Towards a Pairwise Ranking Model with Orderliness and Monotonicity for Label Enhancement

Yunan Lu[1,2], **Xixi Zhang**[1], **Yaojin Lin**[3], **Weiwei Li**[4], **Lei Yang**[2], **Xiuyi Jia**[1*]

[1] Nanjing University of Science and Technology, Nanjing, China
[2] The Hong Kong Polytechnic University, Hong Kong, China
[3] Minnan Normal University, Fujian, China
[4] Nanjing University of Aeronautics and Astronautics, Nanjing, China
{luyn,zhangxixi,jiaxy}@njust.edu.cn, yjlin@mnnu.edu.cn,
liweiwei@nuaa.edu.cn, ray.yang@polyu.edu.hk

Abstract

Label distribution in recent years has been applied in a diverse array of complex decision-making tasks. To address the availability of label distributions, label enhancement has been established as an effective learning paradigm that aims to automatically infer label distributions from readily available multi-label data, e.g., logical labels. Recently, numerous works have demonstrated that the label ranking is significantly beneficial to label enhancement. However, these works still exhibit deficiencies in representing the probabilistic relationships between label distribution and label rankings, or fail to accommodate scenarios where multiple labels are equally important for a given instance. Therefore, we propose PROM, a pairwise ranking model with orderliness and monotonicity, to explain the probabilistic relationship between label distributions and label rankings. Specifically, we propose the monotonicity and orderliness assumptions for the probabilities of different ranking relationships and derive the mass functions for PROM, which are theoretically ensured to preserve the monotonicity and orderliness. Further, we propose a generative label enhancement algorithm based on PROM, which directly learns a label distribution predictor from the readily available multi-label data. Finally, extensive experiments demonstrate the efficacy of our proposed model.

1 Introduction

Label polysemy, i.e., the cases where an instance is associated with multiple labels simultaneously, is a common phenomenon in real-world tasks. To preserve label polysemy, multi-label learning (Tsoumakas and Katakis, 2006) employs binary values to indicate the presence or absence of each label for a given instance. However, multi-label learning cannot directly handle a further question with more polysemy: How much is each label associated with an instance? Hence, Geng (2016) introduced LDL (Label Distribution Learning). Unlike multi-label learning, LDL assigns a real-valued vector to each instance, resembling probability distributions, where each element, called label description degree, represents the extent to which the label describes the instance. By providing richer label information, LDL has been applied in various fields, including age estimation (Gao et al., 2018, Geng et al., 2013) and affective analysis (Jia et al., 2019, Machajdik and Hanbury, 2010).

A significant bottleneck hindering the broader application of LDL is the challenge of acquiring ground-truth label distributions, as accurately quantifying these distributions can be costly. To address

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

this issue, LE (Label Enhancement) has been proposed to automatically infer label distributions from more readily available multi-label data by mining the underlying label polysemy information (Xu et al., 2018, 2020). Initially, most LE methods mainly focused on mining the underlying label polysemy information from instance correlation and label correlation. Recently, some novel findings (Jia et al., 2023b, 2024, Wang and Geng, 2021) that the label ranking is beneficial for improving the generalization of LDL have inspired a portion of LE research to explore the label polysemy information from the perspective of label ranking. For example, Jia et al. (2023a), Lu et al. (2023a,b) proposed to incorporate the ranking relationship between positive and negative labels into the loss function or the generation model of label distributions. Although these works have made headway in designing methods for regularizing LE processes by label rankings, there remain the following two research gaps. 1) Current works ignore the ranking relation of tie, i.e. the cases where multiple labels simultaneously describe the instance to almost the same degree, which is ubiquitous in real-world tasks. 2) Current works neglect the qualitative and quantitative modeling of the probabilistic relationships between label distributions and label rankings, which is essential for the interpretable and generative modeling of label enhancement. Besides, it should be noted that while some classic ranking models, such as Bradley-Terry (Hunter, 2004) and Plackett-Luce (Guiver and Snelson, 2009) models, can be employed as a quantitative model, they also cannot explicitly model the tie relation.

Therefore, we propose PROM (a Pairwise Ranking model with Orderliness and Monotonicity) to qualitatively and quantitatively explain the probabilistic relationship between label distributions and label rankings, which can serve as a generative distribution of label rankings in generative label enhancement or as a loss function for measuring the inconsistency between label rankings and label distributions. Specifically, we first conduct a qualitative analysis of the probabilistic principles governing the generation process of label rankings from label distributions, and formalize these principles as the assumptions on the monotonicity and orderliness of the probabilities of different ranking relationships. Second, we design parameterized probability mass functions for the ranking model, and derive the conditions under which the model adheres to probabilistic monotonicity and orderliness. Third, we propose a generative label enhancement algorithm based on the PROM model, called LE-PROM, which integrates the LE process and the LDL process into a unified framework, and directly learns an LDL mapping on the training instances with multi-label data. Finally, we validate our proposal through extensive experiments on real-world datasets. The experimental results demonstrate the superiority of our proposed method.

2 Related Work

To tackle the challenge of acquiring accurate ground-truth label distributions, LE (Label Enhancement) was proposed to automatically infer label distributions from more readily available multi-label data, such as logical labels (Xu et al., 2018, Kou et al., 2025), ternary labels Lu and Jia (2024), multi-label rankings (Lu and Jia, 2022, Lu et al., 2023c), or inaccurate label distributions Kou et al. (2024, 2023), Lu et al. (2025). Most LE methods mine the label polysemy from instance and label correlations. For example, in order to capture the instance correlation, the prototype-based LE algorithms (El Gayar et al., 2006, Jiang et al., 2006, Wang et al., 2023, Fan et al., 2024) identify the representative points and subsequently estimate the label distribution based on the representative points of each label. The graph-based LE algorithms (Xu et al., 2018, Zhang et al., 2021, Xu et al., 2019, Liu et al., 2021) directly calculate affinities between instances based on the features, which is further utilized to regularize the label distribution matrix. The manifold-based LE algorithms (Hou et al., 2016, Wen et al., 2021, Tang et al., 2020, Zhang et al., 2018) learn coefficients for each instance, by which the feature vector of each instance can be linearly reconstructed from its neighbors, and then maintain the reconstruction within label distributions. In order to capture the label correlation, Luo et al. (2021) utilized the confusion matrix to estimate global label correlation. Besides, several works address the LE task under special data distribution. For example, VIB-ILE Song et al. (2024) addresses imbalanced label information in LE by adopting a variational information bottleneck mechanism and introducing consistency regularization.

Recently, a portion of LE research attempts to explore the label polysemy information from the perspective of label ranking. For example, Jia et al. (2023a) utilized a margin-based loss function to penalize the cases where the description degree of negative labels exceeds that of positive labels; Lu et al. (2023a) designs a conditional distribution of logical labels given label distributions, whose

parameters are regularized by the ranking between negative and positive labels. Lu et al. (2023b) derives a variational approximation for label distribution in generative label enhancement, which is theoretically assured to uphold the ranking relation between negative and positive labels. Besides, Lu and Jia (2022), Lu et al. (2023c) theoretically and methodologically investigated the task of predicting label distributions directly from multi-label rankings. Despite the success of these efforts, there remains a gap in the qualitative and quantitative modeling of the probabilistic relationships between label distributions and the tie-allowed label rankings.

3 Methodology

In this section, we first introduce the notations commonly used in this paper, secondly elaborate our proposed PROM model, and finally illustrate the generative LE framework.

