Enhancing Few-Shot Class-Incremental Learning via Training-Free Bi-Level Modality Calibration - Supplementary Material -

Yiyang Chen[1], Tianyu Ding[2], Lei Wang[3], Jing Huo[1], Yang Gao[1], Wenbin Li[1,4*]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Applied Sciences Group, Microsoft, USA [3]University of Wollongong, Australia
[4]Shenzhen Research Institute of Nanjing University, Shenzhen, China

A. Detailed Experimental Results

We provide detailed comparative results for the CIFAR100 [1] and CUB200 [9] datasets in Tables 1 and 2. From these results, we can draw the following conclusions: First, in terms of absolute performance, our methods, BiMC or BiMC†, surpass all comparison methods. On the CIFAR100 dataset, our BiMC† achieves an improvement of 4.25 over the best comparison method, and on the CUB200 dataset, our method improves by 3.56 over the best comparator. Furthermore, our methods also achieve competitive results on the forgetting metrics (PD), being the best except for CLIP Zero-Shot on CIFAR100, while on CUB200, our BiMC achieves the lowest PD metric. Both in terms of absolute performance and PD, our framework demonstrates outstanding performance. We also present the complete performance curves in the ablation analysis of the semantic and covariance-enhanced metric in Figure 1.

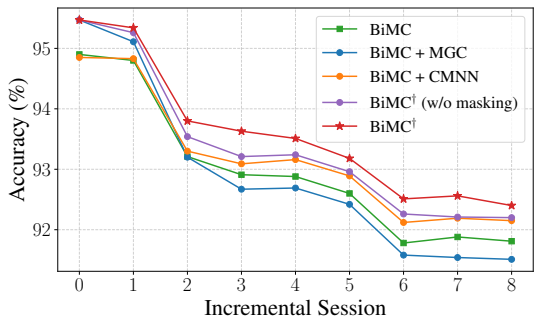
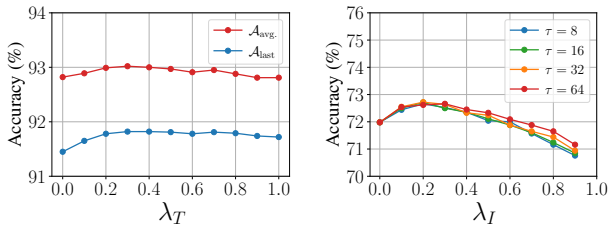


Figure 1. Influence of different shot on incremental sessions

B. More Analyses on Hyper-parameter

In this subsection, we further analyze the impact of the hyperparameters λ_T, λ_I, and τ. For λ_T, the results are shown in Figure 2a. When $\lambda_T = 0$, the method degenerates to using only the CLIP Zero-Shot classifier results, without fine-grained dynamic category information. As λ_T increases, both the average performance and the performance of the last session climb and gradually level off. This indicates that through calibration within the textual modality, the classifiers for each category can absorb category-specific knowledge, thereby enhancing accuracy. Regarding λ_I and τ, as shown in Figure 2b, the method's performance is relatively robust to the effects of τ. When $\lambda_I = 0$, the method degenerates to having no visual intra-modal calibration. As λ_I increases, the overall performance first increases and then decreases, suggesting that an appropriate λ_I can effectively calibrate for new categories.



(a) Analysis on λ_T on *mini*ImageNet (b) Analysis on λ_I, τ on CIFAR100

Figure 2. Analysis on hyper-parameters λ_T, λ_I and τ.

C. LLM generated description

For the CIFAR100 dataset, we use the LLM descriptions provided by [4]. For the *mini*ImageNet and CUB200 datasets, we use the LLM descriptions provided by [7]. In [7], the original descriptions generated by the large language model start with `"An object"`, similar to `"An object which has a pair of long, spiny antennae."` In our experiments, we found that directly using these descriptions for calibration was not effective. Consequently, we replaced `"object"` with the specific real category names, result-

*Corresponding author

Table 1. Detailed session-wise accuracy, average accuracy Avg, and performance degradation PD comparison on the CIFAR100 dataset. **V** and **L** denote the visual and language modalities, respectively. BiMC shows results from the bi-level calibration framework, while BiMC† includes the ensemble classifier strategy. The highest scores in each session are highlighted in **bold**, and the second-highest are underlined. An upward arrow (↑) indicates higher is better, and a downward arrow (↓) indicates lower is better.

Method	Modality	Accuracy in each session(%) ↑									Avg ↑	PD ↓
		0	1	2	3	4	5	6	7	8		
CLIP Zero-Shot [5]	**L**	74.77	73.12	72.56	71.08	70.46	70.41	70.58	70.06	68.93	71.33	**5.84**
Visual Prototype [8]	**V**	75.88	72.78	71.01	68.52	66.55	64.84	64.12	62.71	61.03	67.49	14.85
TEEN [10]	**V**	75.88	73.00	71.39	68.88	67.14	65.41	64.94	63.44	61.84	67.99	14.04
FeCAM [2]	**V**	81.38	78.28	76.93	74.55	72.91	71.81	71.26	69.80	68.32	73.92	13.06
BiMC	**V-L**	79.70	77.97	77.29	75.35	74.76	74.39	74.34	73.55	72.50	75.54	7.20
BiMC†	**V-L**	**81.98**	**79.74**	**78.84**	**77.00**	**76.11**	**75.68**	**75.33**	**74.61**	**73.18**	**76.94**	8.80

Table 2. Detailed session-wise accuracy, average accuracy Avg, and performance degradation PD comparison on the CUB200 dataset. **V** and **L** denote the visual and language modalities, respectively. BiMC shows results from the bi-level calibration framework, while BiMC† includes the ensemble classifier strategy. The highest scores in each session are highlighted in **bold**, and the second-highest are underlined. An upward arrow (↑) indicates higher is better, and a downward arrow (↓) indicates lower is better.

