Rethinking BCE Loss for Multi-Label Image Recognition with Fine-Tuning

Ao Zhou, Zhiwei Jiang*, Zifeng Cheng, Cong Wang, Yafeng Yin, Shufan Yang, Qing Gu
State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
za@smail.nju.edu.cn; {jzw, chengzf, wang.c, yafeng}@nju.edu.cn
sfyang@smail.nju.edu.cn; guq@nju.edu.cn

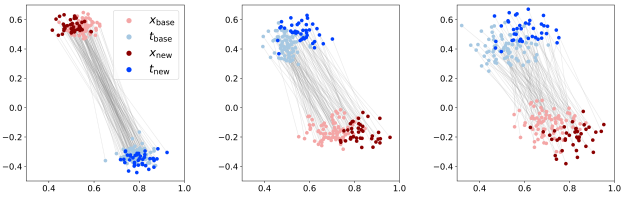
Abstract

Fine-tuning vision–language models such as CLIP has become the mainstream paradigm for multi-label image recognition, and prompt tuning is widely adopted due to its lightweight parameter cost and strong transferability. However, we find that when these methods use Binary Cross-entropy as the supervision loss, the model's confidence structure becomes systematically distorted, leading to pronounced miscalibration. Existing calibration techniques, such as temperature scaling or regularization-based methods, largely fail in multi-label settings because they cannot capture inherent semantic dependencies between classes, nor can they correct the global structural shifts introduced during fine-tuning. To address this issue, we propose Class-wise Covariance Regularization, which aligns the predicted covariance structure of class confidences with the semantic correlations encoded in pretrained text embeddings. This alignment preserves the geometric consistency of the class space throughout fine-tuning, resulting in more stable and interpretable confidence distributions across categories. Experiments on multi-label benchmarks show that CCR significantly reduces calibration errors while maintaining or even improving recognition performance.

1. Introduction

Vision–Language Models (VLMs), such as CLIP [26], have emerged as powerful foundations for Multi-Label Image Recognition (MIR) [19, 46]. In particular, they exhibit strong performance across MIR sub-tasks, including open-vocabulary multi-label recognition [28, 30, 48] and long-tailed multi-label recognition [11, 41]. To further adapt VLMs to specific multi-label downstream tasks, a variety of parameter-efficient fine-tuning techniques have been proposed, such as multi-label prompt tuning [12, 29, 39] as well as prompt-tuning methods that incorporate adapter mechanisms [8, 34].

*Corresponding author



(a) Zero shot CLIP (b) TaI-DPT Ranking (c) TaI-DPT BCE

Figure 1. Paired image (x) and text (t) embeddings from MS-COCO visualized via SVD. *Base* classes are those seen during fine-tuning, whereas *new* classes are unseen. Fine-tuning CLIP induces a larger text–modality gap between classes.

Among these fine-tuning strategies [8, 12, 39], we observe that most methods designed for multi-label data have gradually moved away from the traditional Binary Cross-Entropy (BCE) loss, and instead predominantly adopt Ranking loss [45] as their optimization objective. However, evidence from traditional deep neural network models designed specifically for multi-label learning suggests that binary cross-entropy loss and its variants are often comparatively more effective objectives [2, 36, 47]. This observation naturally leads to a key question: **What makes BCE unsuitable for multi-label CLIP fine-tuning, and why does this limitation motivate the widespread adoption of Ranking-based objectives?** Prior studies attribute this phenomenon to the semantic gap between the vision and language modalities. As noted in [12], "Although BCE loss and its variants (e.g., Asymmetric loss [27]) perform strongly in conventional deep neural network models and are typically paired with a sigmoid function to map outputs into probabilities, directly optimizing probabilities in vision–language models such as CLIP exacerbates the distribution mismatch between training texts and testing images. Consequently, Ranking loss is regarded as a more flexible and suitable supervisory signal under modality discrepancies." To further examine this claim, we follow the *modality gap* analysis framework of [16], feeding paired image–text samples from MS-COCO into both zero-shot CLIP and its fine-tuning counterparts, and projecting their

embeddings into a 2D space using singular value decomposition (SVD). As shown in Figure 1, images and texts exhibit a clear "arm length" separation in the shared representation space, illustrating a typical modality misalignment. In zero-shot CLIP, embeddings within each modality remain relatively compact; however, as illustrated in Figures 1(b) and 1(c), fine-tuning significantly alters the spatial distribution of class text embeddings, with BCE-based tuning inducing a far more pronounced structural shift than ranking-based tuning.

Our work investigates the structural drift in the text embedding space caused by BCE loss during multi-label fine-tuning of CLIP. Our study shows that this structural distortion is tightly linked to the model's confidence behavior: BCE optimization introduces a systematic confidence bias, causing the model to become *under-confident* on base classes while becoming markedly *over-confident* on novel ones. Further examining the frequency structure within base classes, we find that head classes are more prone to under-confidence, whereas tail classes tend to exhibit over-confidence. This imbalance in predicted confidence fundamentally undermines the generalization ability of the fine-tuned model in MIR tasks. In addition, we find that this drift manifests differently when using a Ranking loss versus a BCE loss. Therefore, effective calibration of confidence predictions in multi-label settings becomes particularly crucial. Although numerous temperature scaling–based [7, 44] and regularization-based calibration [37, 38, 43] approaches have shown strong effectiveness in single-label tasks (including binary classification [25], multi-class classification [10], and open-vocabulary recognition [38]). However, our experiments reveal that these methods often fail to correct the global confidence bias introduced during fine-tuning: they cannot simultaneously calibrate both head and tail labels, nor can they balance the calibration trade-off between base and novel classes under open-vocabulary settings. To address this issue, we propose **C**lass-wise **C**ovariance **R**egularization (CCR). CCR aligns the predicted class-wise covariance structure with semantic correlations derived from the text embedding space, thereby preserving semantic topology stability during model optimization. This regularization stabilizes confidence across categories and significantly improves both calibration and overall performance. Moreover, CCR provides a unified and interpretable structural perspective for confidence modeling in multi-label vision–language learning.Our main contributions are summarized as follows:

- We demonstrate through systematic analysis that the independence assumption of BCE overlooks semantic correlations among classes, thereby disrupting the structural consistency of the text embedding space.
- We propose Class-wise Covariance Regularization, a structural calibration method that preserves class-space geometry by aligning predicted confidence covariance with textual semantic correlations.
- We demonstrate the effectiveness of CCR on multiple multi-label benchmarks, showing that it mitigates both over- and under-confidence, achieving more balanced calibration and stronger generalization.

2. Related Work

2.1. CLIP

CLIP is a vision-language model that aligns images and texts in a shared embedding space [26]. In Multi-label Image Recognition (MIR) task, CLIP encodes an image x and a class-related text t_c through its image and text encoders f_{img} and f_{text} [13, 20, 23]. The logit for class c is computed as:

$$z_c = \tau \cdot \text{sim}(f_{\text{img}}(x), f_{\text{text}}(t_c)), \tag{1}$$

where t_c is a prompt such as "a photo of a class {c}", and τ is a scaling factor. For prediction, each class is treated independently by applying a sigmoid activation:

$$p(c|x) = \sigma(z_c) = \frac{1}{1 + \exp(-z_c)} \tag{2}$$

where $p(c|x)$ denotes the probability confidence that image x is associated with class c. Final predictions are obtained by applying per-class thresholds [1] or selecting top-k confident class labels [3, 42].

2.2. Prompt Tuning

Prompt tuning is a parameter-efficient adaptation strategy that updates only a small set of learnable prompts, allowing downstream tasks to benefit from the rich representations of large vision–language models without full-model fine-tuning [9, 15, 35]. For example, in CLIP-based prompt tuning, CoOp [50] learns soft prompts by minimizing the classification loss, while CoCoOp [49] further conditions the prompts on image features to improve generalization to unseen classes. CLIP-based prompt tuning methods have achieved excellent performance in MIR tasks. For example, TaI-DPT [12] extracts both coarse-grained and fine-grained embeddings by treating textual descriptions as visual inputs within a prompt tuning framework. T2I-PAL [8] introduces a dual-path framework which couples prompt tuning and a shared adapter, enabling the model to jointly leverage CLIP's pre-trained knowledge and learn task-specific features, with a tunable fusion mechanism to bridge the modality gap.

2.3. CLIP Calibration

A well-calibrated model aligns its predicted confidence with the actual accuracy. The Expected Calibration Error (ECE) [10] quantifies calibration performance

by partitioning N predictions into K confidence bins $\{B_1, B_2, \ldots, B_K\}$ and computing:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|, \tag{3}$$

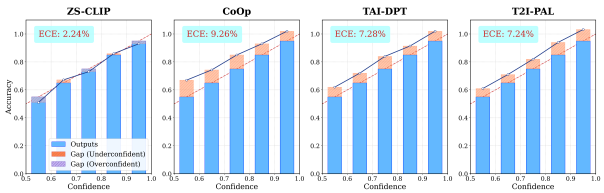
where $\mathrm{acc}(B_k)$ and $\mathrm{conf}(B_k)$ denote the average accuracy and confidence within bin B_k, respectively. While CLIP demonstrates strong calibration properties under zero-shot inference, existing work shows that fine-tuning can introduce substantial calibration shifts [10, 22]. To mitigate these issues, Distance-Aware Calibration (DAC) [38] adjusts logits using text-dependent bias, and Dynamic Outlier Regularization (DOR) [37] addresses the overconfidence and underconfidence behaviors introduced by prompt-based fine-tuning. Further details are provided in Appendix A. However, existing research on CLIP calibration remains largely confined to single-label classification tasks [31, 32, 44], while its calibration in multi-label settings has yet to be systematically explored. The few methods specifically designed for MIR, such as DCLR [4], are primarily proposed to mitigate over-confidence. As CLIP becomes the mainstream framework for MIR tasks [18, 21] and open-vocabulary multi-label recognition emerges as a more prevalent task, the calibration behavior of CLIP in such scenarios remains poorly understood.

3. Motivation and Analysis

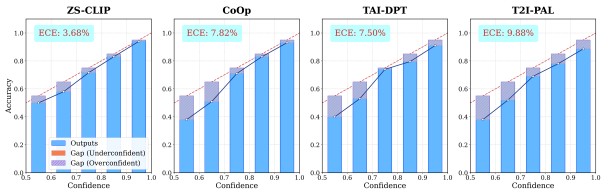
Why does CLIP fine-tuned with BCE loss underperform in multi-label scenarios? We hypothesize that the fine-tuning process disrupts CLIP's original semantic structure, causing systematic miscalibration in confidence scores. To investigate this issue, we conduct a comprehensive empirical study using the OpenAI pre-trained ViT-B/16 model [26] on the MS-COCO dataset. Our analysis examines confidence distributions from a class-level perspective, focusing on how confidence varies across head classes, tail classes, and new classes. Our experiments include the zero-shot CLIP baseline and three representative fine-tuning paradigms: the classical prompt-tuning approach CoOp [50], the multi-label–specific method TaI-DPT [12], and the adapter-based approach T2I-PAL [8], which integrates prompt tuning with lightweight visual adaptation.