3.1 Notations

We denote the D-dimensional feature space by $\mathcal{X}^D = \mathbb{R}^D$, denote the M-dimensional label distribution space by $\Delta^M = \{\boldsymbol{v} \in \mathbb{R}_+^M : \sum_{m=1}^M v_m = 1\}$, denote the label space by $\mathcal{Y} = \{\ell_m\}_{m=1}^M$. We cope with the training datasets that appear as data pairs $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$, where $\boldsymbol{x}_n \in \mathcal{X}^D$ and $\boldsymbol{y}_n$ denote the feature vector and the vector of the easily available label values of the n-th instance, respectively. The data of $\boldsymbol{y}_n$ can directly yield pairwise ranking relation among labels. We denote the ranking relation between label ℓ_i and label ℓ_j by $\xi_{ij} \in \{\ell_i \prec \ell_j, \ell_i \succ \ell_j, \ell_i \simeq \ell_j\}$, where $\ell_i \prec \ell_j$, $\ell_i \succ \ell_j$, and $\ell_i \simeq \ell_j$ denote that the label ℓ_i, compared to the label ℓ_j, describes the instance to a higher, lower, and approximately the same degree, respectively. The goal of LE is to learn a label distribution predictor $f : \mathcal{X}^D \to \Delta^M$, i.e., a mapping from feature space to label distribution space, by the training dataset $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$.

3.2 PROM

In this subsection, we first propose assumptions to formalize the monotonicity and orderliness of ranking probability. Then, we derive the parametric form of PROM.

3.2.1 Fundamental Assumptions of PROM

On the one hand, we study how the label description degree affects the probability of the label ranking, focusing on the monotonicity of the probability of the label ranking. Obviously, given any two labels ℓ_i, ℓ_j of an instance, the label ℓ_i is more likely to rank above the label ℓ_j (or the label ℓ_j is more likely to rank below the label ℓ_i alternatively) if the description degree of ℓ_i exceeds that of ℓ_j by a larger margin. Besides, if the description degrees of two labels are close, then the strict ranking relation between the two labels will be difficult to distinguish, i.e., it is more likely that the two labels are tied. We formalize the above intuition as the probability monotonicity assumption.

Assumption 3.1 (Probability monotonicity)**.** Given any two labels ℓ_i, ℓ_j of an instance and their respective description degrees z_i, z_j, $p(\ell_i \prec \ell_j | z_i = u_i, z_j = u_j) < p(\ell_i \prec \ell_j | z_i = v_i, z_j = v_j)$ and $p(\ell_i \succ \ell_j | z_i = u_i, z_j = u_j) > p(\ell_i \succ \ell_j | z_i = v_i, z_j = v_j)$ holds for any $u_i - u_j < v_i - v_j$. $p(\ell_i \simeq \ell_j | z_i = u_i, z_j = u_j) > p(\ell_i \simeq \ell_j | z_i = v_i, z_j = v_j)$ holds for any $|u_i - u_j| < |v_i - v_j|$, where $u_i, u_j, v_i, v_j \in [0, 1]$.

On the other hand, we explore how the label description degree affects the orderliness among the probabilities of different rankings. Obviously, given any two labels ℓ_i, ℓ_j of an instance, these two labels are most likely to be tied if the description degrees of the two labels are sufficiently close; the label ℓ_i is most likely to rank above the label ℓ_j (or the label ℓ_j is most likely to rank below the label ℓ_i alternatively) if the label description degree of ℓ_i exceeds that of ℓ_j by a sufficiently large margin. We formalize the above intuition as the probability orderliness assumption.

Assumption 3.2 (Probability orderliness)**.** There exists two real thresholds $-1 < z_- < 0 < z_+ < 1$, $p(\ell_i \succ \ell_j | z_i, z_j) > p(\ell_i \simeq \ell_j | z_i, z_j) > p(\ell_i \prec \ell_j | z_i, z_j)$ holds for any $z_i - z_j < z_-$; $p(\ell_i \succ \ell_j | z_i, z_j) < p(\ell_i \simeq \ell_j | z_i, z_j) < p(\ell_i \prec \ell_j | z_i, z_j)$ holds for any $z_i - z_j > z_+$; $p(\ell_i \simeq \ell_j | z_i, z_j) > \max\{p(\ell_i \succ \ell_j | z_i, z_j), p(\ell_i \prec \ell_j | z_i, z_j)\}$ holds for any $z_- < z_i - z_j < z_+$.

The intuition of probability monotonicity and orderliness is visualized in Figure 1. Actually, the above discussion also implicitly adheres to the following two assumptions. The one is that the ranking relation between two labels depends solely on the difference in their description degrees, and is independent of the description degrees themselves. We refer to this assumption as the translational invariance assumption. The other is that the ranking relation between two labels is symmetric, i.e., $\ell_i \prec \ell_j$ is actually $\ell_j \succ \ell_i$. We refer to this assumption as the ranking symmetry assumption. We formalize these two assumptions as follows.

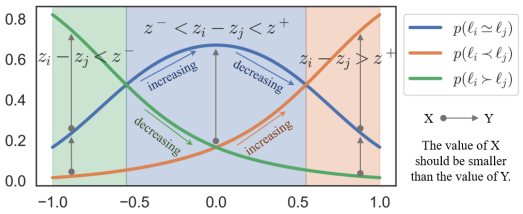


Figure 1: Schematic of probability monotonicity and probability orderliness.

Assumption 3.3 (Translational invariance)**.** Given any two labels ℓ_i, ℓ_j of an instance and their respective description degrees z_i, z_j, $p(\ell_i \prec \ell_j | z_i = u_i, z_j = u_j) = p(\ell_i \prec \ell_j | z_i = v_i, z_j = v_j)$ and $p(\ell_i \succ r_j | z_i = u_i, z_j = u_j) = p(\ell_i \succ \ell_j | z_i = v_i, z_j = v_j)$ hold for any $u_i - u_j = v_i - v_j$; $p(\ell_i \simeq \ell_j | z_i = u_i, z_j = u_j) = p(\ell_i \simeq \ell_j | z_i = v_i, z_j = v_j)$ holds for any $|u_i - u_j| = |v_i - v_j|$, where $u_i, u_j, v_i, v_j \in [0, 1]$.

Assumption 3.4 (Ranking symmetry)**.** Given any two labels ℓ_i, ℓ_j of an instance and their respective description degrees z_i, z_j, $p(\ell_i \prec \ell_j | z_i, z_j) = p(\ell_j \succ \ell_i | z_j, z_i)$ and $p(\ell_i \simeq \ell_j | z_i, z_j) = p(\ell_j \simeq \ell_i | z_j, z_i)$.

