# Appendix

## Contents



## A. Text Encoder is Better for Cross-Domain Task

In the main text, we described the discovery of the lost layer and proposed the VtT model to address this issue. In this section, we provide additional insights into our previous findings and the design of our method. We conducted an in-depth comparison between the text encoder and the vision encoder of CLIP, examining their differences and their impact in cross-domain scenarios.

### A.1. Different Thinking Patterns

We examined the attention distribution corresponding to the CLS/EOS token in each layer of the CLIP's vision and text encoder , as illustrated in Figure 1a and 1c. We have two discoveries: (1) In the vision encoder, shallow layers focus on specific semantic parts, while deeper layers shift to focusing on a small number of background non-semantic tokens. (2) In the text encoder, shallow layers concentrate on contextual content, whereas deeper layers shift to focusing on semantically rich category tokens.

We also quantitatively demonstrated this phenomenon. We measured the attention weight ratio of category tokens to all text tokens in each layer of the text encoder, as depicted by the red line in Figure 1d. The attention weight of the EOS token to category tokens in shallow layers is nearly zero, indicating that these blocks predominantly focus on context rather than task-relevant semantic information. As the layers deepen, this attention ratio increases, with over 60% of attention in the final layers directed towards aggregating category-related token information. Similarly, in the vision branch, we tracked the attention weight of the top 10 tokens with the highest attention in the final layer across all layers, shown by the red line in Figure 1b. As the layers deepen, these top 5% tokens capture approximately 60% of the attention score.

### A.2. How This Difference Influences Cross-Domain Tasks

An intuitive hypothesis is that a text encoder model, which focuses more on semantic information in the final layers, should extract more domain-independent features. For instance, the phrases "a photo of a dog" and "a photo of a cartoon dog" should emphasize the semantic information "dog" thereby mitigating the impact of domain-specific information like "cartoon". To validate this hypothesis, we measured the similarity of text and visual features from different layers of the encoder for the same category but from different domains on ImageNet-R. We use different templates containing domain information to simulate cross-domain conditions in the text branch, such as "a cartoon photo of a dog" and "a photo of a painting-style dog". The results are shown by the blue line in Figure 1b and 1d.

In Figure 1d, the similarity of text features containing different domain information declines in the initial layers but increases in the final layers, supporting our hypothesis. The text encoder in the final layers (believed to emphasize classification tasks) stresses category-related semantic information. Consequently, the similarity of text descriptions with the same semantics but different domain information increases in these layers. The final extracted text characteristics, which are **dominated by semantic information, are less sensitive to domain differences**. The similarity between text features of the same category but different domains remains above 0.9, indicating that the text encoder extracts domain-independent features.

Conversely, the vision encoder reduces its focus on semantic parts in the final layers and shifts attention to a small number of background tokens (Figure 1b, red line). The

