past candidate prediction margins enter the current pseudo-label margin through the momentum coefficients. If $\alpha^{(s)} \in [0, 1]$ for all s, then $M_{i,uv}^{(t)}$ is a convex combination of $M_{i,uv}^{(1)}$ and $\Delta_{i,uv}^{(2)}, \ldots, \Delta_{i,uv}^{(t)}$.

Proof. Applying Eq. (A59) to u and v, respectively, and subtracting the two identities, we obtain

$$w_{i,u}^{(t)} - w_{i,v}^{(t)} = \alpha^{(t)} \left(w_{i,u}^{(t-1)} - w_{i,v}^{(t-1)} \right) + \left(1 - \alpha^{(t)}\right) \left(\tilde{p}_{i,u}^{(t)} - \tilde{p}_{i,v}^{(t)} \right). \tag{A63}$$

By the definitions of $M_{i,uv}^{(t)}$ and $\Delta_{i,uv}^{(t)}$, this proves Eq. (A60). Iterating Eq. (A60) from epoch 2 to epoch t gives

$$M_{i,uv}^{(t)} = \left(\prod_{\tau=2}^{t} \alpha^{(\tau)} \right) M_{i,uv}^{(1)} + \sum_{s=2}^{t} \left[\left(1 - \alpha^{(s)}\right) \left(\prod_{\tau=s+1}^{t} \alpha^{(\tau)} \right) \Delta_{i,uv}^{(s)} \right], \tag{A64}$$

which is Eq. (A61). Moreover, the identity Eq. (A62) follows from Eq. (A58):

$$\Delta_{i,uv}^{(s)} = \tilde{p}_{i,u}^{(s)} - \tilde{p}_{i,v}^{(s)} = \frac{\hat{p}_{i,u}^{(s)} - \hat{p}_{i,v}^{(s)}}{\sum_{a \in \mathcal{S}_i} \hat{p}_{i,a}^{(s)}}. \tag{A65}$$

Finally, if $\alpha^{(s)} \in [0, 1]$, all coefficients in (A61) are nonnegative. Their sum is

$$\left(\prod_{\tau=2}^{t} \alpha^{(\tau)} \right) + \sum_{s=2}^{t} \left[\left(1 - \alpha^{(s)}\right) \left(\prod_{\tau=s+1}^{t} \alpha^{(\tau)} \right) \right] = 1. \tag{A66}$$

Hence (A61) is a convex combination, and the proof is complete. □

APPENDIX B
FURTHER EXPERIMENTAL ANALYSIS

B.1 Mechanisms Underlying the Effectiveness of CDL

The calibratable disambiguation loss (CDL) significantly improves both classification and calibration performance compared to baseline methods DEMIPL, ELIMIPL, and MIPLMA. To elucidate the reasons behind its effectiveness, we analyze its impact on two critical components of MIPL approaches based on the embedded-space paradigm: feature aggregation and label disambiguation.

B.1.1 Enhancing Feature Aggregation

To evaluate the impact on feature aggregation, we visualize the bag-level feature representations, i.e., z_i from Eq. (13), for the baseline methods DEMIPL, ELIMIPL, MIPLMA, and our proposed methods on the MNIST-MIPL dataset. Unlike the standard MNIST dataset, which comprises 10 target classes, the MNIST-MIPL dataset focuses on 5 target classes, with negative instances sourced from the remaining 5 classes [4]. For $r = 1$ and $r = 2$, DEMIPL, ELIMIPL, and MIPLMA produce reasonably good results. However, when $r = 3$, the feature representations aggregated by the three methods become increasingly disordered, with features from different classes becoming mixed. In contrast, our methods consistently produce well-separated feature representations for $r \in \{1, 2, 3\}$. Specifically, for $r = 1$ and $r = 2$, the class clusters are compact and distinctly separable. Although the clusters are somewhat dispersed at $r = 3$, our methods remain notably more distinct and separable compared to those generated by the baseline methods DEMIPL, ELIMIPL, and MIPLMA.

Therefore, the visualizations in Fig. A1 demonstrate that our proposed CDL significantly improves the aggregation of bag-level feature representations, resulting in more compact and distinguishable clusters. This enhancement directly contributes to superior classification performance.

B.1.2 Optimizing Label Disambiguation and Calibration

The disambiguation performance and calibration performance of a model are directly influenced by its predicted probabilities. As shown in Fig. 2, the baseline methods DEMIPL, ELIMIPL, and MIPLMA generate low predicted probabilities for true labels on the training set, leading to poor calibration performance. In contrast, our methods yield significantly higher predicted probabilities for true labels compared to these baseline methods. To further investigate the effect of CDL on predicted probabilities, we present the average predicted probabilities for true labels (T-labels), false-positive labels (FP-labels), and non-candidate labels (NC-labels) on the training set of the C-KMeans dataset.

As illustrated in Fig. A2, for DEMIPL, ELIMIPL, and MIPLMA, the probabilities for true labels are comparable to those for false-positive labels. In contrast, our methods achieve markedly higher probabilities for true labels compared to false-positive labels. Although our methods also show higher probabilities for false-positive labels relative to DEMIPL, ELIMIPL, and MIPLMA, the proportion of probabilities for false-positive labels relative to candidate labels is substantially lower. Additionally, our methods achieve significantly lower predicted probabilities for non-candidate labels compared to the baseline methods.