Method	Modality	Accuracy in each session(%) ↑											Avg ↑	PD ↓
		0	1	2	3	4	5	6	7	8	9	10		
CLIP Zero-Shot [5]	**L**	66.38	64.05	62.71	59.07	59.24	59.27	57.76	56.69	55.46	55.19	55.73	59.23	10.65
Visual Prototype [8]	**V**	81.42	78.78	77.43	74.22	72.91	71.59	70.65	69.97	68.79	68.93	68.80	73.04	12.62
TEEN [10]	**V**	81.42	79.16	77.49	74.60	73.26	71.75	70.68	70.21	68.90	69.28	69.12	73.26	12.30
FeCAM [2]	**V**	82.26	79.48	77.76	74.68	72.86	71.10	69.92	69.17	67.61	67.89	67.40	72.74	14.86
BiMC	**V-L**	82.16	79.99	79.04	76.10	75.07	74.04	73.13	73.08	71.88	72.26	72.25	75.36	**9.91**
BiMC†	**V-L**	**83.00**	**81.01**	**79.83**	**76.96**	**75.84**	**74.73**	**73.86**	**73.47**	**72.26**	**72.61**	**72.68**	**76.02**	10.32

ing in descriptions like "`A king crab which has a pair of long, spiny antennae.`" This modified description is then used as a descriptor for the category. We provide detailed examples of the descriptions generated by the LLM and the category distribution for each dataset in Table 3.

D. Analyses of Different CLIP Backbones

We analyze the comparative results under different CLIP [5] backbone networks (ResNet-101, ViT-B/32, ViT-B/16 and ViT-L/14) to demonstrate the broad applicability of our framework. In Table 4, we present the experimental results on the *mini*ImageNet [6] dataset, where our framework (BiMC and BiMC†) consistently outperforms comparison methods across various CLIP backbones.

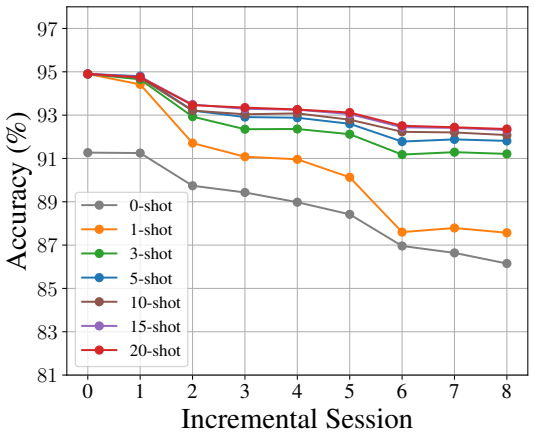


Figure 3. Influence of different shot on incremental sessions

E. Analyses of Incremental Shots

To demonstrate the effectiveness of our method with varying numbers of incremental samples, we vary the k-shot values of the incremental sessions on the *mini*ImageNet dataset. The results are presented in Figure 3, where the gray line represents the CLIP-Zero Shot classifier, and the colored lines indicate the outcomes with different numbers of samples used during the incremental phase. Our method proves effective across various shot numbers. It is evident that starting from 3-shot, our method's performance reaches a significantly high level, proving its effectiveness even with a minimal number of samples. As the number of shots increases, the performance also improves. This improvement is attributed to the increasingly accurate domain priors represented by the visual prototypes, which in turn better calibrate the domain-agnostic linguistic knowledge.

Table 3. Comparison of the number of base classes ($|\mathcal{C}^{\text{base}}|$), number of total novel classes ($|\mathcal{C}^{\text{inc}}|$), and the total number of incremental tasks (T) across various datasets, alongside examples of category descriptions generated by LLMs.