3.2.2 Probability Mass Functions of PROM

Since the ranking relation between two labels can be $\ell_i \prec \ell_j$, $\ell_i \succ \ell_j$, or $\ell_i \simeq \ell_j$, we model the ranking relation by a categorical distribution, which can be formalized as:

$$\xi_{ij} \mid z_i, z_j \sim \text{Categorical}(\xi_{ij} \mid [\underline{\phi}(z_i, z_j), \tilde{\phi}(z_i, z_j), \overline{\phi}(z_i, z_j)]), \tag{1}$$

where $\underline{\phi}(z_i, z_j)$, $\overline{\phi}(z_i, z_j)$, and $\tilde{\phi}(z_i, z_j)$ capture the rules of generating the label ranking relation $\ell_i \succ \ell_j$, $\ell_i \prec \ell_j$, and $\ell_i \simeq \ell_j$, respectively. Considering that the softmax function is most commonly used to model the probability of discrete events, we preliminarily assume the parametric form of $\boldsymbol{\phi}(z_i, z_j)$ as follows:

$$\begin{aligned}
\underline{\phi}(z_i, z_j) &= p(\ell_i \succ \ell_j | z_i, z_j) = Z^{-1} \exp(a_{\underline{\phi}}(z_i - z_j) + b_{\underline{\phi}}), \\
\tilde{\phi}(z_i, z_j) &= p(\ell_i \simeq \ell_j | z_i, z_j) = Z^{-1} \exp(a_{\tilde{\phi}}(z_i - z_j)^2 + b_{\tilde{\phi}}), \\
\overline{\phi}(z_i, z_j) &= p(\ell_i \prec \ell_j | z_i, z_j) = Z^{-1} \exp(a_{\overline{\phi}}(z_i - z_j) + b_{\overline{\phi}}),
\end{aligned} \tag{2}$$

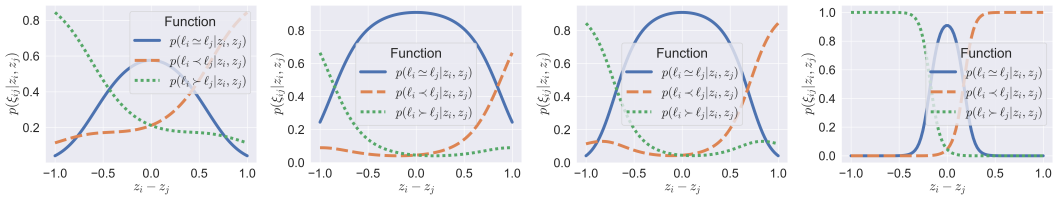
where Z is a normalization factor that ensures $\underline{\phi}(z_i, z_j) + \tilde{\phi}(z_i, z_j) + \overline{\phi}(z_i, z_j) = 1$, $a_{\underline{\phi}} < 0$, $a_{\tilde{\phi}} < 0$, $a_{\overline{\phi}} > 0$, and $b_{\underline{\phi}}$, $b_{\tilde{\phi}}$, and $b_{\overline{\phi}}$ are arbitrary real numbers.

Figure 2 visualizes PROM model at different values of parameters. It is evident that not any value of parameter can adhere to the probability monotonicity assumption and the probability orderliness assumption. For example, Figure 2(a) violates the probability monotonicity assumption, Figure 2(b) violates the probability orderliness assumption, and Figure 2(c) violates both. Figure 2(d) shows a PROM model that we expect. Therefore, we next propose four theorems to ensure that the probability mass functions defined by Equation (2) adhere to the proposed assumptions.

Theorem 3.5. *$p(\xi_{ij} | z_i, z_j)$ defined by Equation* (1) *and Equation* (2) *adheres to the translation invariance assumption, i.e. Assumption 3.3.*

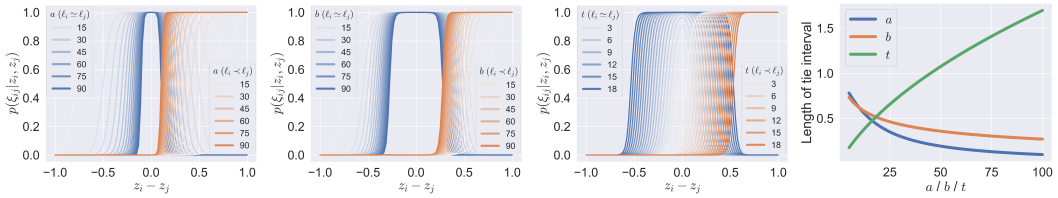
Theorem 3.6. *$p(\xi_{ij} | z_i, z_j)$ defined by Equation* (1) *and Equation* (2) *adheres to the ranking symmetry assumption, i.e. Assumption 3.4, iff $a_{\overline{\phi}} = -a_{\underline{\phi}}$ and $b_{\overline{\phi}} = b_{\underline{\phi}}$.*

Theorem 3.7. *Given $a_{\overline{\phi}} > 0$, $a_{\underline{\phi}} < 0$, $a_{\tilde{\phi}} < 0$, $\Delta = z_i - z_j$. $\tilde{\phi}$ adheres to the monotonicity assumption of $p(\ell_i \simeq \ell_j | z_i, z_j)$ if $a_{\overline{\phi}} = -a_{\underline{\phi}} \exp(b_{\underline{\phi}} - b_{\overline{\phi}})$; $\overline{\phi}$ adheres to the monotonicity assumption of $p(\ell_i \prec \ell_j | z_i, z_j)$ if $(a_{\overline{\phi}}(2a_{\tilde{\phi}})^{-1} < -1) \vee (\overline{\Delta}^\star \leq -1 \wedge h_{\overline{\phi}}(-1) > 0) \vee (\overline{\Delta}^\star > -1 \wedge h_{\overline{\phi}}(\overline{\Delta}^\star) > 0)$; $\underline{\phi}$ adheres to the monotonicity assumption of $p(\ell_i \succ \ell_j | z_i, z_j)$*



(a) $a = 1, b = 3, t = 1.$ (b) $a = 1, b = 3, t = 3.$ (c) $a = 1, b = 5, t = 3.$ (d) $a = 16, b = 8, t = 3.$

Figure 2: The shape of PROM with varying values of parameters.



(a) $b = 10, t = 10.$ (b) $a = 10, t = 10.$ (c) $a = 25, b = 20.$ (d) Length of tie interval.

Figure 3: Marginal effect of parameters on the shape of PROM.

if $(a_{\underline{\phi}}(2a_{\tilde{\phi}})^{-1} > 1) \vee (\underline{\Delta}^\star \geq 1 \wedge h_{\underline{\phi}}(1) > 0) \vee (\underline{\Delta}^\star < 1 \wedge h_{\underline{\phi}}(\underline{\Delta}^\star) > 0)$ *where* $\overline{\Delta}^\star$, $\underline{\Delta}^\star$, $h_{\overline{\phi}}(\Delta)$, *and* $h_{\underline{\phi}}(\Delta)$ *are defined by* $\overline{\Delta}^\star := (a_{\overline{\phi}} + a_{\underline{\phi}} + \sqrt{(a_{\overline{\phi}} - a_{\underline{\phi}})^2 - 8a_{\tilde{\phi}}})(4a_{\tilde{\phi}})^{-1}$, $\underline{\Delta}^\star := (a_{\overline{\phi}} + a_{\underline{\phi}} - \sqrt{(a_{\overline{\phi}} - a_{\underline{\phi}})^2 - 8a_{\tilde{\phi}}})(4a_{\tilde{\phi}})^{-1}$, $h_{\overline{\phi}}(\Delta) := a_{\underline{\phi}}\Delta - a_{\tilde{\phi}}\Delta^2 + b_{\underline{\phi}} - b_{\tilde{\phi}} + \log \frac{a_{\overline{\phi}} - a_{\underline{\phi}}}{2a_{\tilde{\phi}}\Delta - a_{\overline{\phi}}}$, $h_{\underline{\phi}}(\Delta) := -a_{\tilde{\phi}}\Delta^2 + a_{\overline{\phi}}\Delta + b_{\overline{\phi}} - b_{\tilde{\phi}} + \log \frac{a_{\overline{\phi}} - a_{\underline{\phi}}}{a_{\underline{\phi}} - 2a_{\tilde{\phi}}\Delta}$.