Empirical study on ECE To evaluate the model's calibration error on each class, we first decompose each multi-label sample (x, Y) (where Y denotes the set of relevant classes) into $|Y|$ independent pairs $\{(x, y_c) : y_c \in Y\}$. For each class c, we compute its ECE based on all associated sample pairs, following Eq. (3).

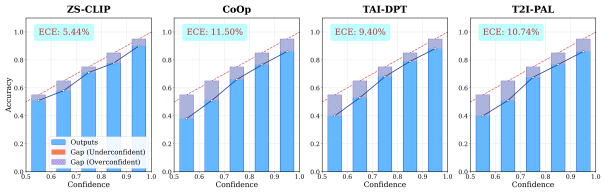
Figures 2 present the fine-tuning results based on BCE loss. Compared with zero-shot CLIP (ZS-CLIP), BCE fine-tuning induces a systematic shift in confidence: base classes



(a) Head



(b) Tail



(c) New

Figure 2. Expected Calibration Error (lower is better). Miscalibration is depicted in orange for underconfidence and purple for overconfidence.

become under-confident, while new classes become over-confident. We analyze the reasons as follows: given a sample x, the BCE loss is

$$\mathcal{L} = -\frac{1}{C} \sum_{c=1}^{C} \Big[y_c \log p(c|x) + \big(1 - y_c\big) \log \big(1 - p(c|x)\big) \Big] \tag{4}$$

and its gradients are

$$\frac{\partial \mathcal{L}}{\partial z_c(x)} = p(c|x) - y_c \rightarrow \nabla_{t_c} \mathcal{L} \propto \big(p(c|x) - y_c\big)\, f_{\text{img}}(x) \tag{5}$$

when $y_c = 0$, $p(c|x) > 0$ forces $z_c(x)$ down, and when $y_c = 1$, $p(c|x) - 1 < 0$ pulls it up. For base classes, the dominance of irrelevant samples for each label during gradient updates creates a global "cooling effect," leading to generally conservative predictions. Specifically, head classes experience consistent gradient corrections of $p(c|x) - 1 < 0$, which progressively suppress their initially high zero-shot confidence levels, resulting in significant under-confidence. In contrast, tail classes, despite having sparse positive samples, receive strong gradient signals when updated. Through multi-label feature sharing mechanisms, their representations become broadly aligned with high-frequency visual components, ultimately producing over-confident predictions on many irrelevant samples. Mathematically, in the gradient expectation $\mathbb{E}[\partial \mathcal{L}/\partial z_c] = \mathbb{E}[p(c|x)|y_c = 0] - \pi_c$, where $\pi_c = \Pr(y_c = 1)$ denotes

the class prior of class c. The small π_c values maintain persistently negative gradients, while sparse positive updates induce excessively generalized matching patterns. For new classes, the complete absence of supervision signals ($y_c \equiv 0$) during fine-tuning means their text prompt optimization is driven solely by negative sample gradients.

Unlike probabilistic objective functions, ranking-based fine-tuning does not explicitly define confidence. The score z_c produced by the model for each class reflects only relative ordering preferences rather than probabilistic outputs suitable for calibration evaluation. Consequently, standard ECE metrics cannot be directly computed for models fine-tuned with ranking loss. To enable a comparable analysis, following Appendix B, we construct a form of "pseudo-confidence" by applying a scaling transformation to the output scores [1]. Using this aligned measure, we observe that ranking-loss fine-tuning leads to a consistent increase in such pseudo-confidence across head, tail, and new classes. Therefore, additional analysis from alternative perspectives is required to obtain a more complete understanding.

Embedding Divergence Given the above observations, we further investigate how different loss functions in prompt tuning lead to confidence bias. Since the visual features remain fixed during prompt tuning, we hypothesize that the textual features play a critical role in shaping confidence calibration. We define a class-aware text Embedding Divergence *(ED)* metric to quantify the structural dispersion of class text embeddings after fine-tuning. Let $\mathcal{T} = \{f_{\text{text}}(t_c)\}_{c=1}^{C}$ be the set of class text embeddings. The *ED* score for class $f_{\text{text}}(t_i)$ is given by:

$$ED(t_i) = \frac{1}{k} \sum_{f_{\text{text}}(t_j) \in \mathcal{N}_k} \text{dist}\langle f_{\text{text}}(t_i), f_{\text{text}}(t_j) \rangle \quad (6)$$

where $\mathcal{N}_k$ denotes the set of k most semantically similar class embedding to $f_{\text{text}}(t_i)$, and $\text{dist}\langle \cdot, \cdot \rangle$ measures the Euclidean distance between two embeddings. A higher *ED* value indicates greater local dispersion and weaker semantic compactness. In addition, to assess the impact of the fine-tuning process on the semantic topology, we define a Neighborhood rank Preservation *(NP)* metric. For a class c, let $\mathcal{R}^{\text{zs}}(c) = [c_1^{\text{zs}}, c_2^{\text{zs}}, \ldots, c_{C-1}^{\text{zs}}]$ denote the sequence of all other classes ranked in descending order of their distance to c in the zero-shot prompt embedding. Similarly, let $\mathcal{R}^{\text{ft}}(c)$ denote the neighborhood rank sequence after fine-tuning. The top k NP score for class c is defined as:

$$NP@K(c) = \frac{\left|\mathcal{R}^{\text{zs}}_{1:K}(c) \cap \mathcal{R}^{\text{ft}}_{1:K}(c)\right|}{k} \quad (7)$$

where $\mathcal{R}_{1:K}(c)$ represents the top k nearest neighbors in the ranking sequence $\mathcal{R}(c)$, and $|\cdot|$ denotes the cardinality of the set. The global semantic topology preservation is obtained by averaging over all classes:

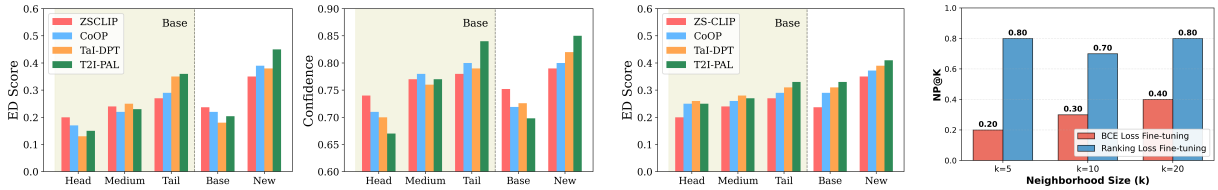
$$NP@K = \frac{1}{C} \sum_{c=1}^{C} NP@K(c) \quad (8)$$

Figure 3 [a] illustrates the relationship between the *Embedding Divergence (ED)* scores and model confidence across base and novel classes under BCE fine-tuning. A clear positive correlation is observed: models with higher embedding divergence tend to exhibit stronger prediction confidence. Under BCE fine-tuning, the text embedding space of base classes becomes more compact (lower *ED*), corresponding to conservative and under-confident predictions. In contrast, the embedding space of novel classes tends to be more dispersed, which aligns with the model's over-confident behavior on those categories. As shown in Figure 3 [b], ranking-based fine-tuning significantly expands the embedding space (higher *ED*), making the semantic distribution of all classes more dispersed and consistent with the previously observed systematic over-confidence (i.e., uniform logit elevation at the sample level). These findings indicate that the structural divergence of the text embedding space reflects the model's confidence distribution: greater semantic dispersion in the embedding manifold is generally associated with stronger over-confidence, whereas a more compact structure corresponds to conservative predictions. Notably, despite the global expansion introduced by ranking-based fine-tuning, its semantic neighborhood structure remains largely preserved. As shown in Figure 3 [c], the higher *NP@K* scores suggest that semantically similar classes in the zero-shot CLIP space remain close after fine-tuning. This demonstrates that ranking-based fine-tuning, while enlarging the overall semantic margins between classes, still maintains a consistent global semantic topology. Such structural retention helps stabilize inter-class relationships and explains why Ranking loss based fine-tuning generally achieves better overall performance than BCE loss. Therefore, we require a method that constrains the relative relationships among class embeddings from a global, class-level perspective.

4. Class-wise Covariance Regularization

During fine-tuning, our goal is to recover reliable inter-class semantic relationships, since these structural cues are often distorted by the BCE loss. A natural way to probe these relationships is through the model's confidence structure. However, positive samples are extremely sparse in multi-label datasets, making co-occurrence–based estimation unreliable and biased toward head classes. This motivates us to shift perspective. Instead of relying on scarce positive activations (y=1), we estimate inter-class dependencies by examining samples in which two classes are simultaneously

[1] This approach is quite unrigorous; our aim is simply to observe the changes in relative trends before and after fine-tuning.



(a) BCE Loss Embedding Divergence and Confidence

(b) Ranking Loss Embedding Divergence

(c) Neighbor rank Preservation

Figure 3. Evaluation of *ED* and *NP@K* for zero-shot CLIP and multiple prompt-tuning fine-tuned models on the MS-COCO dataset.

inactive (y=0). This yields a crucial advantage: for any class pair, the amount of negative evidence is orders of magnitude larger than positive evidence, providing a dense, stable statistical signal. Conceptually, this reframes the estimation problem. Rather than asking " How confidently does the model believe the image contains class c? " we instead ask " How certain is the model that the image does not contain class c? " This shared "semantic background" region of the label space produces robust co-variation patterns that reflect how the model internally perceives classes. Based on this insight, we construct a symmetric covariance matrix $\mathbf{C}_{\text{pred}}$, which captures how the model jointly suppresses or co-activates classes across a mini-batch. Despite being computed locally at the batch level, this covariance reliably reflects global structural tendencies because inactive predictions dominate in multi-label settings. Since the predicted covariance $\mathbf{C}_{\text{pred}}$ measures raw confidence co-variation, its scale is not directly comparable to the text-based semantic similarity $\mathbf{\Sigma}_{\text{text}}(i,j) = \text{sim}\langle \mathbf{t}_i, \mathbf{t}_j \rangle$, which is computed from the zero-shot CLIP text embeddings. To ensure scale consistency, we normalize $\mathbf{C}_{\text{pred}}$ into a correlation matrix:

$$\tilde{\mathbf{C}}_{\text{pred}}(i,j) = \frac{\mathbf{C}_{\text{pred}}(i,j)}{\sqrt{\mathbf{C}_{\text{pred}}(i,i)}\sqrt{\mathbf{C}_{\text{pred}}(j,j)}} \tag{9}$$

This normalization removes magnitude bias and preserves only the relational structure among classes, making it directly comparable to the semantic similarity matrix. Finally, we align the normalized predicted correlation $\tilde{\mathbf{C}}_{\text{pred}}$ with the text-derived semantic correlation $\mathbf{\Sigma}_{\text{text}}$ through the following regularization loss:

$$\mathcal{L}_{\text{cov}} = \left\| \tilde{\mathbf{C}}_{\text{pred}} - \mathbf{\Sigma}_{\text{text}} \right\|_F^2 \tag{10}$$

This regularization explicitly constrains the relational structure between any two classes i and j based on their joint inactive behavior, preserving the semantic geometry encoded in the CLIP text space. By focusing on the shared "background" region, it provides a dense and stable signal that maintains semantic consistency and enhances calibration robustness. Finally, the proposed Class-wise Covariance Regularization (CCR) is integrated with the base BCE loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{cov}} \tag{11}$$

where λ is a balancing coefficient that controls the strength of structural regularization. By explicitly constraining the label-space covariance, CCR serves as a structural calibration prior that effectively counteracts the over-cooling effect in standard BCE fine-tuning.

