Attn-Adapter: Attention Is All You Need for Online Few-shot Learner of Vision-Language Model

Phuoc-Nguyen Bui*
Convergence Research Institute
Sungkyunkwan University, Suwon, Korea
phuocnguyen@skku.edu

Khanh-Binh Nguyen*
School of Information Technology
Deakin University, Geelong, Australia
binh.nguyen@deakin.edu.au

Hyunseung Choo†
Department of Electrical and Computer Engineering
Sungkyunkwan University, Suwon, Korea
choo@skku.edu

Abstract

Contrastive vision-language models excel in zero-shot image recognition but face challenges in few-shot scenarios due to computationally intensive offline fine-tuning using prompt learning, which risks overfitting. To overcome these limitations, we propose Attn-Adapter, a novel online few-shot learning framework that enhances CLIP's adaptability via a dual attention mechanism. Our design incorporates dataset-specific information through two components: the Memory Attn-Adapter, which refines category embeddings using support examples, and the Local-Global Attn-Adapter, which enriches image embeddings by integrating local and global features. This architecture enables dynamic adaptation from a few labeled samples without retraining the base model. Attn-Adapter outperforms state-of-the-art methods in cross-category and cross-dataset generalization, maintaining efficient inference and scaling across CLIP backbones.

1. Introduction

Vision-language models (VLMs) unify visual and textual understanding for multimodal tasks. CLIP [25], a prominent example, enables zero-shot image recognition via large-scale contrastive learning between images and text. This allows CLIP to generalize across diverse visual concepts without category-specific supervision. However, many real-world tasks in medical imaging or robotics, require domain adaptation with limited labeled data [18, 19, 21]. Few-shot learning addresses this by adapting models to novel classes with minimal supervision. Offline methods like CoOp [35], CoCoOp [34], ProMIM [1] and CLIP-Adapter [7] fine-tune prompts or models using support data, but they are compute intensive and prone to overfitting. Online methods such as Tip-Adapter [33], Proto-CLIP [23], and Meta-Adapter [26] avoid full fine-tuning by via support features. Meta-Adapter improves generalization using a lightweight residual adapter, but still struggles to capture dataset-specific nuances due to reliance on zero-shot CLIP features [20].

To address these challenges, we propose Attn-Adapter, a novel online few-shot learner for vision-language models that leverages attention mechanisms to dynamically refine both category and image embeddings. Our approach introduces two key components: (1) a Memory Attn-Adapter that applies cross-attention to refine category embeddings using support embeddings as keys and values, and (2) a Local-Global Attn-Adapter that enhances image embeddings by integrating local and global features through attention mechanisms. Unlike previous methods that only fine-tune the affinity matrix from few-shot support samples, Attn-Adapter imposes dataset-specific information during fine-tuning through its dual attention mechanism, enabling more effective generalization across diverse datasets and tasks. Our contributions can be summarized as follows:

- We propose Attn-Adapter, a lightweight online few-shot learner that leverages attention mechanisms to dynamically refine CLIP features guided by few-shot samples.
- We introduce a novel dual attention architecture consisting of Memory Attn-Adapter and Local-Global Attn-Adapter, which effectively captures dataset-specific information and enhances both category and image embeddings.
- Extensive experiments show that Attn-Adapter outperforms SOTA online methods across different configura-

*Equal contribution
†Corresponding author

tions while maintaining higher inference speed.

2. Related Work

2.1. Few-Shot Learning

Offline methods Prompt learning is a popular offline strategy. CoOp [35] optimizes learnable prompt vectors for CLIP, improving dataset-specific performance. CoCoOp [34] enhances generalization by conditioning prompts on image features. CLIP-Adapter [7] uses lightweight adapters for feature-level adaptation. However, these methods are computationally expensive, prone to overfitting, and require retraining for new tasks, limiting their practicality.