According to the definition of CDL, there exists an inverse relationship between the maximum predicted probabilities on candidate labels and the outcome of $\Phi(\cdot)$. For samples with high maximum predicted probabilities on candidate labels, the outcome of $\Phi(\cdot)$ is relatively small, resulting in lower loss values for these samples. Samples with low maximum predicted probabilities produce a larger outcome of $\Phi(\cdot)$, leading to higher loss values. As a result, the model focuses more on these high-loss, low-margin cases and increases their separability, which alleviates under-confidence. When a bag becomes well separated, the modulation decreases and automatically limits further sharpening, which helps mitigate over-confidence.

B.2 Parameter Sensitivity

In our proposed CDL, the only hyperparameter is the exponential factor γ. Consistent with focal loss [5], [6], we treat γ as a constant throughout our experiments. To evaluate the sensitivity of our methods to γ, Fig. A3 presents the accuracy and expected calibration error of DEMIPL, ELIMIPL, MIPLMA, and our six methods with γ ranging over $\{1, 2, 3, 4, 5\}$ on the C-R34-25 dataset.

Across all γ settings, our methods consistently outperform DEMIPL, ELIMIPL, and MIPLMA in both classification and calibration performance. As γ increases from 1 to 5, we observe a general improvement in calibration performance, though the gains diminish at higher values. Notably, ELIMIPL and MIPLMA exhibit greater robustness to variations in γ compared to DEMIPL, likely due to their more sophisticated attention mechanisms,

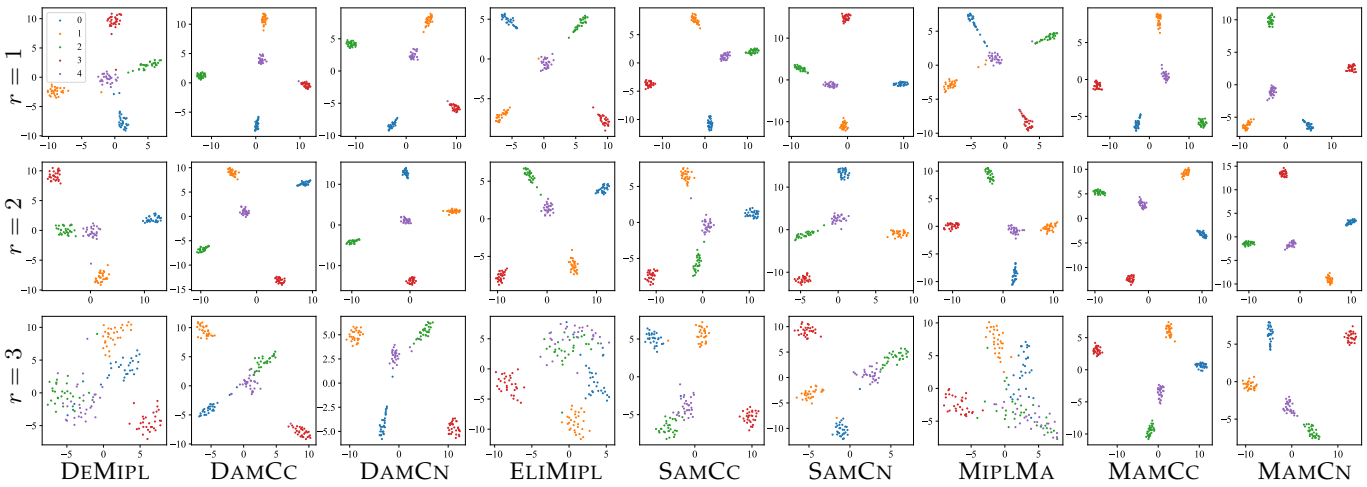


Fig. A1: t-SNE visualization of aggregated bag-level feature representations produced by the attention mechanisms in DeMIPL [1], EliMIPL [2], MIPLMA [3], and ours on the test set of the MNIST-MIPL dataset, which comprises five classes.

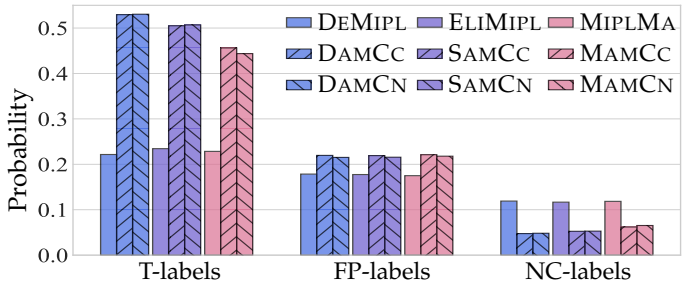


Fig. A2: Probabilities at the last epoch on the training set of the C-KMeans dataset. T-labels, FP-labels, and NC-labels denote the average probabilities for true, false positive, and non-candidate labels, respectively.

which provide better resistance to hyperparameter fluctuations. Furthermore, the two CDL instantiations yield comparable results, indicating that CDL's effectiveness is largely independent of the specific instantiation.

