A. Proofs

To facilitate the proofs of our results, we first present some established results from the prior works.

Lemma A.1. *([68, Theorem 1], [60, Theorem 1]) Suppose that* $\ell : \Delta \times \{-1, +1\} \to \mathbb{R}$ *is a CPE loss and its partial losses* ℓ_{-1}, ℓ_{+1} *are differentiable.* ℓ *is proper if and only if for all* $\widehat{\eta} \in (0, 1)$

$$\frac{-\ell'_{+1}(\widehat{\eta})}{1-\widehat{\eta}} = \frac{\ell'_{-1}(\widehat{\eta})}{\widehat{\eta}} = w(\widehat{\eta}) \tag{A.1}$$

for a weight function $w : (0, 1) \to \mathbb{R}_+$ *such that* $\int_\epsilon^{1-\epsilon} w(c)dc < \infty$ *for all* $\epsilon > 0$.

Lemma A.2. *([4]) A proper loss* ℓ *is strictly proper if and only if its weight function* w *of* (A.1) *satisfies that* $w(c) > 0$ *almost everywhere on* $(0, 1)$.

Lemma A.3. *([68, Corollary 1]) Suppose that* $\ell_{-1,\Psi}, \ell_{+1,\Psi}$ *are the partial losses of a composite CPE loss and differentiable. The CPE loss is proper if and only if the link* Ψ *satisfies*

$$[\Psi]^{-1}(v) = \frac{\ell'_{-1,\Psi}(v)}{\ell'_{-1,\Psi}(v) - \ell'_{+1,\Psi}(v)}, \forall v \in \mathbb{R}. \tag{A.2}$$

A.1. Proof of Theorem 1

Proof. The ASY loss of (8),

$$\begin{cases} \ell_{+1}^{ASY}(\widehat{\eta}) = -(1-\widehat{\eta})^{\gamma^+}\log(\widehat{\eta}) \\ \ell_{-1}^{ASY}(\widehat{\eta}) = -(\widehat{\eta}-m)_+{}^{\gamma^-}\log(1-(\widehat{\eta}-m)_+) \end{cases} \tag{A.3}$$

has derivatives

$$\ell'_{+1}(\widehat{\eta}) = \frac{(1-\widehat{\eta})^{\gamma^+-1}(\gamma^+\widehat{\eta}\log(\widehat{\eta}) - (1-\widehat{\eta}))}{\widehat{\eta}}, \tag{A.4}$$

$$\ell'_{-1}(\widehat{\eta}) = \frac{(\widehat{\eta}-m)_+^{\gamma^--1}((\widehat{\eta}-m)_+ - \gamma^-(1-(\widehat{\eta}-m)_+)\log(1-(\widehat{\eta}-m)_+))}{1-(\widehat{\eta}-m)_+}. \tag{A.5}$$

It follows that

$$\frac{-\ell'_{+1}(\widehat{\eta})}{1-\widehat{\eta}} = \frac{(1-\widehat{\eta})^{\gamma^+-1}((1-\widehat{\eta}) - \gamma^+\widehat{\eta}\log(\widehat{\eta}))}{\widehat{\eta}(1-\widehat{\eta})}, \tag{A.6}$$

$$\frac{\ell'_{-1}(\widehat{\eta})}{\widehat{\eta}} = \frac{(\widehat{\eta}-m)_+^{\gamma^--1}((\widehat{\eta}-m)_+ - \gamma^-(1-(\widehat{\eta}-m)_+)\log(1-(\widehat{\eta}-m)_+))}{\widehat{\eta}(1-(\widehat{\eta}-m)_+)}. \tag{A.7}$$

When $m \in \mathbb{R}_+$ and $\gamma^+, \gamma^- \in \mathbb{R}_{++}$, the RHS of (A.6) and the RHS of (A.7) cannot be equivalent for all $\widehat{\eta} \in (0, 1)$ and, from Lemma A.1, the ASY loss is not proper and thus not strictly proper.