Online methods Online methods [22, 23, 26, 33] enable adaptation without backbone updates. Tip-Adapter [33] uses a training-free cache-based approach for efficient inference but requires dataset-specific hyperparameter tuning. Meta-Adapter [26] improves robustness with gated attention trained through meta-learning, though it relies on static CLIP features. Proto-CLIP [23] aligns image-text prototypes but lacks flexibility due to per-dataset optimization.

2.2. Attention Mechanisms in VLMs

Attention mechanisms are central to VLMs, enabling focus on relevant input features. They operate via self-attention within modalities and cross-attention across modalities. Transformer-based VLMs, like CLIP, use multi-head self-attention to model relationships within image patches or text tokens. Cross-attention, as in LXMERT [28] and ViLBERT [15], aligns visual and textual features for multimodal reasoning. Recent advances, such as Gallop [13], enhance few-shot learning by using attention to integrate local and global visual features in prompt learning. Our proposed Attn-Adapter leverages attention to dynamically refine category and image embeddings with dataset-specific cues for online few-shot learning.

3. Methodology

3.1. Revisiting CLIP, Tip- and Meta-Adapter

CLIP [25] achieves strong zero-shot performance by contrastively training on large-scale noisy image-text pairs [5, 8]. For zero-shot classification, CLIP computes cosine similarities between an image feature $f \in \mathbb{R}^{D\times 1}$ and a set of text-derived class embeddings $\{w_i\}_{i=1}^{N}$, where $w_i \in \mathbb{R}^{D\times 1}$ and N is the number of classes. Text features are generated from templates like "a photo of [CLS]". The predicted logits of the given image, y, belonging to the i-th class can be formulated as:

$$\text{logits}(y_c = i) = \frac{w_i^\top f}{\|w_i\| \, \|f\|}, \tag{1}$$

Tip-Adapter [33] proposes an online method for few-shot adaptation using a simple modulation function with two hyper-parameters, α and β, and stochastic hyper-parameter search technique. Given support images $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{N}$ from N classes with K shots each, the predicted logits are:

$$\text{logits}(y_c = i|\mathbf{x}, \alpha, \beta) = \frac{w_i^\top f}{\|w_i\| \, \|f\|} + \alpha \cdot \exp(-\beta(1 - \frac{\mathbf{F}_j^\top f}{\|\mathbf{F}_j\| \, \|f\|}))\mathbf{L}_j \tag{2}$$

Here, $\mathbf{F}_i \in \mathbb{R}^{D\times K}$ represents the embeddings of few-shot support samples, while $\mathbf{L}_i \in \mathbb{R}^{N\times K}$ denotes the corresponding one-hot labels for the i-th class. While effective, Tip-Adapter relies on dataset-specific hyper-parameter tuning, limiting generalization.

Meta-Adapter [26] improves Tip-Adapter's poor generalization with a lightweight residual adapter that refines CLIP features using few-shot samples, replacing Tip-Adapter's manual modulation. However, its reliance on zero-shot CLIP features limits dataset-specific adaptability, reducing performance on diverse tasks. Incorporating such information during training could enhance adaptability.

3.2. Attn-Adapter

In contrast with previous methods [26, 33], which only fine-tuned the affinity matrix from few-shot support samples, Attn-Adapter proposes a new update strategy to fully leverage the trainable framework. As shown in Figure 1, we first extract support and category embeddings using CLIP encoders. Afterward, two proposed adapters process few-shot and test samples separately. The Memory Attn-Adapter refines category embeddings by applying them as queries over support embeddings with multi-head attention. For test samples, global and local features (g, l) are passed through the Local-Global Attn-Adapter to obtain refined image embeddings f. The final logits are computed as:

$$\text{logits}(y_c = i|\mathbf{x}) = \frac{\hat{w}_i^\top f}{\|\hat{w}_i\| \, \|f\|}, \tag{3}$$

where $\hat{w} = \text{Memory Attn-Adapter}(w, \mathbf{F})$ and $f = \text{Local-Global Attn-Adapter}(g, f)$. The $\hat{w}$ and f are the refined category embeddings and image embeddings.

