INSTANCE-DEPENDENT COMPLEMENTARY MULTI-LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Complementary multi-label learning (CMLL) is a prevalent weakly supervised multi-label classification paradigm where each training instance is annotated with a complementary label set containing only irrelevant labels. Existing CMLL methods commonly assume that complementary labels are selected uniformly. However, this assumption is often violated in real-world scenarios, where the selection of complementary labels naturally depends on the characteristics of individual instances. In this paper, we study instance-dependent CMLL and propose ID-CMLL, a propensity-weighted framework for explicitly modeling and mitigating instance-dependent selection bias. We first formalize the instance-dependent complementary label generation mechanism and derive a selection-weighted unbiased risk estimator. Since the instance-conditioned selection probabilities are unknown, we introduce an identifiable double-logistic model over a fixed representation to factorize the observed complementary label probability, together with a semantic orientation mechanism for reliable propensity estimation. Empirical results demonstrate that ID-CMLL achieves superior performance against state-of-the-art CMLL and partial multi-label learning (PML) baselines.

1 INTRODUCTION

Supervised learning has achieved great success in various traditional tasks. Unfortunately, collecting strong supervision is extremely expensive in many real-world applications, and only weakly supervised data are available. Weakly supervised learning aims to learn effective models with such data, and many specific paradigms have been well studied, *e.g.*, complementary label learning (Ishida et al., 2017; 2019; Wang et al., 2024; Gao et al., 2023; 2024; 2025b; 2026), partial label learning (Zhang & Yu, 2015; Feng et al., 2020; Xie & Huang, 2018; Wang et al., 2025; Yu et al., 2018a; Sun et al., 2019; Li et al., 2024), positive and unlabeled learning (du Plessis et al., 2015; Kiryo et al., 2017; Elkan & Noto, 2008; Wang et al., 2026b; Xinrui et al., 2023; Li et al., 2022), and noisy label learning (Natarajan et al., 2013; Li et al., 2020; Song et al., 2022), to name just a few. Among these, complementary multi-label learning (CMLL) (Gao et al., 2025b) addresses the multi-label setting (Zhang & Zhou, 2013; Tsoumakas & Katakis, 2007) and considers the scenario where each instance is annotated with labels definitively irrelevant to the ground-truth ones.

To recover the multi-label classifier from complementary information, existing efforts predominantly focus on the single complementary label scenario, assuming a uniform selection process or modeling label-dependent bias through transition matrices (Gao et al., 2023; 2024). For the more general complementary multi-label learning (CMLL), where each instance is associated with multiple complementary labels, research primarily investigates risk consistency under simplified random generation processes (Gao et al., 2025b). However, these approaches generally do not account for the instance-dependent nature of complementary labels, where the selection process is conditioned on the distinctive features of each instance. In real-world scenarios, annotators are more likely to identify labels that are obviously irrelevant for a specific instance, leading to inherent selection bias.

In this paper, we propose ID-CMLL, a propensity-weighted framework that explicitly models and mitigates instance-dependent selection bias in CMLL. We first formalize the instance-dependent complementary label generation mechanism and derive a selection-weighted unbiased risk estimator. To handle unknown selection probabilities, we construct a double-logistic model over a fixed representation to factorize the observed complementary label probability, and establish identifiability and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of complementary-label learning methods. "Est." denotes learned selection, transition, or propensity correction; "RCR", corrected or reweighted classification risk; "URE", exact or asymptotic unbiasedness of the risk estimator; and "C.", consistency with the supervised population risk minimizer within the hypothesis class, as defined in the corresponding reference. For ID-CMLL, Eq. (7) is exactly unbiased, while Eq. (19) is asymptotically unbiased.

Method	Est.	RCR	URE	C.	Core mechanism
Foundational CLL (not directly benchmarked)					
CLL-URE (Ishida et al., 2019)	×	✓	✓	✓	Uniform-selection URE
FWD (Yu et al., 2018b)	✓	✓	×	✓	Estimated transition + forward correction
Complementary(-multi-)label learning baselines					
GDF (Gao et al., 2023)	×	✓	×	×	Gradient-friendly complementary loss
MLCL (Gao et al., 2024)	✓	✓	×	✓	Estimated transition + label-correlation correction
NMCB (Gao et al., 2026)	×	×	×	×	Mixup + label-correlation consistency
ComRank (Zhu et al., 2025)	×	×	–	✓	Bayes-consistent ranking surrogate
CTL (Gao et al., 2025b)	×	✓	×	×	Risk-consistent estimator + confidence truncation
ID-CMLL (proposed)	✓	✓	✓*	✓	Instance-dependent propensity + IPW

a semantic orientation mechanism for reliable propensity estimation. Finally, we incorporate Hájek self-normalization (Hájek, 1971) and non-negative correction (Ishida et al., 2019) to stabilize the inverse propensity weighted risk and improve the training stability of the final classifier. Our main contributions of this work can be summarized as follows.

- We formalize the instance-dependent selection mechanism for CMLL and derive a propensity-weighted unbiased risk estimator.
- We propose ID-CMLL, a unified framework that combines identifiable double-logistic factorization, semantic orientation, Hájek self-normalization, and non-negative correction for stable classifier training.
- We verify the effectiveness of our method through extensive experiments on eleven benchmark datasets under multiple complementary-label selection mechanisms.

2 Methodology

In this section, we introduce the proposed ID-CMLL method for instance-dependent complementary multi-label learning.

2.1 Problem Setup

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the feature space and $\mathcal{Y} = \{1, \ldots, K\}$ the label space with K possible labels ($K > 2$). In MLL, an instance $\mathbf{x} \in \mathcal{X}$ is associated with a true label vector $\mathbf{y} \in \{0, 1\}^K$, where $y_j = 1$ indicates that label j is relevant and $y_j = 0$ indicates that it is irrelevant, drawn from an unknown joint distribution with density $p(\mathbf{x}, \mathbf{y})$. We denote by $\eta_j(\mathbf{x}) = p(y_j = 1 \mid \mathbf{x})$ the ground-truth posterior for label j. The goal of MLL is to learn a classifier $f \in \mathcal{F}$, where $\mathcal{F}$ denotes the hypothesis class of functions $f : \mathcal{X} \to [0, 1]^K$. Considering label-wise binary losses, with binary cross-entropy (BCE) as a standard example (Zhang & Zhou, 2013; Tsoumakas & Katakis, 2007), the expected classification risk is

$$R(f) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \left[\sum_{j=1}^{K} y_j \, \ell(f_j(\mathbf{x})) + \sum_{j=1}^{K} (1 - y_j) \, \bar{\ell}(f_j(\mathbf{x})) \right], \tag{1}$$

where $\ell(f_j(\mathbf{x}))$ and $\bar{\ell}(f_j(\mathbf{x}))$ denote the loss incurred when predicting label j as relevant and irrelevant, respectively. For the BCE loss, $\ell(f_j(\mathbf{x})) = -\log f_j(\mathbf{x})$ and $\bar{\ell}(f_j(\mathbf{x})) = -\log(1 - f_j(\mathbf{x}))$, where $f_j(\mathbf{x})$ estimates the posterior $\eta_j(\mathbf{x})$.

In CMLL, each training instance is equipped with a set of complementary labels rather than relevant ones, where a complementary label is a label confirmed to be irrelevant to the instance (e.g., "car" for an image of a cat). We represent the complementary label information for instance $\mathbf{x}_i$ as a binary vector $\bar{\mathbf{y}}_i \in \{0,1\}^K$, where the j-th element $\bar{y}_{ij} = 1$ indicates that label j is observed as complementary, i.e., $y_{ij} = 0$, whereas $\bar{y}_{ij} = 0$ does not tell whether label j is relevant or irrelevant but unselected. The training set is $\bar{\mathcal{D}} = \{(\mathbf{x}_i, \bar{\mathbf{y}}_i)\}_{i=1}^n$, and at least one complementary label is observed per instance.

2.2 Instance-Dependent Complementary Label Generation

Prior studies (*e.g.*, GDF (Gao et al., 2023)) assume that each complementary label is drawn uniformly and independently of the instance from the pool of irrelevant labels at the level of method design. While operationally convenient, this uniform selection process fails to capture the instance-dependent selection bias arising in human annotation, where annotators more readily identify labels that are obviously absent given the distinctive features $\mathbf{x}$ of a specific instance. We formalize this selection mechanism and extend it to the instance-dependent setting through a per-dimension label transition framework (Cheng et al., 2022).

For each label dimension $j \in \mathcal{Y}$, the relationship between the true label $y_j \in \{0,1\}$ and the observed complementary indicator $\bar{y}_j \in \{0,1\}$ is governed by a 2×2 row-stochastic transition matrix $T_j(\mathbf{x})$, whose (a,b)-entry equals $p(\bar{y}_j = b \mid \mathbf{x}, y_j = a)$. Two entries of this matrix are structurally fixed for all j and all $\mathbf{x}$ by the definition of complementary labels: a truly relevant label ($y_j = 1$) can never appear as a complementary label, so $\bar{y}_j$ must equal 0 with certainty. The top row, corresponding to $y_j = 0$, contains the single free parameter that governs how often a truly irrelevant label is selected as complementary. We denote this parameter $e_j(\mathbf{x})$ and define it formally as follows.

Definition 2.1 (Instance-Dependent Selection Probability)**.** The instance-dependent selection probability for label j at instance $\mathbf{x}$ is

$$e_j(\mathbf{x}) \;=\; p(\bar{y}_j = 1 \mid \mathbf{x},\; y_j = 0) \in (0,\, 1], \tag{2}$$

the probability that label j is selected as a complementary label given that it is truly irrelevant to $\mathbf{x}$.

With $e_j(\mathbf{x})$ in hand, the complete per-dimension transition matrix is

$$T_j(\mathbf{x}) \;=\; \begin{array}{c} \\ y_j = 0 \\ y_j = 1 \end{array} \begin{array}{c} \begin{array}{cc} \bar{y}_j = 0 & \bar{y}_j = 1 \end{array} \\ \begin{pmatrix} 1 - e_j(\mathbf{x}) & e_j(\mathbf{x}) \\ 1 & 0 \end{pmatrix} \end{array}. \tag{3}$$

The bottom row is fixed at $(1,\, 0)$ for every j and $\mathbf{x}$, reflecting the hard constraint that a relevant label is never annotated as complementary. The top row sums to one by row normalization, so $e_j(\mathbf{x})$ is the only quantity that needs to be estimated per dimension. In the class-conditional (instance-independent) special case, $e_j(\mathbf{x}) = e_j$ is a constant and $T_j(\mathbf{x})$ reduces to a fixed matrix T_j. The present work relaxes this restriction by allowing $e_j(\mathbf{x})$ to vary with the instance, turning each T_j into a matrix-valued function of $\mathbf{x}$ while preserving the same structural constraints. From Eq. (3), we have $\mathbb{E}[\bar{y}_j \mid x, y_j] = (1 - y_j)\, e_j(x)$.

To extend from a single dimension to the full label vector, we impose the following label-wise marginal selection condition.

Assumption 2.2 (Instance-Dependent Selected At Random, ID-SAR)**.** For each label dimension $j \in \mathcal{Y}$, an irrelevant label is selected as complementary with an instance-dependent probability $e_j(\mathbf{x})$, *i.e.*, $p(\bar{y}_j = 1 \mid \mathbf{x}, y_j = 0) = e_j(\mathbf{x})$. In contrast, a relevant label is never selected as complementary, *i.e.*, $p(\bar{y}_j = 1 \mid \mathbf{x}, y_j = 1) = 0$.

Assumption 2.2 generalizes the label-dependent yet instance-independent selection model of Gao et al. (2024), in which $e_j(\mathbf{x}) = e_j$ is a label-specific constant, to the instance-dependent setting.

2.3 Selection-Weighted Unbiased Risk Estimator

We address the following question: *given the selection probabilities* $\{e_j(\mathbf{x})\}$*, how can* $R(f)$ *be estimated unbiasedly from* $\bar{\mathcal{D}}$*?* Estimation of these selection probabilities when they are unknown is deferred to Section 2.4.

Lemma 2.3 (Irrelevance-Selection Factorization)**.** *Under Assumption 2.2, for any label $j \in \mathcal{Y}$ and measurable function $f_j : \mathcal{X} \to \mathbb{R}$,*

$$\mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\frac{\bar{y}_j \cdot f_j(\mathbf{x})}{e_j(\mathbf{x})}\right] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1-y_j)\cdot f_j(\mathbf{x})]\,. \tag{4}$$

Lemma 2.4 (Relevance-Irrelevance Decomposition)**.** *For any label $j \in \mathcal{Y}$ and measurable function $f_j : \mathcal{X} \to \mathbb{R}$,*

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[y_j \cdot f_j(\mathbf{x})] = \mathbb{E}_{p(\mathbf{x})}[f_j(\mathbf{x})] - \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1-y_j)\cdot f_j(\mathbf{x})]\,. \tag{5}$$

Theorem 2.5 (Selection-Weighted Risk Equivalence)**.** *Under Assumption 2.2,*

$$R(f) = \mathbb{E}_{p(\mathbf{x})}\left[\sum_{j=1}^{K}\ell(f_j(\mathbf{x}))\right] + \mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\sum_{j=1}^{K}\frac{\bar{y}_j}{e_j(\mathbf{x})}\Big(\bar{\ell}(f_j(\mathbf{x})) - \ell(f_j(\mathbf{x}))\Big)\right]. \tag{6}$$

The first term is evaluated over all K labels for each instance and the second is nonzero only at observed complementary positions $\bar{y}_{ij} = 1$, where the inverse weight $1/e_j(\mathbf{x})$ corrects for the under-representation of rarely selected labels. Replacing population expectations with sample averages gives the empirical estimator

$$\hat{R}(f;e) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\ell(f_j(\mathbf{x}_i)) + \sum_{j=1}^{K}\frac{\bar{y}_{ij}}{e_j(\mathbf{x}_i)}\Big(\bar{\ell}(f_j(\mathbf{x}_i)) - \ell(f_j(\mathbf{x}_i))\Big)\right]. \tag{7}$$

2.4 Identifiable Propensity Estimation under a Fixed Representation

The risk estimator in Eq. (7) requires the instance-dependent selection probabilities $\{e_j(\mathbf{x})\}_{j=1}^{K}$, which are not directly observed. We introduce a fixed-representation double-logistic model to identify and estimate these probabilities from complementary labels. Because $\bar{y}_j = 1$ can occur only when $y_j = 0$, applying the law of total probability to the complementary-label generation process gives

$$q_j(\mathbf{x}) = p(\bar{y}_j = 1 \mid \mathbf{x}) = p(y_j = 0 \mid \mathbf{x})\, p(\bar{y}_j = 1 \mid \mathbf{x}, y_j = 0) = \big(1-\eta_j(\mathbf{x})\big)e_j(\mathbf{x}), \tag{8}$$

where $1-\eta_j(\mathbf{x})$ denotes the irrelevance probability, and $e_j(\mathbf{x})$ denotes the selection probability. Although $q_j(\mathbf{x})$ can be estimated from $\bar{\mathcal{D}}$, its split into $1-\eta_j(\mathbf{x})$ and $e_j(\mathbf{x})$ is not identified without further restrictions. For any measurable a_j with $a_j(\mathbf{x})$ satisfying $q_j(\mathbf{x}) \le a_j(\mathbf{x}) \le 1, a_j(\mathbf{x}) > 0$, the pair $e'_j(\mathbf{x}) = a_j(\mathbf{x})$, $\eta'_j(\mathbf{x}) = 1 - \frac{q_j(\mathbf{x})}{a_j(\mathbf{x})}$ also satisfies $q_j(\mathbf{x}) = \big(1-\eta'_j(\mathbf{x})\big)e'_j(\mathbf{x})$, so the observed complementary-label distribution alone cannot separate $\eta_j(\mathbf{x})$ from $e_j(\mathbf{x})$.