5. Experiments

5.1. Experimental Setup

Following recent works [10, 37], we conduct evaluations under two standard benchmark settings: (1) Generalization from Base to New Classes, where the downstream dataset is evenly split into base and new classes, the model is trained exclusively on base classes in a few-shot setting and evaluated on both, with primary focus on the harmonic mean of their performance to investigate whether BCE loss combined with our calibration method outperforms Ranking loss; and (2) Domain Generalization, where the model undergoes few-shot training on the MS-COCO dataset and is then evaluated on four COCO-derived test sets encompassing various domain shifts.

Datasets We conduct evaluations on six standard multi-label benchmark datasets: MS-COCO (General annotations) [17], PASCAL-VOC (General annotations) [33], NUS-WIDE (General annotations) [6], COCO-LT (Long-tail) [11], VOC-LT (Long-tail) [41], and Open-Images-V6 (Open-Vocabulary) [14].

Calibration Metrics To comprehensively evaluate confidence reliability, we employ four standard calibration metrics: Expected Calibration Error (ECE) [10], Maximum Calibration Error (MCE) [10], Adaptive Calibration Error (ACE) [24], and Proximity-Informed Expected Calibration Error (PIECE) [40]. These metrics respectively measure the average, worst-case, adaptive, and proximity-aware calibration performance of the model. We list the details of the calibration metrics in Appendix C.

Implementation Details For each dataset, we perform few-shot fine-tuning (16 samples per class) respectively

Table 1. Average calibration performance across six datasets. "Conf" represents the origin performance on base and new classes with existing tuning methods. Two calibration methods (DAC, DOR) and our method (CCR) to existing tuning methods, ↓ indicates smaller values are better. Calibration error (ECE)is given by $\times 10^{-2}$. **Bold** numbers are significantly superior results.

Method	ECE(↓)				ACE(↓)				MCE(↓)				PIECE(↓)			
	Conf	DAC	DOR	CCR	Conf	DAC	DOR	CCR	Conf	DAC	DOR	CCR	Conf	DAC	DOR	CCR
CoOp	13.25	10.85	9.92	**7.35**	14.18	10.78	9.85	**7.26**	3.95	2.95	2.41	**1.81**	15.12	12.36	10.95	**9.42**
CoCoOp	6.64	5.75	5.42	**5.07**	6.56	5.68	5.35	**5.02**	1.89	1.68	1.59	**1.50**	8.42	7.88	7.66	**7.40**
KgCoOp	5.52	4.32	4.25	**3.94**	4.60	4.43	4.32	4.71	1.37	1.39	1.64	**1.22**	6.82	6.65	6.60	**6.50**
DualCoOp	4.53	4.61	4.51	4.57	4.68	4.60	4.58	**4.53**	1.38	1.25	1.24	**1.07**	6.82	6.84	6.92	**6.78**
TaI-DPT	6.02	5.35	5.08	**4.76**	5.96	5.30	5.02	**4.79**	1.92	1.70	1.61	**1.52**	7.84	7.45	7.32	**7.13**
TaI++	4.47	4.16	4.05	**3.89**	4.21	4.52	4.09	**3.94**	1.32	1.25	1.22	**1.19**	7.00	6.83	6.78	**6.70**
T2I-PAL	4.09	3.93	3.88	**3.62**	4.17	4.00	3.95	**3.75**	1.37	1.18	1.26	**1.04**	7.51	6.41	6.39	**6.15**

Table 2. Average calibration across six datasets. "+CCR" to our method applied to standard tuning methods. ↓ indicates smaller values are better. Calibration error is given by $\times 10^{-2}$. "HM" denotes the harmonic mean.

Method	ECE(↓)			ACE(↓)			MCE(↓)			PIECE(↓)		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
ZSCLIP	3.58	4.61	4.10	3.62	4.58	4.10	0.97	1.21	1.09	6.35	6.55	6.45
CoOp	6.92	21.58	13.25	6.85	21.51	13.18	1.97	5.93	3.95	7.56	22.68	15.12
+CCR	**3.67**	**11.03**	**7.35**	**3.63**	**10.89**	**7.26**	**0.90**	**2.72**	**1.81**	**4.71**	**14.13**	**9.42**
TaI-DPT	3.01	9.03	6.02	2.98	8.94	5.96	0.96	2.88	1.92	3.92	11.76	7.84
+CCR	**2.38**	**7.14**	**4.76**	**2.39**	**7.19**	**4.79**	**0.76**	**2.28**	**1.52**	**3.56**	**10.70**	**7.13**
T2I-PAL	2.04	6.14	4.09	2.08	6.26	4.17	0.68	2.06	1.37	3.75	11.27	7.51
+CCR	**1.81**	**5.43**	**3.62**	**1.87**	**5.63**	**3.75**	**0.52**	**1.56**	**1.04**	**3.07**	**9.23**	**6.15**

[37]. In the few-shot setting, we adopt five different data splits to ensure statistical significance of results [18]. All experiments are conducted using the pre-trained CLIP-ViT-B/16 model [26], following a unified training strategy and hyperparameter configuration consistent with [12]. We fine-tune each model for 10 epochs with a batch size of 32. We use accuracy (0.5 as the threshold) as the performance metric because it reflects performance under well-calibrated conditions. Similarly, we report other metrics such as mAP in the Appendix D.