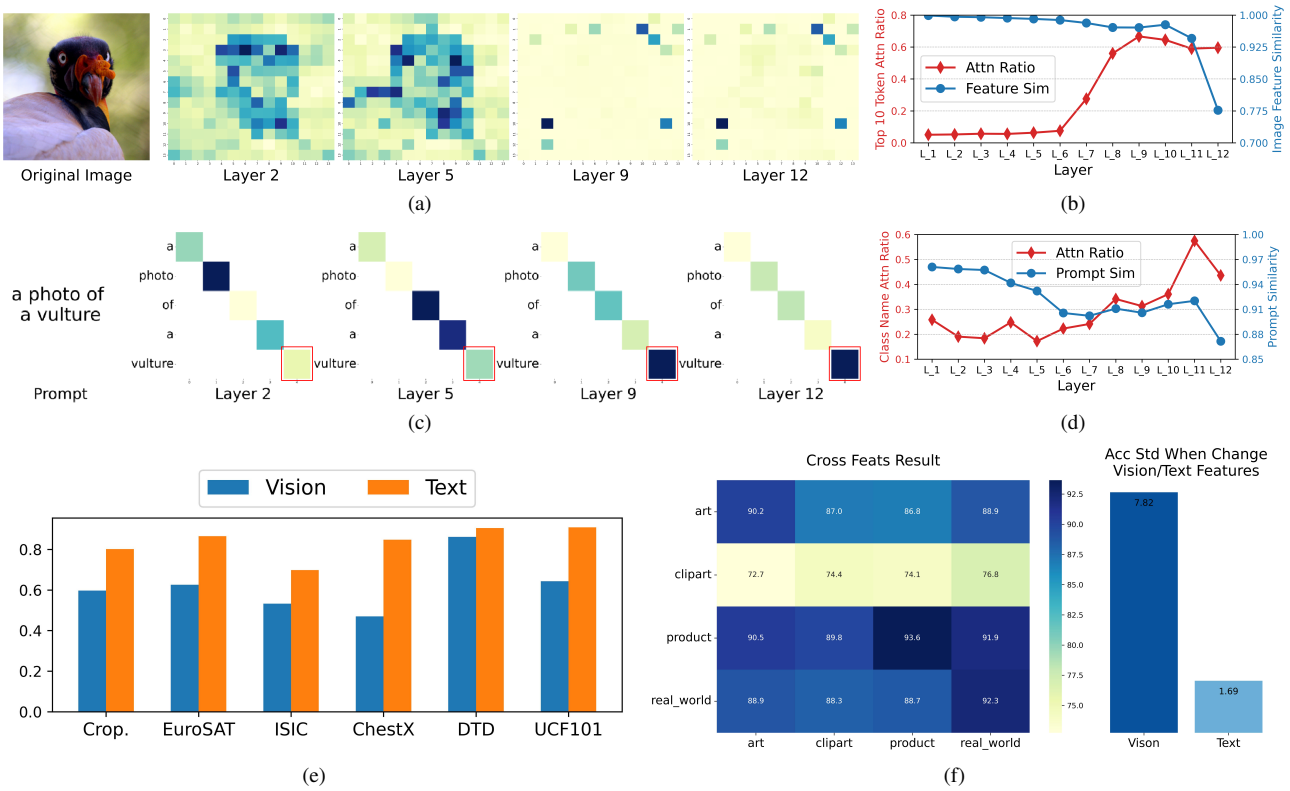



Figure 1. (a) The attention distribution across different layers of the vision encoder. (b) The red line quantitatively shows the attention weight percentage of the top 10 tokens with the highest attention weights in each layer, while the blue line shows the similarity of features extracted by different layers of the vision encoder for visual samples from different domains but of the same category. A higher similarity indicates more domain-independent features. (c) Visualization of the attention weight distribution across different layers of the text encoder. (d) The red line quantitatively shows the attention weight percentage of the category name token in each layer of the text encoder, while the blue line shows the similarity of features extracted by different layers of the text encoder for textual descriptions with different domain information but containing the same category name. A higher similarity indicates more domain-independent features. (e) Domain distance between visual features and text features extracted by the CLIP model when transferred from the ImageNet dataset (source domain) to downstream cross-domain datasets. (f) Cross-domain transfer experiments of visual and textual features from different domains.

similarity between features of the same category but different domains is nearly 1 in the initial layers but significantly decreases in the final layers. This indicates that the vision encoder's "domain-sensitive" approach in the final layers amplifies the impact of domain differences, resulting in a similarity of around 0.77, much lower than the text encoder's 0.9.

## A.3. Experimental Proof

We also conducted specific experiments to demonstrate that, compared to the vision encoder, the text encoder in CLIP contains more domain-independent information, making it easier to transfer to cross-domain tasks.

We utilized the CKA [22] (Centered Kernel Alignment) similarity metric to measure domain distance between visual features and text features extracted by the CLIP model when transferred from the ImageNet dataset (source domain) to downstream cross-domain datasets. A higher CKA similarity indicates a smaller domain distance, suggesting that the encoder contains less domain-specific information. As shown in Figure 1e, the CKA metric for text features is consistently higher than that for visual features across all downstream datasets, indicating that the text encoder extracts more domain-independent text features.