To obtain identification, we restrict the two latent factors to a parametric double-logistic family over a fixed representation. Specifically, we first learn a shared representation $\mathbf{z} = h(\mathbf{x};\boldsymbol{\theta}_0) \in \mathbb{R}^d$ during a representation-learning stage. After this stage, the backbone parameters $\boldsymbol{\theta}_0$ are frozen throughout propensity estimation. Thus, the two factor heads share the same representation, while the representation itself is not updated by the observed-data objective. We make the following parametric correct-specification assumption.

Assumption 2.6 (Fixed-Representation Double-Logistic Specification)**.** Conditional on the fixed representation $h(\mathbf{x};\boldsymbol{\theta}_0)$, for each label $j \in \mathcal{Y}$, there exist population parameters $\mathbf{w}^*_{j1}, c^*_{j1}, \mathbf{w}^*_{j2}$, and c^*_{j2} such that

$$1-\eta_j(\mathbf{x}) = \sigma\big(\mathbf{w}^{*\top}_{j1}h(\mathbf{x};\boldsymbol{\theta}_0) + c^*_{j1}\big)\,, \qquad e_j(\mathbf{x}) = \sigma\big(\mathbf{w}^{*\top}_{j2}h(\mathbf{x};\boldsymbol{\theta}_0) + c^*_{j2}\big)\,, \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Under Assumption 2.6, the observed marginal probability in Eq. (8) is the product of two affine-logistic functions over the same fixed representation. Accordingly, for each label j, we fit two unordered affine-logistic factors

$$g_{j1}(\mathbf{x}) = \sigma\big(\mathbf{w}^{\top}_{j1}h(\mathbf{x};\boldsymbol{\theta}_0) + c_{j1}\big)\,, \qquad g_{j2}(\mathbf{x}) = \sigma\big(\mathbf{w}^{\top}_{j2}h(\mathbf{x};\boldsymbol{\theta}_0) + c_{j2}\big)\,. \tag{10}$$

Before semantic orientation, neither fitted factor is assigned the role of the irrelevance probability $1-\eta_j(\mathbf{x})$ or the selection probability $e_j(\mathbf{x})$. Their product defines the model-implied complementary-label probability

$$q_j(\mathbf{x};\boldsymbol{\Theta}) = g_{j1}(\mathbf{x})g_{j2}(\mathbf{x}), \tag{11}$$

where $\mathbf{\Theta} = \{\mathbf{w}_{j1}, c_{j1}, \mathbf{w}_{j2}, c_{j2}\}_{j=1}^{K}$.

Because $\bar{y}_{ij}$ is observed with marginal conditional success probability $q_j(\mathbf{x}_i)$, we estimate the two factors by minimizing the label-wise Bernoulli negative log-likelihood

$$\widehat{\mathcal{L}}_{\text{obs}}(\mathbf{\Theta}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\left[\bar{y}_{ij}\log q_{ij} + (1-\bar{y}_{ij})\log\left(1-q_{ij}\right)\right] + \lambda_n \sum_{j=1}^{K}\sum_{k=1}^{2}\left(\|\mathbf{w}_{jk}\|_2^2 + c_{jk}^2\right), \tag{12}$$

where $q_{ij} = q_j(\mathbf{x}_i; \mathbf{\Theta})$ and $\lambda_n \geq 0$ is a symmetric ridge penalty used for numerical stability rather than identification. We use a fixed value in experiments, while the asymptotic analysis in Section Section 2.5 assumes $\lambda_n \to 0$ as $n \to \infty$. Since Eq. (12) is constructed from label-wise marginals, it remains a valid composite likelihood when complementary indicators are dependent across labels. The two factors are initialized from the warm-up classifier (Appendix D), which provides a starting point, and their semantic roles are fixed by Eq. (15).

Under Assumption 2.6, a relative-open support condition on the fixed representation, and the generic non-degeneracy conditions stated in Section 2.5, the observed function $q_j(\mathbf{x})$ identifies the unordered pair $\{1-\eta_j(\mathbf{x}), e_j(\mathbf{x})\}$ up to permutation. Thus, the parametric double-logistic restriction reduces the original nonparametric ambiguity to a two-fold permutation ambiguity, namely which of the two fitted factors corresponds to $1-\eta_j(\mathbf{x})$ and which to $e_j(\mathbf{x})$. To determine the semantic orientation of the two identified factors, we impose the following condition.

Assumption 2.7 (Average Irrelevance Dominance)**.** For each label $j \in \mathcal{Y}$, there exists a constant $\delta_j > 0$ such that

$$\mathbb{E}_{p(\mathbf{x})}\left[1-\eta_j(\mathbf{x})\right] \geq \mathbb{E}_{p(\mathbf{x})}\left[e_j(\mathbf{x})\right] + \delta_j. \tag{13}$$

Assumption 2.7 requires label irrelevance to be more prevalent on average than complementary-label selection. It does not require $1-\eta_j(\mathbf{x}) \geq e_j(\mathbf{x})$ for every instance. This condition is plausible for multi-label annotation. Multi-label data are typically sparse, so each label is irrelevant to most instances, whereas annotators mark only a portion of the irrelevant labels, so the average selection probability is usually smaller than the average irrelevance probability.

Let $\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \widehat{\mathcal{L}}_{\text{obs}}(\mathbf{\Theta})$ and let $\hat{g}_{j1}$ and $\hat{g}_{j2}$ denote the corresponding fitted factors. We compute their empirical means as

$$\hat{m}_{j1} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{j1}(\mathbf{x}_i), \quad \hat{m}_{j2} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_{j2}(\mathbf{x}_i). \tag{14}$$

The factor with the larger empirical mean is assigned to the irrelevance probability. Accordingly, the estimated relevance and selection probabilities are defined by

$$\left(\hat{\eta}_j(\mathbf{x}), \hat{e}_j(\mathbf{x})\right) = \begin{cases} \left(1-\hat{g}_{j1}(\mathbf{x}), \hat{g}_{j2}(\mathbf{x})\right), & \text{if } \hat{m}_{j1} \geq \hat{m}_{j2}, \\ \left(1-\hat{g}_{j2}(\mathbf{x}), \hat{g}_{j1}(\mathbf{x})\right), & \text{otherwise.} \end{cases} \tag{15}$$

Under Assumption 2.7 and consistent estimation of the unordered factors, this empirical comparison recovers their semantic ordering with probability approaching one as the sample size increases.

After semantic orientation, let $\pi_j := \mathbb{E}_{p(\mathbf{x})}[1-\eta_j(\mathbf{x})]$ denote the marginal irrelevance probability of label j. We estimate it by

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\left(1-\hat{\eta}_j(\mathbf{x}_i)\right). \tag{16}$$

The quantity $\hat{\pi}_j$ is not an additional learned model component. It is the empirical mean of the estimated irrelevance probability after factor orientation.

Replacing the unknown selection probability $e_j(\mathbf{x}_i)$ in Eq. (7) with its oriented estimate $\hat{e}_j(\mathbf{x}_i)$ gives the feasible plug-in estimator

$$\hat{R}(f;\hat{e}) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\ell(f_j(\mathbf{x}_i)) + \sum_{j=1}^{K}\frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\left(\bar{\ell}(f_j(\mathbf{x}_i)) - \ell(f_j(\mathbf{x}_i))\right)\right]. \tag{17}$$

Unlike the oracle estimator in Eq. (7), the feasible estimator in Eq. (17) is not generally unbiased in finite samples because the true selection probabilities are replaced by estimated quantities.

Non-negative and Hájek corrections. First, small values of $\hat{e}_j(\mathbf{x}_i)$ may produce large inverse weights. We therefore define the label-specific normalizing factor $\hat{W}_j = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{ij}/\hat{e}_j(\mathbf{x}_i)$, and set the corresponding normalized correction to zero when $\hat{W}_j = 0$. The Hájek-type risk is

$$\hat{R}^{\mathrm{H}}(f;\hat{\pi},\hat{e}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\ell(f_j(\mathbf{x}_i)) + \frac{1}{n}\sum_{j=1}^{K}\frac{\hat{\pi}_j}{\hat{W}_j}\sum_{i=1}^{n}\frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\Big(\bar{\ell}(f_j(\mathbf{x}_i)) - \ell(f_j(\mathbf{x}_i))\Big). \tag{18}$$

Dividing by $\hat{W}_j$ normalizes the inverse-propensity-weighted complementary observations within label j, while $\hat{\pi}_j$ restores the estimated marginal probability mass of the irrelevant-label population.

Second, to prevent the complete empirical objective from becoming negative during optimization, we apply a label-wise non-negative correction to the Hájek-type risk

$$\hat{R}^{\mathrm{H}+}(f;\hat{\pi},\hat{e}) = \frac{1}{n}\sum_{j=1}^{K}\left|\sum_{i=1}^{n}\ell(f_j(\mathbf{x}_i)) - \frac{\hat{\pi}_j}{\hat{W}_j}\sum_{i=1}^{n}\frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\ell(f_j(\mathbf{x}_i))\right| + \frac{1}{n}\sum_{j=1}^{K}\frac{\hat{\pi}_j}{\hat{W}_j}\sum_{i=1}^{n}\frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\bar{\ell}(f_j(\mathbf{x}_i)). \tag{19}$$

The Hájek-type normalization and the non-negative correction are introduced for numerical stability and do not preserve exact finite-sample unbiasedness. Their consistency properties under consistent factor estimation, positivity, and the remaining regularity conditions are established in Section 2.5.

After estimating and orienting the two factors, we treat $\hat{e}_j(\mathbf{x})$ and $\hat{\pi}_j$ as fixed quantities and optimize the final classifier according to

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}^{\mathrm{H}+}(f;\hat{\pi},\hat{e}). \tag{20}$$

The estimated factors are treated as fixed constants, with no gradients propagated back to the propensity-estimation module.

2.5 Theoretical Guarantees

Assumption 2.8 (Representation Stability)**.** Let $\hat{\boldsymbol{\theta}}_0$ denote the warm-up backbone parameters estimated from $\bar{\mathcal{D}}$ and frozen for propensity estimation. We assume $\hat{\boldsymbol{\theta}}_0 \xrightarrow{p} \boldsymbol{\theta}_0^{\dagger}$ for some deterministic $\boldsymbol{\theta}_0^{\dagger}$, and that $h(\mathbf{x};\boldsymbol{\theta})$ is Lipschitz in $\boldsymbol{\theta}$, uniformly over $\mathbf{x}\in\mathrm{supp}(X)$, in a neighborhood of $\boldsymbol{\theta}_0^{\dagger}$.

Assumption 2.8 is a standard condition for two-step estimation with an estimated nuisance representation: it lets the finite-sample objective evaluated at the plugged-in $\hat{\boldsymbol{\theta}}_0$ inherit, via a Lipschitz argument, the uniform law of large numbers that holds at the limiting $\boldsymbol{\theta}_0^{\dagger}$. We do not derive the existence of $\boldsymbol{\theta}_0^{\dagger}$ from the GDF warm-up objective itself (see Appendix C.1).

Theorem 2.9 (Identification and Consistent Orientation)**.** *Suppose Assumption 2.8 holds, let $\boldsymbol{\theta}_0^{\dagger}$ denote the resulting limiting backbone, $Z = h(X;\boldsymbol{\theta}_0^{\dagger})$, and suppose Assumption 2.6 holds with $\boldsymbol{\theta}_0$ replaced by $\boldsymbol{\theta}_0^{\dagger}$. For each $j\in\mathcal{Y}$, let $\Theta_j^* = (\mathbf{w}_{j1}^*, c_{j1}^*, \mathbf{w}_{j2}^*, c_{j2}^*)$ denote the true factor parameters and define the permutation-equivalence class $[\Theta_j^*] = \{(\mathbf{w}_{j1}^*, c_{j1}^*, \mathbf{w}_{j2}^*, c_{j2}^*), (\mathbf{w}_{j2}^*, c_{j2}^*, \mathbf{w}_{j1}^*, c_{j1}^*)\}$. Assume that: (i) the support of Z is compact and contains a non-empty open set; (ii) $\mathbf{w}_{j1}^*$, $\mathbf{w}_{j2}^*$, and $\mathbf{w}_{j1}^* + \mathbf{w}_{j2}^*$ are nonzero and pairwise distinct; and (iii) the parameter space is compact and $\lambda_n \to 0$. Then the observed marginal probability identifies the two affine-logistic factors up to permutation. Specifically, regarding q_j as a function of the limiting representation, if $q_j(\mathbf{z};\Theta_j) = q_j(\mathbf{z};\Theta_j^*)$ for every $\mathbf{z}\in\mathrm{supp}(Z)$, then $\Theta_j \in [\Theta_j^*]$. Moreover, let $\widehat{\boldsymbol{\Theta}}$ be any global minimizer of Eq. (12) computed on the estimated representation $h(\cdot;\hat{\boldsymbol{\theta}}_0)$, and let $\widehat{\Theta}_j$ denote its parameters for label j. Then*

$$d\Big(\widehat{\Theta}_j, [\Theta_j^*]\Big) \xrightarrow{p} 0, \qquad d\Big(\widehat{\Theta}_j, [\Theta_j^*]\Big) = \min_{\Theta\in[\Theta_j^*]} \|\widehat{\Theta}_j - \Theta\|_2. \tag{21}$$

If Assumption 2.7 also holds, the empirical-mean rule in Eq. (15) recovers the correct semantic ordering with probability approaching one. Consequently,

$$\|\hat{e}_j - e_j\|_{\infty} \xrightarrow{p} 0, \qquad \|\hat{\eta}_j - \eta_j\|_{\infty} \xrightarrow{p} 0. \tag{22}$$

Table 2: Statistics of the eleven benchmark datasets. n: number of instances; d: feature dimensionality; K: number of labels (capped at 15 for large-label datasets); AvgCLs: average number of complementary labels per instance; AvgGLs: average number of ground-truth labels per instance.

Domain	Dataset	n	d	K	AvgGLs	AvgCLs
Standard benchmarks	scene	2,407	294	6	1.07	2.96
	yeast	2,417	103	14	4.24	6.97
EU legislation	eurlex_dc	8,636	5,000	15	1.02	7.44
	eurlex_sm	13,270	5,000	15	1.75	7.50
News (Reuters)	rcv1-s1	5,815	944	15	1.66	7.47
	rcv1-s2	5,252	944	15	1.67	7.51
	rcv1-s3	5,410	944	15	1.61	7.51
	rcv1-s4	5,761	944	15	1.48	7.46
	rcv1-s5	5,532	944	15	1.65	7.51
Social tagging	bookmark	38,912	2,150	15	1.25	7.49
	delicious	14,784	500	15	4.33	7.46

Theorem 2.9 first identifies the unordered factors and then resolves their semantic ordering through Assumption 2.7.

Assumption 2.10 (Bounded Classifier Range)**.** There exists $\tau \in (0, \frac{1}{2})$ such that $\mathcal{F} \subseteq \{f : \mathcal{X} \to [\tau, 1-\tau]^K\}$. Consequently, the BCE loss pair is bounded by $C_{\mathrm{L}} = -\log \tau$ and ρ-Lipschitz with $\rho = 1/\tau$ on this range. We also assume $e_j(\mathbf{x}) \geq e_{\min} > 0$ and $\pi_j \geq \pi_{\min} > 0$ for all j.