B.3 Comparison with FL and IFL

To evaluate the effectiveness of CDL compared to focal loss (FL) and inverse focal loss (IFL), we integrate FL and IFL into MAM, resulting in two variants: MAMFL and MAMIFL. The corresponding loss functions are defined in Eqs. (5) and (6), respectively, where the weights $w_{i,c}^{(t)}$ of candidate labels are initialized as $w_{i,c}^{(1)} = \frac{1}{|\mathcal{S}_i|}$ at $t = 1$ and updated using Eq. (15). Therefore, the only distinction between MAMFL/MAMIFL and our proposed MAMCC/MAMCN lies in the exponential term of the loss functions.

Fig. A4 presents the classification accuracy and expected calibration error of MIPLMA, MAMCC, MAMCN, MAMFL, and MAMIFL on the Birdsong-MIPL and SIVAL-MIPL datasets, with the number of false-positive labels $r \in \{1, 2, 3\}$. As shown in Fig. A4, MAMFL and MAMIFL improve the calibration of MIPLMA by reducing ECE. However, this improvement comes at the expense of classification accuracy, which declines more significantly as the number of false-positive labels increases. Furthermore, both MAMFL and MAMIFL underperform compared to our proposed CDL in both classification and calibration. These findings indicate that FL and IFL are insufficient for handling classification and calibration challenges in MIPL.

B.4 Comparison with PLL Methods

We compare our methods with three popular PLL methods: PRODEN [7], LWS [8], and POP [9]. PRODEN is a classical PLL method based on deep learning that employs progressive disambiguation loss to identify true labels from candidate label sets. LWS introduces a weighted disambiguation loss to refine the true labels. POP is an instance-dependent PLL method that trains the classifier using progressive purification of candidate labels. To adapt MIPL data for PLL methods, we employ two strategies from the existing MIPL literature [4] to convert MIPL data into PLL data. 1) Mean: For each bag, we compute the average value of instances in each feature dimension as the representative feature value for that dimension. This strategy generates a holistic feature vector that retains the same dimensionality as the original instances. 2) MaxMin: We derive the maximum and minimum values across all instances for each feature dimension and concatenate these values. This strategy produces a holistic feature vector with a combined dimensionality of $2d$.

Fig. A5 presents the classification accuracy and expected calibration error of our six methods and the three PLL methods on the CRC-MIPL dataset using ResNet-34 features. Our methods consistently achieve the highest classification accuracy and the lowest expected calibration error across all three datasets. The results demonstrate that our methods substantially outperform the comparative PLL methods.

B.5 CDL for PLL Methods

The proposed CDL is a plug-and-play loss function that can be seamlessly incorporated into both MIPL and PLL frameworks. To adapt CDL for PLL, we propose two variants for POP [9], namely POP-CC and POP-CN. The POP-CC variant integrates the first instantiation $\mathcal{L}_{\text{CDL-CC}}$ into POP, whereas POP-CN incorporates the second instantiation $\mathcal{L}_{\text{CDL-CN}}$.

Table A1 reports the classification accuracy and ECE of these three methods on the CRC-MIPL datasets, using

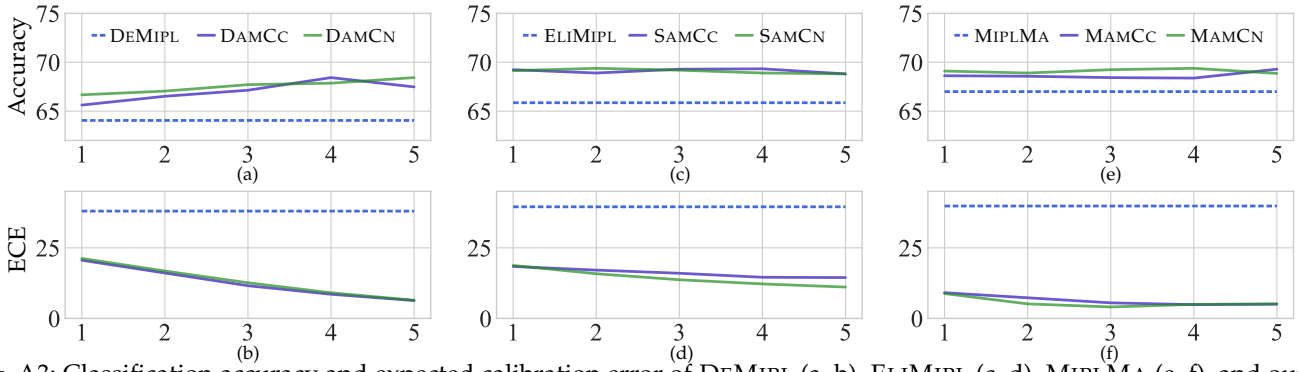


Fig. A3: Classification accuracy and expected calibration error of DEMIPL (a, b), ELIMIPL (c, d), MIPLMA (e, f), and our six methods on the C-R34-25 dataset for $\gamma \in \{1, 2, 3, 4, 5\}$. The horizontal axis represents γ.

