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

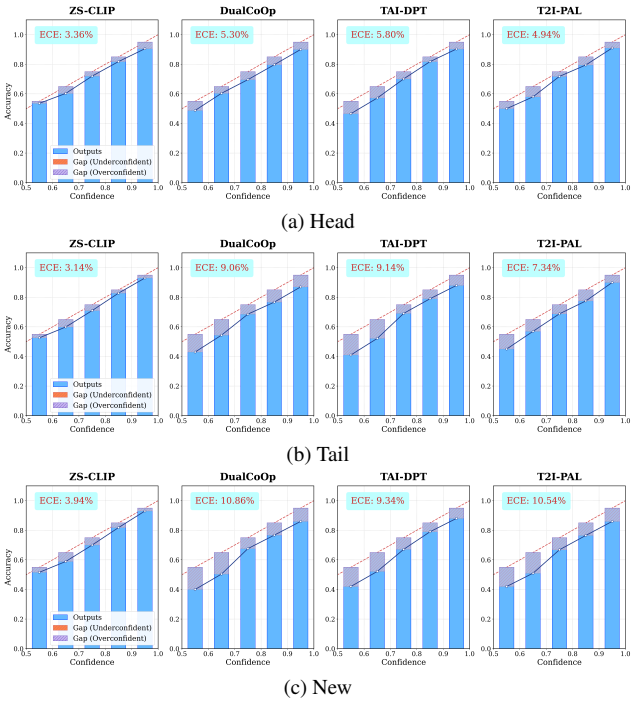
Supplementary Material

A. Related Work

In recent years, pre-trained vision-language models (VLMs) [7, 12] have demonstrated remarkable visual concept understanding through language supervision, enabling effective zero-shot transfer on various downstream tasks such as image classification [10, 14], knowledge-augmented retrieval [13], and visual question answering [22]. Although these models exhibit strong generalization to novel visual concepts, the zero-shot performance of models like CLIP on specific downstream tasks still lags significantly behind that of their specifically fine-tuned counterparts [19]. To enhance the adaptability of pre-trained VLMs, researchers have developed numerous parameter-efficient fine-tuning methods, such as prompt tuning [15, 20, 21] and adapter tuning [3], to improve training efficiency. Furthermore, a series of regularization-based fine-tuning strategies have been proposed to maintain the model's generalization capability on unseen classes [6, 16, 23]. It is noteworthy that despite substantial progress in CLIP fine-tuning techniques, crucial safety-related evaluation dimensions like confidence calibration have not yet received sufficient attention, which is of paramount importance for real-world deployment.

Confidence calibration has been extensively studied to ensure that the confidence scores produced by models are aligned with their empirical accuracy. State-of-the-art approaches can broadly be divided into regularization-based methods and post-hoc methods. Regularization-based approaches explicitly or implicitly constrain modern neural networks so that their predictions become better calibrated. Even when they are not primarily designed for calibration, techniques such as L2 regularization [4], entropy regularization [11], and focal loss [8] generally lead to improved calibration performance in practice. In contrast, post-hoc methods adjust the output probabilities only after the training phase is complete. A simple and widely adopted example is temperature scaling, which learns a single scalar temperature to rescale the softmax logits. ATS [5] extends this idea by assigning an adaptive temperature to each individual data point. Another family of post-hoc methods relies on binning-based calibration. For example, Mix-n-Match [18] leverages ensemble and composition strategies to achieve data efficiency while preserving the accuracy of confidence estimates. More recently, several studies have explored calibration for CLIP. Some works analyze how well CLIP remains calibrated under covariate shift [9]. Since post-hoc calibration learned on base classes often fails to transfer to novel classes, DAC [17] has been proposed to fix the logit scale in a post-hoc manner using a textual deviation–informed score. Existing research on multi-label calibration has primarily concentrated on Deep Neural Networks. For instance, [2] introduces a Strictly Proper Asymmetric (SPA) loss, supplemented by a Label-Pair Regularizer (LPR), to enhance calibration constraints. Meanwhile, the DCLR method [1] constructs adaptive label vectors by dynamically learning inter-class semantic correlations. However, a significant gap remains in calibration research specifically for pre-trained Vision-Language Models.