Theorem 2.11 (Asymptotic Unbiasedness of the Non-Negative Hájek Risk)**.** *Under Assumption 2.10, suppose further that* $\max_j \|\hat{e}_j - e_j\|_\infty \xrightarrow{p} 0$ *and* $\max_j |\hat{\pi}_j - \pi_j| \xrightarrow{p} 0$*. Then, for every fixed* $f \in \mathcal{F}$*,*

$$\hat{R}^{H+}(f; \hat{\pi}, \hat{e}) \xrightarrow{p} R(f), \qquad \mathbb{E}\big[\hat{R}^{H+}(f; \hat{\pi}, \hat{e})\big] \to R(f). \tag{23}$$

As shown in Appendix I, the label-wise correction in Eq. (19) is applied to the estimated component whose population counterpart is $\mathbb{E}[y_j \ell(f_j(\mathbf{x}))] \geq 0$ for every $f \in \mathcal{F}$. Therefore, $\hat{R}^{H+}(f; \hat{\pi}, \hat{e})$ converges to the original classification risk $R(f)$ without requiring any additional sign condition.

Theorem 2.12 (Estimation Error Bound)**.** *Let* $\hat{f}$ *be defined in Eq. (20) and let* $f^* = \arg\min_{f \in \mathcal{F}} R(f)$*. Under Assumption 2.10, define* $\varepsilon_n = \sup_{f \in \mathcal{F}} \big|\hat{R}^{H+}(f; \hat{\pi}, \hat{e}) - \hat{R}^{H+}(f; \pi, e)\big|$ *and* $\mathcal{G}_j = \{\mathbf{x} \mapsto f_j(\mathbf{x}) \mid f \in \mathcal{F}\}$*. Then there exists a universal constant* $c > 0$ *such that, for any* $\delta > 0$ *and sufficiently large* n*, with probability at least* $1 - \delta$*,*

$$0 \leq R(\hat{f}) - R(f^*) \leq \frac{c\rho}{e_{\min}\pi_{\min}} \sum_{j=1}^{K} \mathcal{R}_n(\mathcal{G}_j) + \frac{cKC_{\mathrm{L}}}{e_{\min}\pi_{\min}} \sqrt{\frac{\log(4K/\delta)}{n}} + 2\varepsilon_n, \tag{24}$$

where $\mathfrak{R}_n(\mathcal{G}_j)$ *denotes the expected Rademacher complexity of* $\mathcal{G}_j$*. If* $\varepsilon_n = o_p(1)$ *and* $\sum_{j=1}^{K} \mathcal{R}_n(\mathcal{G}_j) \to 0$*, then* $R(\hat{f}) - R(f^*) \xrightarrow{p} 0$*.*

Eq. (24) decomposes the excess risk into the standard finite-sample estimation terms and the additional error ε_n induced by estimating (π, e). Together with Theorem 2.11, this establishes consistency with respect to the original classification risk $R(f)$. Moreover, Theorem K.1 in Appendix K shows that $\varepsilon_n \xrightarrow{p} 0$ under consistent estimation of the nuisance quantities.

3 Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of ID-CMLL.

3.1 Experimental Settings

Datasets. We evaluate on eleven multi-label datasets spanning image, text, biology, and multimedia domains, covering diverse scales, label-space sizes, and average label cardinalities. The benchmarks are summarized in Table 2. For datasets whose original label space exceeds fifteen labels, we retain only the fifteen most frequent labels following the pre-processing convention of Gao et al. (2023).

Under the *instance-dependent* (ID-SAR) condition, the complementary-label selection probabilities are constructed from the relevance scores predicted by an auxiliary classifier, such that labels that

Table 3: Experimental results (mean ± std) on benchmark datasets.

Methods	bookmark	delicious	eurlex_dc	eurlex_sm	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	scene	yeast
Coverage Error ↓											
PML-MD	.241±.030	.595±.015	.134±.017	.279±.013	.219±.025	.236±.024	.244±.024	.220±.010	.227±.019	.096±.010	.492±.028
GLC	**.135±.003**	.588±.002	.204±.006	.408±.007	.225±.004	.246±.009	.238±.008	.225±.009	.229±.009	.249±.017	.485±.010
PARD	.162±.004	.536±.012	.187±.010	.353±.008	.178±.010	.186±.012	.186±.010	**.159±.012**	.172±.006	.094±.012	.476±.008
PML-FSMIR	.176±.010	.616±.013	.133±.012	.310±.025	.249±.016	.244±.018	.254±.013	.230±.023	.230±.023	.098±.014	.475±.017
GDF	.160±.006	.545±.010	.080±.005	.219±.026	.186±.010	.245±.014	.246±.014	.214±.014	.189±.014	.099±.007	.476±.005
MLCL	.143±.007	.569±.006	.240±.009	.382±.009	.354±.030	.397±.036	.404±.048	.342±.032	.348±.017	.117±.051	**.443±.009**
NMCB	.728±.326	.898±.001	.866±.003	.868±.001	.877±.002	.875±.003	.876±.003	.876±.003	.879±.002	.588±.263	.904±.002
ComRank	.164±.005	.539±.011	.048±.012	.202±.021	.188±.013	.250±.008	.245±.014	.218±.015	.184±.014	.093±.006	.462±.008
CTL	.145±.005	.564±.008	.082±.006	.232±.009	.172±.009	.244±.012	.239±.015	.216±.019	.180±.008	.098±.005	.483±.007
Ours	.145±.008	**.530±.009**	**.043±.004**	**.171±.008**	**.132±.010**	**.189±.007**	**.185±.019**	.179±.016	**.169±.019**	**.092±.005**	.457±.007
One Error ↓											
PML-MD	.484±.027	.458±.024	.385±.012	.413±.023	.518±.020	.500±.035	.501±.020	.461±.027	.474±.024	**.280±.017**	.257±.033
GLC	.378±.005	.455±.009	.556±.015	.618±.010	.613±.011	.563±.025	.639±.009	.594±.013	.559±.012	.616±.025	.253±.015
PARD	.375±.010	.354±.009	.507±.017	.637±.015	.453±.024	.417±.018	.455±.013	.395±.026	.411±.021	.299±.036	.247±.028
PML-FSMIR	.412±.034	.425±.028	.457±.038	.479±.054	.527±.029	.557±.038	.529±.029	.526±.049	.505±.053	.290±.035	.248±.025
GDF	.365±.007	.365±.010	.309±.013	.379±.139	.546±.028	.590±.028	.621±.052	.552±.039	.524±.016	.308±.019	**.245±.011**
MLCL	**.346±.009**	.426±.013	.498±.017	.419±.040	.644±.021	.681±.081	.859±.088	.575±.069	.664±.030	.296±.038	.276±.004
NMCB	.599±.268	.778±.008	.496±.011	.502±.015	.649±.030	.639±.028	.633±.020	.673±.030	.660±.020	.517±.240	.865±.022
ComRank	.367±.007	.353±.013	.203±.057	.290±.061	.525±.032	.540±.026	.579±.046	.514±.032	.513±.015	.291±.019	**.245±.011**
CTL	.361±.010	.376±.014	.320±.019	.316±.018	.606±.029	.653±.031	.693±.068	.606±.036	.582±.012	.313±.016	**.245±.011**
Ours	.353±.010	**.348±.006**	**.183±.014**	**.272±.019**	**.208±.031**	**.209±.041**	**.254±.085**	**.257±.032**	**.320±.024**	.283±.024	.246±.010
Ranking Loss ↓											
PML-MD	.217±.031	.300±.009	.140±.018	.189±.010	.152±.021	.158±.016	.165±.021	.157±.014	.151±.017	.098±.011	.203±.018
GLC	.131±.002	.288±.003	.216±.006	.307±.006	.159±.003	.161±.008	.170±.005	.166±.008	.161±.008	.281±.019	.211±.007
PARD	.136±.004	.237±.004	.198±.010	.270±.008	.112±.008	.112±.008	.118±.006	.105±.011	.103±.005	.097±.014	.200±.006
PML-FSMIR	.155±.004	.307±.013	.140±.013	.226±.025	.175±.015	.168±.010	.180±.011	.173±.023	.154±.020	.101±.017	.190±.013
GDF	.134±.006	.245±.006	.083±.005	.143±.028	.126±.009	.156±.010	.163±.014	.143±.012	.133±.010	.103±.007	.194±.006
MLCL	.121±.006	.286±.005	.255±.011	.298±.005	.285±.026	.304±.038	.330±.050	.245±.027	.277±.011	.127±.064	.187±.005
NMCB	.450±.201	.672±.003	.496±.004	.492±.002	.554±.006	.562±.004	.553±.006	.554±.008	.551±.007	.472±.212	.743±.002
ComRank	.138±.005	.236±.007	.049±.012	.126±.024	.128±.011	.159±.007	.163±.015	.146±.012	.128±.009	**.095±.007**	.188±.007
CTL	.121±.005	.262±.005	.086±.006	.149±.008	.117±.008	.150±.008	.156±.015	.142±.014	.127±.004	.101±.005	.201±.007
Ours	**.120±.008**	**.234±.007**	**.044±.004**	**.099±.006**	**.068±.005**	**.087±.006**	**.090±.020**	**.101±.011**	**.109±.013**	**.095±.005**	**.180±.006**
Average Precision ↑											
PML-MD	.610±.026	.580±.011	.717±.012	.641±.014	.617±.018	.626±.023	.619±.013	.644±.012	.639±.018	**.832±.011**	.722±.020
GLC	.719±.003	.590±.005	.564±.010	.462±.010	.560±.007	.566±.016	.547±.008	.563±.011	.579±.011	.588±.018	.703±.010
PARD	.707±.007	.647±.005	.601±.013	.481±.011	.675±.015	.694±.013	.674±.009	.713±.017	.703±.013	.826±.022	.719±.011
PML-FSMIR	.668±.014	.583±.014	.669±.026	.579±.044	.583±.023	.580±.022	.587±.024	.598±.041	.615±.037	.826±.022	.732±.016
GDF	.714±.006	.636±.006	.783±.010	.679±.081	.626±.014	.588±.018	.571±.034	.614±.025	.635±.010	.818±.011	.728±.008
MLCL	.728±.008	.572±.007	.587±.015	.546±.021	.489±.023	.434±.053	.350±.068	.528±.045	.477±.016	.814±.041	.721±.008
NMCB	.451±.202	.469±.002	.583±.005	.592±.001	.541±.004	.531±.004	.535±.004	.531±.007	.543±.008	.498±.223	.431±.006
ComRank	.710±.005	**.649±.007**	.864±.034	.741±.045	.633±.018	.606±.015	.587±.031	.628±.023	.644±.007	.829±.011	.736±.010
CTL	.723±.008	.617±.006	.776±.013	.706±.013	.607±.014	.563±.018	.544±.040	.585±.029	.615±.006	.816±.009	.720±.009
Ours	**.730±.009**	.648±.006	**.877±.010**	**.767±.013**	**.821±.011**	**.781±.022**	**.772±.057**	**.756±.015**	**.733±.019**	**.832±.012**	**.744±.009**
Hamming Loss ↓											
PML-MD	.081±.004	.309±.003	.052±.002	.105±.003	.113±.003	.112±.004	.108±.003	**.093±.004**	.106±.003	**.106±.006**	.230±.011
GLC	.067±.001	.306±.001	.075±.002	.132±.001	.126±.001	.120±.004	.126±.002	.111±.002	.118±.002	.218±.009	.255±.005
PARD	.345±.013	.329±.003	.293±.016	.354±.004	.340±.012	.327±.007	.349±.009	.353±.013	.329±.008	.321±.017	.308±.009
PML-FSMIR	.079±.006	.303±.008	.062±.005	.114±.008	.114±.004	.119±.005	.111±.003	.102±.006	.110±.008	.109±.011	.225±.008
GDF	.101±.001	.262±.004	.075±.004	.122±.027	.145±.003	.155±.005	.162±.008	.140±.007	.134±.002	.166±.009	.235±.004
MLCL	**.058±.001**	.250±.003	.054±.002	.093±.003	.117±.003	.132±.008	.142±.012	.103±.008	.110±.004	.116±.005	.288±.003
NMCB	.417±.187	.505±.001	.504±.003	.500±.002	.498±.002	.506±.004	.507±.002	.506±.003	.495±.007	.430±.193	.506±.005
ComRank	.185±.003	.287±.006	.090±.096	.239±.126	.254±.010	.256±.005	.276±.011	.241±.011	.237±.007	.265±.006	.220±.002
CTL	.234±.002	.308±.005	.162±.019	.181±.016	.246±.004	.229±.009	.256±.009	.222±.006	.236±.010	.278±.006	.261±.004
Ours	.067±.004	**.244±.004**	**.039±.002**	**.070±.002**	**.081±.004**	**.099±.003**	**.100±.008**	**.093±.005**	**.088±.008**	.124±.008	**.218±.004**

are more clearly irrelevant to a specific instance are more likely to be selected as complementary labels. The detailed generation procedure is provided in Appendix L. All experiments use 5-fold cross-validation, and we report mean and standard deviation across folds on five metrics: Hamming Loss (↓), One Error (↓), Coverage Error (↓), Ranking Loss (↓), and Average Precision (↑).

Table 4: Experimental results of ablation study, where the best performance is shown in bold.

(a)	(b)	(c)	bookmark	delicious	eurlex_dc	eurlex_sm	rcv1-s1	scene	yeast
Ranking Loss ↓									
	✓	✓	.121±.008	.250±.007	.427±.005	.529±.006	.124±.008	.101±.006	.181±.008
✓		✓	**.120±.007**	.239±.006	.048±.004	.104±.006	.070±.003	.096±.007	**.180±.007**
✓	✓		.121±.009	.254±.007	.427±.005	.529±.006	.110±.007	.097±.007	.182±.007
✓	✓	✓	**.120±.008**	**.234±.007**	**.044±.004**	**.099±.006**	**.068±.005**	**.095±.005**	**.180±.006**
Average Precision ↑									
	✓	✓	.728±.009	.627±.009	.221±.005	.222±.005	.601±.011	.822±.009	.743±.010
✓		✓	.725±.008	.643±.006	.863±.009	.764±.011	.816±.011	.830±.014	.743±.009
✓	✓		.721±.009	.622±.006	.221±.005	.222±.005	.647±.012	.826±.010	.742±.010
✓	✓	✓	**.730±.009**	**.648±.006**	**.877±.010**	**.767±.013**	**.821±.011**	**.832±.012**	**.744±.009**

Baselines. We compare against three groups of methods. The first group consists of *weakly supervised MLL* methods that do not assume full label annotation. We include four partial multi-label learning (PML) methods that receive the complement of the complementary label set as candidate labels: **PML-MD** (Xie et al., 2021), **GLC** (Sun et al., 2021), **PARD** (Hang & Zhang, 2023), and **PML-FSMIR** (Gao et al., 2025a). The second group consists of dedicated *multi-labeled complementary label learning (MLCLL)* methods assuming a single complementary label. We include the gradient-descent-friendly risk estimator **GDF** (Gao et al., 2023), the correlation-aware transition matrix method **MLCL** (Gao et al., 2024), the mixup-based noise-mitigation framework **NMCB** (Gao et al., 2026), and the Bayes-consistent ranking loss framework **ComRank** (Zhu et al., 2025). The third group focuses on CMLL with multiple complementary labels, including the stability-enhanced truncated loss method **CTL** (Gao et al., 2025b).

3.2 Main Results

Table 3 reports the results on eleven benchmark datasets. Overall, ID-CMLL achieves consistently strong performance, with the most notable advantage appearing on ranking-oriented metrics. It obtains the best or tied-best Ranking Loss on ten out of eleven datasets and remains highly competitive in Average Precision. The gains are particularly pronounced on *eurlex_dc*, *eurlex_sm*, and *rcv1-s1*, where the complementary-label selection bias is more effectively corrected by the learned instance-dependent propensities. This suggests that explicitly modeling the selection mechanism is especially beneficial for recovering the relative relevance among labels. On *bookmark*, *scene*, and *yeast*, several baselines remain competitive on individual metrics, but ID-CMLL maintains a favorable overall balance across metrics. Overall, these observations show that the proposed correction is not limited to a particular dataset and consistently improves the quality of multi-label ranking.