TABLE A1: Classification accuracy and expected calibration error (mean±std%) of POP, POP-CC, and POP-CN on the CRC-MIPL datasets using ResNet-34.

	C-R34-9		C-R34-16		C-R34-25	
	Mean	MaxMin	Mean	MaxMin	Mean	MaxMin
			Accuracy			
POP	49.83±1.28	43.23±1.08	54.86±1.37	44.96±0.93	59.06±1.41	46.16±1.49
POP-CC	61.85±1.67	**53.74±1.61**	66.36±1.43	**56.11±0.78**	69.17±1.07	57.51±1.21
POP-CN	**62.25±1.55**	53.15±1.66	**66.75±1.05**	56.00±1.25	**69.18±1.01**	**57.74±1.04**
			ECE			
POP	25.01±1.43	29.74±2.01	22.00±1.22	28.87±1.23	19.49±1.52	27.26±1.41
POP-CC	16.91±2.14	**18.50±2.21**	13.90±1.38	**16.98±1.54**	**12.94±1.27**	**16.29±1.53**
POP-CN	**16.66±1.45**	19.53±1.83	**13.70±1.31**	17.13±1.71	13.08±1.05	16.35±1.60

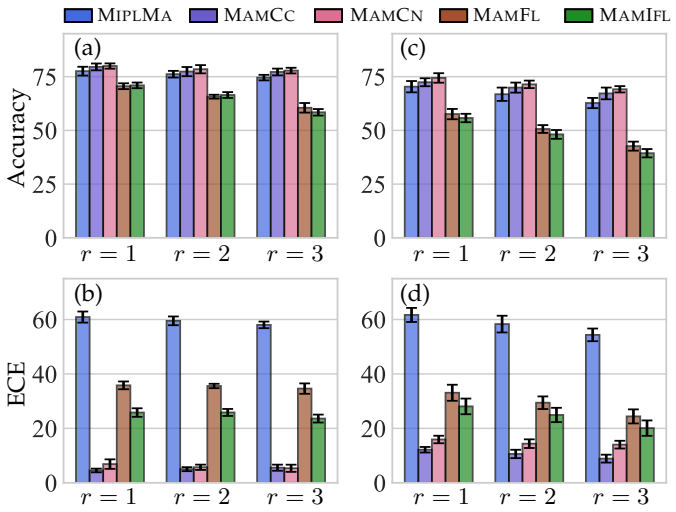


Fig. A4: Classification accuracy and expected calibration error (mean and std) of MIPLMA, MAMCC, MAMCN, MAMFL, and MAMIFL on the Birdsong-MIPL (a, b) and SIVAL-MIPL (c, d) datasets.

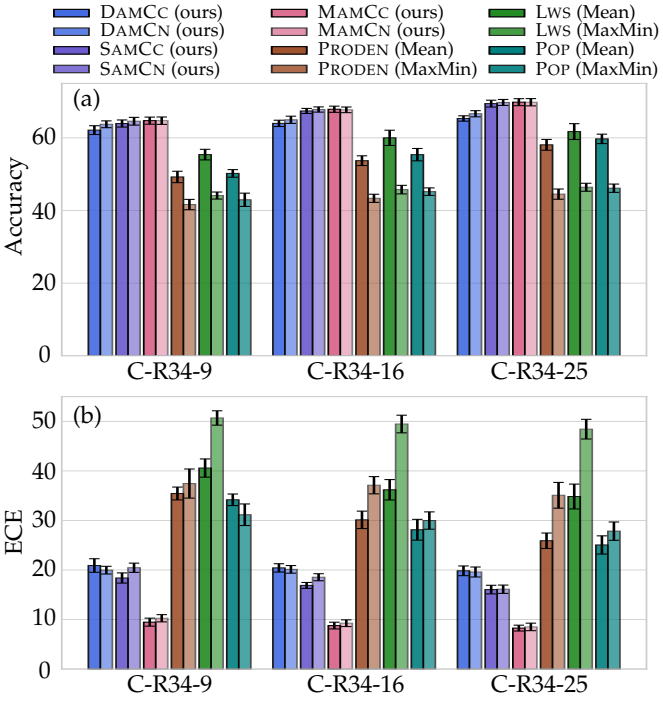


Fig. A5: Classification accuracy and expected calibration error (mean and std) of our six methods and three PLL methods on the CRC-MIPL datasets with ResNet-34 as the feature extractor.

features extracted by ResNet-34. Both variants significantly outperform the vanilla POP, achieving higher classification accuracy and lower expected calibration error, highlighting the effectiveness of CDL in improving both classification and calibration. Notably, the accuracy and expected calibration error of POP-CC and POP-CN are highly similar, indicating that both CDL instantiations are equally effective in enhancing POP's classification and calibration capability.

B.6 Additional Robustness Analysis

Following the taxonomy of [10], we explicitly distinguish *class noise* and *attribute noise* in MIPL, and we include both a discussion and controlled experiments to substantiate the robustness of our method.