To further demonstrate the robustness of the text encoder against domain information, we conducted cross-domain experiments with visual and text features. We selected the Office-Home dataset [42], which contains category names and visual images of the same categories from four different domains. Using LoRA, we fine-tuned the vision and text branches separately on the datasets from the four domains, recorded the resulting visual and text features, and performed cross-combination CLIP classification tasks. The results, shown in Figure 1f, indicate that when we fix the visual features (vertical axis) and vary the text features (horizontal axis), the performance change is minimal. This implies that the text features extracted from different do-

Table 1. In various backbone versions of the CLIP model, masking a specific layer of the text encoder can lead to significant performance improvements in zero-shot SF-CDFSL tasks.

| Backbone | CropDisease | | EuroSAT | | ISIC | | ChestX | |
|---|---|---|---|---|---|---|---|---|
| | Full | Masked | Full | Masked | Full | Masked | Full | Masked |
| VIT-RN50 | 34.7 | 37.8+3.1 | 38.0 | 39.4+1.4 | 24.7 | 28.1+3.4 | 19.9 | 20.1+0.2 |
| VIT-RN101 | 34.1 | 35.6+1.5 | 38.1 | 40.3+2.2 | 27.6 | 27.7+0.1 | 19.6 | 20.0+0.4 |
| VIT-B16 | 39.8 | 44.0+4.2 | 54.8 | 63.2+8.4 | 27.3 | 29.2+1.9 | 21.8 | 21.9+0.1 |
| VIT-L14 | 52.2 | 55.2+3.0 | 69.7 | 71.2+1.5 | 27 | 29.8+2.8 | 21.7 | 22.5+0.8 |

Table 2. When employing different PEFT methods for 5-way 1-shot SF-CDFSL tasks, training after removing a specific layer of the text encoder can achieve better performance compared to using the complete text encoder.

| Method | CropDisease | | EuroSAT | | ISIC | | ChestX | |
|---|---|---|---|---|---|---|---|---|
| | Full | Masked | Full | Masked | Full | Masked | Full | Masked |
| Lora-Vison | 82.5 | 85.0+2.5 | 81.8 | 82.4+0.6 | 35.4 | 36.3+0.9 | 21.5 | 21.8+0.3 |
| Lora-Text | 83.1 | 83.3+0.2 | 74.1 | 74.6+0.5 | 34.4 | 34.5+0.1 | 21.4 | 22.2+0.8 |
| Lora-Both | 84.3 | 84.6+0.3 | 81.4 | 82.9+1.5 | 33.6 | 34.3+0.7 | 22.4 | 22.7+0.3 |
| Maple | 81.8 | 82.6+0.8 | 76.5 | 77.2+0.7 | 33.6 | 33.8+0.2 | 21.3 | 22.2+0.9 |

Table 3. Further ablation study for V-T Fusion module on 5-way 1-shot task.

| Method | Crop. | EuroSAT | ISIC | ChestX | Avg |
|---|---|---|---|---|---|
| **V-T Fusion** | 87.0 | 85.0 | 38.2 | 22.7 | 58.2 |
| V-T Fusion(R.) | 86.2 | 84.5 | 38.1 | 22.2 | 57.7 |
| V-T Fusion(2D) | 86.9 | 84.7 | 38.4 | 22.7 | 58.2 |
| V Fusion | 86.0 | 83.4 | 37.3 | 21.8 | 57.1 |
| T Fusion | 86.2 | 83.9 | 37.9 | 22.1 | 57.5 |
| V-T Mean | 85.9 | 83.5 | 36.9 | 21.6 | 57.0 |
| T Mean | 85.7 | 83.5 | 37.5 | 21.7 | 57.1 |
| V Mean | 85.8 | 83.6 | 37.3 | 21.5 | 57.0 |

