

Beyond Graph Model: Reliable VLM Fine-Tuning via Random Graph Adapter

Supplementary Material

A. Datasets

Following previous works [18, 22, 23], we conduct extensive experiments on 11 diverse downstream tasks: ImageNet-1K [3], StanfordCars [9], Caltech101 [5], UCF101 [17], Flowers102 [12], Food101 [1], DTD [2], EuroSAT [8], FGVCaircraft [11], OxfordPets [13], and SUN397 [20]. These datasets span a wide range of visual domains and classification tasks. Among them, OxfordPets (37 dog and cat breeds), Food101 (101 food classes), StanfordCars (196 car classes), Flowers102 (102 flower species), and FGVCaircraft (100 aircraft models) datasets represent fine-grained classification tasks that require the model to distinguish between visually similar subclasses. EuroSAT (10 land use classes) focuses on remote sensing image classification, while DTD (47 texture classes) evaluates texture recognition capabilities. The remaining datasets: Caltech101 (101 object classes), UCF101 (101 action classes), and SUN397 (397 scene classes) cover general object recognition, action recognition, and scene understanding, respectively. ImageNet-1K serves as our primary benchmark with its comprehensive collection of 1,000 diverse object classes. To assess the model’s robustness to out-of-distribution shifts, we also conduct generalization experiments on ImageNet-V2 [16] and ImageNet-Sketch [19] datasets. ImageNet-V2 represents a natural distribution shift from ImageNet-1K, while ImageNet-Sketch introduces style variations by replacing natural images with sketch-style images, presenting more challenging evaluation scenarios.

B. More Experimental Results

B.1. Impact of Hyperparameters

We investigate the sensitivity of four main hyperparameters in our framework on ImageNet-1K [3] under 16-shot setting, including the number of descriptions M , trade-off parameter α , uncertainty scale λ , and fusion weight β . As shown in Fig. 1, performance improves as M increases from 2 to 50, peaking at $M = 50$ before saturation. This indicates that diverse descriptions capture richer semantics but a moderate number is sufficient. For trade-off parameter α , our model achieves the optimal balance when $\alpha = 0.7$, while further increasing α leads to performance degradation. This demonstrates the importance of properly combining initial semantic features and adapter-enhanced information. The uncertainty scale λ affects the sensitivity of confidence measurement, and we find that a moderate value of $\lambda = 0.4$ works best, as extreme values under- or over-

emphasize prediction uncertainty. The fusion weight β is relatively poor without fusion, and setting $\beta = 0.5$ achieves the best results, while larger values show slight degradation. This suggests that properly weighting different model predictions is crucial for an effective ensemble.

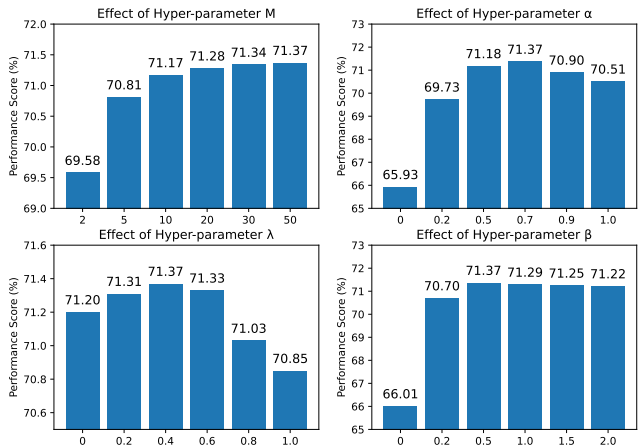


Figure 1. Ablation studies for hyperparameters on ImageNet-1K.

B.2. CLIP Visual Encoders

Furthermore, we evaluate our method on the ImageNet-1K [3] dataset with various CLIP visual encoders, covering both CNN-based architectures, including ResNet-50 [7] and ResNet-101 [7], and transformer-based architectures, including ViT-B/32 [4] and ViT-B/16 [4]. Such a diverse evaluation protocol allows us to thoroughly examine whether the effectiveness of our method is preserved under different backbone designs. As shown in Table 1, our method consistently outperforms existing approaches across all four backbone variants. Notably, the performance gains remain stable for both CNN-based and transformer-based encoders, indicating that the proposed method is not tied to a specific architecture. This strong performance demonstrates the superior generalization ability, robustness, and scalability of our method when applied to different CLIP visual encoders.

B.3. Analysis of Different Prompt Strategies

Table 2 presents the performance of VRGAdapter under different prompt strategies, including simple prompts [15], domain-specific hand-crafted prompts [15], and CuPL prompts [14]. This comparison allows us to examine whether the effectiveness of our method depends on a particular prompt design or can generalize across diverse textual descriptions. From the results, we observe that VRGAdapter consistently outperforms Vanilla CLIP [15] un-

Table 1. Ablation Study (%) of different CLIP visual encoders on 16-shot ImageNet-1K.