Dataset	$\lvert\mathcal{C}^{\text{base}}\rvert$	$\lvert\mathcal{C}^{\text{inc}}\rvert$	T	LLM-generated description examples
CIFAR100	60	40	8	"A bicycle has two wheels, a frame, handlebars, and pedals." "A camel is a four-legged mammal with a long neck." "A forest is a large area of land covered with trees and other plants." "A plate is a flat, round piece of tableware on which food can be served." "A possum is a small, furry mammal with a long snout and a tail." "A streetcar typically has a low profile and runs on tracks in the roadway."
*mini*ImageNet	60	40	8	"A yawl which has a boom on both the main and mizzen sails." "A komondor which has a robust and muscular body structure." "A Tibetan mastiff which has a double coat with a heavy mane around the neck." "A ladybug which may have varying numbers of spots, from zero to more than twenty." "A catamaran which is typically wider than a traditional monohull boat." "A chime which may vary in size, from small handheld bells to large, floor-standing gongs." "An iPod which may have a clip on the back side in some models." "A scoreboard which may have a section for displaying timeouts left in a game."
CUB200	100	100	10	"A Laysan Albatross with a large, pinkish beak that has a dark tip." "A Groove billed Ani with a white patch on the wing, visible in flight." "A Rhinoceros Auklet with red legs and webbed feet." "A Yellow headed Blackbird that is often seen perched on reeds and cattails." "A Cardinal with a loud, clear whistle, which is a common sound made by Cardinals." "A Gray Kingbird with a strong, direct flight with rapid wing beats." "A White breasted Kingfisher with a large, red, dagger-like beak."

Table 4. Performance comparison across CLIP backbones: Best results are bolded, second-best are underlined.

Method	Backbone	$\mathcal{A}_{\text{base}}$	$\mathcal{A}_{\text{last}}$	$\mathcal{A}_{\text{avg.}}$	PD
CLIP Zero-Shot [5]	ResNet-101	85.98	81.24	83.56	4.84
Visual Prototype [8]		88.88	82.52	85.18	7.31
TEEN [10]		89.14	83.39	85.73	6.44
FeCAM [2]		91.35	85.02	87.78	6.90
BiMC		<u>91.89</u>	<u>88.59</u>	<u>89.96</u>	**3.44**
BiMC†		**92.72**	**89.03**	**90.65**	<u>3.89</u>
CLIP Zero-Shot [5]	ViT-B/32	87.77	84.13	85.59	3.77
Visual Prototype [8]		88.18	82.54	84.91	6.76
TEEN [10]		88.55	83.40	85.52	5.90
FeCAM [2]		91.17	85.36	87.89	6.66
BiMC		<u>92.18</u>	<u>89.03</u>	<u>90.25</u>	<u>3.39</u>
BiMC†		**92.63**	**89.67**	**90.91**	**3.21**
CLIP Zero-Shot [5]	ViT-B/16	91.25	86.15	88.76	5.12
Visual Prototype [8]		91.88	86.09	88.55	6.49
TEEN [10]		92.06	86.66	88.92	5.92
FeCAM [2]		93.71	88.20	90.66	6.22
BiMC		<u>94.80</u>	<u>91.81</u>	<u>92.97</u>	<u>3.09</u>
BiMC†		**95.34**	**92.40**	**93.60**	**3.07**
CLIP Zero-Shot [5]	ViT-L/14	94.17	89.68	91.78	4.44
Visual Prototype [8]		95.12	90.41	92.30	5.06
TEEN [10]		95.26	91.07	92.71	4.40
FeCAM [2]		95.49	91.44	93.14	4.61
BiMC		<u>96.72</u>	**94.46**	<u>95.39</u>	**2.19**
BiMC†		**96.78**	**94.46**	**95.41**	<u>2.32</u>

F. Analysis of Classifier Calibration

In the main paper, we explore a property of the calibrated classifier: *the bi-level calibrated classifier enhances prediction confidence*. One might question whether the proposed cross-modal calibrated framework truly "calibrates" the classifier. Specifically, whether it prevents overconfident predictions. To investigate the calibration behavior of the classifier, we plot the calibration curves of three classifiers on the CIFAR-100 dataset, as shown in Figure 4. Intuitively, the closer the curve is to the diagonal, the better the calibration. It can be observed that the Bi-level Calibrated Classifier achieves better calibration compared to unimodal classifiers.

Furthermore, we quantitatively analyze the calibration using two standard calibration metrics: the Expected Calibration Error (ECE) and the Maximum Calibration Error (MCE) [3], which quantify the discrepancy between the model's predicted confidence and the actual accuracy. It can be observed that the classifier calibrated using the bi-level framework achieves a higher degree of calibration.

Method	ECE	MCE
Textual Classifier	0.121	0.207
Visual Classifier	0.092	0.165
Bi-level Calibrated Classifier	**0.062**	**0.136**

Table 5. Comparison of ECE and MCE for three classifiers.

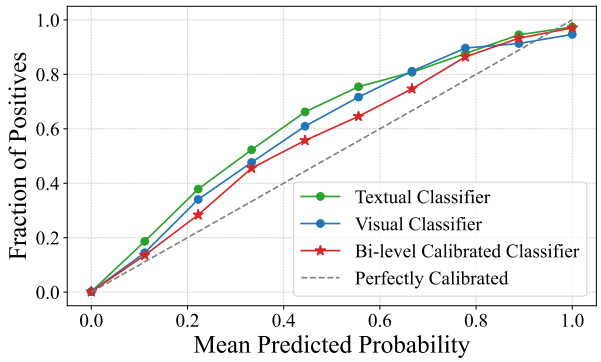


Figure 4. Calibration curve on CIFAR100.

References

[1] Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009. 1

[2] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 4

[4] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1

[5] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[6] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017. 2

[7] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17542–17552, 2024. 1

[8] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[9] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[10] Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3