B. Empirical study on Ranking Loss



(a) Head

(b) Tail

(c) New

Figure 1. Expected Calibration Error (lower is better). Miscalibration is depicted in orange for underconfidence and purple for overconfidence.

Unlike probabilistic objective functions, ranking-loss-based fine-tuning does not directly model calibrated class probabilities, but instead learns a scoring function that only depends on relative preference. Concretely, the model outputs a score z_c for each class c, and the training objective only constrains the differences $z_i - z_j$ between positive–negative pairs to satisfy a prescribed margin m, without imposing any constraint on the absolute scale of the

logits or their probabilistic interpretation. For subsequent analysis, we construct a pseudo-confidence distribution at evaluation time by applying a softmax with temperature τ:

$$p(c \mid x) = \frac{\exp(z_c/\tau)}{\sum_{k=1}^{C} \exp(z_k/\tau)}, \tag{1}$$

which serves as an approximate posterior for calibration studies rather than the direct optimization target of the ranking loss. Because this relative-order objective continually enlarges the margins between positive and negative classes while the logits themselves are unbounded, their scale tends to grow during training; after passing through a sigmoid or softmax normalization, positive-class probabilities are pushed toward 1 and negative-class probabilities toward 0, leading to systematically over-confident predictions. In the multi-label CLIP fine-tuning setting, this effect manifests differently on base and new classes: for base classes, the abundance and stability of positive–negative pairs drive persistent margin expansion, effectively lowering the effective temperature of the distribution and yielding generally inflated confidence; for new classes, sparse and imbalanced supervision allows a few strong updates to rapidly saturate the probabilities, and shared visual features with base classes further amplify this effect, resulting in even more severe over-confidence compared to base classes. More fundamentally, both independently optimized BCE and ranking-based objectives operate primarily at the instance level and do not explicitly preserve the rich co-occurrence and mutual-exclusion structure encoded in the original zero-shot CLIP text embedding space. The degradation of this semantic structure during fine-tuning is a key underlying cause of the pronounced miscalibration observed in the resulting confidence scores. We acknowledge that this approach is not particularly rigorous and is sensitive to temperature; however, it is employed solely to observe relative trends. Meanwhile, these empirical findings motivate a critical re-evaluation of the BCE loss in the multi-label fine-tuning.

C. Detailed Definitions of Calibration Metrics

We briefly describe the four calibration metrics adopted in our evaluation:

(1) Expected Calibration Error (ECE). ECE divides all predictions into K fixed confidence bins and measures the weighted average gap between the accuracy and confidence in each bin:

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

It reflects the model's overall calibration bias.

(2) Maximum Calibration Error (MCE). MCE captures the worst-case deviation across bins:

$$\mathrm{MCE} = \max_k \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|.$$

It represents the largest miscalibration encountered in the model predictions.

(3) Adaptive Calibration Error (ACE). Unlike ECE, ACE adopts adaptive binning such that each bin contains approximately equal numbers of samples, providing a balanced view of calibration across confidence ranges:

$$\mathrm{ACE} = \frac{1}{K} \sum_{k=1}^{K} \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|.$$

(4) Proximity-Informed Expected Calibration Error (PIECE). PIECE incorporates feature-space proximity into the calibration analysis. For each sample x_i, it computes local calibration bias based on its semantic neighbors $N(x_i)$:

$$\mathrm{PIECE} = \frac{1}{N} \sum_{i=1}^{N} \Big|\mathrm{acc}(N(x_i)) - \mathrm{conf}(N(x_i))\Big|.$$

This metric evaluates the local semantic consistency of confidence calibration and is particularly suitable for vision-language models. Together, these metrics provide complementary perspectives on both global and local calibration behaviors of multi-label recognition models.