Theorem 3.8. $p(\xi_{ij}|z_i, z_j)$ *defined by Equation* (1) *and Equation* (2) *adheres to the probability orderliness assumption, i.e. Assumption 3.2, iff* $\max\{b_{\overline{\phi}} - a_{\overline{\phi}}, b_{\underline{\phi}} + a_{\underline{\phi}}\} < a_{\tilde{\phi}} + b_{\tilde{\phi}} < \min\{b_{\overline{\phi}} + a_{\overline{\phi}}, b_{\underline{\phi}} - a_{\underline{\phi}}\}$ *and* $b_{\tilde{\phi}} > \max\{b_{\overline{\phi}}, b_{\underline{\phi}}\}$.

The proof of the above theorems can be found in Appendix A. According to the above theorems, let $a = a_{\overline{\phi}} = -a_{\underline{\phi}}$, $b = -a_{\tilde{\phi}}$, $b_{\overline{\phi}} = b_{\underline{\phi}}$, $t = b_{\tilde{\phi}} - b_{\overline{\phi}}$, we finally define the probability mass function of PROM model as follows:

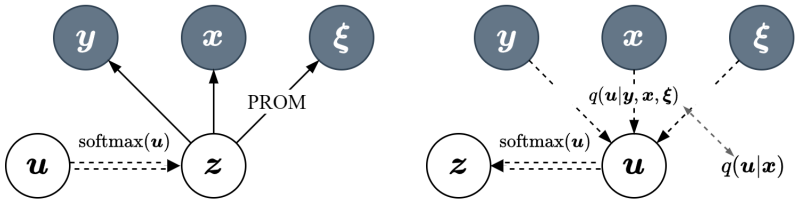
$$\underline{\phi}(z_i, z_j) = Z^{-1}e^{-a(z_i - z_j)},\ \tilde{\phi}(z_i, z_j) = Z^{-1}e^{-b(z_i - z_j)^2 + t},\ \overline{\phi}(z_i, z_j) = Z^{-1}e^{a(z_i - z_j)}, \quad (3)$$

where the parameters are bounded by $(a, b, t) \in \mathcal{F}$, where the feasible region $\mathcal{F}$ is defined by Equation (4):

$$\begin{aligned}\mathcal{F} = \big\{(a, b, t) \mid (a > |b - t| \wedge b > 0 \wedge t > 0) \wedge \Big(\big(a^2 \geq 4b^2 - 2b \wedge t < b + a - \log\frac{2b - a}{2a}\big) \\ \vee \big(a \geq 2b\big) \vee \ \big(t < \frac{2a\sqrt{a^2 + 2b} + a^2}{4b} - \log\frac{\sqrt{a^2 + 2b} - a}{2a} + \frac{1}{2} \wedge\ a^2 < 4b^2 - 2b\big)\Big)\big\}.\end{aligned} \quad (4)$$

3.2.3 Marginal Effect of Parameters on PROM

Intuitively, the parameters a and b control the uncertainty of $\ell_i \prec \ell_j$ (or $\ell_i \succ \ell_j$) and $\ell_i \simeq \ell_j$, respectively. As shown in Figure 3(a) and Figure 3(b), larger a and b correspond to the steeper density curve of the probabilities of $\ell_i \prec \ell_j$ (or $\ell_i \succ \ell_j$) and $\ell_i \simeq \ell_j$, respectively. More results are shown in Appendix B. We do not show $p(\ell_i \succ \ell_j | z_i, z_j)$ for the sake of readability. The parameter t is similar to the temperature coefficient in the softmax function, whose effect on the PROM distribution is shown in Figure 3(c). Overall, these three hyper-parameters directly affect the length of tie interval of PROM distribution, i.e., the length of the interval of $z_i - z_j$ with $p(\ell_i \simeq \ell_j) > \max\{p(\ell_i \prec \ell_j), p(\ell_i \succ \ell_j)\}$, which is shown in Figure 3(d).



(a) Generation Process (b) Inference Process

Figure 4: Schematic diagram of LE-PROM.

3.3 LE-PROM

LE-PROM consists of a probabilistic generation process for observed variables (features, available labels, pairwise rankings) and a posterior inference process for latent variables (label distribution), which will be illustrated in the following two subsections.

3.3.1 Generation Process

Here we illustrate the process of generating the observations, which is shown in Figure 4(a) and formalized as follows.

1. Generate a sample of label distribution:
   (a) Generate a real-valued vector $\boldsymbol{u}$ from a standard multivariate normal distribution, i.e., $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{u} \mid \mathbf{0}_M, \boldsymbol{I}_M)$, where $\mathbf{0}_M$ is an M-dimensional all-zero vector, $\boldsymbol{I}_M$ is an $M \times M$ identity matrix.
   (b) Transform the real-valued vector $\boldsymbol{u}$ into the label distribution $\boldsymbol{z}$, i.e., $\boldsymbol{z} = \text{softmax}(\boldsymbol{u})$.
2. Generate the available label $\boldsymbol{y}$ from a task-specific distribution given $\boldsymbol{z}$: $\boldsymbol{y} \mid \boldsymbol{z} \sim \tilde{p}(\boldsymbol{y}|\boldsymbol{z})$[2].
3. Generate a sample of feature variables $\boldsymbol{x}$ from a multivariate normal distribution conditioned on the label distribution[3]. Formally, $\boldsymbol{x} \mid \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{x} \mid \mu_{\boldsymbol{x}}(\boldsymbol{z}), \text{diag}(\sigma^2_{\boldsymbol{x}}(\boldsymbol{z})))$, where $\mu_{\boldsymbol{x}}(\cdot)$ and $\sigma_{\boldsymbol{x}}(\cdot)$ are two neural networks, and $\text{diag}(\sigma)$ is a diagonal matrix whose (i, i) entry is the i-th element of σ.
4. Generate a sample of pairwise rankings $\boldsymbol{\xi}$ from our designed PROM distribution conditioned on the label distribution $\boldsymbol{z}$. Formally, $\boldsymbol{\xi} \mid \boldsymbol{z} \sim \prod_{i<j} \text{PROM}(\xi_{ij} \mid z_i, z_j)$.

According to the above generation process, the joint distribution of the complete-data can be factorized as $p(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}) = p(\boldsymbol{u})p(\boldsymbol{x}|\boldsymbol{u})p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{\xi}|\boldsymbol{u})$. It should be noted that we omit label distribution $\boldsymbol{z}$ since $\boldsymbol{z}$ can be deterministically obtained from $\boldsymbol{u}$.