Baseline Methods We conduct comprehensive experiments on a range of prompt tuning approaches, including standard prompt tuning (CoOp [50], CoCoOp [49]), its regularized extension KgCoOp [43], multi-label specific methods (DualCoOp [29], TAI-DPT [12], TAI++ [39]), as well as the adapter tuning model T2I-PAL [8]. Based on existing fine-tuning methods, we select two state-of-the-art calibration approaches as comparative baselines: Distance-Aware Calibration (DAC) [38], which adjusts logits through a text bias-aware mechanism, and Dynamic Outlier Regularization (DOR) [37], which employs a dynamic outlier regularization strategy.

5.2. Results

Our evaluation spans multiple perspective, including calibration performance on base and new classes under open-vocabulary settings (Table 1-3), base-to-new generalization (Table 4), robustness under domain shifts (Table 5), and sensitivity to hyperparameters (Figure 4). Due to space limitations, we report the main results under the 16-shot setting in the main paper, additional results (different shot) are provided in Appendix D.

CCR addresses structural miscalibration. As shown in Table 1, CCR consistently achieves the lowest calibration errors across the vast majority of prompt-tuning models, a result that holds for all prompt-tuning methods and all four evaluation metrics. This consistent superiority, compared to methods like DAC and DOR which only improve certain models, demonstrates the strong generalization capability of our approach. CCR improves not only the average calibration error (ECE) but also the maximum calibration error (MCE), demonstrating its effectiveness in mitigating severe overconfidence. The consistent improvements across seven tuning frameworks further indicate that CCR functions as a general and robust structural calibration prior, independent

Table 3. Average ECE (%) of regularization-based methods across six datasets. "Vanilla" denotes the baseline with Ranking Loss. Red indicates an increase in ECE (worse) after calibration.

Metric	CoOp				TaI-DPT				T2I-PAL			
	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	6.92	5.83	7.45	**3.67**	2.75	2.52	2.98	**2.38**	2.04	1.89	2.21	**1.81**
Medium	5.25	4.91	4.65	**3.12**	2.95	2.82	2.71	**2.51**	2.35	2.54	2.42	**2.12**
Tail	8.37	9.28	6.92	**4.89**	3.33	3.25	2.99	**2.76**	2.78	3.12	2.65	**2.43**
New	21.58	16.42	19.75	**11.03**	9.03	9.58	8.86	**7.14**	6.14	5.78	5.87	**5.43**

Table 4. Average accuracy (%) across six base-to-new datasets. CCR can improve the generalization capacity on unseen classes while maintaining the performance on base classes. Blue indicates the original domain accuracy.

Class	ZSCLIP	CoOp		CoCoOp		KgCoOp		DualCoOp		TaI-DPT		TaI++		T2I-PAL	
		Vanilla	+CCR	Vanilla	+CCR	Vanilla	+CCR	Vanilla	+CCR	Vanilla	+CCR	Vanilla	+CCR	Vanilla	+CCR
Head	80.15	81.23	**82.76**	79.84	**81.95**	81.67	81.42	81.92	**83.18**	82.45	**84.07**	81.58	**83.29**	84.91	**86.24**
Tail	63.83	64.92	64.45	71.36	**75.68**	72.54	**76.91**	71.08	**74.82**	64.27	**72.64**	72.43	**75.86**	78.95	**81.73**
New	72.46	71.15	**73.82**	75.28	**78.43**	77.13	**79.84**	76.04	**78.67**	73.68	**78.25**	76.92	**79.15**	81.86	**83.97**

of specific prompt-tuning strategies.

CCR achieves a balanced calibration trade-off As shown in Table 2 and 3, existing regularizers such as DAC and DOR create trade-offs within base classes, often improving one group (e.g., tail) at the expense of another (e.g., head). In contrast, CCR simultaneously improves calibration across all Head, Medium, and Tail classes. This indicates that CCR captures the intrinsic link between class frequency and calibration difficulty, balancing confidence distributions via covariance regularization rather than applying uniform shifts or scalings.

CCR improves generalization on both base and new classes As shown in Table 4, CCR consistently improves performance on both base and new classes across all prompt-tuning methods, resulting in an average 4.8% increase in their harmonic mean. Unlike prior approaches such as DOR, which often improve new-class accuracy at the cost of base-class performance, CCR achieves simultaneous gains in both settings, demonstrating a more stable and effective form of generalization. These improvements arise because CCR preserves the semantic geometry of the text embedding space by aligning the predicted inter-class covariance with semantic correlations, while simultaneously stabilizing class-wise confidence relationships and preventing the structural drift induced by BCE fine-tuning. As a result, CCR supports more reliable predictions on previously unseen classes without compromising accuracy on base classes, enabling consistently stronger base-to-new generalization.

CCR exhibits inherent stability across different evaluation distributions Table 5 shows that CCR consistently reduces calibration error across both source and target distributions for CoOp and TaI-DPT. For example, CCR lowers the ECE of CoOp from 5.10% to 2.92% on the source domain and yields similar improvements across all COCO-derived target sets. Importantly, CCR achieves these calibration gains without sacrificing accuracy; in fact, accuracy on both source and target domains improves slightly. These results indicate that CCR enhances the model's ability to maintain reliable confidence estimates when evaluated on shifted data distributions, while also supporting better recognition performance.

CCR integrates local supervision with global semantic structures Unlike instance-focused regularizers such as DAC or outlier-based embedding constraints like DOR, CCR incorporates a global semantic prior derived from text embeddings. By aligning the batch-level predicted covariance matrix with the semantic similarity structure, the model learns both co-occurrence and exclusion relationships among classes. This enables more coherent and interpretable confidence estimates for unseen class combinations, highlighting the importance of structured confidence modeling in open-vocabulary multi-label settings.