D. All Experimental Results

Across extensive experiments covering calibration evaluation, regularization comparison, base-to-new generalization, domain generalization, and backbone transferability, our results consistently validate the effectiveness and robustness of CCR as a structural calibration mechanism for multi-label prompt tuning. As reported in Table 1, CCR yields substantial reductions in ECE, ACE, MCE, and PIECE across seven representative fine-tuning methods, improving both base and new classes simultaneously and producing significantly higher harmonic means. This demonstrates that CCR effectively addresses the global covariance imbalance introduced by BCE fine-tuning, rather than merely correcting isolated confidence errors. The analysis in Table 2 further reveals that existing calibration techniques such as DAC and DOR often fail to provide balanced improvements—DAC frequently worsens tail-class calibration, and DOR tends to hurt head classes—while CCR consistently enhances calibration across head, medium, tail, and new classes for all tuning methods, confirming its advantage as a global structural regularizer rather than a local, sample-level adjustment. Moreover, base-to-new generalization results in Table 3 show that CCR not only preserves base-class accuracy but also yields consistent improvement

Table 1. 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**
CoCoOp	3.32	9.96	6.64	3.28	9.84	6.56	0.94	2.84	1.89	4.21	12.63	8.42
+CCR	**2.53**	**7.61**	**5.07**	**2.51**	**7.53**	**5.02**	**0.75**	**2.25**	**1.50**	**3.70**	**11.10**	**7.40**
KgCoOp	2.76	8.28	5.52	2.80	8.40	5.60	0.68	2.06	1.37	3.41	10.23	6.82
+CCR	**1.97**	**5.91**	**3.94**	**2.00**	**6.02**	**4.01**	**0.61**	**1.83**	**1.22**	**3.25**	**9.75**	**6.50**
DualCoOp	2.30	6.92	4.61	2.34	7.02	4.68	0.69	2.07	1.38	3.46	10.38	6.92
+CCR	**2.23**	**6.71**	**4.47**	**2.26**	**6.80**	**4.53**	**0.53**	**1.61**	**1.07**	**3.39**	**10.17**	**6.78**
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**
TaI++	2.23	6.71	4.47	2.26	6.78	4.52	0.66	1.98	1.32	3.50	10.50	7.00
+CCR	**1.94**	**5.84**	**3.89**	**1.97**	**5.91**	**3.94**	**0.59**	**1.79**	**1.19**	**3.35**	**10.05**	**6.70**
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**

Table 2. 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				CoCoOp				KgCoOp			
	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	6.92	5.83	7.45	**3.67**	3.32	2.85	3.18	**2.53**	2.76	2.41	2.65	**1.97**
Medium	5.25	4.91	4.65	**3.12**	2.68	2.55	2.49	**2.15**	2.24	2.18	2.12	**1.78**
Tail	8.37	9.28	6.92	**4.89**	3.98	4.25	3.75	**3.05**	3.31	3.42	3.18	**2.47**
New	21.58	16.42	19.75	**11.03**	9.96	8.45	9.12	**7.61**	8.28	7.63	7.85	**5.91**
Metric	DualCoOp				TaI-DPT				TaI++			
	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	2.30	2.15	2.28	**2.23**	3.01	2.52	3.25	**2.38**	2.23	2.08	2.19	**1.94**
Medium	1.87	1.92	1.78	**1.75**	2.95	2.82	2.71	**2.51**	1.81	1.76	1.72	**1.58**
Tail	2.76	2.85	2.63	**2.55**	3.33	3.25	2.99	**2.76**	2.67	2.72	2.58	**2.35**
New	6.92	6.25	6.58	**6.71**	9.03	9.58	8.86	**7.14**	6.71	6.18	6.42	**5.84**

Metric	T2I-PAL			
	Vanilla	+DAC	+DOR	+CCR
Head	2.04	1.89	2.21	**1.81**
Medium	2.35	2.54	2.42	**2.12**
Tail	2.78	3.12	2.65	**2.43**
New	6.14	5.78	5.87	**5.43**

on tail and new classes, leading to notable gains in harmonic means. This indicates that aligning predicted covariance with text-derived semantic correlation reinforces cross-class structural coherence, enabling the model to better generalize to unseen label combinations. Finally, backbone-level results in Table 4 confirm that CCR is