Table 4. Ablation study for SSM network.

| Method | Crop. | Euro. | ISIC | ChestX | Avg |
|---|---|---|---|---|---|
| SSM (OURS) | 87.0 | 85.0 | 38.2 | 22.7 | 58.2 |
| MH Attention | 85.3 | 84.2 | 37.3 | 22.1 | 57.2 |
| RNN | 85.9 | 83.9 | 36.9 | 22.2 | 57.2 |
| LSTM | 85.8 | 84.3 | 37.6 | 22.1 | 57.4 |

mains are relatively consistent and have good transferability. Conversely, when we fix the text features and vary the visual features (viewed from top to bottom), the performance varies significantly, indicating that the visual features learned from different domains are vastly different and less transferable. The right part of Figure 1f illustrates the standard deviation in classification accuracy as visual and text features change.

### A.4. Conclusion and Discussion

Through our analysis, we found that compared to the vision encoder, the text encoder in CLIP: (1) operates in a semantically-driven manner, making it less sensitive to domain information, and (2) its information transfers better to cross-domain tasks. These findings further validate our proposed idea of "teaching the vision encoder to think like the text encoder", emphasizing its significance and necessity in cross-domain scenarios.

## B. Widespread Occurrence of the Lost Layer

We find that the phenomenon of shortcut layers is prevalent in CLIP of different backbones, as shown in Table 1. Moreover, this phenomenon persists even after fine-tuning with a few samples, as seen in Table 2. It is evident that removing a certain layer from the text encoder (Masked) consistently improves classification performance across four cross-domain few-shot datasets.

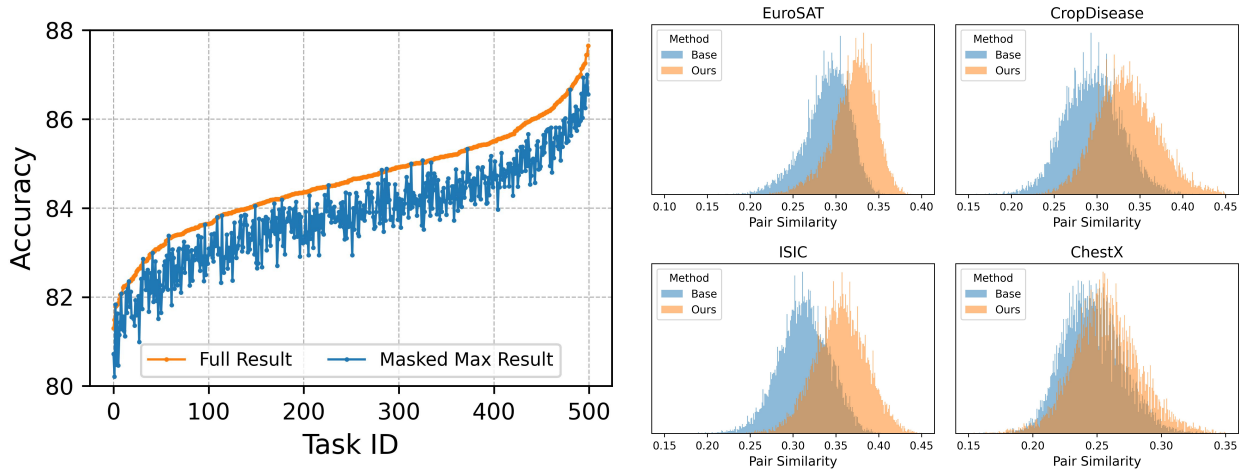
## C. Detailed V-T Fusion Module Ablation

Additionally, we demonstrate the performance of various methods for integrating the outputs of the text encoder and visual encoder to highlight the rationality and effectiveness of our V-T Fusion module. First, we compare the performance of different scanning strategies used in the V-T Fusion module to generate $L_i$, as shown in the first three rows of Table 3. Our approach, V-T Fusion, uses a deep-to-shallow scanning strategy. In contrast, V-T Fusion(R.) represents the reverse, using a shallow-to-deep scanning strategy. V-T Fusion(2D) involves scanning in both directions, with both sequences fed into the subsequent SSM network, and their outputs combined. As shown, V-T Fusion achieves the best performance. Although V-T Fusion(2D) yields similar results, considering computational complexity, we opt for the single-scan V-T Fusion method.