3.3.2 Model Inference

Here we derive the inference method of our proposed generative model. The schematic is shown in Figure 4(b). We employ variational inference to approximate the true posterior, since the exact inference of the posterior of $\boldsymbol{z}$ is intractable due to the nonlinear and non-conjugate relationships among variables. First, we assume that the variational posterior of $\boldsymbol{u}$ can be decomposed as

$$q(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{u} \mid \mu_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}), \text{diag}(\sigma_{\boldsymbol{u}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}))). \tag{5}$$

Then, we find the variational posterior distribution of $\boldsymbol{u}$ that is closest to the true posterior distribution. It has been proven that the Kullback-Leibler (KL) divergence between the variational posterior and the true posterior can be minimized by maximizing the ELBO (evidence lower bound):

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})}[\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}|\boldsymbol{u})] - \mathcal{D}_{\text{KL}}(q(\boldsymbol{u}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}) \| p(\boldsymbol{u})). \tag{6}$$

In the right-hand side of Equation (6), the first term is also known as the reconstruction term, and the second term is also known as the prior regularization term. The reconstruction term serves to

[2]The task-specific distribution can be conditioned on $\boldsymbol{u}$ or $\boldsymbol{z}$ since $\boldsymbol{z}$ can be deterministically obtained by $\boldsymbol{u}$.
[3]Analogous to $\tilde{p}(\boldsymbol{y}|\boldsymbol{z})$, the multivariate normal distribution for features can also be conditioned on $\boldsymbol{z}$ or $\boldsymbol{u}$.

ensure that the label distribution can accurately reconstruct the observed variables. It is evident that an analytic form of reconstruction term is intractable due to the integral of a complicated likelihood function. Therefore, we turn to SGVB estimator. Since $q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})$ is multivariate normal, it can be reparameterized by $q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) = \mu_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) + \sigma_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is standard normal noise. Besides, benefiting from (Lu et al., 2023a), we utilize hyperparameters α and β to re-weight the reconstruction quality of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{u}$ to re-balance the magnitudes of different likelihood functions. Finally, the reconstruction term can be estimated by:

$$\mathbb{E}_{q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})}[\log p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}|\boldsymbol{u})] \approx \frac{1}{(1+\alpha+\beta)L}\sum_{t=1}^{L}\log p\left(\boldsymbol{x}|\mu_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) + \sigma_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) \odot \boldsymbol{\epsilon}^{(t)}\right)$$
$$+\alpha \log p\left(\boldsymbol{y}|\mu_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) + \sigma_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) \odot \boldsymbol{\epsilon}^{(t)}\right) + \beta \log p\left(\boldsymbol{\xi}|\mu_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) + \sigma_{\boldsymbol{u}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) \odot \boldsymbol{\epsilon}^{(t)}\right), \tag{7}$$

where L is the number of Monte Carlo samples, $\boldsymbol{\epsilon}^{(t)}$ is a sample from the standard normal distribution. The prior regularization term encourages the posterior to approach the prior. Since both the $\boldsymbol{u}$-posterior and $\boldsymbol{u}$-prior are multivariate normal, their KL divergence can be computed analytically. Besides, in order to train a feature-conditioned label distribution predictor $\hat{p}(\boldsymbol{u}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{u} \mid \mu(\boldsymbol{x}), \mathrm{diag}(\sigma(\boldsymbol{x})))$ during the model inference phase, we add a prediction loss term to ELBO. Finally, the optimization objective can be formalized as follows:

$$\arg\max\ \mathbb{E}_{q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})}[\log p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}|\boldsymbol{u})] - \mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})\|p(\boldsymbol{u})) - \lambda\mathcal{D}_{\mathrm{KL}}(\hat{p}(\boldsymbol{u}|\boldsymbol{x})\|q(\boldsymbol{u}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi})), \tag{8}$$

where λ is a trade-off hyperparameter. As suggested by Daunizeau (2017), since $\boldsymbol{z}$ is the softmax normalization of a Gaussian random vector, the expected label distribution based on $\hat{p}(\boldsymbol{u}|\boldsymbol{x})$ can be approximated as follows:

$$\mathbb{E}_{\hat{p}(\boldsymbol{u}|\boldsymbol{x})}[z_m] \approx \frac{1}{\sum_{i=1}^{M}\exp(\psi_{mi})}, \quad \psi_{mi} = \frac{\mu_i - \mu_m}{\sqrt{1+3\pi^{-2}(\sigma_m^2+\sigma_i^2)}}, \tag{9}$$

where μ_i and σ_i denote the i-th elements of $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$, respectively.

4 Experiments

4.1 Datasets and Evaluation Measures

We conduct experiments on 16 real-world datasets. These datasets are collected from diverse real-world tasks, including emotion analysis, movie rating prediction, and bioinformatics. Specifically, "Painting" (Machajdik and Hanbury, 2010), "Emotion6" (Peng et al., 2015), "Music" (Lee et al., 2021), "BU-3DFE" (Yin et al., 2006), and "JAFFE" (Lyons et al., 1998) are collected from emotion analysis tasks. "Movie" (Geng, 2016) is a dataset from movie rating prediction task. "Alpha", "Cdc", "Cold", "Diau", "Dtt", "Elu", "Heat", "Spo", "Spo5", and "Spoem" (Geng, 2016) are collected from a gene expression analysis task. To accelerate convergence, we use min-max normalization to preprocess the feature data. We evaluate the model performance by two representative LDL measures: KL (Kullback-Leibler divergence) and Cosine (cosine similarity). The KL (or Cosine) metric with smaller values represents the better (or worse) performance. More details of the datasets and the results on more metrics can be found in Appendix C.

4.2 Experiments on Logical Label Enhancement

Here, we aim to experimentally answer the question of *whether our proposed LE-PROM is superior to the state-of-the-art logical label enhancement algorithms*.

4.2.1 Experimental Configurations

Experimental Method. Initially, we randomly divide the dataset, allocating 70% of the dataset to the training set and the remaining 30% to the testing set. Then we reduce the label distributions of the training instances to logical labels, and the detailed reduction algorithm can be found in Appendix C. Subsequently, we apply LE algorithms to transform the logical labels into the label distribution for each training instance, and utilize the recovered label distributions to train an LDL