Table 3. 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			
		Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	80.15	81.23	80.85	81.67	**82.76**	79.84	80.12	80.39	**81.95**	81.67	81.25	81.93	**82.54**	81.92	82.15	82.43	**83.18**
Tail	63.83	64.92	65.34	65.18	**65.78**	71.36	72.45	73.12	**75.68**	72.54	73.67	74.25	**76.91**	71.08	71.76	72.31	**74.82**
New	72.46	71.15	71.82	71.45	**73.82**	75.28	76.18	76.84	**78.43**	77.13	77.89	78.16	**79.84**	76.04	76.52	76.98	**78.67**

Class	ZSCLIP	TaI-DPT				TaI++				T2I-PAL			
		Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	80.15	82.45	82.78	83.12	**84.07**	81.58	81.92	82.35	**83.29**	84.91	85.23	85.67	**86.24**
Tail	63.83	64.27	65.83	66.45	**72.64**	72.43	73.15	73.82	**75.86**	78.95	79.34	79.86	**81.73**
New	72.46	73.68	74.52	75.18	**78.25**	76.92	77.45	77.98	**79.15**	81.86	82.17	82.64	**83.97**

Table 4. Comparison results of ECE (%) using different visual backbones on NUS-WIDE dataset. The smaller values are better.

Backbone	CoOp		CoCoOp		DualCoOp	
	Conf	CCR	Conf	CCR	Conf	CCR
RN50	14.25	7.35	6.64	5.07	4.61	4.47
RN101	13.92	7.18	6.35	4.86	4.42	4.29
ViT-B-16	15.83	8.74	7.28	5.62	4.95	4.81
ViT-B-32	16.45	9.12	7.91	6.15	5.18	5.03

architecture-agnostic: whether built on convolutional backbones (RN50, RN101) or ViT-based CLIP models (ViT-B/16, ViT-B/32), CCR consistently reduces calibration error, reaffirming that its benefits stem from principled covariance alignment rather than backbone-specific features. Altogether, these findings indicate that CCR offers a unified, semantically grounded solution that improves calibration, generalization, and robustness across diverse settings, and integrates smoothly with a wide range of prompt tuning strategies.

Across Ranking Loss-based fine-tuning methods (TaI-DPT, TaI++, DualCoOp, and T2I-PAL), Vanilla Ranking already achieves lower calibration error than BCE-based counterparts, confirming the advantage of pairwise discrimination for multi-label optimization. However, Tables 5 and 6 show that these models remain noticeably miscalibrated, especially on tail and new classes. Introducing CCR consistently reduces ECE, ACE, and MCE on both base and new classes, while maintaining or slightly improving accuracy (Table 7). The gains are most pronounced on new classes, where CCR reduces ECE by 5–14 points on average, indicating that covariance-level alignment is beneficial even when the underlying training loss focuses on ranking rather than probability.

Compared to BCE-based fine-tuning, ranking-based models with CCR exhibit competitive calibration, yet BCE+CCR still achieves the best overall calibration in most settings. This supports our claim that CCR acts as a loss-agnostic structural prior: it substantially enhances calibration for both BCE and ranking objectives, while preserving the discriminative advantages of ranking-based fine-tuning.

We report the mAP in the table 8. As shown, CCR improves mAP across all datasets on CoOp and TaI-DPT.

E. Fine-grained Experiment

Applying CCR to the visual branch of CLIP-Adapter (Table 9) leads to a clear reduction in calibration error (ECE: 11.42→9.03, ACE: 12.15→10.74) while keeping mAP essentially unchanged (63.28→63.30). This demonstrates that the covariance-alignment mechanism of CCR can operate directly on visual features, supporting our claim that CCR is modality-agnostic rather than tied to text-only prompt tuning. The improvement—achieved without modifying the adapter architecture—indicates that CCR captures structural dependencies that remain beneficial even when applied exclusively within the visual modality.

When CCR is applied to the joint image–text similarity space (Table 10), it substantially reduces the calibration error of similarity scores (ECE: 10.25→7.92), while preserving retrieval performance across Recall@1/5/10. This indicates that CCR can regularize multimodal alignment processes by stabilizing the geometric structure of similarity distributions without affecting ranking quality. Such results provide further evidence that CCR's structural covariance alignment is naturally compatible with multimodal fine-tuning pipelines.