Models	Backbone			
	RN50	RN101	ViT-B/32	ViT-B/16
ZS-CLIP [15]	58.18	61.62	62.05	66.73
CoOp [24]	62.95	66.60	66.85	71.92
CLIP-Adapter [6]	63.59	65.39	66.19	71.13
TaskRes [21]	64.75	67.70	68.20	73.07
Tip-Adapter-F [22]	65.51	68.56	68.65	73.69
GraphAdapter [10]	65.70	68.23	68.80	73.68
CaFo [23]	68.79	70.82	70.82	74.48
AMU-Tuning [18]	70.02	71.58	71.65	74.98
VRGAdapter	71.37	73.28	73.34	76.78

der all prompt settings on both DTD and ImageNet. This suggests that our method can better exploit the semantic information provided by prompts and effectively enhance the quality of text representations.

Table 2. Results on the 16-shot setting. Vanilla CLIP denotes the original CLIP text embeddings. Simple: “a photo of a [class]”. Hand-crafted: standard CLIP domain-specific prompts.

Method	Prompting	DTD	ImageNet
Vanilla CLIP	Simple (M=1)	69.68	70.05
VRGAdapter	Simple (M=1)	72.03	70.81
Vanilla CLIP	Hand-crafted (M=8)	70.35	70.23
VRGAdapter	Hand-crafted (M=8)	72.58	71.02
Vanilla CLIP	CuPL (M=50)	71.39	70.41
VRGAdapter	CuPL (M=50)	73.48	71.37

B.4. Core VRGAdapter without Ensemble

To further examine the intrinsic effectiveness of VRGAdapter, we remove the multi-branch ensemble and evaluate a simplified version that uses only the CLIP branch, denoted as VRGAdapter*. This setting allows us to isolate the contribution of the core design of VRGAdapter, independent of additional gains brought by model ensembling. As shown in Table 3, even without the CLIP, MoCo, and DINO ensemble, the VRGAdapter* consistently outperforms both the Linear-probe CLIP [15] and the CLIP-Adapter [6] across all shot settings. In particular, in the 16-shot setting, VRGAdapter* achieves 66.03% accuracy, surpassing CLIP-Adapter by 2.44%. The improvements are also stable under lower-shot setting, demonstrating that the proposed adaptation strategy itself is highly effective even in the absence of auxiliary visual branches.

B.5. Few-Shot Classification

Table 4 presents the detailed numerical results of our method across 11 diverse downstream tasks under various few-shot settings (1-shot, 2-shot, 4-shot, 8-shot, and 16-shot). Following standard practice, we conduct all ex-

Table 3. Results on ImageNet-1K. *: only CLIP, no ensemble.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
Linear-probe CLIP [15]	22.17	31.90	41.20	49.52	56.13
CLIP-Adapter [6]	61.20	61.52	61.84	62.68	63.59
VRGAdapter*	62.49	62.92	63.57	64.60	66.03

periments with three random seeds and report the mean accuracy together with the standard deviation. The results demonstrate that our proposed VRGAdapter consistently outperforms state-of-the-art methods including Tip-Adapter-F [22], GraphAdapter [10], CaFo [23], and AMU-Tuning [18] across most datasets and settings. These results clearly validate the superior adaptability, robustness, and generalization ability of VRGAdapter across diverse downstream benchmarks.

References

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—Mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1
- [2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1
- [5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 1
- [6] 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*, pages 1–15, 2023. 2
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1
- [8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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*, 12(7):2217–2226, 2019. 1
- [9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV workshops*, pages 554–561, 2013. 1
- [10] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2023. 2, 3

Table 4. Comparison (%) of different SOTA methods under various few-shot classification. The best performances are marked in bold.