Dataset	LE-PROM	CWLD	VIB-ILE	GLLE	KMLE	FCMLE
			KL (↓)			
Painting	(1) $0.567_{\pm 0.022}$	(5) $0.991_{\pm 0.068}$•	(6) $27.118_{\pm 2.496}$•	(2) $0.570_{\pm 0.025}$	(4) $0.978_{\pm 0.217}$•	(3) $0.589_{\pm 0.030}$•
Emotion6	(1) $0.616_{\pm 0.013}$	(5) $0.686_{\pm 0.014}$•	(6) $0.737_{\pm 0.017}$•	(2) $0.627_{\pm 0.013}$•	(3) $0.646_{\pm 0.026}$•	(4) $0.682_{\pm 0.012}$•
Movie	(1) $0.114_{\pm 0.002}$	(5) $0.211_{\pm 0.016}$•	(4) $0.193_{\pm 0.021}$•	(2) $0.115_{\pm 0.003}$•	(6) $0.216_{\pm 0.015}$•	(3) $0.178_{\pm 0.003}$•
Music	(1) $0.113_{\pm 0.007}$	(4) $0.147_{\pm 0.002}$•	(6) $0.225_{\pm 0.012}$•	(2) $0.135_{\pm 0.008}$•	(5) $0.187_{\pm 0.094}$•	(3) $0.136_{\pm 0.009}$•
BU-3DFE	(1) $0.072_{\pm 0.001}$	(5) $0.101_{\pm 0.003}$•	(4) $0.096_{\pm 0.003}$•	(1) $0.072_{\pm 0.002}$	(6) $0.111_{\pm 0.009}$•	(3) $0.079_{\pm 0.002}$•
JAFFE	(1) $0.054_{\pm 0.005}$	(5) $0.152_{\pm 0.039}$•	(4) $0.084_{\pm 0.006}$•	(2) $0.064_{\pm 0.007}$•	(6) $0.409_{\pm 0.168}$•	(3) $0.071_{\pm 0.005}$•
Alpha	(1) $0.006_{\pm 0.000}$	(5) $0.012_{\pm 0.001}$•	(1) $0.006_{\pm 0.000}$	(4) $0.010_{\pm 0.001}$•	(6) $0.024_{\pm 0.003}$•	(1) $0.006_{\pm 0.000}$
Cdc	(1) $0.007_{\pm 0.000}$	(5) $0.012_{\pm 0.000}$•	(3) $0.008_{\pm 0.000}$•	(4) $0.011_{\pm 0.001}$•	(6) $0.019_{\pm 0.001}$•	(1) $0.007_{\pm 0.000}$
Cold	(1) $0.012_{\pm 0.001}$	(5) $0.017_{\pm 0.000}$•	(2) $0.013_{\pm 0.001}$	(4) $0.016_{\pm 0.001}$•	(6) $0.027_{\pm 0.002}$•	(2) $0.013_{\pm 0.000}$•
Diau	(1) $0.014_{\pm 0.001}$	(3) $0.020_{\pm 0.001}$•	(4) $0.022_{\pm 0.001}$•	(4) $0.022_{\pm 0.001}$•	(6) $0.075_{\pm 0.003}$•	(2) $0.015_{\pm 0.000}$•
Dtt	(2) $0.007_{\pm 0.000}$	(5) $0.010_{\pm 0.000}$•	(1) $0.006_{\pm 0.000}$◦	(4) $0.009_{\pm 0.000}$•	(6) $0.014_{\pm 0.002}$•	(2) $0.007_{\pm 0.001}$
Elu	(1) $0.006_{\pm 0.000}$	(5) $0.012_{\pm 0.000}$•	(3) $0.007_{\pm 0.001}$•	(4) $0.011_{\pm 0.001}$•	(6) $0.019_{\pm 0.001}$•	(1) $0.006_{\pm 0.000}$
Heat	(1) $0.013_{\pm 0.000}$	(5) $0.017_{\pm 0.000}$•	(1) $0.013_{\pm 0.000}$•	(4) $0.014_{\pm 0.001}$•	(6) $0.018_{\pm 0.001}$•	(1) $0.013_{\pm 0.000}$
Spo	(1) $0.025_{\pm 0.001}$	(5) $0.030_{\pm 0.001}$•	(3) $0.027_{\pm 0.001}$•	(2) $0.026_{\pm 0.001}$•	(6) $0.034_{\pm 0.002}$•	(3) $0.027_{\pm 0.001}$•
Spo5	(1) $0.030_{\pm 0.001}$	(6) $0.054_{\pm 0.002}$•	(4) $0.034_{\pm 0.001}$•	(3) $0.033_{\pm 0.001}$•	(4) $0.034_{\pm 0.001}$•	(1) $0.030_{\pm 0.001}$◦
Spoem	(1) $0.026_{\pm 0.001}$	(6) $0.057_{\pm 0.003}$•	(1) $0.026_{\pm 0.001}$•	(1) $0.026_{\pm 0.002}$•	(5) $0.029_{\pm 0.005}$•	(4) $0.027_{\pm 0.001}$•
			Cosine (↑)			
Painting	(1) $0.718_{\pm 0.009}$	(4) $0.648_{\pm 0.013}$•	(6) $0.281_{\pm 0.076}$•	(1) $0.718_{\pm 0.010}$	(5) $0.633_{\pm 0.025}$•	(3) $0.705_{\pm 0.010}$•
Emotion6	(1) $0.707_{\pm 0.006}$	(4) $0.668_{\pm 0.007}$•	(4) $0.668_{\pm 0.009}$•	(2) $0.701_{\pm 0.005}$•	(3) $0.693_{\pm 0.013}$•	(6) $0.665_{\pm 0.004}$•
Movie	(1) $0.926_{\pm 0.001}$	(5) $0.892_{\pm 0.003}$•	(3) $0.902_{\pm 0.003}$•	(2) $0.925_{\pm 0.002}$•	(4) $0.901_{\pm 0.003}$•	(6) $0.871_{\pm 0.002}$•
Music	(1) $0.915_{\pm 0.005}$	(4) $0.897_{\pm 0.001}$•	(6) $0.832_{\pm 0.008}$•	(2) $0.903_{\pm 0.005}$•	(5) $0.877_{\pm 0.041}$•	(3) $0.902_{\pm 0.005}$•
BU-3DFE	(2) $0.929_{\pm 0.001}$	(5) $0.905_{\pm 0.002}$•	(4) $0.910_{\pm 0.002}$•	(1) $0.930_{\pm 0.002}$	(6) $0.902_{\pm 0.002}$•	(3) $0.923_{\pm 0.002}$•
JAFFE	(1) $0.948_{\pm 0.006}$	(5) $0.909_{\pm 0.008}$•	(4) $0.920_{\pm 0.006}$•	(2) $0.940_{\pm 0.006}$•	(6) $0.816_{\pm 0.047}$•	(3) $0.933_{\pm 0.005}$•
Alpha	(1) $0.994_{\pm 0.000}$	(4) $0.990_{\pm 0.000}$•	(1) $0.994_{\pm 0.000}$	(4) $0.990_{\pm 0.001}$•	(6) $0.979_{\pm 0.001}$•	(1) $0.994_{\pm 0.001}$•
Cdc	(1) $0.993_{\pm 0.000}$	(5) $0.988_{\pm 0.000}$•	(1) $0.993_{\pm 0.001}$•	(4) $0.990_{\pm 0.000}$•	(6) $0.981_{\pm 0.001}$•	(1) $0.993_{\pm 0.000}$
Cold	(1) $0.988_{\pm 0.000}$	(5) $0.984_{\pm 0.000}$•	(1) $0.988_{\pm 0.000}$	(4) $0.985_{\pm 0.001}$•	(6) $0.973_{\pm 0.002}$•	(1) $0.988_{\pm 0.001}$•
Diau	(1) $0.987_{\pm 0.000}$	(3) $0.982_{\pm 0.000}$•	(4) $0.980_{\pm 0.001}$•	(5) $0.979_{\pm 0.001}$•	(6) $0.939_{\pm 0.002}$•	(2) $0.986_{\pm 0.000}$•
Dtt	(1) $0.994_{\pm 0.000}$	(5) $0.990_{\pm 0.000}$•	(1) $0.994_{\pm 0.000}$	(4) $0.992_{\pm 0.001}$•	(6) $0.987_{\pm 0.002}$•	(1) $0.994_{\pm 0.000}$
Elu	(1) $0.994_{\pm 0.000}$	(5) $0.988_{\pm 0.000}$•	(1) $0.994_{\pm 0.000}$•	(4) $0.990_{\pm 0.000}$•	(6) $0.982_{\pm 0.001}$•	(1) $0.994_{\pm 0.000}$
Heat	(1) $0.988_{\pm 0.000}$	(5) $0.983_{\pm 0.000}$•	(1) $0.988_{\pm 0.000}$	(4) $0.986_{\pm 0.001}$•	(6) $0.982_{\pm 0.001}$•	(1) $0.988_{\pm 0.000}$
Spo	(1) $0.977_{\pm 0.001}$	(5) $0.972_{\pm 0.001}$•	(2) $0.975_{\pm 0.001}$•	(2) $0.975_{\pm 0.001}$•	(6) $0.968_{\pm 0.002}$•	(2) $0.975_{\pm 0.001}$•
Spo5	(2) $0.973_{\pm 0.001}$	(6) $0.956_{\pm 0.001}$•	(4) $0.970_{\pm 0.001}$•	(3) $0.971_{\pm 0.001}$•	(4) $0.970_{\pm 0.001}$•	(1) $0.974_{\pm 0.000}$◦
Spoem	(1) $0.979_{\pm 0.001}$	(6) $0.953_{\pm 0.002}$•	(2) $0.978_{\pm 0.001}$•	(2) $0.978_{\pm 0.001}$•	(5) $0.976_{\pm 0.004}$•	(4) $0.977_{\pm 0.001}$•

Table 1: Prediction performance measured by KL and Cosine on logical label enhancement.

model (SABFGS (Geng, 2016) is utilized in this paper). Then, we evaluate the prediction performance of SABFGS on the testing set. Finally, we repeat the above process ten times and report the mean and standard deviation.