3.2.1. Memory Attn-Adapter

In the Memory Attn-Adapter, the introduced approach dynamically combines the support embeddings based on the relation between categories and few-shot images. This approach employs a cross-attention mechanism:

$$\hat{\mathbf{F}} = \mathbf{F}^\top \sigma(\text{MLP}_K(\mathbf{F})\text{MLP}_Q(w)^\top / \sqrt{D}), \tag{4}$$

where MLP_K and MLP_Q denote the MLP layers for key and query. The softmax is represented by σ, D is the scaling factor, and $\hat{\mathbf{F}}$ stands for the aggregated support embeddings. Analogous to non-local filters, the Memory Attn-Adapter can ignore certain outlier samples while focusing more on

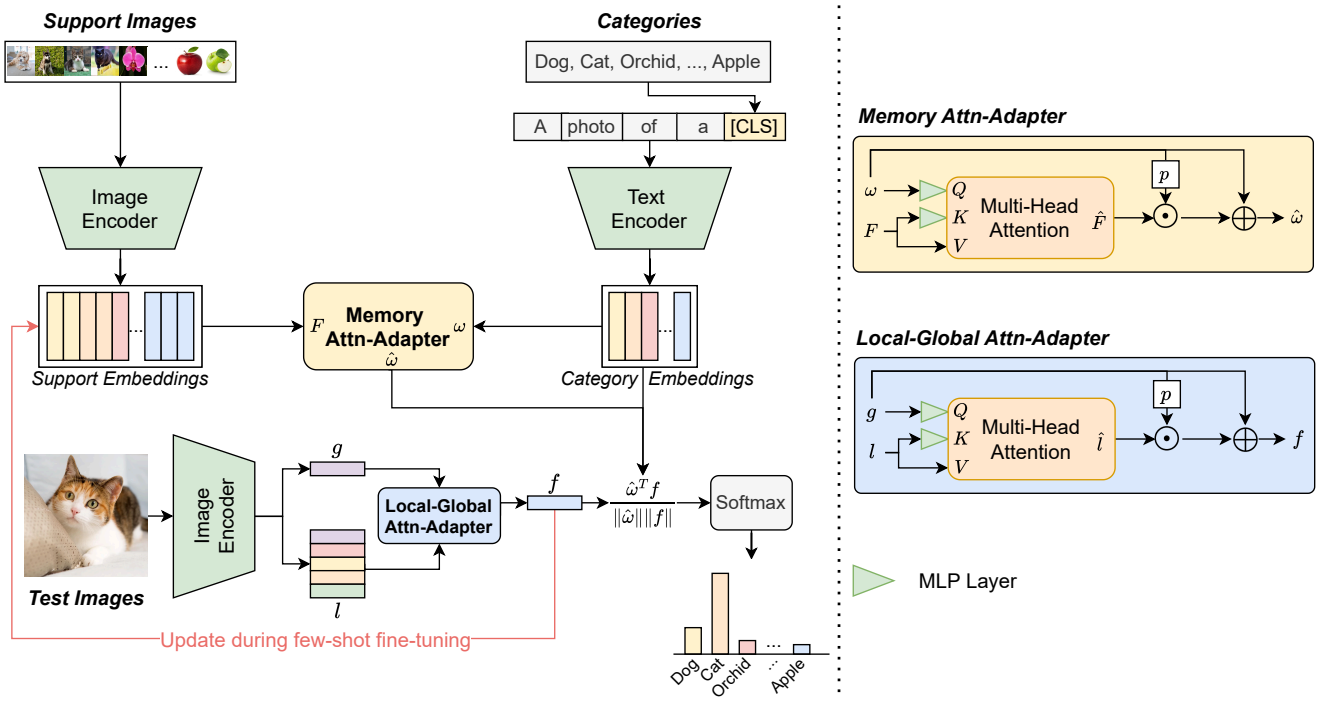

Figure 1. Illustration of the Attn-Adapter model, which utilizes a trainable network with two attention-based components to adjust the category embeddings using few-shot images as guidance.

samples closely aligned with the category description [32], thus achieving robust feature representations.