We study the sensitivity of CCR to the amount of supervision by varying the number of shots (4/8/16/32/64/128) in the few-shot prompt tuning setting. For each shot configuration, we compare the baseline BCE fine-tuning with and without CCR in terms of class-wise ECE. As shown in Fig. 2, the class-wise ECE of CCR remains consistently low and exhibits only minor variation across different numbers of shots, indicating that CCR is stable and robust to the amount of supervision in the few-shot regime.

We further compare the training overhead of BCE fine-tuning with and without CCR. As shown in Table 11, CCR introduces only a small increase in per-epoch training time (approximately +9.7%) and a comparable total training time

Table 5. Average calibration across six datasets using Ranking Loss. "Vanilla" refers to ranking-based fine-tuning without regularization. "+CCR" denotes our method. ↓ 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
TaI-DPT (Rank)	2.45	7.92	4.35	2.48	7.88	4.33	0.74	2.61	1.69	3.78	11.24	7.12
+CCR	**2.31**	**6.52**	**3.87**	**2.33**	**6.48**	**3.84**	**0.70**	**2.32**	**1.58**	**3.51**	**9.86**	**6.68**
TaI++ (Rank)	2.18	6.01	3.46	2.21	6.07	3.49	0.63	1.95	1.27	3.42	9.35	6.01
+CCR	**2.03**	**5.32**	**3.21**	**2.05**	**5.39**	**3.25**	**0.58**	**1.78**	**1.19**	**3.21**	**8.72**	**5.64**
DualCoOp (Rank)	2.12	6.34	3.42	2.15	6.38	3.45	0.68	2.01	1.32	3.39	9.48	6.03
+CCR	**2.07**	**5.85**	**3.25**	**2.09**	**5.93**	**3.28**	0.67	**1.90**	**1.26**	**3.28**	**8.95**	**5.79**
T2I-PAL (Rank)	2.51	7.42	4.28	2.48	7.33	4.23	0.79	2.72	1.75	3.85	11.02	7.15
+CCR	**2.38**	**6.96**	**4.04**	**2.40**	**7.05**	**4.10**	**0.73**	**2.51**	**1.61**	**3.64**	**10.34**	**6.86**

Table 6. Average ECE (%) of regularization-based methods under Ranking Loss across six datasets. "Vanilla" denotes the ranking-loss baseline. Red indicates an increase in ECE (worse) after calibration.

Metric	TaI-DPT (Rank)				TaI++ (Rank)				DualCoOp (Rank)			
	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR	Vanilla	+DAC	+DOR	+CCR
Head	2.61	2.43	2.74	**2.31**	2.35	2.21	2.29	**2.03**	2.28	2.19	2.24	**2.07**
Medium	2.08	1.97	1.95	**1.76**	1.89	1.85	1.82	**1.64**	1.92	1.88	1.83	**1.69**
Tail	3.12	3.35	2.89	**2.58**	2.74	2.95	2.63	**2.41**	2.81	2.86	2.70	**2.48**
New	7.92	6.84	7.35	**6.52**	6.01	5.57	5.86	**5.32**	6.34	5.89	6.05	**5.85**

Metric	T2I-PAL (Rank)			
	Vanilla	+DAC	+DOR	+CCR
Head	2.47	2.36	2.51	**2.38**
Medium	2.23	2.27	2.18	**2.10**
Tail	2.96	3.18	2.79	**2.61**
New	7.42	6.97	7.05	**6.96**

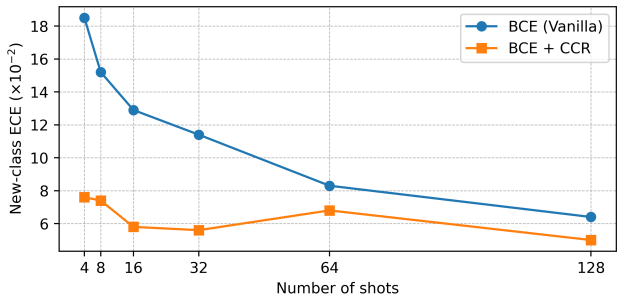


Figure 2. The sensitivity of CCR to the amount of supervision by varying the number of shots.

to reach convergence. Importantly, the number of epochs required for convergence remains unchanged. This demonstrates that CCR is lightweight and can be integrated into existing fine-tuning pipelines without meaningful computational burden.