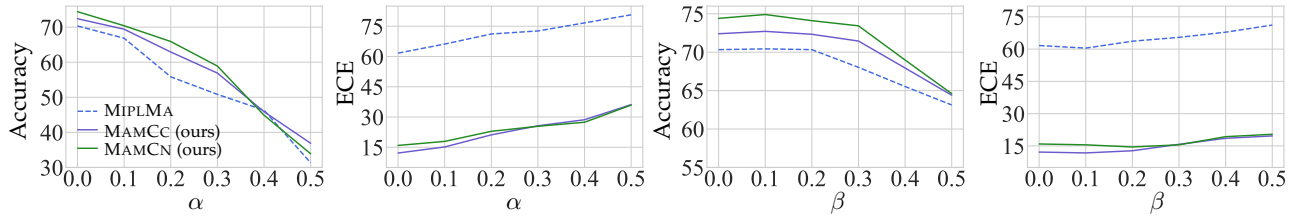


Fig. A6: Classification accuracy and expected calibration error of MIPLMA, MAMCC, and MAMCN on the SIVAL-MIPL dataset ($r = 1$) with varying class noise and attribute noise.

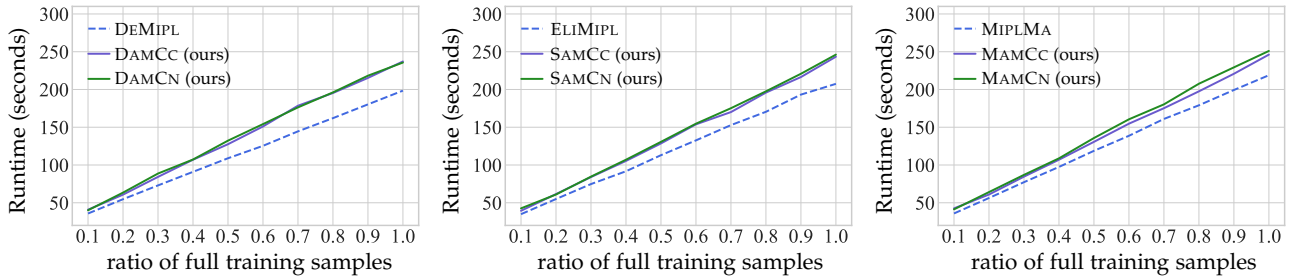


Fig. A7: Runtime of training and testing on the MNIST-MIPL dataset ($r = 1$) versus the fraction of full training samples.

MIPL naturally contains noise in both spaces: the candidate label set is ambiguous and contains false positives; and each bag contains many irrelevant/background instances, and the informative instances are unknown and can be sparse, which introduces instance-level outliers. To quantitatively evaluate robustness under different noise sources, we conduct experiments on the SIVAL-MIPL dataset with $r = 1$ and consider two perturbations: **(i) Class noise.** We randomly select an α proportion of bags and remove the ground-truth label from their candidate label sets, so that these bags are associated with only false-positive labels. We vary $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Note that this is a stringent test that violates the standard MIPL assumption that the true label is included in the candidate label set, and it simulates severe label noise. **(ii) Attribute noise.** We randomly drop a β proportion of instances in each bag, which may remove informative instances (including positive ones) and thus increases the difficulty of the bag. We vary $\beta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Fig. A6 reports the accuracy and ECE of MIPLMA, MAMCC, and MAMCN. Under both class noise and attribute noise, our methods consistently achieve higher or comparable accuracy and much lower ECE than the baselines. For example, under severe class noise ($\alpha = 0.5$), all methods degrade in accuracy, but MAMCC and MAMCN remain noticeably better calibrated (ECE $\approx 35\%$ vs. $\approx 80\%$ for MIPLMA) and retain slightly higher accuracy. Under severe attribute noise ($\beta = 0.5$), MAMCC and MAMCN maintain accuracy around 65% while keeping ECE around 20%, whereas MIPLMA exhibits substantially worse calibration (ECE $\approx 70\%$). Interestingly, removing a small number of instances may improve performance or leave it unchanged. We speculate that this is because randomly removing some uninformative instances allows the model to focus more on learning from informative ones. Since informative instances are sparse in MIPL, increasing the removal ratio degrades the model's performance. Overall, these results support our robustness claim: CDL improves not only classification performance but also, importantly, calibration when the data contain class noise and attribute noise. This observation is also consistent with our theoretical analysis that CDL acts as an adaptive regularizer and alleviates poor calibration arising from ambiguous supervision.

B.7 Scalability Experiments

We conduct additional scalability experiments by progressively increasing the training set size from 10% to 100% of the full training samples. As shown in Fig. A7, the runtime of all methods increases in an approximately linear manner with respect to the proportion of training samples. Moreover, the curves of our methods remain nearly parallel to those of their corresponding baselines, suggesting that our methods preserve the same asymptotic scaling trend with respect to data size as the comparative methods. In addition to the scalability curves, we summarize the computational and resource costs in Table A2, including floating point operations (FLOPs), number of parameters (Params), maximum GPU memory usage (MM), total runtime (Runtime) for training and testing, and the average accuracy (Acc) over 10 runs. Notably, all methods have identical FLOPs (95.58M) and parameter counts (0.13M), indicating that our methods do not increase the model size or theoretical computational complexity. In practice, our methods incur only modest overhead relative to the corresponding baselines, the maximum GPU memory increases by approximately 1.2%–2.1%, and the total runtime increases by about 12.5%–19.5% when using the full training set. Overall, these results demonstrate that our method scales favorably with increasing training data and exhibits good scalability in terms of both runtime growth and resource consumption.