□

A.2. Proof of Theorem 2

Proof. The function

$$f_\gamma(q) = -q^\gamma \log(1-q), \forall q \in (0, 1) \tag{A.8}$$

is convex on $(0, 1)$ and has first-order derivative

$$f'_\gamma(q) = \frac{q^{\gamma-1}(q - \gamma(1-q)\log(1-q))}{1-q}, \tag{A.9}$$

which is positive and strictly increasing on $q \in (0, 1)$.

Consider the minimization of the conditional risk

$$\min_{\widehat{\eta}\in\Delta} \quad C^{\text{ASY}}(\eta,\widehat{\eta}) = \eta\ell_{+1}(\widehat{\eta}) + (1-\eta)\ell_{-1}(\widehat{\eta}) \tag{A.10}$$

for any $\eta \in (0,1)$. This is a convex optimization problem with a bounded feasible set $\{\widehat{\eta}|\widehat{\eta} \in \Delta\}$. Note that $\frac{\partial C^{\text{ASY}}(\eta,\widehat{\eta})}{\partial\widehat{\eta}}\Big|_{\widehat{\eta}=0+} = -\infty$, and $\frac{\partial C^{\text{ASY}}(\eta,\widehat{\eta})}{\partial\widehat{\eta}}\Big|_{\widehat{\eta}=1-} > 0$ since

$$\frac{\partial C^{\text{ASY}}(\eta,\widehat{\eta})}{\partial\widehat{\eta}}\Bigg|_{\widehat{\eta}=1-} = \begin{cases} +\infty, & m=0, \\ \frac{(1-m)^{\gamma^- -1}((1-m)-\gamma^- m\log(m)}{m} > 0, & 0<m<1. \end{cases} \tag{A.11}$$

Because $\frac{\partial C^{\text{ASY}}(\eta,\widehat{\eta})}{\partial\widehat{\eta}}$ is continuous on $\widehat{\eta} \in (0,1)$, the minimum is attained at the critical point $\widehat{\eta} = \widehat{\eta}^*$ such that $\frac{\partial C^{\text{ASY}}(\eta,\widehat{\eta})}{\partial\widehat{\eta}}\Big|_{\widehat{\eta}=\widehat{\eta}^*} = 0$, *i.e.*

$$\begin{aligned} &- \quad \eta\frac{(1-\widehat{\eta}^*)^{\gamma^+-1}(\gamma^+\widehat{\eta}^*\log(\widehat{\eta}) - (1-\widehat{\eta}^*))}{\widehat{\eta}^*} \\ &- \quad (1-\eta)\frac{(\widehat{\eta}^*-m)_+^{\gamma^- -1}((\widehat{\eta}^*-m)_+ - \gamma^-(1-(\widehat{\eta}^*-m)_+)\log(1-(\widehat{\eta}^*-m)_+))}{(1-(\widehat{\eta}^*-m)_+)} = 0. \end{aligned} \tag{A.12}$$

Using the definitions of $h(z;\gamma)$ in (15),

$$h(z;\gamma) = \frac{z^\gamma - \gamma z^{\gamma-1}(1-z)\log(1-z)}{1-z}, \tag{A.13}$$

the above equation can be rewritten as

$$\eta h(1-\widehat{\eta}^*;\gamma^+) - (1-\eta)h((\widehat{\eta}^*-m)_+;\gamma^-) = 0. \tag{A.14}$$

It follows that

$$\eta = \frac{h((\widehat{\eta}^*-m)_+;\gamma^-)}{h((\widehat{\eta}^*-m)_+;\gamma^-) + h(1-\widehat{\eta}^*;\gamma^+)} = \phi(\widehat{\eta}^*;\gamma^-,\gamma^+,m), \tag{A.15}$$

where we have used the definition of $\phi(z;\gamma^-,\gamma^+,m)$ in (14), *i.e.*

$$\phi(z;\gamma^-,\gamma^+,m) = \frac{h((z-m)_+;\gamma^-)}{h((z-m)_+;\gamma^-) + h(1-z;\gamma^+)}. \tag{A.16}$$