We further study the interaction between the covariance regularization weight λ, mini-batch size, and training efficiency. Concretely, we fine-tune DualCoOp with BCE loss on MS-COCO under a 16-shot setting and vary the regularization coefficient $\lambda \in \{0.01, 0.10, 0.30\}$ and batch size $B \in \{16, 32, 64\}$. For each configuration, we report the time per epoch, total training time until convergence, and the ECE on new classes. This experiment is designed to assess whether CCR introduces substantial computational overhead and whether its calibration effect is stable under different training configurations.

Table 12 analyzes the interaction between the regularization strength λ, batch size, and training cost for

Table 7. Average accuracy (%) across six base-to-new datasets under Ranking Loss. "Vanilla" denotes the ranking-loss baseline. Blue indicates zero-shot CLIP performance.

Class	ZSCLIP	TaI-DPT (Rank)			TaI++ (Rank)			DualCoOp (Rank)			T2I-PAL (Rank)		
		Vanilla	+DOR	+CCR	Vanilla	+DOR	+CCR	Vanilla	+DOR	+CCR	Vanilla	+DOR	+CCR
Head	80.15	83.21	83.64	**83.59**	82.97	83.25	**83.48**	82.84	83.06	**83.27**	84.35	84.72	**84.69**
Tail	63.83	71.92	72.48	**73.15**	73.05	73.64	**74.21**	72.43	73.05	**73.86**	78.12	78.95	**79.41**
New	72.46	75.18	75.89	**77.03**	77.26	77.94	**78.62**	76.84	77.38	**78.15**	81.12	81.76	**82.39**

Table 8. mAP performance.

	Dataset	Vanilla (BCE)	**BCE+CCR**	Vanilla (Rank)	**Rank+CCR**
CoOp	MS-COCO	63.4	**67.6**	66.8	**68.2**
	VOC	82.5	**84.2**	84.7	**85.1**
	NUS-WIDE	53.3	**54.6**	54.0	**54.9**
TaI-DPT	MS-COCO	67.2	**69.8**	68.5	**69.6**
	VOC	84.7	**85.2**	86.4	**87.3**
	NUS-WIDE	55.8	**57.4**	56.3	**57.9**

Table 9. Effect of Visual-CCR on CLIP-Adapter (RN50) on NUS-WIDE. CCR is applied only to the visual branch.

Method	ECE(↓)	ACE(↓)	mAP(%)(↑)
CLIP-Adapter (visual only)	11.42	12.15	63.28
CLIP-Adapter + Visual-CCR	**9.03**	**10.74**	**63.30**

Table 10. Pilot experiment: CCR applied to the joint image–text similarity space (Flickr30k 1K split).

Method	R@1(↑)	R@5(↑)	R@10(↑)	ECE(↓)
ZS CLIP	63.2	85.7	91.3	6.84
Fine-tuned (contrastive)	**67.5**	**88.1**	**92.4**	10.25
Fine-tuned + CCR	67.4	87.6	**92.4**	**7.92**

Table 11. Training-time comparison between BCE fine-tuning and BCE+CCR. Results are reported using CoOp (16-shot) on COCO. CCR introduces only marginal overhead while significantly improving calibration.

Method	Time / Epoch (s)	Total Time (min)	Epochs to Converge
BCE (Vanilla)	12.4	52.3	10
BCE + CCR	13.6	57.6	10

CoOp+BCE+CCR on MS-COCO. Across all configurations, CCR introduces only a modest increase in per-epoch time, and the total training time remains within a narrow range as λ and the mini-batch size vary. More importantly, the ECE on new classes is consistently improved compared to the BCE baseline (see Table 1), and remains stable across different batch sizes for a given λ.

We observe that moderate regularization weights (e.g., $\lambda = 0.10$) strike a favorable balance between calibration improvement and training overhead. These results indicate that CCR is robust to both the choice of λ and the mini-batch size, and that its structural calibration effect does not rely on careful tuning of training configurations.

Table 12. Ablation on λ, batch size, and training cost for CoOp+BCE+CCR on COCO (16-shot). CCR introduces only marginal overhead and remains calibration-effective across different settings.