CCR achieves hyperparameter robustness through structural constraints As shown in Figure 4, CCR remains stable across a wide range of the regularization coefficient λ, owing to its moment-based regularization mechanism. By constraining second-order statistics (covariance) instead of first-order statistics (means), CCR avoids the sensitivity issues seen in methods like DAC and DOR, which

Table 5. Accuracy comparison on domain generalization datasets. CCR boosts the calibration and generalization of existing methods.

Method	ECE (↓)					Accuracy (↑)				
	Source	Target				Source	Target			
	MS-COCO	COCO-2014	COCO-2015	COCO-2017	COCO-LT	MS-COCO	COCO-2014	COCO-2015	COCO-2017	COCO-LT
ZSCLIP	3.86	4.54	4.88	4.51	5.79	66.73	60.87	66.09	70.98	57.19
CoOp	5.10	5.19	5.40	4.80	5.18	69.44	63.55	75.76	74.81	65.72
+CCR	**3.92**	**3.95**	**3.17**	**2.58**	**4.89**	**71.47**	**72.47**	**76.28**	**75.12**	**66.73**
TaI-DPT	4.13	4.56	4.88	4.06	5.23	69.05	69.57	76.78	76.66	66.41
+CCR	**2.86**	**3.89**	**3.96**	4.37	5.58	**71.93**	**74.94**	**78.77**	**78.29**	**69.55**

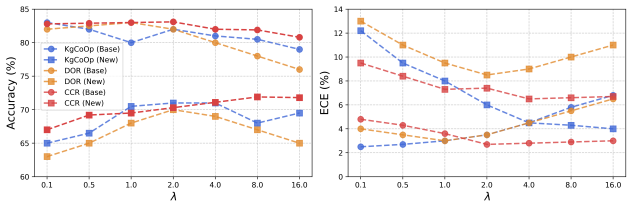
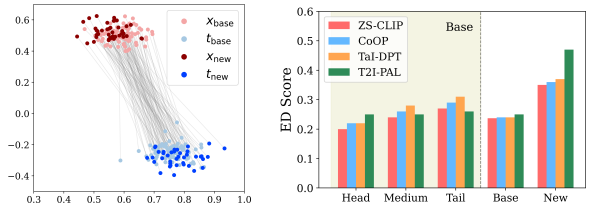


Figure 4. Parameter sensitivity analysis shows our method achieves stable Accuracy and ECE on base classes across λ values, outperforming KgCoOp and DOR in robustness (with CoOp backbone).



(a) Modality gap before CCR (b) ED score before CCR

Figure 5. Compared to the results in Figure 1 and Figure 3, the introduction of CCR effectively reduces the modality gap and the ED score.

often over-regularize when λ is improperly set. Moreover, CCR is also robust to the choice of batch size (Appendix E), since in multi-label settings its covariance structure is primarily determined by the large number of negative (inactive) class pairs. As negative samples dominate in any mini-batch, the resulting covariance remains highly consistent across different batch sizes, allowing CCR to maintain stable behavior under varying training configurations.

6. Discussion

Firstly, while CCR is developed within a BCE framework, its formulation as a covariance regularizer is orthogonal to the choice of supervision loss. CCR operates on the structure of the confidence distribution, a level of abstraction distinct from the pairwise ordering constraints of ranking losses. We posit that integrating CCR could imbue ranking-based fine-tuning with improved probabilistic interpretability, yielding models that maintain discriminative margins while providing better-calibrated confidences. Preliminary results (see Appendix E) indicate that incorporating CCR helps stabilize confidence distributions while maintaining recall performance.

Secondly, the principle of CCR is modality-agnostic, as it regularizes inter-class correlation structures independent of input modality. Although our study focuses on prompt tuning, CCR can be extended to visual fine-tuning or multimodal alignment. For instance, applying CCR to align the covariance of visual features with a semantic prior could act as a cross-modal consistency constraint. CCR introduces a structural level regularizer, contrasting with sample level methods like weight decay or label smoothing [4, 5]. Instead of perturbing individual outputs, CCR preserves the global semantic topology across classes, guiding the model to learn a semantically coherent confidence manifold.

7. Conclusion

In this paper, we propose Class-wise Covariance Regularization, a lightweight structural regularizer designed to improve confidence calibration in multi-label prompt tuning. CCR addresses this issue by aligning the predicted inter-class covariance with the semantic correlations encoded in CLIP's text embeddings, effectively restoring the underlying semantic geometry. As a result, the model attains more balanced and reliable confidence estimates across both base and new classes, while further enhancing the performance of existing prompt-tuning methods. While CCR consistently improves calibration and generalization, it depends on reliable semantic priors from the CLIP text space, which may limit its applicability when textual embeddings are weak. Moreover, CCR models only pairwise linear correlations.

Future work may explore richer dependency structures to better connect discriminative learning with probabilistic calibration, while also exploring more rigorous approaches to calibrating ranking loss. We hope our findings inspire further studies on structured and geometry-aware calibration for multi-label vision–language learning.