Comparison Algorithms. We choose two state-of-the-art LE algorithms: CWLD (Fan et al., 2024) and VIB-ILE (Song et al., 2024), and three baseline LE algorithms: GLLE, KMLE, and FCMLE (Xu et al., 2018). The hyperparameter configurations adhere to the respective papers. For CWLD, α is selected among $\{0.1, 0.2, \ldots, 1.0\}$. For VIB-ILE, β is selected among $\{10^{-3}, 10^{-2}, \ldots, 10^{0}\}$, γ is selected among $\{2, 20\}$, and λ is set to 0.1. For GLLE, λ is selected among $\{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$. For KMLE, δ is selected among $\{1, 2, 3, 4, 5\}$. For FCMLE, β is selected among $\{1, 1.1, 1.2, \ldots, 3\}$. For our method, λ and β are selected among $\{10^{-1}, 10^{0}, \ldots, 10^{5}\}$, α is selected among $\{10^{-2}, 10^{-1}, \ldots, 10^{4}\}$, the probability distribution $\tilde{p}(\boldsymbol{y}|\boldsymbol{z})$ is modeled by Bernoulli distribution, the parameters of PROM are set as $a = 16, b = 8, t = 3$.

4.2.2 Results and Discussions

The prediction performance is shown in Table 1. The rankings of each method are provided in the parentheses before the mean value of performance. Additionally, we use a pairwise two-tailed t-test to perform a more thorough comparative analysis. We denote $\circ/\bullet$ as the cases that our proposed

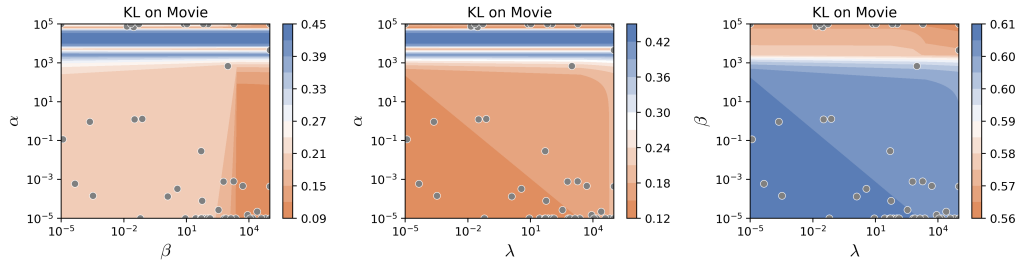


Figure 5: KL-based prediction performance on different values of hyperparameters α, β, and λ.

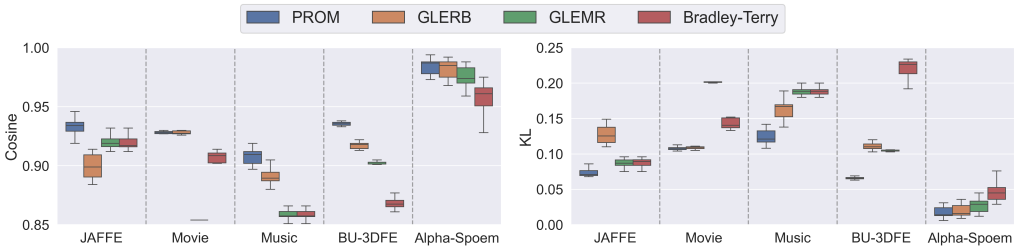


Figure 6: Prediction performance measured by KL and Cosine on label ranking enhancement.

LE-PROM is significantly inferior/superior to the corresponding comparison algorithms. If there is no $\circ$ nor $\bullet$, it indicates that there is no significant performance difference between LE-PROM and the corresponding algorithm. Overall, our algorithm achieves competitive prediction performance. Our algorithm achieves an average ranking of 1.09 and wins 138 times, ties 19 times, and loses 3 times, out of 160 statistical significance comparisons.

4.2.3 Hyperparameter Analysis

In this subsection, our objective is to demonstrate the effectiveness of each component by varying the values of hyperparameters, i.e. the weights of different components, in LE-PROM. In Figure 5, we show the prediction performance measured by KL on "Movie" dataset with varying values of hyperparameters α, β, and λ. The experimental results show that both α (the weight of logical label information) and β (the weight of label ranking information) have significant impact on the performance of our model. It can be found that an increase in α or β usually leads to a performance improvement. However, their values cannot be arbitrarily increased, and when these two hyperparameters reach extremely high values, the performance of the model also decreases. This observation implies that features, logical labels, and label rankings all have an impact on prediction performance, and a proper assignment of weights to these loss terms is critical to model performance. The last two figures in Figure 5 show that λ affects the performance to some extent. However, regardless of λ, optimal performance can typically be achieved by adjusting the α and β.

4.3 Experiments on Label Ranking Enhancement

In this subsection, our objective is to experimentally answer the question of *whether our proposed PROM is superior to current ranking models for the label ranking enhancement task*. The experimental procedure is analogous to Section 4.2.1, with the sole distinction lying in the input label data: one is the logical label, while the other is the label ranking. The training set with label distributions can be directly transformed into the one with label rankings. We choose three comparison algorithms, GLERB (Lu et al., 2023b) and GLEMR (Lu et al., 2023a) are the two state-of-the-art label-enhancement oriented ranking models and Bradley-Terry is a classic ranking model. Since comparison algorithms cannot be directly applied for tie-allowed label rankings, we preprocess the label rankings to accommodate the comparison algorithms. Due to the page limit, we put the details of label preprocessing in Appendix C. From the results in Figure 6, it can be seen that LE-PROM outperforms the comparison algorithms in most cases.

5 Conclusion and Limitation Discussion

Conclusion. In this paper, we qualitatively and quantitatively study the probabilistic relationship between label distributions and label rankings. We propose assumptions to characterize the probabilistic monotonicity and orderliness of label rankings. Subsequently, we derive a pairwise ranking model PROM that theoretically preserves the probabilistic monotonicity and orderliness. Besides, we propose a generative label enhancement algorithm based on the PROM model to directly learn an LDL mapping on the training instances with multi-label data. The experimental results on extensive real-world datasets demonstrate the superiority of our proposed method.

Limitation. Despite the contributions of this paper, it is imperative to acknowledge that the feasible region of PROM exhibits considerable complexity. This complexity may potentially impede the hyperparameter selection. Besides, if users attempt to adaptively learn the parameters a, b, t, the feasible region may lead to a difficult optimization problem. Hence, future research efforts will be directed towards simplifying the boundaries of the feasible region, with the aim of enhancing its compatibility with more established optimization methods.

6 Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (62176123, 62476130, 62576166), the Natural Science Foundation of Jiangsu Province (BK20242045), the Innovation and Technology Fund (GHP/079/22SZ, ITS/034/23FP), and the UGC/GRF (No. 15211024, 15215421).

References

Jean Daunizeau. Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables. *arXiv*, 2017.

Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of KNN classifiers trained using soft labels. In *Artificial Neural Networks in Pattern Recognition*, pages 67–80, 2006.