Methods	Setting	Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet-1K	OxfordPets	StanfordCars	SUN397	UCF101	Avg.
Zero-shot CLIP [15]	0-shot	86.29	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.95
Tip-Adapter-F [22]	1-shot	89.33	49.65	59.53	20.22	79.98	77.51	61.32	87.01	58.86	62.50	64.87	64.62
GraphAdapter [10]		88.90	51.77	63.30	20.93	80.00	75.43	61.50	84.40	59.70	61.93	64.93	64.80
CaFo [23]		91.85	53.43	69.01	24.96	80.88	77.99	63.80	89.21	61.98	64.89	68.60	67.87
AMU-Tuning [18]		91.32	53.33	69.53	22.28	80.33	77.88	62.60	88.34	59.19	63.65	67.54	66.88
VRGAdapter (std)		92.01	56.97	72.92	24.35	85.80	77.79	63.93	88.59	60.39	64.99	67.55	68.66
		±0.81	±1.63	±1.75	±0.28	±1.45	±0.07	±0.05	±0.36	±0.11	±0.28	±0.96	±0.70
Tip-Adapter-F [22]	2-shot	89.74	53.72	66.15	23.16	82.30	77.81	61.69	87.03	61.50	63.64	66.43	66.65
GraphAdapter [10]		90.20	55.75	67.27	23.80	85.63	76.27	62.32	86.30	63.23	64.60	69.47	67.71
CaFo [23]		92.37	56.32	72.86	26.04	84.94	78.10	64.64	89.10	63.36	66.81	70.45	69.54
AMU-Tuning [18]		92.30	55.21	73.03	25.70	84.76	78.00	64.25	88.69	61.18	65.80	70.16	69.01
VRGAdapter (std)		92.76	59.91	75.85	29.31	91.18	78.24	65.43	89.22	64.39	67.44	73.07	71.53
		±0.84	±1.78	±1.00	±0.78	±1.16	±0.14	±0.33	±0.36	±0.15	±0.09	±0.62	±0.66
Tip-Adapter-F [22]	4-shot	90.56	57.39	74.12	25.80	88.83	78.24	62.52	87.54	64.57	66.21	70.55	69.67
GraphAdapter [10]		90.97	59.63	75.20	26.97	89.90	76.77	63.12	86.57	66.53	66.70	71.47	70.35
CaFo [23]		93.14	60.99	83.90	32.94	90.95	78.32	65.64	90.11	65.69	69.17	72.96	73.07
AMU-Tuning [18]		94.85	61.70	84.40	32.69	90.61	78.19	65.92	91.00	65.09	68.08	72.65	73.19
VRGAdapter (std)		95.02	65.45	85.33	34.84	94.67	78.81	67.45	90.29	68.39	69.82	75.97	75.09
		±0.30	±0.09	±0.72	±0.64	±0.44	±0.09	±0.07	±0.56	±0.07	±0.41	±0.68	±0.37
Tip-Adapter-F [22]	8-shot	91.44	62.71	77.93	30.21	91.51	78.64	64.00	88.09	69.25	68.87	74.25	72.45
GraphAdapter [10]		92.45	64.50	80.17	31.37	94.07	77.73	64.23	87.63	70.57	68.97	75.73	73.40
CaFo [23]		93.83	66.19	86.48	40.38	92.98	78.84	66.86	90.52	70.31	70.34	78.06	75.89
AMU-Tuning [18]		95.75	65.80	87.97	39.85	94.95	78.75	68.25	91.25	70.54	69.60	78.30	76.46
VRGAdapter (std)		95.76	68.85	88.76	41.98	97.35	79.53	69.37	91.30	73.40	71.80	80.49	78.05
		±0.34	±0.71	±0.32	±0.40	±0.05	±0.18	±0.15	±0.13	±0.56	±0.15	±0.63	±0.33
Tip-Adapter-F [22]	16-shot	92.86	66.55	84.54	35.55	94.80	79.43	65.51	89.70	75.74	71.47	78.03	75.83
GraphAdapter [10]		93.33	67.57	85.27	36.87	96.23	78.63	65.70	88.57	76.23	71.20	78.80	76.22
CaFo [23]		94.60	69.62	88.68	49.05	95.86	79.30	68.79	91.55	76.73	72.60	79.94	78.79
AMU-Tuning [18]		96.28	70.82	90.22	48.39	97.58	79.28	70.02	92.01	76.72	71.40	80.30	79.37
VRGAdapter (std)		96.31	73.48	91.20	51.10	98.42	80.11	71.37	92.38	80.53	73.72	82.86	81.04
		±0.21	±0.36	±0.51	±0.32	±0.31	±0.03	±0.06	±0.18	±0.21	±0.04	±0.24	±0.22

- [11] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1
- [12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1
- [13] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE Computer Society, 2012. 1
- [14] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15691–15701, 2023. 1
- [15] 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
- [16] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, pages 5389–5400. PMLR, 2019. 1
- [17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1
- [18] Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Amu-tuning: Effective logit bias for clip-based few-shot learning. In *CVPR*, pages 23323–23333, 2024. 1, 2, 3
- [19] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1
- [20] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1
- [21] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *CVPR*, pages 10899–10909, 2023. 2
- [22] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-

Adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, pages 493–510. Springer, 2022. [1](#), [2](#), [3](#)

- [23] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *CVPR*, pages 15211–15222, 2023. [1](#), [2](#), [3](#)
- [24] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. [2](#)