Appendix: Quantized Prompt for Efficient Generalization of Vision-Language Models

Tianxiang Hao[1,2], Xiaohan Ding[3](✉), Juexiao Feng[1,2], Yuhong Yang[1,2], Hui Chen[2], and Guiguang Ding[1,2](✉)

[1] School of Software, Tsinghua University, Beijing, China
[2] BNRist, Beijing, China
[3] Bytedance
{beyondhtx,xiaohding,suise.con,jichenhui}@gmail.com
fjx20@mails.tsinghua.edu.cn, dinggg@tsinghua.edu.cn

1 Datasets

Building upon prior research [19,20], we utilize eleven datasets pertaining to image recognition to substantiate the effectiveness of the proposed methodology in addressing the base-to-new generalization task. These datasets encompass two repositories dedicated to generic object classification, namely ImageNet [3] and Caltech101 [4], five repositories catering to fine-grained classification, including OxfordPets [13], StanfordCars [10], Flowers102 [12], Food101 [1], and FGVCAircraft [11], one repository for scene recognition, denoted as SUN397 [18], one repository for action recognition, known as UCF101 [15], one repository for texture classification, termed DTD [2], and one repository for satellite imagery recognition, designated EuroSAT [5]. Consistent with earlier studies [9, 19–21], for each dataset in base-to-new generalization, we evenly partition the classes into two distinct groups that do not overlap, with one group serving as the base classes and the other as the new classes. We train all models only using the base classes and conduct evaluation on both the base and new classes to verify the specialization capability and generalization capability of the models.

In the domain generalization task, we leverage ImageNet-A [7], ImageNet-R [6], ImageNetv2 [14], and ImageNet-S [16] to assess the robustness of the model. In this context, the model is initially trained using ImageNet, followed by direct utilization of images from the aforementioned datasets for inference.

Concerning the cross-dataset transfer task, the datasets mirror those utilized in the base-to-new generalization task. Analogous to domain generalization, the model undergoes initial training on ImageNet followed by inference on the remaining ten distinct datasets.

For the few-shot learning task, the datasets align with those employed in the base-to-new generalization task. The model is trained and assessed with varying numbers of shots, specifically 1, 2, 4, 8, and 16 shots separately.

The dataset partitioning mirrors that of earlier works [19,20]. We present the average model performance over three iterations with distinct random seeds to ensure fair comparisons.

Table 1: Full results in base-to-new generalization. H: harmonic mean [17].

(a) An overview of the size of different methods..

Method	CoOp	CoCoOp	Adapter	LoRA	ProGrad	QCoOp	MaPLe	QMaPLe
size	4.1KB	70.8KB	1051KB	258KB	16.4KB	0.26KB	7096KB	1774KB

(b) Average

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
Adapter	82.62	70.97	76.35
LoRA	84.30	67.33	74.86
ProGrad	82.79	68.55	75.00
QCoOp	80.68	74.44	77.43
MaPLe	82.28	75.14	78.55
QMaPLe	83.02	75.57	79.12

(c) ImageNet

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
Adapter	76.53	66.67	71.26
LoRA	74.77	58.47	65.62
ProGrad	77.03	68.80	72.68
QCoOp	76.17	70.73	73.35
MaPLe	76.66	70.54	73.47
QMaPLe	76.93	70.73	73.70

(d) Caltech101

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
Adapter	98.20	93.20	95.63
LoRA	98.49	90.33	94.24
ProGrad	98.50	91.90	95.09
QCoOp	97.80	95.03	96.40
MaPLe	97.74	94.36	96.02
QMaPLe	97.97	95.00	96.46

(e) OxfordPets

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
Adapter	94.40	94.10	94.25
LoRA	94.90	92.57	93.72
ProGrad	94.40	95.10	94.75
QCoOp	95.17	97.60	96.37
MaPLe	95.43	97.76	96.58
QMaPLe	95.67	97.63	96.64

(f) StanfordCars

	Base	New	H
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
Adapter	77.13	69.23	72.97
LoRA	81.07	65.30	72.34
ProGrad	79.00	67.93	73.05
QCoOp	73.73	72.90	73.31
MaPLe	72.94	74.00	73.47
QMaPLe	75.00	73.67	74.33

(g) Flowers102

	Base	New	H
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
Adapter	97.70	70.83	82.13
LoRA	98.23	60.20	74.65
ProGrad	96.27	71.07	81.77
QCoOp	95.57	74.67	84.22
MaPLe	95.92	72.46	82.56
QMaPLe	96.43	74.33	83.95

(h) Food101

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
Adapter	90.40	90.40	90.40
LoRA	88.57	87.30	87.93
ProGrad	90.17	89.53	89.85
QCoOp	90.87	91.90	91.38
MaPLe	90.71	92.05	91.38
QMaPLe	90.63	92.10	91.36

(i) FGVCAircraft

	Base	New	H
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
Adapter	39.57	32.27	35.55
LoRA	46.27	28.83	35.53
ProGrad	42.63	26.97	33.04
QCoOp	37.50	34.03	35.68
MaPLe	37.44	35.61	36.50
QMaPLe	39.10	34.90	36.88