When $m=0$,

$$\phi(\widehat{\eta}^*;\gamma^-,\gamma^+,0) = \frac{1}{1+\frac{h(1-\widehat{\eta}^*;\gamma^+)}{h(\widehat{\eta}^*;\gamma^-)}} = \frac{1}{1+\frac{f'_{\gamma^+}(1-\widehat{\eta}^*)}{f'_{\gamma^-}(\widehat{\eta}^*)}}. \tag{A.17}$$

On $\widehat{\eta}^* \in (0,1)$, $f'_{\gamma^+}(1-\widehat{\eta}^*)$ is positive and strictly decreasing and $f'_{\gamma^-}(\widehat{\eta}^*)$ is positive and strictly increasing, it follows that $\phi(\widehat{\eta}^*;\gamma^-,\gamma^+,0)$ is strictly increasing. Furthermore, $\lim_{\eta\to 0+}\phi(\eta;\gamma^-,\gamma^+,0) = 0$ and $\lim_{\eta\to 1-}\phi(\eta;\gamma^-,\gamma^+,0) = 1$, hence $\phi(\widehat{\eta}^*;\gamma^-,\gamma^+,0)$ is a bijective mapping between $(0,1)$ and $(0,1)$.

When $0<m<1$, we have $\phi(\widehat{\eta}^*;\gamma^-,\gamma^+,m) = 0, \forall\widehat{\eta}^* \in (0,m)$, and thus ϕ is not a bijective map. □

A.3. Proof of Theorem 3

Proof. The first-order derivatives of $\ell'_{-1,\Psi}(v)$ and $\ell'_{+1,\Psi}(v)$ are

$$\begin{cases} \ell'_{+1,\Psi}(v) = -\frac{k^+}{\zeta^+}\sigma(-k^+(v-b^+)), \\ \ell'_{-1,\Psi}(v) = \frac{k^-}{\zeta^-}\sigma(k^-(v-b^-)), \end{cases} \tag{A.18}$$

from which

$$\frac{\ell'_{-1,\Psi}(v)}{\ell'_{-1,\Psi}(v) - \ell'_{+1,\Psi}(v)} = \frac{\frac{k^-}{\zeta^-}\sigma(k^-(v-b^-))}{\frac{k^-}{\zeta^-}\sigma(k^-(v-b^-)) + \frac{k^+}{\zeta^+}\sigma(-k^+(v-b^+))} = \Psi^{-1}(v). \tag{A.19}$$

It follows, by Lemma A.3, that the CPE loss defined by (16) and (17) is proper. Using (A.1) and the chain rule, we have

$$w(\widehat{\eta}) = \frac{\ell'_{-1}(\widehat{\eta})}{\widehat{\eta}} \tag{A.20}$$

$$= \frac{\ell'_{-1,\Psi}(\Psi(\widehat{\eta}))}{\widehat{\eta}}\Psi'(\widehat{\eta}). \tag{A.21}$$

Since $[\Psi]^{-1}(v) = \frac{1}{1+\frac{k^+\zeta^-\sigma(-k^+(v-b^+))}{k^-\zeta^+\sigma(k^-(v-b^-))}}$ is positive and strictly increasing on $v \in (-\infty, +\infty)$, Ψ is strictly increasing and $\Psi'(\widehat{\eta}) > 0, \forall\widehat{\eta} \in (0,1)$. Also, $\ell'_{-1,\Psi}(v) > 0, \forall v \in (-\infty, +\infty)$. It follows that $w(\widehat{\eta}) > 0, \forall\widehat{\eta} \in (0,1)$, and the loss is strictly proper by Lemma A.2. □