Jun Fan, Heng-Ru Zhang, and Fan Min. Learning cluster-wise label distribution for label enhancement. *International Journal of Machine Learning and Cybernetics*, 2024.

Binbin Gao, Hongyu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *International Joint Conference on Artificial Intelligence*, pages 712–718, 2018.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28 (7):1734–1748, 2016.

Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:2401–2412, 2013.

John Guiver and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. In *International Conference on Machine Learning*, page 377–384, 2009.

Peng Hou, Xin Geng, and Min-Ling Zhang. Multi-label manifold learning. In *AAAI Conference on Artificial Intelligence*, pages 1680–1686, 2016.

David R Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32 (1):384–406, 2004.

Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9841–9850, 2019.

Xiuyi Jia, Yunan Lu, and Fangwen Zhang. Label enhancement by maintaining positive and negative label relation. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1708–1720, 2023a.

Xiuyi Jia, Xiaoxia Shen, Weiwei Li, Yunan Lu, and Jihua Zhu. Label distribution learning by maintaining label ranking relation. *IEEE Transactions on Knowledge and Data Engineering*, 35 (2):1695–1707, 2023b.

Xiuyi Jia, Tian Qin, Yunan Lu, and Weiwei Li. Adaptive weighted ranking-oriented label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11302–11316, 2024.

Xiufeng Jiang, Zhang Yi, and Jian Cheng Lv. Fuzzy SVM with a new fuzzy membership function. *Neural Computing & Applications*, 15(3):268–276, 2006.

Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36: 1425–1437, 2023.

Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Inaccurate label distribution learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10237–10249, 2024.

Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Progressive label enhancement. *Pattern Recognition*, 160:111172, 2025.

Harin Lee, Frank Hoeger, Marc Schoenwiesner, Minsu Park, and Nori Jacoby. Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. In *International Society for Music Information Retrieval Conference*, pages 366–373, 2021.

Xinyuan Liu, Jihua Zhu, Qinghai Zheng, Zhongyu Li, Ruixin Liu, and Jun Wang. Bidirectional loss function for label enhancement and distribution learning. *Knowledge-Based System*, 213:106690, 2021.

Yunan Lu and Xiuyi Jia. Predicting label distribution from multi-label ranking. In *Advances in Neural Information Processing Systems*, pages 36931–36943, 2022.

Yunan Lu and Xiuyi Jia. Predicting label distribution from ternary labels. In *Advances in Neural Information Processing Systems*, pages 70431–70452, 2024.

Yunan Lu, Liang He, Fan Min, Weiwei Li, and Xiuyi Jia. Generative label enhancement with Gaussian mixture and partial ranking. In *AAAI Conference on Artificial Intelligence*, pages 8975–8983, 2023a.

Yunan Lu, Weiwei Li, Huaxiong Li, and Xiuyi Jia. Ranking-preserved generative label enhancement. *Machine Learning*, 112:4693–4721, 2023b.

Yunan Lu, Weiwei Li, Huaxiong Li, and Xiuyi Jia. Predicting label distribution from tie-allowed multi-label ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15364–15379, 2023c.

Yunan Lu, Weiwei Li, Dun Liu, Huaxiong Li, and Xiuyi Jia. Adaptive-grained label distribution learning. In *AAAI Conference on Artificial Intelligence*, pages 19161–19169, 2025.

Jianqiao Luo, Yihan Wang, Yang Ou, Biao He, and Bailin Li. Neighbor-based label distribution learning to model label ambiguity for aerial scene classification. *Remote Sensing*, 13(4):755, 2021.

Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with Gabor wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.

Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, page 83–92, 2010.

Kuan Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015.

Anning Song, Chao Tan, Jiaxi Zhang, and Zilong Xu. Imbalanced label enhancement based on variational information bottleneck. In *International Joint Conference on Neural Networks*, pages 1–8, 2024.

Haoyu Tang, Jihua Zhu, Qinghai Zheng, Jun Wang, Shanmin Pang, and Zhongyu Li. Label enhancement with sample correlations via low-rank representation. In *AAAI Conference on Artificial Intelligence*, pages 5932–5939, 2020.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2006.

Jing Wang and Xin Geng. Label distribution learning machine. In *International Conference on Machine Learning*, pages 10749–10759, 2021.

Ke Wang, Ning Xu, Miaogen Ling, and Xin Geng. Fast label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1502–1514, 2023.

Tao Wen, Weiwei Li, Lei Chen, and Xiuyi Jia. Semi-supervised label enhancement via structured semantic extraction. *International Journal of Machine Learning and Cybernetics*, 13:1131–1144, 2021.

Ning Xu, An Tao, and Xin Geng. Label enhancement for label distribution learning. In *International Joint Conference on Artificial Intelligence*, pages 1632–1643, 2018.

Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *AAAI Conference on Artificial Intelligence*, pages 5557–5564, 2019.

Ning Xu, Jun Shu, Yun-Peng Liu, and Xin Geng. Variational label enhancement. In *International Conference on Machine Learning*, volume 119, pages 10597–10606, 2020.

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

Min-Ling Zhang, Qian-Wen Zhang, Jun-Peng Fang, Yukun Li, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33:2057–2070, 2021.

Qian-Wen Zhang, Yun Zhong, and Min-Ling Zhang. Feature-induced labeling information enrichment for multi-label learning. In *AAAI Conference on Artificial Intelligence*, pages 4446–4453, 2018.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract accurately reflects the three contributions of the paper, which are further explained in paragraph 3 of the introduction. That is, monotonicity and orderliness assumptions are proposed for the probabilities of different ranking relationships, the quality function of PROM is derived, and a generative LE algorithm based on PROM is proposed.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: In Section 3.2.2, the assumptions underlying the theoretical results are explained in Assumptions 3.1, 3.2, 3.3, and 3.4. The details of the proofs of Theorems 3.5, 3.6, 3.7, and 3.8 are provided in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: In Sections 4.2.1 and 4.2.2, for the comparison algorithms, the hyperparameter configurations and dataset division methods of each comparison method are fully disclosed. For our method, the parameters of the PROM model are specifically set to $a = 16$, $b = 8$, and $t = 3$.

Guidelines:

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

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

5. **Open access to data and code**

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

   Answer: [No]

   Justification: In terms of data access, the datasets used in this article strictly follow the data usage agreement of the original creator, and this article does not authorize their distribution. In terms of code implementation, PROM is very easy to implement, and functions can be reproduced without relying on additional code files. Nevertheless, the code will be accessible after publication.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: In Section 4.2.1, the dataset partitioning, selection method, and label distribution learning model are described in detail. In Section 4.2.2, the hyperparameter settings are described in detail. In Appendix C.3, the algorithm for reducing the label distribution of training instances to logical labels is described in detail.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Table 1 displays the statistical significance, mean, and standard deviation of the prediction performance.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The difference in computing resource configuration has no significant effect on the experimental results.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this article complies in all respects with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [NA]

Justification: There is no significant social impact from the work done for this paper.

Guidelines:

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

11. **Safeguards**

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

Answer: [NA]

Justification: There is no such risk associated with the work described in this paper.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The original papers of five comparative algorithm models and sixteen datasets are fully cited in this article.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [NA]

    Justification: There are no new assets released in this article.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [NA]

    Justification: There is no research involving crowdsourcing or human subjects in this article.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The research of this paper did not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: In this research core methodology, LLM is not a significant, unique, or non-standard component.

    Guidelines:

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