(j) SUN397

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
Adapter	81.67	73.93	77.61
LoRA	79.73	69.00	73.98
ProGrad	80.70	71.03	75.56
QCoOp	79.20	77.93	78.56
MaPLe	80.82	78.70	79.75
QMaPLe	81.33	78.27	79.77

(k) DTD

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
Adapter	80.47	52.23	63.35
LoRA	82.93	54.90	66.06
ProGrad	76.70	46.67	58.03
QCoOp	74.97	58.37	65.63
MaPLe	80.36	59.18	68.16
QMaPLe	80.77	57.63	67.27

(l) EuroSAT

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
Adapter	86.93	64.20	73.86
LoRA	94.90	65.67	77.62
ProGrad	91.37	56.53	69.85
QCoOp	83.53	69.80	76.05
MaPLe	94.07	73.23	82.35
QMaPLe	94.30	79.47	86.25

(m) UCF101

	Base	New	H
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
Adapter	85.80	73.63	79.25
LoRA	87.47	68.03	76.53
ProGrad	83.90	68.50	75.42
QCoOp	81.87	75.93	78.79
MaPLe	83.00	78.66	80.77
QMaPLe	85.10	77.50	81.12

Table 2: Results of deep prompts for QCoOp.

total depth	size	base	new	H
1	0.26KB	79.49	72.65	75.92
2	0.52KB	78.19	73.67	75.86
3	0.78KB	79.82	72.27	75.86
4	1.04KB	80.77	72.31	76.31
5	1.30KB	81.14	72.67	76.67
6	1.56KB	81.65	72.68	76.90

2 Training Configuration

Following the conventional setup outlined in [19], we employ ViT-B/16 as the image encoder within CLIP. Prior to feeding into the image encoder, each training image is resized to 224×224. To augment the data, standard techniques such as random cropping and flipping are applied, consistent with the methodology described in [19]. During training, a batch size of 32 is utilized, and stochastic gradient descent (SGD) is employed to optimize the learnable parameters. Similar to the approach detailed in [20], a warm-up scheme is implemented during the first epoch, which proves crucial for prompt tuning. All other baselines are configured strictly according to the specifications provided in their respective original papers.

Hyperparameter tuning is performed via a grid search methodology, guided by the parameter configurations reported in previous studies [9, 20]. For experiments involving QCoOp, the quantization bit of the prompts is set to 1 by default unless otherwise specified. In the case of QMaPLe experiments, the parameters in the projection layer, responsible for transforming text prompts into image prompts, and the parameters in the prompts are all subjected to quantization, with a quantization bit of 4. Additionally, following the structure of MaPLe, nine transformer layers are typically modified within QMaPLe experiments.

3 Full Base-To-New Generalization Results

As shown in Tab. 1, QCoOp and QMaPLe significantly improve the generalization capability represented by the accuracy on the new classes. Compared with CoOp, QCoOp earns 11.22% accuracy gain on the new classes and 2.10% accuracy drop on the base classes. Among all the lightweight SOTA methods, QCoOp gets strongest harmonic mean accuracy on 7 out of 11 datasets including ImageNet, Caltech101, StanfordCars, Flowers102, Food101, FGVCAircraft, and SUN397. Compared with a heavy method MaPLe, QMaPLe achieves better harmonic mean accuracy on 9 out of 11 datasets, including ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, FGVCAircraft, SUN397, EuroSAT and UCF101.

Table 3: Comparisons of using quantized prompts in textual encoder and image encoder at minimal size.

	size	base	new	H
QCoOp	0.26KB	79.49	72.65	75.92
QVPT	0.39KB	73.15	68.93	70.98

4 Experiments on Deep Prompt Configurations

In this paragraph, we investigate the impact of deep prompts, as described in Eq. 4 of the main text. By default, QCoOp employs prompts with a depth of 1, which means the prompts are only added to the input layer. We systematically increase the number of tunable prompts in subsequent transformer layers. Considering that the text transformer in CLIP consists of 12 layers, we add prompts to a maximum of 6 layers for verification. As illustrated in Table Tab. 2, incorporating more prompts across multiple layers tends to improve the specialized capability, albeit at the expense of a slight reduction in generalized capability. In summary, at the expense of increased dimensionality, using more prompts across multiple layers can improve the harmonic mean accuracy to some extent.

5 Prompt Modality Considerations for Minimizing Storage Cost

In this paragraph, we conduct a brief comparison of quantizing prompts in the visual transformer, following the approach outlined in VPT [8]. The results are presented in Tab. 3. It is evident that when operating under a stringent storage constraint, QCoOp outperforms QVPT in terms of both accuracy and size efficiency. Adding prompts to the textual transformer is better.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV (2014) 1
2. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014) 1
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 1
4. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR-W (2004) 1
5. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019) 1
6. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV (2021) 1

7. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. CVPR (2021) 1
8. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. arXiv preprint arXiv:2203.12119 (2022) 4
9. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023) 1, 3
10. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV-W (2013) 1
11. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 1
12. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008) 1
13. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR (2012) 1
14. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019) 1
15. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 1
16. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. pp. 10506–10518 (2019) 1
17. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR (2017) 2
18. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010) 1
19. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022) 1, 3
20. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 1, 3
21. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15659–15669 (2023) 1