3.3 Ablation Study

To validate the contribution of different components, we conduct an ablation study on several representative multi-label datasets. Table 4 evaluates three key components: (a) non-negative correction (NN), (b) Hájek normalization, and (c) warm-up initialization. The results show that all three components contribute to the overall performance. Due to space limitations, complete results on all five evaluation metrics are provided in Appendix A.3. Removing NN or warm-up leads to substantial degradation on datasets such as *eurlex_dc* and *eurlex_sm*. For example, without warm-up, Average Precision on *eurlex_dc* drops from 0.877 to 0.221. Removing Hájek normalization causes milder but consistent degradation on several datasets. Combining all three components achieves the best or competitive results across most settings, validating their complementary roles in improving training stability and predictive performance.

4 Conclusion

This paper revisited CMLL by considering instance-dependent complementary-label selection. We derived an unbiased inverse-propensity-weighted risk estimator under known selection propensities and further developed a practical approach for the unknown-propensity setting using estimates obtained from observed complementary labels. Based on these estimates, we constructed a stabilized learning objective with theoretical guarantees on identifiability, consistency, and estimation error.

AI USE STATEMENT

We used generative AI tools to polish the language of the manuscript, to draft non-technical sections such as the conclusion and the statements, and to check the paper for inconsistencies in notation and tables. We did not use them to generate research ideas, theoretical results, proofs or experimental results. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the final content of this work, including its text, claims and artifacts.

REPRODUCIBILITY STATEMENT

The datasets and their statistics are summarized in Table 2, and the pre-processing convention is stated in Section 3.1. The generation of complementary labels under different selection mechanisms is detailed in Appendix L. The assumptions and complete proofs of the theoretical results are provided in Appendices E–K. In addition, we provide the implementation code in the supplementary materials.

REFERENCES

De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16609–16618, 2022.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pp. 1386–1394, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. *Neural Information Processing Systems*, 33:10948–10960, 2020.

Wanfu Gao, Hanlin Pan, Qingqi Han, and Kunpeng Liu. Noise-resistant label reconstruction feature selection for partial multi-label learning. In *International Joint Conference on Artificial Intelligence*, pp. 5172–5180, 2025a.

Yi Gao, Miao Xu, and Min-Ling Zhang. Unbiased risk estimator to multi-labeled complementary label learning. In *International Joint Conference on Artificial Intelligence*, pp. 3732–3740, 2023.

Yi Gao, Miao Xu, and Min-Ling Zhang. Complementary to multiple labels: A correlation-aware correction approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9179–9191, 2024.

Yi Gao, Jing-Yi Zhu, Miao Xu, and Min-Ling Zhang. Multi-label learning with multiple complementary labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8013–8024, 2025b.

Yi Gao, Yuan-Yuan Meng, Miao Xu, and Min-Ling Zhang. Data augmentation with regularization for multi-labeled complementary label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2368–2379, 2026.

Jaroslav Hájek. Discussion of 'an essay on the logical foundations of survey sampling, part i', by d. basu. *Foundations of statistical inference*, 326, 1971.

Jun-Yi Hang and Min-Ling Zhang. Partial multi-label learning with probabilistic graphical disambiguation. *Neural Information Processing Systems*, 36:1339–1351, 2023.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *Neural Information Processing Systems*, 30, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International Conference on Machine Learning*, pp. 2971–2980, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Neural Information Processing Systems*, 30, 2017.

Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *International Conference on Learning Representations*, 2022.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *International Conference on Learning Representations*, 2020.

Ximing Li, Yuanchao Dai, Bing Wang, Changchun Li, Renchu Guan, Fangming Gu, and Jihong Ouyang. Wpml^{3}cp: Wasserstein partial multi-label learning with dual label correlation perspectives. In *International Joint Conference on Artificial Intelligence*, pp. 4479–4487, 2024.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Neural Information Processing Systems*, 26, 2013.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.

Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning with low-rank and sparse decomposition. In *Association for the Advancement of Artificial Intelligence*, pp. 5016–5023, 2019.

Lijuan Sun, Songhe Feng, Jun Liu, Gengyu Lyu, and Congyan Lang. Global-local label correlation for partial multi-label learning. *IEEE Transactions on Multimedia*, 24:581–593, 2021.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

Wei Wang, Takashi Ishida, Yu-Jie Zhang, Gang Niu, and Masashi Sugiyama. Learning with complementary labels revisited: The selected-completely-at-random setting is more practical. In *International Conference on Machine Learning*, pp. 50683–50710, 2024.

Wei Wang, Dong-Dong Wu, Jindong Wang, Gang Niu, Min-Ling Zhang, and Masashi Sugiyama. Realistic evaluation of deep partial-label learning algorithms. *International Conference on Learning Representations*, 2025.

Wei Wang, Tianhao Ma, Ming-Kun Xie, Gang Niu, and Masashi Sugiyama. Rethinking consistent multi-label classification under inexact supervision. In *International Conference on Learning Representations*, volume 2026, pp. 102638–102663, 2026a.

Wei Wang, Dong-Dong Wu, Ming Li, Jingxiong Zhang, Gang Niu, and Masashi Sugiyama. Accessible, realistic, and fair evaluation of positive-unlabeled learning algorithms. In *International Conference on Learning Representations*, 2026b.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2022.

Ming-Kun Xie, Feng Sun, and Sheng-Jun Huang. Partial multi-label learning with meta disambiguation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1904–1912, 2021.

Wang Xinrui, Wenhai Wan, Chuanxing Geng, Shao-Yuan Li, and Songcan Chen. Beyond myopia: Learning from positive and unlabeled data through holistic predictive trends. *Neural Information Processing Systems*, 36:67589–67602, 2023.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *IEEE International Conference on Data Mining*, pp. 1398–1403, 2018a.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *European Conference on Computer Vision*, pp. 68–83, 2018b.

Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *International Joint Conference on Artificial Intelligence*, pp. 4048–4054, 2015.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

Jing-Yi Zhu, Yi Gao, Miao Xu, and Min-Ling Zhang. Comrank: Ranking loss for multi-label complementary label learning. In *Neural Information Processing Systems*, volume 38, pp. 120011–120037, 2025.

A Related Work

A.1 Complementary Multi-label Learning

Complementary label learning was originally proposed for the multi-class single-label setting, where each training instance is paired with exactly one label confirmed to be incorrect. Early formulations build an unbiased risk estimator under the assumption that the complementary label is drawn uniformly from the incorrect classes (Ishida et al., 2017; 2019). More recent work questions this uniform assumption and argues that a selected completely at random formulation better reflects how annotators actually behave (Wang et al., 2024), which motivates treating the label selection process itself as an object worth modeling rather than something to assume away. Complementary multi-label learning extends this idea to the multi-label regime, where an instance may possess several relevant labels while the annotation marks only a subset of the irrelevant labels as complementary. Since the number of relevant labels per instance is unknown, risk estimators designed for the single-label case cannot be applied directly. An early unbiased risk estimator treats each label dimension independently and corrects the training loss with an estimated class prior (Gao et al., 2023). A subsequent study introduces a label-specific transition matrix together with label correlation modeling to reduce the variance of this correction (Gao et al., 2024). When an instance carries many complementary labels at once, a truncated loss has been proposed to keep the resulting estimator numerically stable (Gao et al., 2025b), while a parallel line of work relies on mixup-style augmentation and consistency regularization to compensate for the sparsity of negative supervision (Gao et al., 2026). ComRank instead optimizes a ranking-oriented objective and derives a Bayes-consistent surrogate loss under a biased complementary label distribution (Zhu et al., 2025).

These methods share a common premise, namely that the complementary label is produced either uniformly or through a transition process that remains fixed across instances sharing the same true label. This premise is difficult to justify in practice, since an annotator naturally gravitates toward labels that look obviously wrong for the specific instance at hand, which introduces an instance-level bias that existing risk estimators do not account for. The present work departs from this premise by treating the selection probability as an unknown function of the instance, deriving a propensity-weighted risk under this more general mechanism, and characterizing the residual bias that remains once the true propensity is replaced by an estimated one.

A.2 Partial Multi-label Learning

Partial multi-label learning studies a related but distinct form of weak supervision in which each instance receives a candidate label set mixing relevant and irrelevant labels, rather than a set of confirmed complementary labels (Xie & Huang, 2018). The main difficulty is disambiguation, recovering the true relevant labels hidden inside this noisy candidate set. Prior solutions approach disambiguation from different angles, such as estimating a confidence score for every candidate label (Xie & Huang, 2018), decomposing the label matrix into low-rank and sparse components (Sun et al., 2019), propagating label correlations over a similarity graph (Sun et al., 2021), learning the disambiguation process itself through a meta objective (Xie et al., 2021), formulating disambiguation as inference in a probabilistic graphical model (Hang & Zhang, 2023), identifying noisy candidates at the feature level (Yu et al., 2018a; Xie & Huang, 2022), and coupling feature selection with noise-resistant label reconstruction (Gao et al., 2025a). A Wasserstein-distance-based formulation further models label correlation from a distributional perspective to improve disambiguation quality (Li et al., 2024).

Partial multi-label learning and complementary multi-label learning are connected through a simple complementation argument, since the complement of an observed complementary label set can technically be treated as a candidate label set. Adopting this conversion in practice is problematic, because the resulting candidate set covers nearly the entire label space, leaving the few true relevant labels buried among an overwhelming number of irrelevant ones. Existing partial multi-label learning methods are built around candidate sets with a much more balanced mixture of relevant and irrelevant labels, so their disambiguation assumptions do not transfer well to this extreme regime. A recent study revisits this connection directly and shows that treating inexact multi-label supervision, including the partial and complementary variants, under a single consistency framework requires assumptions that neither setting satisfies in isolation (Wang et al., 2026a), reinforcing the view that a method calibrated for one form of weak supervision cannot be ported to the other without an explicit account of how its

Table 5: Experimental results of ablation study, where the best performance is shown in bold.

(a)	(b)	(c)	bookmark	delicious	eurlex_dc	eurlex_sm	rcv1-s1	scene	yeast
Coverage Error ↓									
	✓	✓	.145±.008	.547±.011	.401±.005	.590±.005	.181±.009	.098±.006	**.457±.009**
✓		✓	**.144±.007**	.535±.009	.047±.003	.178±.008	.134±.006	.093±.007	.458±.008
✓	✓		.146±.008	.551±.009	.401±.005	.590±.005	.168±.009	.095±.007	.458±.008
✓	✓	✓	.145±.008	**.530±.009**	**.043±.004**	**.171±.008**	**.132±.010**	**.092±.005**	**.457±.007**
One Error ↓									
	✓	✓	.354±.010	.378±.018	.778±.013	.674±.010	.615±.025	.297±.018	**.246±.010**
✓		✓	.358±.009	.356±.005	.205±.012	.275±.016	.219±.029	.286±.025	**.246±.010**
✓	✓		.361±.010	.384±.009	.778±.013	.674±.010	.526±.026	.292±.016	**.246±.010**
✓	✓	✓	**.353±.010**	**.348±.006**	**.183±.014**	**.272±.019**	**.208±.031**	**.280±.024**	**.246±.010**
Ranking Loss ↓									
	✓	✓	.121±.008	.250±.007	.427±.005	.529±.006	.124±.008	.101±.006	.181±.008
✓		✓	**.120±.007**	.239±.006	.048±.004	.104±.006	.070±.003	.096±.007	**.180±.007**
✓	✓		.121±.009	.254±.007	.427±.005	.529±.006	.110±.007	.097±.007	.182±.007
✓	✓	✓	**.120±.008**	**.234±.007**	**.044±.004**	**.099±.006**	**.068±.005**	**.095±.005**	**.180±.006**
Average Precision ↑									
	✓	✓	.728±.009	.627±.009	.221±.005	.222±.005	.601±.011	.822±.009	.743±.010
✓		✓	.725±.008	.643±.006	.863±.009	.764±.011	.816±.011	.830±.014	.743±.009
✓	✓		.721±.009	.622±.006	.221±.005	.222±.005	.647±.012	.826±.010	.742±.010
✓	✓	✓	**.730±.009**	**.648±.006**	**.877±.010**	**.767±.013**	**.821±.011**	**.832±.012**	**.744±.009**
Hamming Loss ↓									
	✓	✓	**.063±.001**	.263±.004	1.000±.000	1.000±.000	.110±.002	**.122±.007**	.229±.003
✓		✓	**.063±.001**	.246±.003	.034±.002	**.070±.002**	.085±.005	.125±.005	**.218±.003**
✓	✓		.071±.002	.262±.004	1.000±.000	1.000±.000	.107±.002	.125±.004	.219±.004
✓	✓	✓	.067±.004	**.244±.004**	**.039±.002**	**.070±.002**	**.081±.004**	.124±.008	**.218±.004**

labels are generated. For this reason, complementary multi-label learning is treated in this paper as a distinct learning problem with its own supervision signal, rather than as a direct application of partial multi-label learning techniques.

A.3 Complete Ablation Results

Table 5 reports the complete ablation results on all five evaluation metrics, including Coverage Error, One Error, Ranking Loss, Average Precision, and Hamming Loss. The main text reports Ranking Loss and Average Precision due to space limitations. The complete results further confirm the observations in Section 3.3, showing that the three components contribute complementarily to the overall performance of ID-CMLL.

B Parameter Sensitivity

Figure 1 presents the sensitivity analysis of the regularization coefficient λ_n across eight representative multi-label datasets, where λ_n varies from 10^{-3} to 10^{3}. Overall, the performance is relatively insensitive to λ_n over a wide range. Most datasets show slight improvements as λ_n increases from 10^{-3}, and the performance remains stable within $\lambda_n \in [10^{-1}, 10^{2}]$. Only minor fluctuations are observed at extremely small or large values, indicating that the proposed method is robust to the choice of λ_n. Based on these observations, we set $\lambda_n = 10^{-1}$ as the default value in our experiments.

C Role of the Warm-up Initialization

The warm-up stage plays two distinct roles. First, it provides a stable and semantically informative initialization for optimizing Eq. (12) (detailed below). This role is immaterial to Theorem 2.9, since the theorem characterizes the global minimizer of Eq. (12) regardless of how optimization is initialized. Second, it defines the frozen backbone $\hat{\boldsymbol{\theta}}_0$ on which the representation $h(\cdot; \hat{\boldsymbol{\theta}}_0)$, and hence the entire propensity-estimation stage, is built. This second role is governed by Assumption 2.8, discussed in Appendix C.1. The semantic roles of the two factors are, in either case, determined only after estimation, through the empirical orientation rule in Eq. (15). The detailed initialization of the two propensity factors is provided in Appendix D.

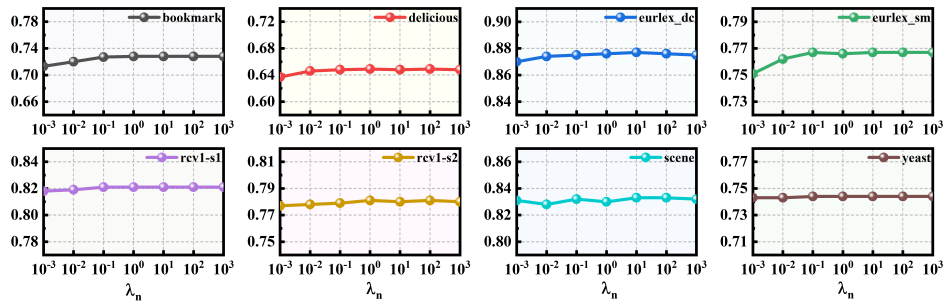

Figure 1: Parameter sensitivity analysis of parameters λ_n.