8. Acknowledgments

This work is supported by the JiangSu Natural Science Foundation under Grant No. BK20251989; the National Natural Science Foundation of China under Grants Nos. 62172208, 62441225, 61972192; the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No.JYB2025XDXM118). This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Yuexuan An, Hui Xue, Xingyu Zhao, Ning Xu, Pengfei Fang, and Xin Geng. Leveraging bilateral correlations for multi-label few-shot learning. *IEEE TNNLS*, 2024. 2

[2] Haoru Chen, Tianjiao Wan, Zhimin Lin, Kele Xu, Jin Wang, and Huaimin Wang. Vtqagen: Bart-based generative model for visual text question answering. In *ACM MM*, pages 9456–9461, 2023. 1

[3] Tianshui Chen, Liang Lin, Riquan Chen, Xiaolu Hui, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE TPAMI*, 44(3): 1371–1384, 2020. 2

[4] Tianshui Chen, Weihang Wang, Tao Pu, Jinghui Qin, Zhijing Yang, Jie Liu, and Liang Lin. Dynamic correlation learning and regularization for multi-label confidence calibration. *IEEE TIP*, 2024. 3, 8

[5] Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. In *CVPR*, pages 27589–27599, 2024. 8

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, pages 1–9, 2009. 5

[7] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *ICCV*, pages 6889–6899, 2021. 2

[8] Chun-Mei Feng, Kai Yu, Xinxing Xu, Salman Khan, Rick Siow Mong Goh, Wangmeng Zuo, and Yong Liu. Text to image for multi-label image recognition with joint prompt-adapter learning. *IEEE TPAMI*, 2025. 1, 2, 3, 6

[9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 2

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017. 2, 3, 5

[11] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *CVPR*, pages 15089–15098, 2021. 1, 5

[12] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *CVPR*, pages 2808–2817, 2023. 1, 2, 3, 6

[13] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost Van De Weijer. Generative multi-label zero-shot learning. *IEEE TPAMI*, 2023. 2

[14] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 5

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 2

[16] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35:17612–17625, 2022. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5

[18] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *AAAI*, pages 3513–3521, 2024. 3, 6

[19] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE TPAMI*, 44(11):7955–7974, 2021. 1

[20] Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *ICML*, 2024. 2

[21] Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Sparc: Score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *CVPR*, pages 4313–4321, 2025. 3

[22] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *ECCV*, pages 147–165, 2024. 3

[23] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *ICCV*, pages 8731–8740, 2021. 2

[24] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, 2019. 5

[25] Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *NeurIPS*, 37:12677–12707, 2024. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 3, 6

[27] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 1

[28] Samuel Scheele, Katherine Picchione, and Jeffrey Liu. Ladi v2: Multi-label dataset and classifiers for low-altitude disaster imagery. In *CVPR*, pages 2235–2243, 2025. 1

[29] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *NeurIPS*, 35:30569–30582, 2022. 1, 6

[30] Hao Tan, Zichang Tan, Jun Li, Ajian Liu, Jun Wan, and Zhen Lei. Recover and match: Open-vocabulary multi-label recognition through knowledge-constrained optimal transport. In *CVPR*, pages 4650–4660, 2025. 1

[31] Christian Tomani, Futa Kai Waseda, Yuesong Shen, and Daniel Cremers. Beyond in-domain scenarios: Robust density-aware calibration. In *ICML*, pages 34344–34368, 2023. 3

[32] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *ICML*, 2024. 3

[33] Sara Vicente, Joao Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing pascal voc. In *CVPR*, pages 41–48, 2014. 5

[34] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 1

[35] Cong Wang, Zexuan Deng, Zhiwei Jiang, Fei Shen, Yafeng Yin, Shiwei Gan, Zifeng Cheng, Shiping Ge, and Qing Gu. Advanced sign language video generation with compressed and quantized multi-condition tokenization. *NeurIPS*, 2025. 2

[36] Cong Wang, Kuan Tian, Yonghang Guan, Fei Shen, Zhiwei Jiang, Qing Gu, and Jun Zhang. Ensembling diffusion models via adaptive feature aggregation. In *ICLR*, 2025. 1

[37] Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *ICML*, 2024. 2, 3, 5, 6

[38] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned clip. In *ICML*, pages 51734–51754, 2024. 2, 3, 6

[39] Xiangyu Wu, Qing-Yuan Jiang, Yang Yang, Yi-Feng Wu, Qing-Guo Chen, and Jianfeng Lu. Tai++ text as image for multi-label image classification by co-learning transferable prompt. In *IJCAI*, pages 5226–5234, 2024. 1, 6

[40] Miao Xiong, Ailin Deng, Pang Wei W Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. Proximity-informed calibration for deep neural networks. *NeurIPS*, 36:68511–68538, 2023. 5

[41] Jiexuan Yan, Sheng Huang, NanKun Mu, Luwen Huangfu, and Bo Liu. Category-prompt refined feature learning for long-tailed multi-label image classification. In *ACM MM*, pages 2146–2155, 2024. 1, 5

[42] Kun Yan, Chenbin Zhang, Jun Hou, Ping Wang, Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. In *AAAI*, pages 2991–2999, 2022. 2

[43] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 2, 6

[44] Yaodong Yu, Stephen Bates, Yi Ma, and Michael Jordan. Robust calibration with multi-domain temperature scaling. *NeurIPS*, 35:27510–27523, 2022. 2, 3

[45] Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *SIGKDD*, pages 999–1008, 2010. 1

[46] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE TKDE*, 26(8):1819–1837, 2013. 1

[47] Ao Zhou, Zebo Gu, Tenghao Sun, Jiawen Chen, Mingsheng Tu, Zifeng Cheng, Yafeng Yin, Zhiwei Jiang, and Qing Gu. Hierarchical vision-language reasoning for multimodal multiple-choice question answering. In *ACM MM*, pages 13784–13790, 2025. 1

[48] Ao Zhou, Bin Liu, Jin Wang, and Grigorios Tsoumakas. Batch selection for multi-label classification guided by uncertainty and dynamic label correlations. In *AAAI*, pages 22902–22909, 2025. 1

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2, 6

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2, 3, 6