We also compare other methods of integrating visual and text features from each layer, with results shown in rows 4 to 8 of Table 3. T Fusion and V Fusion denote learning with only the text output and visual input, respectively, into the SSM network. V-T Mean, T Mean, and V Mean denote averaging the outputs of all layers from both branches, only the text branch, and only the visual branch, respectively. It is evident that none of these methods outperform our proposed V-T Fusion approach.

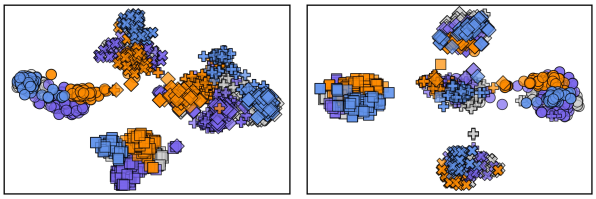
## D. Replacing SSM Network

In the V-T Fusion module, we use the SSM network to serialize and fuse information from the visual-text feature sequences (see Equation 8 in the main text). In this section, we further evaluate the performance of other methods that replace the SSM network. The comparison methods mainly fall into two categories: the first category includes sequential modeling methods such as RNN [54] and LSTM [34]. The second category includes attention mechanism networks like Multi-Head Attention [41]. The results




(a) (b)

Figure 2. (a) In the source domain (ImageNet), performance on sampled sub-tasks containing 100 different categories shows that optimal performance is almost always achieved using the complete text encoder, suggesting that the category composition is not the cause of the shortcut layer. (b) Cosine similarity between the visual features and the corresponding textual features extracted by the model before and after applying our method.



(a) base (b) ours

Figure 3. t-SNE [30] visualization results for the ImageNet-R [13] (5 classes: glodfish, hammerhead shark, tiger shark, stingray and hen). Different colors denote samples from different domain, and different shapes indicate samples of different classes.

are shown in Table 4. It can be seen that none of these methods outperforms the SSM network used by us.

## E. Semantic Information Does Not Cause the Lost Layer

The results in Figure 2(d) of the main text demonstrate that changes in the visual domain are the primary factor causing the lost layer. To further confirm that category information is not the primary factor for the shortcut layer phenomenon, we sampled 100 categories from ImageNet each time to create 500 sub-tasks, each containing different category information. We recorded the classification performance of using the full text encoder (Full Result) and best performance achieved after removing a specific layer from the text encoder (Masked Max Result). The results, as depicted in Figure 2a, indicate that the shortcut layer phenomenon is absent in nearly all sub-tasks on ImageNet.

## F. Specific Effects of the VtT Model

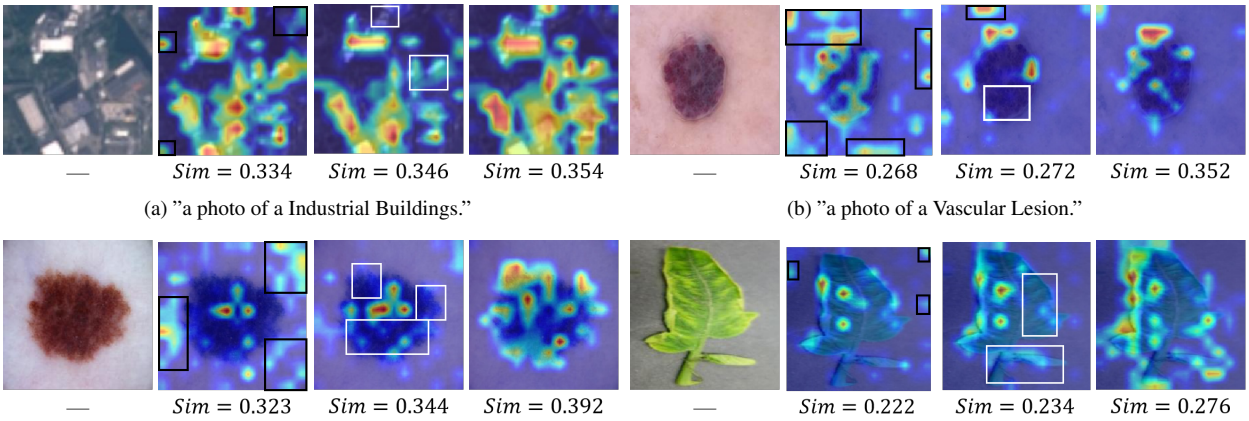
The results in Figure 1(c) of the main text demonstrate that the lost layer disappears when our method is applied, indicating that our approach effectively leverages the information from the text encoder. We further elucidate the impact of our method through additional experiments.