C.1 On the Coupling Between the Warm-up Representation and Propensity Estimation

In practice, $\hat{\boldsymbol{\theta}}_0$ is estimated from the same dataset $\bar{\mathcal{D}}$ that is subsequently used, through $h(\cdot;\hat{\boldsymbol{\theta}}_0)$, to estimate $\boldsymbol{\Theta}$ in Eq. (12). This makes $\{h(\mathbf{x}_i;\hat{\boldsymbol{\theta}}_0)\}_{i=1}^n$ dependent across i through the shared estimate $\hat{\boldsymbol{\theta}}_0$, so the classical argument for Theorem 2.9, which treats the representation as a fixed non-random map, does not apply verbatim to terms indexed by $\hat{\boldsymbol{\theta}}_0$.

Assumption 2.8 resolves this without modifying the training procedure or splitting the sample. Write $\widehat{\mathcal{L}}_{\text{obs}}(\Theta;\boldsymbol{\theta})$ for the empirical criterion in Eq. (12) evaluated at backbone $\boldsymbol{\theta}$. Since $\sigma(\cdot)$ is Lipschitz and, by Assumption 2.8, $h(\mathbf{x};\boldsymbol{\theta})$ is Lipschitz in $\boldsymbol{\theta}$ uniformly over $\mathbf{x}\in\text{supp}(X)$ near $\boldsymbol{\theta}_0^\dagger$, and the parameter space for Θ is compact, there is a constant $L<\infty$ such that

$$\sup_{\Theta}\left|\widehat{\mathcal{L}}_{\text{obs}}(\Theta;\hat{\boldsymbol{\theta}}_0)-\widehat{\mathcal{L}}_{\text{obs}}(\Theta;\boldsymbol{\theta}_0^\dagger)\right|\le L\left\|\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}_0^\dagger\right\|=o_p(1). \tag{25}$$

This deterministic Lipschitz bound composed with $\hat{\boldsymbol{\theta}}_0\xrightarrow{p}\boldsymbol{\theta}_0^\dagger$ is the only place where the coupling between the two stages enters the argument, and it does not require independence between $\hat{\boldsymbol{\theta}}_0$ and $\bar{\mathcal{D}}$. Sample splitting is therefore not required: Eq. (25) holds whether $\hat{\boldsymbol{\theta}}_0$ is estimated on the full sample or a subsample, as long as it converges to a deterministic limit. The full argument is completed in Appendix H.

The Lipschitz part of Assumption 2.8 is mild for the backbone architectures used here, holding automatically for feedforward networks with Lipschitz activations and parameters restricted to a bounded set. The convergence $\hat{\boldsymbol{\theta}}_0\xrightarrow{p}\boldsymbol{\theta}_0^\dagger$ is a separate condition on the GDF warm-up objective that we do not derive here. Establishing it would require its own M-estimation analysis of the GDF procedure, which is outside the scope of this paper.

D Propensity Head Initialization

After the GDF warm-up stage, we freeze the representation parameters at $\boldsymbol{\theta}_0$ and use the warm-up classifier to initialize the two logistic factors in Eq. (10). Let the output probability of the warm-up classifier for label j be

$$\hat{f}_j^{(0)}(\mathbf{x})=\sigma\left(\mathbf{u}_j^{(0)\top}h(\mathbf{x};\boldsymbol{\theta}_0)+b_j^{(0)}\right). \tag{26}$$

where $\mathbf{u}_j^{(0)}$ and $b_j^{(0)}$ denote the weight vector and bias of the classifier head after warm-up.

Initialization of the first factor. The warm-up classifier estimates the relevance probability of label j, so the corresponding irrelevance probability is $1-\hat{f}_j^{(0)}(\mathbf{x})$. We therefore initialize the first logistic factor as

$$\begin{aligned} g_{j1}^{(0)}(\mathbf{x}) &= 1-\hat{f}_j^{(0)}(\mathbf{x}) \\ &= \sigma\left(-\mathbf{u}_j^{(0)\top}h(\mathbf{x};\boldsymbol{\theta}_0)-b_j^{(0)}\right). \end{aligned} \tag{27}$$

where the last equality follows from the identity $1 - \sigma(s) = \sigma(-s)$. Accordingly, the parameters of the first factor are initialized as

$$\mathbf{w}_{j1}^{(0)} = -\mathbf{u}_j^{(0)}, \qquad c_{j1}^{(0)} = -b_j^{(0)}. \tag{28}$$

This initialization transfers the label information learned by the warm-up classifier to the first factor without requiring an additional fitting procedure.

Initialization of the second factor. For each label j, we compute its empirical complementary-label frequency as

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{ij}. \tag{29}$$

We also compute the warm-up estimate of its marginal irrelevance probability:

$$\hat{\pi}_j^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \hat{f}_j^{(0)}(\mathbf{x}_i)\right). \tag{30}$$

Recall that the marginal probability of observing label j as complementary satisfies

$$q_j(\mathbf{x}) = (1 - \eta_j(\mathbf{x}))\, e_j(\mathbf{x}). \tag{31}$$

Taking the expectation over $\mathbf{x}$ gives

$$\mathbb{E}_{p(\mathbf{x})}\left[q_j(\mathbf{x})\right] = \mathbb{E}_{p(\mathbf{x})}\left[(1 - \eta_j(\mathbf{x}))\, e_j(\mathbf{x})\right]. \tag{32}$$

The empirical frequency $\hat{\mu}_j$ estimates the left-hand side of Eq. (32). To obtain a stable initial value, we approximate the selection probability during initialization by a label-specific constant $e_j(\mathbf{x}) \approx e_j^{(0)}$.

Under this approximation, the moment relation becomes $\hat{\mu}_j \approx \hat{\pi}_j^{(0)} e_j^{(0)}$. We therefore define the initial selection probability as

$$e_j^{(0)} = \mathrm{clip}\left(\frac{\hat{\mu}_j}{\max\left\{\hat{\pi}_j^{(0)}, \epsilon_\pi\right\}}, \epsilon_e, 1 - \epsilon_e\right). \tag{33}$$

where $\epsilon_\pi > 0$ prevents unstable division when $\hat{\pi}_j^{(0)}$ is small, and $\epsilon_e > 0$ keeps the initial probability away from the boundaries of the sigmoid function.

The second logistic factor is initialized as the constant function

$$g_{j2}^{(0)}(\mathbf{x}) = e_j^{(0)} \quad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{34}$$

This is implemented by setting its weight vector to zero and its bias to the logit of the initial selection probability:

$$\mathbf{w}_{j2}^{(0)} = \mathbf{0}, \qquad c_{j2}^{(0)} = \mathrm{logit}\left(e_j^{(0)}\right) = \log\left(\frac{e_j^{(0)}}{1 - e_j^{(0)}}\right). \tag{35}$$

In our implementation, we set $\epsilon_\pi = 0.05$ and $\epsilon_e = 0.05$, so that $e_j^{(0)} \in [0.05, 0.95]$.

Optimization and semantic orientation. The first factor initially inherits the instance-dependent irrelevance pattern learned by the warm-up classifier, while the second factor starts from a label-specific average selection probability. These initial values provide a stable starting point for minimizing the observed-data objective in Eq. (12), but they do not determine the final semantic roles of the fitted factors.

After optimization, both factors are evaluated on the complete training set, and their empirical means are computed as

$$\hat{m}_{j1} = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{j1}(\mathbf{x}_i), \qquad \hat{m}_{j2} = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{j2}(\mathbf{x}_i). \tag{36}$$

Following the orientation rule in Eq. (15), the factor with the larger empirical mean is assigned to the irrelevance probability, while the other factor is assigned to the instance-dependent selection probability. Thus, the asymmetric initialization improves numerical optimization without replacing the empirical orientation procedure.

E Proof of Lemma 2.3 (Irrelevance-Selection Factorization)

By Assumption 2.2 and the entries of $T_j(\mathbf{x})$, the conditional distribution of $\bar{y}_j$ given $(\mathbf{x}, y_j)$ is

$$p(\bar{y}_j \mid \mathbf{x}, y_j) = \mathbb{I}[y_j = 1]\,\mathbb{I}[\bar{y}_j = 0] + \mathbb{I}[y_j = 0]\, e_j(\mathbf{x})^{\bar{y}_j}\big(1 - e_j(\mathbf{x})\big)^{1-\bar{y}_j}. \tag{37}$$

Hence

$$\mathbb{E}[\bar{y}_j \mid \mathbf{x}, y_j] = (1 - y_j)\, e_j(\mathbf{x}). \tag{38}$$

Since $f_j(\mathbf{x})$ and $e_j(\mathbf{x})$ are functions of $\mathbf{x}$ alone, the tower property, conditioning on $(\mathbf{x}, y_j)$, gives

$$\mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\frac{\bar{y}_j f_j(\mathbf{x})}{e_j(\mathbf{x})}\right] = \mathbb{E}_{p(\mathbf{x},y_j)}\left[\frac{f_j(\mathbf{x})}{e_j(\mathbf{x})}\,\mathbb{E}[\bar{y}_j \mid \mathbf{x}, y_j]\right]. \tag{39}$$

Substituting Eq. (38) and cancelling $e_j(\mathbf{x}) > 0$ gives

$$\mathbb{E}_{p(\mathbf{x},y_j)}\left[\frac{f_j(\mathbf{x})}{e_j(\mathbf{x})}(1 - y_j)\, e_j(\mathbf{x})\right] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\big[(1 - y_j)\, f_j(\mathbf{x})\big]. \tag{40}$$

F Proof of Lemma 2.4 (Relevance-Irrelevance Decomposition)

Since $y_j + (1 - y_j) = 1$ holds $p(\mathbf{x}, \mathbf{y})$-almost surely, multiplying both sides by $f_j(\mathbf{x})$ and taking expectations gives

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[y_j f_j(\mathbf{x})] + \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1 - y_j) f_j(\mathbf{x})] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[f_j(\mathbf{x})]\,. \tag{41}$$

Since $f_j(\mathbf{x})$ depends only on $\mathbf{x}$, the right-hand side equals $\mathbb{E}_{p(\mathbf{x})}[f_j(\mathbf{x})]$. Rearranging gives the stated identity.

G Proof of Theorem 2.5

Decompose $R(f) = R^+(f) + R^-(f)$, where

$$R^+(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[y_j \ell(f_j(\mathbf{x}))]\,, \tag{42}$$

$$R^-(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\big[(1 - y_j)\bar{\ell}(f_j(\mathbf{x}))\big]\,. \tag{43}$$

Treatment of $R^-(f)$. Applying Lemma 2.3 with $f_j(\mathbf{x}) \leftarrow \bar{\ell}(f_j(\mathbf{x}))$ to each summand gives

$$R^-(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\frac{\bar{y}_j \bar{\ell}(f_j(\mathbf{x}))}{e_j(\mathbf{x})}\right]. \tag{44}$$

Treatment of $R^+(f)$. Applying Lemma 2.4 with $f_j(\mathbf{x}) \leftarrow \ell(f_j(\mathbf{x}))$ gives

$$R^+(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x})}[\ell(f_j(\mathbf{x}))] - \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1 - y_j)\ell(f_j(\mathbf{x}))]\,. \tag{45}$$

Applying Lemma 2.3 with $f_j(\mathbf{x}) \leftarrow \ell(f_j(\mathbf{x}))$ to the second sum yields

$$R^+(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x})}[\ell(f_j(\mathbf{x}))] - \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\frac{\bar{y}_j \ell(f_j(\mathbf{x}))}{e_j(\mathbf{x})}\right]. \tag{46}$$

Combination. Adding $R^+(f)$ and $R^-(f)$ and using linearity of expectation gives

$$R(f) = \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x})}[\ell(f_j(\mathbf{x}))] + \sum_{j=1}^{K} \mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\frac{\bar{y}_j}{e_j(\mathbf{x})}\Big(\bar{\ell}(f_j(\mathbf{x})) - \ell(f_j(\mathbf{x}))\Big)\right]$$
$$= \mathbb{E}_{p(\mathbf{x})}\left[\sum_{j=1}^{K} \ell(f_j(\mathbf{x}))\right] + \mathbb{E}_{p(\mathbf{x},\bar{\mathbf{y}})}\left[\sum_{j=1}^{K} \frac{\bar{y}_j}{e_j(\mathbf{x})}\Big(\bar{\ell}(f_j(\mathbf{x})) - \ell(f_j(\mathbf{x}))\Big)\right], \tag{47}$$

which is the stated identity.

H Proof of Theorem 2.9

We prove the result label by label and suppress the label index j when no confusion arises. Let $a_k(\mathbf{z}) = \mathbf{w}_k^\top \mathbf{z} + c_k$ and $g_k(\mathbf{z}) = \sigma(a_k(\mathbf{z}))$ for $k \in \{1, 2\}$.

Identification up to permutation. Using $\sigma(t)^{-1} = 1 + \exp(-t)$, the product of the two logistic factors satisfies

$$\frac{1}{g_1(\mathbf{z})g_2(\mathbf{z})} - 1 = e^{-c_1}e^{-\mathbf{w}_1^\top \mathbf{z}} + e^{-c_2}e^{-\mathbf{w}_2^\top \mathbf{z}} + e^{-(c_1+c_2)}e^{-(\mathbf{w}_1+\mathbf{w}_2)^\top \mathbf{z}}. \tag{48}$$

Suppose another parameter tuple $\widetilde{\Theta} = (\widetilde{\mathbf{w}}_1, \widetilde{c}_1, \widetilde{\mathbf{w}}_2, \widetilde{c}_2)$ produces the same marginal probability on a non-empty open set $\mathcal{U}$. Eq. (48) then gives

$$e^{-c_1}e^{-\mathbf{w}_1^\top \mathbf{z}} + e^{-c_2}e^{-\mathbf{w}_2^\top \mathbf{z}} + e^{-(c_1+c_2)}e^{-(\mathbf{w}_1+\mathbf{w}_2)^\top \mathbf{z}} = e^{-\widetilde{c}_1}e^{-\widetilde{\mathbf{w}}_1^\top \mathbf{z}} + e^{-\widetilde{c}_2}e^{-\widetilde{\mathbf{w}}_2^\top \mathbf{z}}$$
$$+ e^{-(\widetilde{c}_1+\widetilde{c}_2)}e^{-(\widetilde{\mathbf{w}}_1+\widetilde{\mathbf{w}}_2)^\top \mathbf{z}} \tag{49}$$

for every $\mathbf{z} \in \mathcal{U}$.