Moreover, the significance of textual and visual elements for few-shot learning can differ significantly across various data distributions [7]. To address this, we introduce a learnable projector p designed to dynamically adjust the balance between category embeddings and combined support embeddings. As a result, the enhanced category embedding is derived as follows: $\hat{w} = w + p(w) \odot \hat{\mathbf{F}}$, where $\odot$ represents the Hadamard product. By training with the few-shot samples, p can tailor the proportion based on the category descriptions, which allows our method to successfully merge few-shot learning with zero-shot knowledge.

3.2.2. Local-Global Attn-Adapter

Inspired by LoCoOp [17] and Gallop [13] in using the local features to enhance the global features in prompt learning, the Memory Attn-Adapter aggregates the local and global features to generate image embeddings f, this f is then used to update the support embeddings. Here, we also use the cross-attention mechanism with the global features as query, and the local features as key and value.

$$\hat{l} = l^{\top}\sigma(\mathrm{MLP}_K(l)\mathrm{MLP}_Q(g)^{\top}/\sqrt{D}), \tag{5}$$

where MLP_K and MLP_Q denote the MLP layers for key and query. The enhanced embedding is derived as follows:

$$f = g + p(g) \odot \hat{l}, \tag{6}$$

where $\odot$ represents the Hadamard product.

In contrast to the typical transformer encoder [29], our strategy incorporates Multilayer Perceptron (MLP) layers solely for the query and key components, thus the zero-shot value is unchanged.

3.2.3. Training Objectives

The objective function is a weighted combination of contrastive loss $\mathcal{L}_{ce}$ and regularization loss $\mathcal{L}_{l2}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{l2},$$

where

$$\mathcal{L}_{ce} = -\sum_{\mathbf{x}\in\mathbf{X}} \log\left(\frac{\exp(d(\mathbf{x}, \hat{w}_y)/\tau)}{\sum_{i=1}^{N_c}\exp(d(\mathbf{x}, \hat{w}_i)/\tau)}\right), \tag{7}$$

$$\mathcal{L}_{l2} = \|f - g\|_2^2 \tag{8}$$

here $d(\cdot,\cdot)$ is cosine similarity, and τ is the temperature.

4. Experiments

We evaluate Attn-Adapter on two tasks: cross-dataset generalization and cross-category, comparing it with Zero-shot (ZS) CLIP [25], Tip-Adapter [33], and Meta-Adapter [26].

Datasets For cross-category generalization, we use eight classification benchmarks: ImageNet [3], FGVCAircraft [16], OxfordPets [24] (Pets), SUN397 [31],

Table 1. Comparison of cross-dataset generalization based on ImageNet [3] pre-training. The Tip-Adapter, Meta-Adapter, and Attn-Adapter are fine-tuned on ImageNet and frozen for other datasets. Δ reflects the improvement against the latest SOTA.

Method	FGVC	Pets	SUN397	UCF101	Caltech101	DTD	EuroSAT	Food101	Cars	Flowers	Avg.	Δ
Zero-shot CLIP	0.42	56.25	28.96	21.05	60.62	10.00	4.17	77.40	55.70	66.00	38.06	-
Tip-Adapter*	13.96	68.75	45.16	40.09	68.33	42.92	56.25	79.50	75.20	94.90	58.51	-
Tip-Adapter	13.96	67.19	43.80	39.47	67.08	40.00	56.25	77.80	66.70	89.90	56.22	-
Meta-Adapter*	19.58	72.66	51.25	52.28	71.46	49.17	64.58	81.33	78.15	95.12	63.56	-
Meta-Adapter	15.21	72.66	48.54	47.54	67.92	48.33	62.50	79.00	67.30	93.50	60.25	-
Attn-Adapter*	**31.25**	**74.22**	**62.66**	**55.44**	**75.00**	**60.83**	**79.19**	**84.21**	**78.68**	**95.65**	**69.71**	**+6.15**
Attn-Adapter	**22.92**	**73.97**	**55.02**	**49.93**	**68.33**	**50.67**	**66.25**	**83.12**	**73.51**	**95.13**	**63.89**	**+3.64**

* indicates individually searching hyper-parameters or training for each evaluation dataset.