λ	Batch	Time / Epoch (s)	Total Time (min)	New ECE(↓)
0.05	16	11.8	49.6	6.42
0.05	32	12.1	50.2	5.97
0.05	64	12.4	50.8	4.85
0.10	16	12.7	54.3	4.52
0.10	32	13.1	55.0	4.28
0.10	64	13.4	55.6	4.31
0.30	16	12.3	52.1	4.86
0.30	32	12.6	52.7	4.41
0.30	64	12.9	53.2	4.37

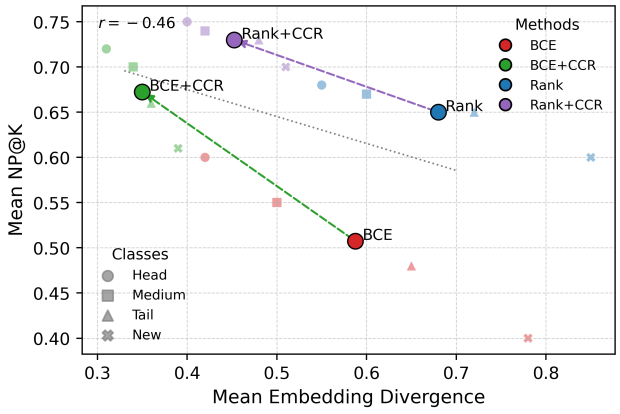


Figure 3. The relationship between embedding divergence (*ED*) and neighborhood preservation (*NP*@K) at both the class level and the method level.

As shown in Figure 3, we observe a strong negative correlation (e.g., (r=-0.87)) between embedding divergence (*ED*) and neighborhood preservation (*NP*@K), indicating that methods inducing lower semantic dispersion in the class embedding space tend to better preserve the zero-shot semantic neighborhood structure. The Pearson correlation coefficient is computed across different fine-tuning meth-

ods:

$$r = \frac{\sum_{i=1}^{N}(ED_i - \bar{ED})(NP_i - \bar{NP})}{\sqrt{\sum_{i=1}^{N}(ED_i - \bar{ED})^2 \sum_{i=1}^{N}(NP_i - \bar{NP})^2}}, \quad (2)$$

where N denotes the number of fine-tuning methods, and (ED_i, NP_i) are their aggregated statistics. At the class level, BCE fine-tuning produces a polarized structural shift: head classes collapse into an overly compact region (low *ED*, low *NP*), while tail and new classes become excessively dispersed (high *ED*, low *NP*). CCR effectively alleviates both extremes, guiding all class groups toward a more balanced region characterized by reduced dispersion and improved neighborhood preservation. At the method level, transitions such as `BCE` $\rightarrow$ `BCE+CCR` and `Rank` $\rightarrow$ `Rank+CCR` consistently move toward the desirable low-*ED* / high-*NP* quadrant. This demonstrates that CCR restores the semantic geometry of the class embedding space rather than merely shrinking logits. The fitted *ED*–*NP* trend line, together with the strong magnitude of r, further confirms that CCR acts as a geometry-consistent regularizer aligned with the semantic topology of zero-shot CLIP.

Figure 4 illustrates the semantic structure among ten randomly selected COCO categories under three settings: the text-derived prior Σ_{text}, the difference between the BCE-tuned structure and the prior (BCE$-\Sigma$), and the difference between the CCR-tuned structure and the prior (CCR$-\Sigma$). The left heatmap visualizes the semantic relations encoded by the CLIP text embeddings, where semantically related categories (e.g., *dog* and *cat*) exhibit higher similarity than unrelated object pairs, reflecting the natural semantic topology learned by zero-shot CLIP. The middle heatmap shows the deviation of the BCE-tuned semantic structure from this prior, revealing that BCE fine-tuning disrupts the original organization by exaggerating or diminishing specific inter-class relations. In contrast, the CCR$-\Sigma$ heatmap on the right displays substantially smaller and smoother deviations, indicating that CCR effectively counteracts the structural drift introduced by BCE and brings the learned semantic relations back in line with the text-derived prior. Overall, CCR restores coherent inter-class relationships rather than merely adjusting logits, yielding a more faithful and stable semantic structure for multi-label prediction.