### F.1. Enhancing Alignment

Firstly, we measure the cosine similarity between the visual features and the corresponding textual features extracted by the model before and after applying our method. The results, shown in Figure 2b, indicate that our method increase the similarity between the visual and text features, enhancing cross-modal alignment.

### F.2. More Domain-Independent Features

We also visualized the features extracted by the model for images from five categories within the ImageNet-R dataset, across different domains, before and after applying our method. The results are presented in Figure 3. In these visualizations, colors represent the domains from which the samples originate, and different shapes denote different categories. Before applying our method, it is evident that the categories were difficult to distinguish (points of different shapes are mixed together), and samples from different domains were clearly separated (points of the same shape but different colors are distinctly divided). This indicates that the features extracted by the model were domain-dependent. However, after applying our method, we observed an improvement in classification results (points of different shapes are separated), and the model extracted




(a) "a photo of a Industrial Buildings." (b) "a photo of a Vascular Lesion."

(c) "a photo of a Benign Keratosi." (d) "a photo of a Tomato Yellow Leaf Curl Virus."

Figure 4. The attention maps of the three models. From left to right: the original image, the baseline result, baseline + remove (see Figure 2(c) in the main text) result, and the result of ours. Black boxes highlight areas of incorrect attention, while white boxes highlight areas of missing attention. $Sim$ represents the cosine similarity between the image features and the text features. A higher similarity indicates better alignment.

consistent features for samples from the same category but different domains (points of the same shape but different colors are nearly indistinguishable). This suggests that our method enables the model to extract more domain-independent features.

### F.3. Reclaiming the Lost Layer Information

We illustrate the specific changes in the model before and after leveraging the information from the lost layer through several examples. Figure 4 shows, from left to right: (1) the original image; (2) when using the full text encoder, the model incorrectly focuses on non-semantic parts (highlighted by the black box); (3) after removing the lost layer, these incorrect focuses disappear, but some effective attention areas (highlighted by the white box) are lost; (4) our method eliminates the incorrect focuses while preserving the effective attention areas, achieving an appropriate focus range and better feature alignment.

## G. Extended Results on CDFSL Task

Table 5 is an extended version of Table 2 in the main text. It includes the performance of various models on four CDFSL datasets under different settings. These settings involve different backbones (CLIP [32], SigLIP2 [39], and PE-Core [3]), the use of a source dataset (Source), and whether fine-tuning on the target domain is applied (FT). Specifically, the methods include ATA [43], AFA [16], wave-SAN [8], StyleAdv [9], StyleAdv-FT (fine-tuned StyleAdv), DARA [58], PMF [15], FLoR [67], CD-CLS [68], AttnTemp [66], VDB [52], IM-DCL [47], Step-STP [48], CoOp [62], Tip-Adapter [56], AMU-Tuning [36], LP++ [17], LDC [25], Maple [18], and CLIP-LoRA [53], which are introduced as our competitors.