UCF101 [27], Caltech101 [6], DTD [2], and EuroSAT [10]. Following [26], categories are split into base (easy) and novel (hard) sets using zero-shot CLIP accuracy to simulate a challenging setup. For cross-dataset generalization, we further evaluate on ImageNet [3] and its variants: ImageNet-A [12], -R [11], -Sketch [30], and -V2 [3].

Implementation details We use ResNet-50 [9] and ViT-B/16 [4] as CLIP backbones. Following prior work [25, 33], we apply prompt ensembling with seven templates. We strictly adhere the training settings in Meta-Adapter [26] including batch size, AdamW [14] optimizer, cosine schedule, and number of epochs on the base dataset.

Table 2. Quantitative results of domain generalization experiments between Tip-Adapter, Meta-Adapter, and Attn-Adapter. The data in parentheses records the changes brought by comparing with Zero-shot CLIP. Adpt stands for Adapter.

Backbone	Model	Target Datasets				
		IN-A	IN-R	IN-S	IN-V2	Avg
RN-50	ZS CLIP	23.88	60.54	35.45	53.25	43.28
	Tip-Adpt	23.25	58.73	34.77	-	-
	Meta-Adpt	23.71	59.96	35.54	-	-
	Attn-Adpt	**37.61**	60.14	**47.88**	**65.47**	**52.78**
ViT-B/16	ZS CLIP	50.65	77.82	48.42	62.30	59.80
	Tip-Adpt	49.89	76.94	48.13	-	-
	Meta-Adpt	51.12	77.54	48.76	-	-
	Attn-Adpt	**62.52**	**78.71**	**59.70**	**73.91**	**68.71**

4.1. Cross-Dataset Generalization

We evaluate cross-dataset generalization by training on ImageNet and testing on other datasets in a zero-shot manner under a 16-shot setup with frozen parameters. Table 1 shows that Attn-Adapter outperforms other baseline methods in all datasets with an average gain of 3.64%. In terms of training time, it matches Meta-Adapter's efficiency while surpassing online methods, as shown in **Supplementary Material**.

We evaluated domain generalization using models trained on ImageNet, tested on ImageNet-A [12], -R [11], -Sketch [30], and -V2, following [34]. Tip-Adapter, with ImageNet-tuned α and β, underperforms Zero-shot CLIP due to overfitting (e.g., -1.81% on ImageNet-R with RN-50). Meta-Adapter slightly improves over Zero-shot CLIP on some variants. Attn-Adapter consistently outperforms baselines, with 10%+ gains on ImageNet-A and -Sketch, and 12%+ on ImageNet-V2, demonstrating superior adaptability to distribution shifts for real-world applications.

Table 3. Quantitative results of in-domain generalization settings on OxfordPets, UCF101, Caltech101 (Caltech), DTD, and FGVCAircraft (FGCV) datasets between Attn-Adapter and other methods.

Model	ImageNet		Pets		SUN397		EuroSAT	
	Base	Novel	Base	Novel	Base	Novel	Base	Novel
ZS CLIP	71.81	32.89	92.89	56.25	71.28	28.96	48.21	4.17
Tip-Adpt	74.16	36.51	94.83	68.75	73.04	45.16	83.04	56.25
Meta-Adpt	66.08	40.19	92.03	72.66	72.95	51.25	68.75	64.58
Attn-Adpt	87.35	**54.99**	92.67	**74.22**	77.89	**62.66**	83.93	**79.19**