B.8 Statistical Comparison

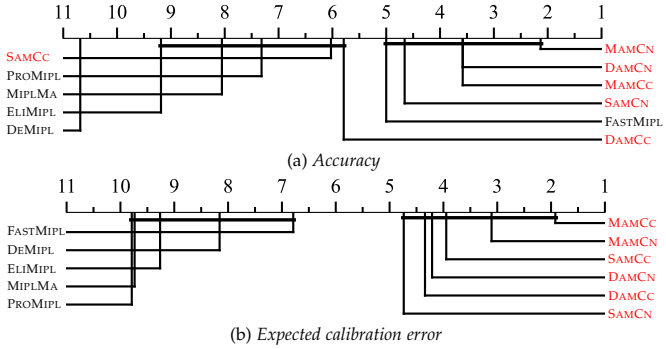
To systematically compare the performance of our methods and the compared methods, we employ the Friedman test [11], a widely used nonparametric procedure for comparing multiple methods across multiple datasets. Given l

TABLE A2: The computational costs on the MNIST-MIPL ($r = 1$) with full training set.

Algorithm	FLOPs (M)	Params (M)	MM (MiB)	Runtime (s)	Acc
DeMipl	95.58	0.13	1822	198.47	97.60±0.80%
DamCc (ours)	95.58	0.13	1858	237.11	99.33±0.52%
DamCn (ours)	95.58	0.13	1858	235.65	99.40±0.55%
EliMipl	95.58	0.13	1824	207.47	99.20±0.65%
SamCc (ours)	95.58	0.13	1863	243.40	99.80±0.43%
SamCn (ours)	95.58	0.13	1863	246.16	99.73±0.44%
MiplMa	95.58	0.13	1848	218.86	98.47±1.03%
MamCc (ours)	95.58	0.13	1870	246.16	99.93±0.20%
MamCn (ours)	95.58	0.13	1870	251.02	99.93±0.20%

TABLE A3: Summary of the Friedman statistics F_F ($l = 11, N = 19$) and the critical value in terms of each evaluation metric (l: # comparing methods; N: # datasets).

Evaluation metric	F_F	critical value ($\alpha = 0.05$)
Accuracy	116.6938	1.8836
Expected calibration error	50.3444	



(a) *Accuracy*

(b) *Expected calibration error*

Fig. A8: Critical difference (CD) diagrams comparing the proposed methods (in red) with the baseline methods using the Nemenyi test. Methods that are not connected in the CD diagram are regarded as having significantly different performance at significance level $\alpha = 0.05$.

methods and N datasets, let $r_{i,j}$ denote the rank of the j-th method on the i-th dataset (mean ranks are assigned in case of ties). The average rank of method j is $R_j = \frac{1}{N}\sum_{i=1}^{N} r_{i,j}$. Under the null hypothesis that all methods have equivalent performance, the Friedman statistic F_F is approximately F-distributed with $(l-1)$ and $(l-1)(N-1)$ degrees of freedom in the numerator and denominator, respectively:

$$F_F = \frac{(N-1)\chi_F^2}{N(l-1)-\chi_F^2}, \tag{A67}$$

where

$$\chi_F^2 = \frac{12N}{l(l+1)}\left[\sum_{j=1}^{l} R_j^2 - \frac{l(l+1)^2}{4}\right]. \tag{A68}$$

Table A3 reports the Friedman statistics F_F and the corresponding critical values for both accuracy and expected calibration error. We exclude MiplGp from the statistical significance analysis because computational constraints prevent its evaluation across multiple datasets. At significance level $\alpha = 0.05$, the null hypothesis of equal performance among the competing methods is rejected for all metrics. This implies that at least one method performs differently from the others, and a post-hoc analysis is required to characterize the relative performance of individual methods.

To this end, we apply the Nemenyi test [12], which compares all pairs of methods based on their average ranks (smaller ranks correspond to better performance). In this test, the absolute difference between the average ranks of two methods is compared with the critical difference (CD):

$$\mathrm{CD} = q_\alpha \sqrt{\frac{l(l+1)}{6N}}, \tag{A69}$$

where q_α is the critical value of the Studentized range statistic for significance level α.

In our experimental setting, we have $q_\alpha = 3.219$ at significance level $\alpha = 0.05$, yielding $\mathrm{CD} = 3.4638$ (with $l = 11$ and $N = 19$). Consequently, the performance of two methods is deemed significantly different if their average ranks over all datasets differ by at least one CD. To visualize the relative performance of the proposed and baseline methods, Fig. A8 shows the CD diagrams for both accuracy and expected calibration error. Each method is positioned on the axis according to its average rank. Any pair of methods whose average ranks differ by less than one CD is connected by a thick horizontal line, indicating that their performances are not significantly different at level $\alpha = 0.05$. Conversely, methods that are not connected are regarded as having significantly different performance.