B. Experiments

B.1. Calibration Metrics

To quantify the calibration error of a CPE DNN, we discretize the interval $[0,1]$ into M bins $\{I_m = [\frac{m-1}{M}, \frac{m}{M})\}_{m=1}^{M}$. Let o_i be the fraction of positive examples in bin I_i, e_i be the mean of estimated posterior probability $\widehat{\eta}$ for examples in bin I_i, and $N = \sum_i^M |I_i|$ be the total number of examples in the test set. Calibration is usually evaluated with three metrics, Expected Calibration Error (ECE), Average Calibration Error (ACE) and Maximum Calibration Error (MCE), defined as follows

$$\text{ECE} = \sum_{m=1}^{M} \frac{|I_m|}{N}|o_m - e_m|. \tag{B.1}$$

$$\text{ACE} = \sum_{m=1}^{M} \frac{1}{M}|o_m - e_m|. \tag{B.2}$$

$$\text{MCE} = \max_{m\in\{1,\cdots,M\}} |o_m - e_m|. \tag{B.3}$$

As pointed out by prior works [45, 51], ECE is unsuited for problems with extreme class imbalance, since it usually assigns disproportionately high weight to the first few bins and fails to capture the calibration error of the remaining. Hence, we adopt ACE and MCE as calibration metrics for the highly imbalanced multi-label learning problems considered in this work.

B.2. VQA

Table B.1 presents a comparison of the strictly proper losses on the VQA task, confirming the observations above. While the two losses have equivalent accuracy, SPA achieves better calibration than BCE, which is widely used for this task. While the gains are not as large as in Table 3, probably because labels are less imbalanced, they are still significant, e.g. a reduction of ACE from 6.4 to 4.2.

	Accuracy↑	ACE↓	MCE↓
BCE	73.27	6.4	11.4
SPA	73.50	4.2	9.3

Table B.1. Evaluation on VQA v2.0.

B.3. Ablation on Hyperparameters for SPA

Since the hyperparamters k^-, b^-, ζ^- of SPA affect the treatment of negative examples, we conduct an ablation study on the effects of k^-, b^-, ζ^- in Table B.2. Compared to the results in the main paper, it is shown that with different choices of hyperparameters, SPA always achieves better accuracy than the symmetric BCE losses and less calibration error than the asymmetric focal and ASY losses that are not strictly proper.

(k^-, b^-, ζ^-)	mAP@y	mAP@x	ACE	MCE
(2,-1,2)	77.1	88.1	5.3	8.0
(2,+1,2)	77.5	88.5	6.6	10.8
(3,-1,2)	76.7	87.8	8.6	12.5
(3,+1,2)	77.1	88.1	5.2	9.5
(4,-1,2)	75.8	87.1	10.2	17.7
(4,+1,2)	76.9	88.0	7.6	13.3
(2,-1,3)	77.5	88.4	5.6	8.2
(2,+1,3)	77.5	88.5	5.7	10.8
(3,-1,3)	77.0	88.1	5.8	10.8
(3,+1,3)	77.2	88.3	6.7	11.6
(4,-1,3)	76.4	87.3	5.3	8.6
(4,+1,3)	76.8	88.0	6.6	13.3
(2,-1,4)	77.6	88.5	3.4	5.8
(2,+1,4)	77.6	88.6	4.6	7.8
(3,-1,4)	77.3	88.0	5.4	8.3
(3,+1,4)	77.3	88.4	5.1	7.7
(4,-1,4)	76.7	87.6	8.9	16.6
(4,+1,4)	76.9	88.0	4.3	7.2
(2,-1,5)	77.6	88.4	5.3	9.8
(2,+1,5)	77.6	88.7	4.6	7.6
(3,-1,5)	77.3	88.3	6.5	11.3
(3,+1,5)	77.8	88.4	5.3	12.0
(4,-1,5)	76.9	87.8	8.2	13.7
(4,+1,5)	77.1	88.2	4.8	8.9
(2,-1,6)	77.6	88.5	5.8	8.8
(2,+1,6)	77.6	88.6	6.0	9.6
(3,-1,6)	77.4	88.1	6.7	11.8
(3,+1,6)	77.5	88.5	6.3	10.6
(4,-1,6)	77.1	88.0	7.8	13.7
(4,+1,6)	77.3	88.3	5.8	11.5

Table B.2. Ablation of SPA hyperparameters k^-, b^-, ζ^-. The hyperparameters of positive examples are fixed as $k^+ = 1, b^+ = 0, \zeta^+ = 1$. All evaluations are made on MS-COCO and ECA-ResNet50-T.