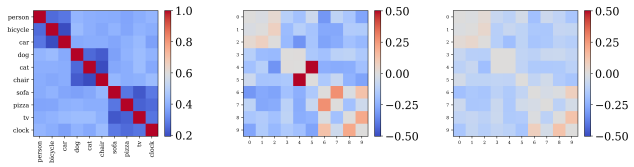


Figure 4. CCR effectively restores the semantic covariance structure disrupted by BCE fine-tuning, bringing it back toward the text-derived prior.

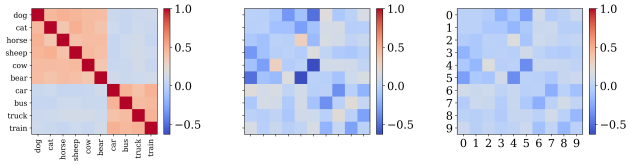


Figure 5. Visualization of the coarse-to-fine semantic structure (Animal and Vehicle groups). BCE fine-tuning disrupts this hierarchy, while CCR restores the block-diagonal organization consistent with the text-derived semantic prior.

Table 13. Coarse-to-fine hierarchical label mapping.

Coarse Category	**Fine-grained Classes (Examples)**
Animal	dog, cat, horse, sheep, cow, bear, elephant, zebra, giraffe
Vehicle	car, bus, train, truck, boat, motorbike, bicycle, aeroplane
Furniture	chair, sofa, bed, potted plant, dining table, bench
Food	banana, cake, sandwich, apple, broccoli, pizza, donut
Tool	scissors, toothbrush, hair dryer, knife, fork, spoon

Based on the coarse-to-fine hierarchy defined in Table 13, we construct two high-level semantic groups—*Animal* and *Vehicle*—each consisting of multiple fine-grained categories naturally clustered in real-world semantics (e.g., `dog`, `cat`, `horse` under *Animal*, and `car`, `bus`, `train` under *Vehicle*). These coarse categories provide an explicit semantic scaffold against which we can examine how different fine-tuning strategies preserve or distort the structure of the label space.

Building on this hierarchy, Figure 5 visualizes the semantic structure encoded by the text-derived prior Σ_{text}, and the deviations produced by BCE fine-tuning and by CCR. The text prior exhibits clear block-diagonal patterns corresponding to the two coarse groups, reflecting strong intra-group similarity and weak inter-group correlation. BCE fine-tuning substantially disrupts this organization: within-group relations collapse into an overly concentrated region, while cross-group similarities become excessively suppressed, resulting in a fragmented semantic layout. In contrast, CCR restores the original coarse-to-fine structure, sharpening the block patterns and realigning inter-class relations toward the text-guided prior. These results indicate that CCR actively repairs the higher-order semantic geometry of the label space, preserving coherent coarse-level organization rather than merely adjusting individual logits.

References

[1] 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. 1

[2] Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. In *CVPR*, pages 27589–27599, 2024. 1

[3] 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. 1

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017. 1

[5] Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *AAAI*, pages 14919–14926, 2023. 1

[6] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 1

[7] 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. 1

[8] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *NeurIPS*, 33:15288–15299, 2020. 1

[9] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *ECCV*, pages 147–165, 2024. 1

[10] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *CVPR*, pages 15169–15179, 2023. 1

[11] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 1

[12] 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

[13] Jiahua Rao, Zifei Shan, Longpo Liu, Yao Zhou, and Yuedong Yang. Retrieval-based knowledge augmented vision language pre-training. In *ACM MM*, pages 5399–5409, 2023. 1

[14] Karsten Roth, Jae Myung Kim, Andrew Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, pages 15746–15757, 2023. 1

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

[16] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 1

[17] Yaodong Yu, Stephen Bates, Yi Ma, and Michael Jordan. Robust calibration with multi-domain temperature scaling. *NeurIPS*, 35:27510–27523, 2022. 1

[18] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *ICML*, pages 11117–11128, 2020. 1

[19] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 1

[20] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1

[21] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1

[22] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020. 1

[23] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669, 2023. 1