Distinct multivariate exponential functions are linearly independent on an open set. Indeed, choose a direction $\mathbf{v}$ whose inner products with all distinct exponent vectors are different and restrict Eq. (49) to the line $\mathbf{z} = \mathbf{z}_0 + t\mathbf{v}$. The resulting univariate exponential functions have distinct exponents and are linearly independent. Hence, the exponent vectors and their coefficients in the two representations coincide up to permutation, which implies

$$\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_1 + \mathbf{w}_2\} = \{\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2, \widetilde{\mathbf{w}}_1 + \widetilde{\mathbf{w}}_2\}. \tag{50}$$

Because $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_1 + \mathbf{w}_2$ are nonzero and pairwise distinct, the member equal to the sum of the other two is uniquely determined. Therefore,

$$\{\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2\} = \{\mathbf{w}_1, \mathbf{w}_2\}. \tag{51}$$

Matching the positive coefficients associated with the identified exponent vectors yields

$$\left\{e^{-\widetilde{c}_1}, e^{-\widetilde{c}_2}\right\} = \left\{e^{-c_1}, e^{-c_2}\right\}, \tag{52}$$

and hence

$$\{(\widetilde{\mathbf{w}}_1, \widetilde{c}_1), (\widetilde{\mathbf{w}}_2, \widetilde{c}_2)\} = \{(\mathbf{w}_1, c_1), (\mathbf{w}_2, c_2)\}. \tag{53}$$

Consistency of the unordered factors. Let $q^*(\mathbf{z}) = q(\mathbf{z}; \Theta^*)$, where Θ^* is defined relative to the limiting representation $Z = h(X; \boldsymbol{\theta}_0^\dagger)$ under Assumption 2.8, and define the label-wise population observed-data criterion by

$$\mathcal{L}_{\text{obs}}^{\text{pop}}(\Theta) = \mathbb{E}\big[-\bar{Y}\log q(Z;\Theta) - (1-\bar{Y})\log\big(1 - q(Z;\Theta)\big)\big]. \tag{54}$$

Since $\mathbb{P}(\bar{Y} = 1 \mid Z) = q^*(Z)$, its excess value satisfies

$$\mathcal{L}_{\text{obs}}^{\text{pop}}(\Theta) - \mathcal{L}_{\text{obs}}^{\text{pop}}(\Theta^*) = \mathbb{E}[\mathrm{KL}(\mathrm{Ber}(q^*(Z)) \,\|\, \mathrm{Ber}(q(Z;\Theta)))] \geq 0. \tag{55}$$

Equality holds if and only if $q(Z;\Theta) = q^*(Z)$ almost surely. Because both sides are continuous and the distribution of Z has the stated support, the equality extends to its support. Eq. (53) then implies that the population minimizer set is exactly

$$[\Theta^*] = \{\Theta^*, (\mathbf{w}_2^*, c_2^*, \mathbf{w}_1^*, c_1^*)\}. \tag{56}$$

Continuity and boundedness of the Bernoulli criterion over the compact support and parameter space give a uniform law of large numbers at the fixed limiting representation $\boldsymbol{\theta}_0^\dagger$:

$$\sup_{\Theta} \left| \widehat{\mathcal{L}}_{\text{obs}}(\Theta; \boldsymbol{\theta}_0^\dagger) - \mathcal{L}_{\text{obs}}^{\text{pop}}(\Theta) \right| \xrightarrow{p} 0, \tag{57}$$

where $\widehat{\mathcal{L}}_{\text{obs}}(\Theta; \boldsymbol{\theta})$ denotes the empirical criterion of Eq. (12) evaluated with backbone $\boldsymbol{\theta}$. Since $\widehat{\boldsymbol{\Theta}}$ is obtained by minimizing $\widehat{\mathcal{L}}_{\text{obs}}(\Theta; \hat{\boldsymbol{\theta}}_0)$ on the estimated backbone, Assumption 2.8 and the Lipschitz bound of Eq. (25) give $\sup_\Theta |\widehat{\mathcal{L}}_{\text{obs}}(\Theta; \hat{\boldsymbol{\theta}}_0) - \widehat{\mathcal{L}}_{\text{obs}}(\Theta; \boldsymbol{\theta}_0^\dagger)| = o_p(1)$, and the triangle inequality yields

$$\sup_{\Theta} \left| \widehat{\mathcal{L}}_{\text{obs}}(\Theta; \hat{\boldsymbol{\theta}}_0) - \mathcal{L}_{\text{obs}}^{\text{pop}}(\Theta) \right| \xrightarrow{p} 0. \tag{58}$$

Since the regularization coefficient in Eq. (12) satisfies $\lambda_n \to 0$, the regularization term vanishes uniformly, and standard argmin consistency gives

$$d\Big(\widehat{\Theta}, [\Theta^*]\Big) \xrightarrow{p} 0. \tag{59}$$

Consistency of semantic orientation. Let $g_{\text{irr}}^*(\mathbf{x}) = 1 - \eta(\mathbf{x})$ and $g_{\text{sel}}^*(\mathbf{x}) = e(\mathbf{x})$, evaluated through the limiting representation $\boldsymbol{\theta}_0^\dagger$, and write $\hat{g}_k(\mathbf{x})$ for the fitted factors evaluated through the estimated representation $\hat{\boldsymbol{\theta}}_0$. By unordered parameter consistency, compactness of the support, and the Lipschitz continuity of $h(\mathbf{x}; \cdot)$ and $\sigma(\cdot)$ together with $\hat{\boldsymbol{\theta}}_0 \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$ (Assumption 2.8), there exists a permutation of the fitted factors such that

$$\max_{k \in \{\text{irr}, \text{sel}\}} \sup_{\mathbf{x} \in \text{supp}(X)} |\hat{g}_k(\mathbf{x}) - g_k^*(\mathbf{x})| \xrightarrow{p} 0. \tag{60}$$

For either $k \in \{\text{irr}, \text{sel}\}$, its empirical mean in Eq. (14) satisfies

$$\left| \frac{1}{n} \sum_{i=1}^n \hat{g}_k(\mathbf{x}_i) - \mathbb{E}[g_k^*(X)] \right| \leq \sup_{\mathbf{x} \in \text{supp}(X)} |\hat{g}_k(\mathbf{x}) - g_k^*(\mathbf{x})| + \left| \frac{1}{n} \sum_{i=1}^n g_k^*(\mathbf{x}_i) - \mathbb{E}[g_k^*(X)] \right| \xrightarrow{p} 0. \tag{61}$$

Under Assumption 2.7,

$$\mathbb{E}[g_{\text{irr}}^*(X)] - \mathbb{E}[g_{\text{sel}}^*(X)] \geq \delta_j > 0. \tag{62}$$

Combining this strict gap with Eq. (61) gives

$$\mathbb{P}(\hat{m}_{\text{irr}} > \hat{m}_{\text{sel}}) \longrightarrow 1. \tag{63}$$

Thus, the rule in Eq. (15) recovers the correct semantic ordering with probability approaching one. After orientation, $\|\hat{e}_j - e_j\|_\infty \xrightarrow{p} 0$ and $\|\hat{\eta}_j - \eta_j\|_\infty \xrightarrow{p} 0$, which proves Eq. (22). Since K is finite, the result holds simultaneously for all labels.

I Proof of Theorem 2.11

As in the original convention, we set the corresponding term to zero whenever $\hat{W}_j = 0$. Since $\hat{e}_j(\mathbf{x}) > 0$ for every $\mathbf{x}$, the event $\hat{W}_j = 0$ is equivalent to $\bar{y}_{ij} = 0$ for all i. Moreover, under Assumption 2.10,

$$\mathbb{P}(\bar{Y}_j = 1) = \mathbb{E}[(1 - Y_j) e_j(\mathbf{X})] \geq e_{\min} \pi_j \geq e_{\min} \pi_{\min}. \tag{64}$$

Since the pairs $(\mathbf{x}_i, \bar{y}_i)$ are i.i.d.,

$$\mathbb{P}(\hat{W}_j = 0) = \big(1 - \mathbb{P}(\bar{Y}_j = 1)\big)^n \leq (1 - e_{\min} \pi_{\min})^n \to 0 \tag{65}$$

exponentially fast, so this convention does not affect the asymptotic results.

For each label j, define the population quantities

$$A_j(f) = \mathbb{E}_{p(\mathbf{x})}[\ell(f_j(\mathbf{x}))], \qquad Q_j(f) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1-y_j)\ell(f_j(\mathbf{x}))],$$
$$P_j(f) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[y_j\ell(f_j(\mathbf{x}))], \qquad N_j(f) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(1-y_j)\bar{\ell}(f_j(\mathbf{x}))]. \tag{66}$$

Since $y_j \in \{0,1\}$ and $\ell, \bar{\ell} \geq 0$, both $P_j(f) \geq 0$ and $N_j(f) \geq 0$ hold for every f. By Lemma 2.4 with $f_j \leftarrow \ell(f_j(\cdot))$, $P_j(f) = A_j(f) - Q_j(f)$. By the decomposition used in the proof of Theorem 2.5, $R(f) = \sum_{j=1}^{K} \big(P_j(f) + N_j(f)\big)$.

Correspondingly, define the empirical quantities

$$\hat{A}_j(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_j(\mathbf{x}_i)), \qquad \hat{Q}_j(f;\hat{\pi},\hat{e}) = \frac{\hat{\pi}_j}{\hat{W}_j} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\ell(f_j(\mathbf{x}_i)),$$
$$\hat{P}_j(f;\hat{\pi},\hat{e}) = \hat{A}_j(f) - \hat{Q}_j(f;\hat{\pi},\hat{e}), \qquad \hat{N}_j(f;\hat{\pi},\hat{e}) = \frac{\hat{\pi}_j}{\hat{W}_j} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\bar{\ell}(f_j(\mathbf{x}_i)), \tag{67}$$

so that, by Eq. (19),

$$\hat{R}^{H+}(f;\hat{\pi},\hat{e}) = \sum_{j=1}^{K} \Big(\big|\hat{P}_j(f;\hat{\pi},\hat{e})\big| + \hat{N}_j(f;\hat{\pi},\hat{e}) \Big). \tag{68}$$

Consistency of the label-wise components. Because $\hat{e}_j$ is estimated from the same sample that defines $\hat{W}_j$, the summands in $\hat{W}_j$ are not themselves i.i.d., so we first make the convergence $\hat{W}_j \xrightarrow{p} \pi_j$ explicit rather than invoking it directly. Define the oracle normalizer

$$W_j^0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_{ij}}{e_j(\mathbf{x}_i)}, \tag{69}$$

an average of i.i.d. terms with $\mathbb{E}[W_j^0] = \pi_j$ by Lemma 2.3 applied with $f_j \leftarrow 1$, so the law of large numbers gives $W_j^0 \xrightarrow{p} \pi_j$. Let $\delta_{e,n} = \max_j \|\hat{e}_j - e_j\|_\infty$, which is $o_p(1)$ by the uniform consistency of $\hat{e}_j$. On the event $\{\delta_{e,n} \leq e_{\min}/2\}$, whose probability tends to one, $\hat{e}_j(\mathbf{x}) \geq e_{\min}/2$ for all $\mathbf{x}$, so $|1/\hat{e}_j(\mathbf{x}) - 1/e_j(\mathbf{x})| \leq 2\delta_{e,n}/e_{\min}^2$, and, using $\frac{1}{n}\sum_i \bar{y}_{ij} \leq 1$,

$$\big|\hat{W}_j - W_j^0\big| \leq \frac{2\delta_{e,n}}{e_{\min}^2} = o_p(1). \tag{70}$$

Hence $\hat{W}_j = W_j^0 + o_p(1) \xrightarrow{p} \pi_j$. Appendix K carries out the identical perturbation in full generality. The same argument, applied with $\bar{y}_{ij}\ell(f_j(\mathbf{x}_i))$ and $\bar{y}_{ij}\bar{\ell}(f_j(\mathbf{x}_i))$ in place of $\bar{y}_{ij}$, together with boundedness of $\ell, \bar{\ell}$, positivity of e_j, Lemma 2.3 applied with $f_j \leftarrow \ell(f_j(\cdot))$ and with $f_j \leftarrow \bar{\ell}(f_j(\cdot))$, and the law of large numbers, gives

$$\hat{W}_j \xrightarrow{p} \pi_j, \qquad \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\ell(f_j(\mathbf{x}_i)) \xrightarrow{p} Q_j(f), \qquad \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{y}_{ij}}{\hat{e}_j(\mathbf{x}_i)}\bar{\ell}(f_j(\mathbf{x}_i)) \xrightarrow{p} N_j(f). \tag{71}$$

Uniform consistency together with $e_j(\mathbf{x}) \geq e_{\min} > 0$ ensures that $\hat{e}_j(\mathbf{x})$ stays bounded away from zero with probability approaching one, so replacing e_j by $\hat{e}_j$ contributes only an $o_p(1)$ term. Since $\hat{\pi}_j \xrightarrow{p} \pi_j$ and $\pi_j \geq \pi_{\min} > 0$, Slutsky's theorem applied to Eq. (71) gives

$$\hat{Q}_j(f;\hat{\pi},\hat{e}) \xrightarrow{p} Q_j(f), \qquad \hat{N}_j(f;\hat{\pi},\hat{e}) \xrightarrow{p} N_j(f). \tag{72}$$

The law of large numbers also gives $\hat{A}_j(f) \xrightarrow{p} A_j(f)$, so

$$\hat{P}_j(f;\hat{\pi},\hat{e}) = \hat{A}_j(f) - \hat{Q}_j(f;\hat{\pi},\hat{e}) \xrightarrow{p} A_j(f) - Q_j(f) = P_j(f). \tag{73}$$

Since $P_j(f) \geq 0$ and $|\cdot|$ is continuous, $\big|\hat{P}_j(f;\hat{\pi},\hat{e})\big| \xrightarrow{p} \big|P_j(f)\big| = P_j(f)$.

Convergence in probability of the risk. Summing Eq. (68) over the finite label set and applying the continuous mapping theorem,

$$\hat{R}^{H+}(f;\hat{\pi},\hat{e}) \xrightarrow{p} \sum_{j=1}^{K} \big(P_j(f) + N_j(f)\big) = R(f), \tag{74}$$

which proves Eq. (23).

Uniform integrability and L^1 convergence. Whenever $\hat{W}_j \neq 0$ and $\hat{\pi}_j > 0$, $\hat{Q}_j(f;\hat{\pi},\hat{e})/\hat{\pi}_j$ is a weighted average of $\{\ell(f_j(\mathbf{x}_i))\}_i$ with non-negative weights, and $\hat{N}_j(f;\hat{\pi},\hat{e})/\hat{\pi}_j$ is a weighted average of $\{\bar{\ell}(f_j(\mathbf{x}_i))\}_i$. Under Assumption 2.10, both ratios lie in $[0, C_L]$, so $\hat{\pi}_j \in [0,1]$ gives $\hat{Q}_j(f;\hat{\pi},\hat{e}), \hat{N}_j(f;\hat{\pi},\hat{e}) \in [0, C_L]$. If instead $\hat{\pi}_j = 0$, then $\hat{Q}_j(f;\hat{\pi},\hat{e}) = \hat{N}_j(f;\hat{\pi},\hat{e}) = 0$ directly by Eq. (67), which again lies in $[0, C_L]$. Either way, $\hat{Q}_j(f;\hat{\pi},\hat{e}), \hat{N}_j(f;\hat{\pi},\hat{e}) \in [0, C_L]$ almost surely, uniformly in f. Together with $\hat{A}_j(f) \in [0, C_L]$, we get $\big|\hat{P}_j(f;\hat{\pi},\hat{e})\big| \leq C_L$, so

$$\hat{R}^{H+}(f;\hat{\pi},\hat{e}) \in [0, 2KC_L] \quad \text{almost surely, for every } n. \tag{75}$$

A sequence bounded uniformly in n is automatically uniformly integrable, so convergence in probability to the finite constant $R(f)$ upgrades to L^1 convergence, and $\mathbb{E}\big[\hat{R}^{H+}(f;\hat{\pi},\hat{e})\big] \to R(f)$.

J Proof of Theorem 2.12

Throughout this proof, the training sample $\{(\mathbf{x}_i, \mathbf{y}_i, \bar{\mathbf{y}}_i)\}_{i=1}^n$ is assumed i.i.d. from $p(\mathbf{x}, \mathbf{y}, \bar{\mathbf{y}})$, which justifies the uniform law of large numbers, Hoeffding's inequality, and the Rademacher contraction argument used below. This is distinct from the conditional selection assumption in Assumption 2.2, which specifies the marginal selection mechanism for each label dimension given a single instance.

Let $\hat{R}_n(f) = \hat{R}^{H+}(f;\hat{\pi},\hat{e})$ and $\hat{R}_n^0(f) = \hat{R}^{H+}(f;\pi,e)$ denote the feasible and oracle empirical objectives, and define $\Gamma_n^0 = \sup_{f\in\mathcal{F}} \big|\hat{R}_n^0(f) - R(f)\big|$.