Model	UCF101		Caltech		DTD		FGVC	
	Base	Novel	Base	Novel	Base	Novel	Base	Novel
ZS CLIP	79.42	21.05	93.39	60.62	59.38	10.00	23.84	0.42
Tip-Adpt	85.17	40.09	95.09	68.33	68.36	42.92	30.27	13.96
Meta-Adpt	82.44	52.28	93.39	71.46	64.26	49.17	27.32	19.58
Attn-Adpt	85.36	**55.44**	94.73	**75.00**	68.36	**60.83**	31.70	**31.25**

4.2. Cross-Category Generalization

As shown in Table 3, Tip-Adapter excels on base datasets (e.g., 94.83% OxfordPets, 95.09% Caltech101) but struggles on novel categories (e.g., 40.09% UCF101). Conversely, Attn-Adapter outperforms baselines on both, achieving 87.35% ImageNet, 83.93% EuroSAT (base), and 79.19% EuroSAT, 62.66% SUN397 (novel). Its attention-based

mechanism dynamically refines embeddings, integrating global and local features for robust generalization. Please refer to **Supplementary Material** for more evaluation.

5. Acknowledgements

This work was supported in part by IITP grant funded by the Korean government (MSIT) under IITP-2025-RS-2020-II201821 (30%), RS-2024-00459512 (30%), RS-2021-II212068 (20%), and RS-2019-II190421 (20%).

6. Conclusion

We present Attn-Adapter, a lightweight online few-shot learning framework enhancing vision-language models utilizing two trainable modules: Memory Attn-Adapter and Local-Global Attn-Adapter. These components refine category and image embeddings using minimal labeled examples without backbone updates by injecting dataset-specific inductive bias during inference. Evaluated on cross-category and cross-dataset generalization, Attn-Adapter outperforms prior methods across image classification benchmarks with low inference cost, generalizing well across backbones (ResNet, ViT). Furthermore, Attn-Adapter generalizes well across backbones (e.g., ResNet, ViT) and is particularly effective in low-shot and domain-shift settings. In future work, we plan to extend Attn-Adapter to more complex downstream tasks such as open-vocabulary object detection and semantic segmentation, further validating its generality and flexibility in broader vision-language applications.

References

[1] Phuoc-Nguyen Bui, Khanh-Binh Nguyen, and Hyunseung Choo. Accelerating conditional prompt learning via masked image modeling for vision-language models. *arXiv preprint arXiv:2508.04942*, 2025. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 3, 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 2

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004. 4

[7] 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. *arXiv preprint arXiv:2110.04544*, 2021. 1, 2, 3

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[10] 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. 4

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 4

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 4

[13] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2025. 2, 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[16] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[17] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] Khanh-Binh Nguyen. Debiasing, calibrating, and improving semi-supervised learning performance via simple ensemble projector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2441–2451, 2024. 1

[19] Khanh-Binh Nguyen. Sequencematch: Revisiting the design of weak-strong augmentations for semi-supervised learning.

In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 96–106, 2024. 1

[20] Khanh-Binh Nguyen and Chae Jung Park. On calibration of prompt learning using temperature scaling. *IEEE Access*, 2025. 1

[21] Khanh-Binh Nguyen and Joon-Sung Yang. Boosting semi-supervised learning by bridging high and low-confidence predictions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1028–1038, 2023. 1

[22] Khanh-Binh Nguyen, Phuoc-Nguyen Bui, Hyunseung Choo, and Duc Thanh Nguyen. Adaptive cache enhancement for test-time adaptation of vision-language models. *arXiv preprint arXiv:2508.07570*, 2025. 2

[23] Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, Yu Xiang, et al. Proto-clip: Vision-language prototypical network for few-shot learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2594–2601. IEEE, 2024. 1, 2

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 3

[25] 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. 1, 2, 3, 4

[26] Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36:55361–55374, 2023. 1, 2, 3, 4

[27] 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. 4

[28] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[30] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[31] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 3

[32] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022. 3

[33] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1, 2, 3, 4

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 4

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2