From the CD diagrams and the accompanying statistics, we make the following observations: 1) The proposed methods obtain the best accuracy in 78.95% of the cases. Most proposed variants appear on the right side of the CD diagram and form a single connected group, which indicates top average ranks with no significant differences within the group. Most baselines lie to the left and are not connected to this group, implying significantly worse accuracy. An exception is FastMipl, whose rank is comparable to the top group. In contrast, SamCc and DamCc do not significantly outperform several baselines. These patterns show that accuracy gains are not uniform across all variants. 2) The proposed methods achieve the lowest ECE in 94.74% of the cases. We can observe a clearer separation in the CD diagram. All proposed methods are grouped on the far right and are disconnected from the baseline cluster on the left. This pattern indicates statistically significant improvements in calibration for every proposed method relative to all baselines, whereas differences within the proposed methods are not statistically significant.

Overall, the CD diagrams show that our methods deliver consistently better calibration and competitive, often supe-

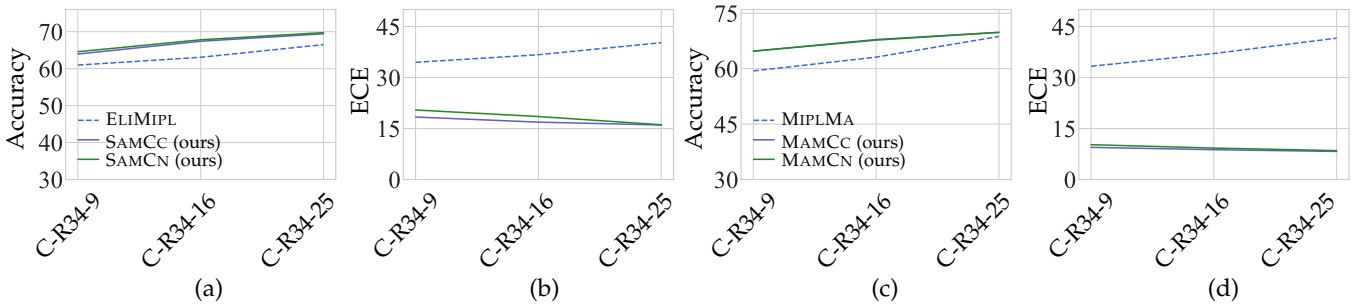


Fig. A9: Effect of patch granularity. (a, b) Accuracy/ECE of EliMipl and our SamCc/SamCn, (c, d) Accuracy/ECE of MiplMa and our MamCc/MamCn.

rior, accuracy compared with the baselines on the benchmark and real-world datasets.

Appendix C
Engineering Case Study

C.1 Engineering Impact of Patch Granularity on the CRC-MIPL

In practical pathology pipelines, a common engineering decision is the patch granularity used to represent a slide as a multi-instance bag: using more patches can capture finer tissue heterogeneity, but also increases the bag size and may change the confidence behavior of the model. To examine the performance on the CRC-mipl dataset, we adopt the ResNet-34 to learn patch-based features and vary the number of non-overlapping patches per image as $N \in \{9, 16, 25\}$, yielding C-R34-9, C-R34-16, and C-R34-25.

Fig. A9 summarizes how both classification and calibration performance evolve as the patch granularity increases. From an engineering perspective, increasing N consistently improves classification accuracy across methods, indicating that finer patch representations provide more informative bag evidence for colorectal cancer (CRC) classification. However, the same increase in granularity can make disambiguation-only baselines less reliable in terms of calibration. As shown in Fig. A9 (b,d), while their accuracies improve, their ECE grows notably as N increases, suggesting that their confidence scores become less suitable for downstream confidence-based decisions. In contrast, our CDL-based variants maintain low and stable ECE across all N while achieving higher accuracy. This behavior is desirable in engineering deployments, where calibrated confidence is needed for confidence-aware operation modes such as triaging low-confidence cases to human review, setting safe decision thresholds, or prioritizing uncertain samples for further inspection. Overall, this study shows that CDL enables practitioners to benefit from finer patch granularity without sacrificing confidence reliability, strengthening its applicability to real CRC pathology pipelines.

We recommend selecting N based on a three-way trade-off among (i) accuracy gain, (ii) calibration reliability (ECE), and (iii) computational budget. A practical rule is to increase N until accuracy improvements begin to saturate while ECE remains acceptable for the intended operating mode. In our CRC pipeline, larger N provides consistent accuracy gains; with CDL, calibration remains stable, so N can be chosen primarily by the desired throughput. On a single NVIDIA RTX 3090 GPU, our training time is approximately 39.97, 41.34, and 43.31 minutes for $N = 9, 16, 25$, respectively, and CDL incurs only about $\sim$10% runtime overhead over the corresponding disambiguation-only baselines. Therefore, when resources allow, $N = 25$ is a strong default to maximize accuracy; $N = 16$ offers a favorable trade-off between performance and cost; and $N = 9$ can be used when throughput is the dominant constraint.