Since $\hat{f}$ minimizes $\hat{R}_n(f)$ over $\mathcal{F}$,

$$R(\hat{f}) - R(f^*) \leq \big(R(\hat{f}) - \hat{R}_n(\hat{f})\big) + \big(\hat{R}_n(f^*) - R(f^*)\big) \leq 2 \sup_{f\in\mathcal{F}} \big|\hat{R}_n(f) - R(f)\big|. \tag{76}$$

By the triangle inequality, $\sup_f |\hat{R}_n(f) - R(f)| \leq \sup_f |\hat{R}_n(f) - \hat{R}_n^0(f)| + \Gamma_n^0 = \varepsilon_n + \Gamma_n^0$, so

$$R(\hat{f}) - R(f^*) \leq 2\varepsilon_n + 2\Gamma_n^0. \tag{77}$$

Bounding Γ_n^0. Write the oracle empirical components, using the true nuisance (π, e),

$$\begin{aligned} Q_j^0(f) &= \frac{\pi_j}{W_j^0} \cdot \frac{1}{n}\sum_{i=1}^n \frac{\bar{y}_{ij}}{e_j(\mathbf{x}_i)} \ell(f_j(\mathbf{x}_i)), & N_j^0(f) &= \frac{\pi_j}{W_j^0} \cdot \frac{1}{n}\sum_{i=1}^n \frac{\bar{y}_{ij}}{e_j(\mathbf{x}_i)} \bar{\ell}(f_j(\mathbf{x}_i)), \\ W_j^0 &= \frac{1}{n}\sum_{i=1}^n \frac{\bar{y}_{ij}}{e_j(\mathbf{x}_i)}, & P_j^0(f) &= \hat{A}_j(f) - Q_j^0(f), \end{aligned} \tag{78}$$

so that $\hat{R}_n^0(f) = \sum_{j=1}^K \big(|P_j^0(f)| + N_j^0(f)\big)$.

Reverse triangle inequality. Since $P_j(f) \geq 0$ for every f, established in Appendix I, the identity $\big||a| - |b|\big| \leq |a - b|$ applied with $a = P_j^0(f)$ and $b = P_j(f)$ gives

$$\big||P_j^0(f)| - P_j(f)\big| \leq \big|P_j^0(f) - P_j(f)\big|, \tag{79}$$

so the transition from the absolute-value correction back to the signed quantity P_j does not enlarge the uniform deviation bound. Combining Eq. (79) with $N_j^0(f) - N_j(f)$ term by term gives

$$\Gamma_n^0 \leq \sum_{j=1}^K \sup_{f\in\mathcal{F}} \big|P_j^0(f) - P_j(f)\big| + \sum_{j=1}^K \sup_{f\in\mathcal{F}} \big|N_j^0(f) - N_j(f)\big|. \tag{80}$$

By Lemma 2.3, $\mathbb{E}[W_j^0] = \pi_j$. Since $0 \le \bar{y}_{ij}/e_j(\mathbf{x}_i) \le 1/e_{\min}$ and $\pi_j \ge \pi_{\min}$, Hoeffding's inequality and a union bound give, for sufficiently large n with probability at least $1 - \delta/4$,

$$\min_{1 \le j \le K} W_j^0 \ge \frac{\pi_{\min}}{2}. \tag{81}$$

On this event, $P_j^0(f) - P_j(f) = \big(\hat{A}_j(f) - A_j(f)\big) - \big(Q_j^0(f) - Q_j(f)\big)$. Writing $Q_j^0(f) = \pi_j \hat{q}_j(f)/W_j^0$ with $\hat{q}_j(f) = \frac{1}{n}\sum_i \bar{y}_{ij}\ell(f_j(\mathbf{x}_i))/e_j(\mathbf{x}_i)$, so that $\mathbb{E}[\hat{q}_j(f)] = Q_j(f)$ by Lemma 2.3, ratio perturbation together with $\pi_j \le 1$ and $Q_j(f) \le C_L$ gives

$$\big|Q_j^0(f) - Q_j(f)\big| \le \frac{2}{\pi_{\min}}\big|\hat{q}_j(f) - Q_j(f)\big| + \frac{2C_L}{\pi_{\min}}\big|W_j^0 - \pi_j\big|, \tag{82}$$

and the analogous bound holds for $|N_j^0(f) - N_j(f)|$ with $\hat{q}_j$ replaced by $\hat{n}_j(f) = \frac{1}{n}\sum_i \bar{y}_{ij}\bar{\ell}(f_j(\mathbf{x}_i))/e_j(\mathbf{x}_i)$.

To make explicit the step from these bounded weighted deviations to the Rademacher complexities $\mathcal{R}_n(\mathcal{G}_j)$, write $w_{ij} = \bar{y}_{ij}/e_j(\mathbf{x}_i) \le 1/e_{\min}$ and, for each label j, define the weighted function class

$$\mathcal{H}_j = \left\{(\mathbf{x}, \bar{y}) \mapsto \frac{\bar{y}_j}{e_j(\mathbf{x})}\ell(f_j(\mathbf{x})) : f \in \mathcal{F}\right\}, \tag{83}$$

and analogously $\bar{\mathcal{H}}_j$ with $\bar{\ell}$ in place of ℓ. Since $w_{ij} \in [0, 1/e_{\min}]$ is a bounded, non-negative multiplier that does not depend on f, and ℓ (respectively $\bar{\ell}$) is ρ-Lipschitz, the multiplier/contraction argument for Rademacher complexity gives

$$\mathcal{R}_n(\mathcal{H}_j) \le \frac{\rho}{e_{\min}}\mathcal{R}_n(\mathcal{G}_j), \qquad \mathcal{R}_n(\bar{\mathcal{H}}_j) \le \frac{\rho}{e_{\min}}\mathcal{R}_n(\mathcal{G}_j). \tag{84}$$

Because ℓ and $\bar{\ell}$ are each ρ-Lipschitz and bounded by C_L, and the weights $\bar{y}_{ij}/e_j(\mathbf{x}_i)$ are bounded by $1/e_{\min}$, symmetrization applied to the function classes $\mathcal{H}_j, \bar{\mathcal{H}}_j$, together with the contraction bound in Eq. (84), controls the two weighted deviation terms in Eq. (80). The remaining unweighted term $\hat{A}_j(f) - A_j(f)$ is controlled by the same contraction argument applied directly to $\ell \circ \mathcal{G}_j$: since ℓ is ρ-Lipschitz, $\mathcal{R}_n(\ell \circ \mathcal{G}_j) \le \rho\,\mathcal{R}_n(\mathcal{G}_j) \le \frac{\rho}{e_{\min}}\mathcal{R}_n(\mathcal{G}_j)$ because $e_{\min} \le 1$, so it is absorbed into the same bound without changing its rate. Combining these bounds with symmetrization, Hoeffding's inequality, and a union bound over the K labels and the two loss terms $\ell, \bar{\ell}$ yields a universal constant $c > 0$ such that, with probability at least $1 - \delta$,

$$\Gamma_n^0 \le \frac{c\rho}{e_{\min}\pi_{\min}}\sum_{j=1}^{K}\mathcal{R}_n(\mathcal{G}_j) + \frac{cKC_L}{e_{\min}\pi_{\min}}\sqrt{\frac{\log(4K/\delta)}{n}}. \tag{85}$$

Substituting Eq. (85) into Eq. (77) and absorbing numerical factors into c proves Eq. (24).

Finally, the optimality of f^* gives $R(\hat{f}) - R(f^*) \ge 0$. If $\varepsilon_n = o_p(1)$ and $\sum_{j=1}^K \mathcal{R}_n(\mathcal{G}_j) \to 0$, every term on the right-hand side of Eq. (24) vanishes, which yields $R(\hat{f}) - R(f^*) \xrightarrow{p} 0$.

K Nuisance-Estimation Error Bound

This appendix states and proves the nuisance-estimation error bound for ε_n referenced in the discussion following Theorem 2.12.

Theorem K.1 (Nuisance-Estimation Error Bound). *Let $\delta_{e,n} = \max_j \|\hat{e}_j - e_j\|_\infty$ and $\delta_{\pi,n} = \max_j |\hat{\pi}_j - \pi_j|$. Under the conditions of Theorems 2.9 and 2.12, together with Assumption 2.7, on an event of probability approaching one,*

$$\varepsilon_n \le C_1\,\delta_{\pi,n} + C_2\,\delta_{e,n}, \tag{86}$$

where $C_1 = 2KC_L$ and $C_2 = 32KC_L/(e_{\min}^2\pi_{\min})$ depend only on $K, C_L, e_{\min}, \pi_{\min}$. Since Theorem 2.9 together with Assumption 2.7 gives $\delta_{e,n} \xrightarrow{p} 0$, and Corollary K.2 gives $\delta_{\pi,n} \xrightarrow{p} 0$, it follows that $\varepsilon_n \xrightarrow{p} 0$.

Corollary K.2. *Under the conditions of Theorem 2.9 and Assumption 2.7, $\max_j |\hat{\pi}_j - \pi_j| \xrightarrow{p} 0$.*

K.1 Proof of Corollary K.2

By definition $\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}(1-\hat{\eta}_j(\mathbf{x}_i))$ and $\pi_j = \mathbb{E}_{p(\mathbf{x})}[1-\eta_j(\mathbf{x})]$. The triangle inequality gives

$$|\hat{\pi}_j - \pi_j| \le \|\hat{\eta}_j - \eta_j\|_\infty + \left|\frac{1}{n}\sum_{i=1}^{n}(1-\eta_j(\mathbf{x}_i)) - \mathbb{E}_{p(\mathbf{x})}[1-\eta_j(\mathbf{x})]\right|. \tag{87}$$

The first term vanishes in probability by Theorem 2.9 together with Assumption 2.7, and the second by the law of large numbers. Since K is finite, $\max_j |\hat{\pi}_j - \pi_j| \xrightarrow{p} 0$.

K.2 Proof of Theorem K.1

Recall from Appendix I and Appendix J the empirical quantities $\hat{A}_j(f)$, $\hat{Q}_j(f;\hat{\pi},\hat{e})$, $\hat{P}_j(f;\hat{\pi},\hat{e}) = \hat{A}_j(f) - \hat{Q}_j(f;\hat{\pi},\hat{e})$, $\hat{N}_j(f;\hat{\pi},\hat{e})$, feasible, using $\hat{e}_j, \hat{\pi}_j$, and their oracle counterparts $Q_j^0(f)$, $P_j^0(f) = \hat{A}_j(f) - Q_j^0(f)$, $N_j^0(f)$, using the true e_j, π_j, so that

$$\hat{R}^{H+}(f;\hat{\pi},\hat{e}) - \hat{R}^{H+}(f;\pi,e) = \sum_{j=1}^{K}\Big(\big|\hat{P}_j(f;\hat{\pi},\hat{e})\big| - \big|P_j^0(f)\big|\Big) + \sum_{j=1}^{K}\Big(\hat{N}_j(f;\hat{\pi},\hat{e}) - N_j^0(f)\Big). \tag{88}$$

Reverse triangle inequality. Since $\hat{A}_j(f)$ does not depend on the choice of $(\hat{\pi},\hat{e})$ versus (π,e), again by $\big||a|-|b|\big| \le |a-b|$, applied with $a = \hat{P}_j(f;\hat{\pi},\hat{e})$ and $b = P_j^0(f)$,

$$\Big|\,\big|\hat{P}_j(f;\hat{\pi},\hat{e})\big| - \big|P_j^0(f)\big|\,\Big| \le \big|\hat{P}_j(f;\hat{\pi},\hat{e}) - P_j^0(f)\big| = \big|\hat{Q}_j(f;\hat{\pi},\hat{e}) - Q_j^0(f)\big|, \tag{89}$$

which reduces the plug-in error of the absolute-value term to the plug-in error of the underlying signed quantity $\hat{Q}_j$. Combining Eq. (89) with the $\hat{N}_j$ term gives

$$\varepsilon_n \le \sum_{j=1}^{K}\sup_{f\in\mathcal{F}}\big|\hat{Q}_j(f;\hat{\pi},\hat{e}) - Q_j^0(f)\big| + \sum_{j=1}^{K}\sup_{f\in\mathcal{F}}\big|\hat{N}_j(f;\hat{\pi},\hat{e}) - N_j^0(f)\big|. \tag{90}$$

We bound each summand by the same argument, applied once with $g = \ell$ and once with $g = \bar{\ell}$. Fix j and write $\hat{h}_j(f;\hat{e}) = \frac{1}{n}\sum_i \bar{y}_{ij} g(f_j(\mathbf{x}_i))/\hat{e}_j(\mathbf{x}_i)$ and $\hat{h}_j(f;e) = \frac{1}{n}\sum_i \bar{y}_{ij} g(f_j(\mathbf{x}_i))/e_j(\mathbf{x}_i)$, so that the quantity to bound is $\big|\hat{\pi}_j\hat{h}_j(f;\hat{e})/\hat{W}_j - \pi_j\hat{h}_j(f;e)/W_j^0\big|$.

Since $g \in \{\ell,\bar{\ell}\}$ is bounded by C_L, and $\hat{h}_j(f;\hat{e})/\hat{W}_j$, respectively $\hat{h}_j(f;e)/W_j^0$, is, whenever the denominator is nonzero, a weighted average of $\{g(f_j(\mathbf{x}_i))\}_i$ with non-negative weights, it cannot leave $[0, C_L]$,

$$\left|\frac{\hat{h}_j(f;\hat{e})}{\hat{W}_j}\right| \le C_L, \qquad \left|\frac{\hat{h}_j(f;e)}{W_j^0}\right| \le C_L, \tag{91}$$

uniformly in $f \in \mathcal{F}$.

Let $A_n = \{\delta_{e,n} \le e_{\min}/2\}$, so $\mathbb{P}(A_n) \to 1$. On A_n, $\hat{e}_j(\mathbf{x}) \ge e_j(\mathbf{x}) - \delta_{e,n} \ge e_{\min}/2$ for all $\mathbf{x}, j$, so $|1/\hat{e}_j(\mathbf{x}) - 1/e_j(\mathbf{x})| \le 2\delta_{e,n}/e_{\min}^2$, and hence

$$\big|\hat{h}_j(f;\hat{e}) - \hat{h}_j(f;e)\big| \le C_L \cdot \frac{2\delta_{e,n}}{e_{\min}^2} =: \alpha, \qquad \big|\hat{W}_j - W_j^0\big| \le \frac{2\delta_{e,n}}{e_{\min}^2}, \tag{92}$$

uniformly in f, using $\frac{1}{n}\sum_i \bar{y}_{ij} \le 1$ for the second bound. By Lemma 2.3, $\mathbb{E}[W_j^0] = \pi_j \ge \pi_{\min}$, so the law of large numbers gives $\mathbb{P}(B_n) \to 1$ for $B_n = \{\min_j W_j^0 \ge \pi_{\min}/2\}$. On $A_n \cap B_n \cap \{2\delta_{e,n}/e_{\min}^2 \le \pi_{\min}/4\}$, a set whose probability tends to one,

$$\hat{W}_j \ge W_j^0 - \big|\hat{W}_j - W_j^0\big| \ge \frac{\pi_{\min}}{2} - \frac{\pi_{\min}}{4} = \frac{\pi_{\min}}{4}. \tag{93}$$

Applying the ratio-perturbation identity $a/b - a'/b' = (a-a')/b + (a'/b') \cdot (b'-b)/b$ with $a = \hat{h}_j(f;\hat{e}), b = \hat{W}_j \geq \pi_{\min}/4, a' = \hat{h}_j(f;e), b' = W_j^0 \geq \pi_{\min}/2, \beta = |\hat{W}_j - W_j^0| \leq 2\delta_{e,n}/e_{\min}^2$, and $L = |a'/b'| \leq C_L$, from Eq. (91), we get, uniformly in f,

$$\left| \frac{\hat{h}_j(f;\hat{e})}{\hat{W}_j} - \frac{\hat{h}_j(f;e)}{W_j^0} \right| \leq \frac{\alpha + L\beta}{\pi_{\min}/4} \leq \frac{16C_L}{e_{\min}^2 \pi_{\min}} \delta_{e,n}. \tag{94}$$

Writing $\hat{\pi}_j \hat{h}_j(f;\hat{e})/\hat{W}_j - \pi_j \hat{h}_j(f;e)/W_j^0 = \hat{\pi}_j\big(\hat{h}_j(f;\hat{e})/\hat{W}_j - \hat{h}_j(f;e)/W_j^0\big) + \big(\hat{h}_j(f;e)/W_j^0\big)(\hat{\pi}_j - \pi_j)$, and bounding $\hat{\pi}_j \leq 1$ together with $|\hat{h}_j(f;e)/W_j^0| \leq C_L$, from Eq. (91), and Eq. (94), gives, on the intersection of the above high-probability events,

$$\sup_{f \in \mathcal{F}} \left| \frac{\hat{\pi}_j \hat{h}_j(f;\hat{e})}{\hat{W}_j} - \frac{\pi_j \hat{h}_j(f;e)}{W_j^0} \right| \leq \frac{16C_L}{e_{\min}^2 \pi_{\min}} \delta_{e,n} + C_L \delta_{\pi,n}. \tag{95}$$

Applying Eq. (95) with $g = \ell$ and with $g = \bar{\ell}$, and summing over $j = 1, \ldots, K$ in Eq. (90), we obtain

$$\varepsilon_n \leq 2K \left(\frac{16C_L}{e_{\min}^2 \pi_{\min}} \delta_{e,n} + C_L \delta_{\pi,n} \right) = 2KC_L \delta_{\pi,n} + \frac{32KC_L}{e_{\min}^2 \pi_{\min}} \delta_{e,n}, \tag{96}$$

which proves Eq. (86) with $C_1 = 2KC_L$ and $C_2 = 32KC_L/(e_{\min}^2 \pi_{\min})$. Since $\delta_{e,n} \xrightarrow{p} 0$ from Theorem 2.9 and $\delta_{\pi,n} \xrightarrow{p} 0$ from Corollary K.2, $\varepsilon_n \xrightarrow{p} 0$.

L Complementary-Label Generation under Different Selection Mechanisms

To further evaluate the robustness of our method to different complementary-label generation mechanisms, we conduct additional experiments under five selection settings. These settings differ in whether the selection probability depends on the instance, the label identity, or only a partial nonlinear transformation of the input features. The resulting mechanisms cover uniform selection, label-dependent selection, instance-difficulty-dependent selection, and nonlinear misspecified selection in addition to the instance-dependent setting considered in the main experiments.

Table 6 reports the complete results under the four additional generation mechanisms. Overall, ID-CMLL performs favorably under the difficulty, part-dependent, and uniform mechanisms, with particularly clear gains on several rcv1 splits and under uniform selection. The advantage is less consistent under ccn, where the selection probability varies across labels but not across instances, leaving little instance-level selection variation for the propensity model to exploit. Under this setting, the two methods are comparable on several datasets, while ID-CMLL also exhibits larger variance on some eurlex datasets. A similar increase in variance is observed for eurlex_dc and eurlex_sm under the part-dependent mechanism, whose nonlinear construction deliberately departs from the affine-logistic propensity specification. Under uniform selection, where complementary labels are selected without instance- or label-dependent preference, ID-CMLL remains competitive and often improves over CTL. Overall, these results show that ID-CMLL is effective across different instance-dependent selection mechanisms and retains competitive performance when the assumed propensity structure is simplified or moderately misspecified.

Instance-dependent and label-dependent selection. The **idsar** setting corresponds to the instance-dependent and label-dependent selection mechanism considered in this work. In this setting, whether an irrelevant label is selected as complementary depends jointly on the instance features and the label identity. The underlying motivation is that labels that are clearly incompatible with a particular instance are more likely to be identified as complementary labels. To construct this mechanism, we first train an MLP classifier using the original fully supervised labels and obtain the predicted relevance probability $f_j(\mathbf{x}_i)$ for each instance-label pair. A smaller value of $f_j(\mathbf{x}_i)$ indicates that label j is less likely to be relevant to instance $\mathbf{x}_i$ and should therefore receive a larger complementary-selection score. We define the corresponding score as

$$s_{ij}^{\text{idsar}} = -\alpha f_j(\mathbf{x}_i), \tag{97}$$

where $\alpha > 0$ controls the strength of the selection bias. The scores are subsequently used to determine the relative selection probabilities among the irrelevant labels of each instance. Since $f_j(\mathbf{x}_i)$ varies across both instances and labels, this mechanism introduces both instance-dependent and label-dependent selection bias.

Uniform random selection. The **uniform** setting follows the standard uniform complementary-label generation protocol and does not depend on either the instance features or the label identity. For each instance, we first sample the number of complementary labels according to the prescribed cardinality distribution. Conditional on the sampled cardinality, the complementary labels are then drawn uniformly without replacement from the set of ground-truth irrelevant labels. Consequently, every irrelevant label of the same instance has the same probability of being selected, and no instance-dependent or label-dependent preference is introduced.

Class-conditional selection. The **ccn** setting introduces label-dependent selection bias while remaining independent of the instance features. We randomly generate a fixed score for each label and use the same label-specific score for every training instance. Let a_j denote the score associated with label j. The selection score is then given by

$$s_{ij}^{\mathrm{ccn}} = a_j. \tag{98}$$

Thus, whenever label j is irrelevant to an instance, its tendency to be selected as a complementary label is determined solely by a_j. Labels with larger scores are systematically more likely to be selected across the entire dataset, while instances sharing the same set of irrelevant labels follow the same label-level selection preference.

Instance-difficulty-dependent selection. The **difficulty** setting introduces instance-level selection bias without explicitly distinguishing among the irrelevant labels of the same instance. We use the auxiliary classifier to measure the difficulty of instance $\mathbf{x}_i$ through its average predicted relevance probability over the ground-truth irrelevant labels. Let $\mathcal{I}_i = \{j : y_{ij} = 0\}$ denote the irrelevant-label set of instance $\mathbf{x}_i$. Its difficulty score is defined as

$$d_i = \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} f_j(\mathbf{x}_i). \tag{99}$$

A larger value of d_i indicates that the classifier assigns relatively high relevance probabilities to labels that are actually irrelevant, suggesting that the instance is more difficult to distinguish. We therefore define the complementary-selection score shared by its irrelevant labels as

$$s_{ij}^{\mathrm{difficulty}} = -\alpha d_i, \quad j \in \mathcal{I}_i. \tag{100}$$

Under this construction, difficult instances receive lower complementary-selection scores as a whole, whereas easier instances are more likely to provide complementary supervision.

Part-dependent nonlinear misspecification. The **partdependent** setting is designed as a robustness test in which the complementary-label generation mechanism depends only on a label-specific subset of the input features through a nonlinear transformation. For each label j, we randomly select a small subset of feature dimensions $\mathcal{S}_j$ and construct a scalar projection from the corresponding partial representation. Denoting this projection by

$$z_{ij} = \mathbf{r}_j^\top \mathbf{x}_{i,\mathcal{S}_j}, \tag{101}$$

we define the selection score through the nonlinear mapping

$$s_{ij}^{\mathrm{part}} = \tanh(z_{ij}) + 0.5 \sin(2 z_{ij}). \tag{102}$$

The resulting mechanism depends jointly on the instance and the label because both $\mathcal{S}_j$ and $\mathbf{r}_j$ are label specific, while its nonlinear form deliberately departs from the affine-logistic specification adopted by our propensity model. This setting therefore evaluates whether the proposed method remains effective when the true complementary-label generation process is only partially represented by the model assumptions.

Table 6: Experimental results (mean ± std) on benchmark datasets.

Methods	ccn		difficulty		part-dependent		uniform	
	CTL	Ours	CTL	Ours	CTL	Ours	CTL	Ours
Hamming Loss ↓								
bookmark	.501±.001	.437±.004	.262±.006	.065±.002	.481±.001	.288±.031	.221±.005	.064±.001
delicious	.390±.002	.365±.005	.310±.004	.329±.016	.366±.002	.268±.005	.311±.003	.251±.003
eurlex_dc	.496±.008	.546±.227	.180±.006	.118±.002	.439±.004	.481±.259	.148±.007	.041±.002
eurlex_sm	.474±.005	.499±.250	.207±.010	.204±.012	.416±.004	.578±.345	.181±.016	.072±.002
rcv1-s1	.446±.003	.383±.014	.233±.007	.075±.002	.368±.007	.114±.005	.243±.006	.085±.007
rcv1-s2	.445±.005	.406±.006	.233±.005	.098±.003	.339±.005	.106±.007	.236±.005	.101±.006
rcv1-s3	.438±.008	.368±.015	.257±.012	.100±.002	.394±.005	.122±.017	.261±.006	.099±.004
rcv1-s4	.512±.007	.347±.117	.219±.005	.086±.004	.363±.005	.111±.004	.223±.005	.085±.004
rcv1-s5	.515±.005	.376±.004	.243±.005	.081±.002	.359±.005	.106±.008	.238±.008	.080±.003
scene	.401±.006	.380±.009	.302±.005	.152±.017	.336±.009	.216±.031	.270±.012	.132±.009
yeast	.370±.003	.303±.006	.283±.006	.244±.005	.279±.004	.230±.003	.250±.006	.230±.002
One Error ↓								
bookmark	.707±.015	.719±.024	.360±.006	.362±.006	.506±.007	.500±.009	.367±.005	.357±.006
delicious	.729±.009	.729±.009	.401±.008	.417±.012	.444±.016	.424±.013	.376±.013	.357±.011
eurlex_dc	.938±.009	.798±.034	.309±.018	.203±.016	.677±.017	.520±.135	.334±.011	.177±.014
eurlex_sm	.920±.006	.823±.074	.304±.016	.303±.026	.588±.019	.548±.093	.323±.011	.280±.017
rcv1-s1	.566±.018	.567±.018	.591±.015	.217±.022	.738±.010	.512±.060	.617±.031	.250±.043
rcv1-s2	.489±.039	.488±.039	.635±.029	.205±.052	.645±.022	.283±.045	.635±.043	.227±.061
rcv1-s3	.514±.054	.515±.053	.676±.068	.238±.056	.737±.033	.349±.060	.625±.066	.238±.084
rcv1-s4	.959±.013	.925±.032	.597±.050	.254±.038	.776±.051	.362±.053	.577±.044	.223±.049
rcv1-s5	.943±.010	.941±.010	.492±.013	.265±.018	.583±.024	.322±.037	.580±.015	.288±.039
scene	.836±.013	.836±.013	.382±.030	.342±.022	.485±.023	.471±.018	.304±.017	.280±.020
yeast	.754±.012	.754±.012	.245±.011	.246±.009	.245±.011	.246±.012	.245±.011	.246±.010
Ranking Loss ↓								
bookmark	.297±.006	.310±.007	.125±.009	.129±.009	.237±.006	.231±.008	.125±.002	.124±.004
delicious	.384±.005	.375±.005	.274±.003	.249±.006	.320±.003	.301±.004	.278±.005	.244±.006
eurlex_dc	.250±.005	.253±.090	.100±.009	.062±.007	.224±.006	.195±.117	.080±.005	.044±.004
eurlex_sm	.313±.003	.324±.101	.139±.008	.142±.014	.285±.011	.343±.148	.160±.010	.114±.006
rcv1-s1	.249±.006	.235±.011	.112±.008	.063±.007	.260±.007	.125±.021	.122±.007	.076±.012
rcv1-s2	.256±.012	.224±.023	.137±.009	.090±.013	.255±.005	.126±.026	.151±.010	.097±.015
rcv1-s3	.259±.016	.249±.038	.157±.013	.093±.015	.306±.023	.136±.036	.148±.017	.087±.017
rcv1-s4	.435±.016	.334±.064	.144±.012	.094±.007	.284±.018	.148±.025	.137±.013	.082±.010
rcv1-s5	.405±.011	.305±.022	.111±.003	.089±.014	.184±.006	.124±.005	.124±.005	.085±.013
scene	.415±.017	.376±.018	.147±.009	.124±.002	.181±.014	.180±.013	.100±.007	.095±.005
yeast	.359±.006	.326±.006	.209±.007	.193±.007	.206±.007	.196±.007	.207±.008	.180±.006
Average Precision ↑								
bookmark	.449±.010	.437±.013	.723±.008	.720±.007	.586±.006	.592±.008	.718±.003	.724±.004
delicious	.468±.001	.475±.001	.609±.004	.623±.006	.563±.006	.582±.005	.610±.006	.639±.007
eurlex_dc	.355±.007	.410±.096	.775±.013	.857±.012	.492±.013	.580±.183	.771±.006	.880±.009
eurlex_sm	.340±.004	.367±.070	.733±.011	.727±.024	.481±.011	.436±.171	.697±.011	.754±.014
rcv1-s1	.461±.011	.473±.013	.619±.014	.821±.017	.425±.009	.634±.044	.598±.018	.796±.029
rcv1-s2	.494±.018	.527±.018	.584±.016	.781±.033	.481±.011	.723±.038	.573±.023	.770±.040
rcv1-s3	.483±.026	.499±.031	.557±.035	.769±.036	.391±.032	.688±.050	.575±.038	.774±.052
rcv1-s4	.227±.011	.320±.035	.589±.031	.759±.022	.410±.035	.661±.037	.597±.028	.788±.026
rcv1-s5	.255±.008	.306±.021	.661±.005	.772±.013	.559±.015	.718±.019	.615±.010	.765±.028
scene	.434±.011	.456±.013	.763±.015	.793±.009	.702±.015	.710±.012	.820±.011	.833±.010
yeast	.522±.005	.535±.006	.708±.009	.724±.011	.719±.009	.731±.010	.718±.009	.743±.009
Coverage Error ↓								
bookmark	.309±.006	.320±.006	.149±.008	.154±.008	.259±.005	.253±.007	.148±.002	.149±.003
delicious	.656±.007	.652±.007	.587±.008	.547±.008	.626±.005	.606±.007	.596±.008	.548±.008
eurlex_dc	.236±.005	.239±.084	.096±.009	.061±.007	.212±.006	.185±.109	.077±.004	.043±.004
eurlex_sm	.394±.005	.406±.091	.225±.012	.224±.021	.380±.012	.424±.134	.244±.013	.188±.009
rcv1-s1	.339±.003	.320±.015	.169±.011	.122±.010	.328±.012	.184±.022	.177±.009	.136±.013
rcv1-s2	.381±.012	.343±.020	.225±.013	.191±.011	.349±.007	.238±.028	.240±.014	.194±.012
rcv1-s3	.371±.017	.356±.034	.240±.013	.186±.013	.388±.023	.235±.032	.227±.019	.182±.014
rcv1-s4	.482±.015	.414±.045	.216±.016	.167±.005	.355±.021	.240±.031	.210±.019	.155±.006
rcv1-s5	.451±.012	.356±.020	.168±.006	.149±.016	.248±.006	.191±.004	.180±.009	.141±.014
scene	.353±.014	.323±.015	.136±.007	.118±.002	.166±.014	.166±.013	.097±.007	.093±.005
yeast	.723±.011	.667±.008	.489±.007	.469±.005	.492±.005	.490±.005	.500±.008	.461±.007