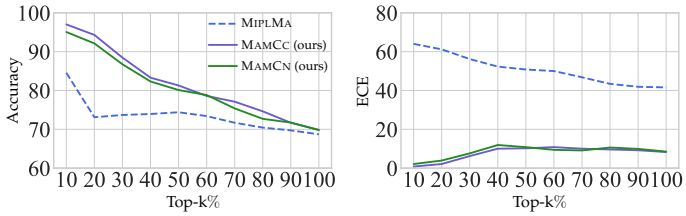


Fig. A10: Classification accuracy and expected calibration error of MiplMa, MamCc, and MamCn on the C-R34-25 dataset with varying top-k% predictions.

C.2 Confidence-Based Triage on the C-R34-25 via Top-$k\%$ Predictions

In engineering deployments of pathology-style MIPL systems, a common operating mode is confidence-aware decision making: the model automatically reports only the most confident predictions, while low-confidence cases are deferred for further review. To assess whether a model's confidence is practically actionable, we analyze classification and calibration on the top-$k\%$ most confident predictions.

For each test bag in the C-R34-25 dataset, we compute the predictive confidence $s = \max_c \hat{p}_c$ (maximum predicted class probability), and rank all test bags by s in descending order. For $k \in \{10, 20, \ldots, 100\}$, we retain the top-$k\%$ predictions and evaluate: (i) *Accuracy@Top-$k\%$* and (ii) *ECE@Top-$k\%$* on the retained subset. This evaluation directly reflects how reliable the model is on uncertain cases.

Fig. A10 compares MiplMa with our CDL-enhanced variants MamCc and MamCn. Two observations are most relevant for triage. First, MamCc and MamCn achieve markedly higher Accuracy@Top-$k\%$ in the low-coverage regime (e.g., top-10% and top-20%) and then decrease smoothly as k increases, suggesting that CDL yields a more informative confidence ranking: the most confident

predictions are substantially more likely to be correct. Second, MAMCC and MAMCN maintain consistently low ECE across k, indicating that their probability estimates remain numerically reliable for thresholding and downstream decision policies. In contrast, MIPLMA exhibits both lower Accuracy@Top-$k\%$ and much larger ECE, implying that its confidence scores are less effective for filtering and are poorly aligned with empirical correctness. This trend aligns with the under-confident behavior observed for MIPL baselines in the reliability analysis (as shown in Fig. 2).

In deployment, one can choose an operating point k or a confidence threshold τ according to the desired safety-throughput trade-off: (i) use Accuracy@Top-$k\%$ to estimate the expected correctness of auto-reported cases (risk $\approx 1 -$ Accuracy@Top-$k\%$), and (ii) use ECE@Top-$k\%$ to assess whether the reported confidence scores can be trusted for thresholding and risk communication. Then, the system auto-reports the top-$k\%$ predictions or those with $s \geq \tau$ and routes the remaining low-confidence cases to human review or further testing. Because MAMCC and MAMCN achieve high accuracy at low coverage and low ECE across operating points, they enable safer and more reliable confidence-based triage on CRC pathology data than MIPLMA.

REFERENCES

[1] W. Tang, W. Zhang, and M.-L. Zhang, "Disambiguated attention embedding for multi-instance partial-label learning," in *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, 2023, pp. 56 756–56 771.

[2] W. Tang, W. Zhang, and M.-L. Zhang, "Exploiting conjugate label information for multi-instance partial-label learning," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, South Korea*, 2024, pp. 4973–4981.

[3] W. Tang, Y.-F. Yang, Z. Wang, W. Zhang, and M.-L. Zhang, "Multi-instance partial-label learning with margin adjustment," in *Advances in Neural Information Processing Systems 37, Vancouver, Canada*, 2024, pp. 26 331–26 354.

[4] W. Tang, W. Zhang, and M.-L. Zhang, "Multi-instance partial-label learning: Towards exploiting dual inexact supervision," *Science China Information Sciences*, vol. 67, no. 3, pp. 132 103:1–132 103:14, 2024.

[5] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[6] D.-B. Wang, L. Feng, and M.-L. Zhang, "Rethinking calibration of deep neural networks: Do not be afraid of overconfidence," in *Advances in Neural Information Processing Systems 34, Virtual Event*, 2021, pp. 11 809–11 820.

[7] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama, "Progressive identification of true labels for partial-label learning," in *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*, 2020, pp. 6500–6510.

[8] H. Wen, J. Cui, H. Hang, J. Liu, Y. Wang, and Z. Lin, "Leveraged weighted loss for partial label learning," in *Proceedings of the 38th International Conference on Machine Learning, Virtual Event*, 2021, pp. 11 091–11 100.

[9] N. Xu, B. Liu, J. Lv, C. Qiao, and X. Geng, "Progressive purification for instance-dependent partial label learning," in *Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 38 551–38 565.

[10] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210, 2004.

[11] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[12] P. B. Nemenyi, "Distribution-free multiple comparisons." Ph.D. dissertation, 1963.