## H. Few Shot Learning Results on Meta-dataset

We further demonstrate the effectiveness of our method on the Meta-dataset [38]. We compare our method with the most recent CLIP-based methods and present the results in Table 6. For results on datasets ISIC, EuroSAT, ChestX, and CropDisease, please refer to Table 5. This comparison includes both fully fine-tuned methods [35, 45] and prompt-tuning methods [4, 18] of CLIP. As shown in Table 6, our method, when added to CLIP-LoRA [53], effectively enhances its performance, achieving an improvement of approximately 3 points in the 1-shot scenario. Additionally, CLIP-LoRA + OURS achieves the highest average performance to date in both 1-shot and 5-shot tasks.

## I. Detailed Implementation Details.

We adopt the ViT-Base/16 network as the backbone, utilizing parameters pre-trained by CLIP [32], SigLIP2 [39] and PE-Core [3]. In every layer of the visual branch, we incorporate the LoRA [14] structure for fine-tuning. For each layer's LoRA component, we set $r = 16$ and $\alpha = 8$. Following the methodology in [48], we perform data augmentation on the support samples in each episode. Subsequently, we train the model for 250 epochs, using two hyperparameters, $\beta$ and $\lambda$ (refer to Method). And for all settings, we fix $\beta = 7$ and $\lambda = 50$. We evaluate each model using 15 query samples per class, randomly selecting 800 episodes, and report the results (in percentage) with a 95% confidence interval. All training and testing procedures are conducted on a single NVIDIA GeForce RTX 3090.

Table 5. The accuracy(%) of four target domain datasets under 5-way 1-shot and 5-way 5-shot tasks. The use of a source dataset (Source), and whether fine-tuning on the target domain is applied (FT)

| Task | Method | backbone | Source | FT | ISIC | EuroSAT | CropDisease | ChestX | Avg |
|---|---|---|---|---|---|---|---|---|---|
| 5-way 1-shot | ATA [43] | RN10 | Y | - | 33.21±0.40 | 61.35±0.50 | 67.47±0.50 | 22.10±0.20 | 46.03 |
| | AFA [16] | RN10 | Y | - | 33.21±0.30 | 63.12±0.50 | 67.61±0.50 | 22.92±0.20 | 46.72 |
| | wave-SAN [8] | RN10 | Y | - | 33.35±0.71 | 69.64±1.09 | 70.80±1.06 | 22.93±0.49 | 49.18 |
| | StyleAdv [9] | RN10 | Y | - | 33.96±0.57 | 70.94±0.82 | 74.13±0.78 | 22.64±0.35 | 50.42 |
| | ATA-FT [43] | RN10 | Y | Y | 34.94±0.40 | 68.62±0.50 | 75.41±0.50 | 22.15±0.20 | 50.28 |
| | DARA [58] | RN10 | Y | Y | 36.42±0.64 | 67.42±0.80 | 80.74±0.76 | 22.92±0.40 | 51.88 |
| | StyleAdv-FT [9] | RN10 | Y | Y | 35.76±0.52 | 72.92±0.75 | 80.69±0.28 | 22.64±0.35 | 53.00 |
| | PMF [15] | ViT/DINO | Y | Y | 30.36±0.36 | 70.74±0.63 | 80.79±0.62 | 21.73±0.30 | 50.91 |
| | StyleAdv-FT [9] | ViT/DINO | Y | Y | 33.99±0.46 | 74.93±0.58 | 84.11±0.57 | 22.92±0.32 | 53.99 |
| | FLoR [67] | ViT/DINO | Y | Y | 35.49 | 73.09 | 83.55 | 23.26 | 53.85 |
| | CD-CLS [68] | ViT/DINO | Y | Y | 35.56 | 74.97 | 84.53 | 23.39 | 54.62 |
| | AttnTemp [66] | ViT/DINO | Y | Y | 38.05 | 75.09 | 84.78 | 23.63 | 55.39 |
| | FN+VDB [52] | RN18 | - | Y | 32.96±0.57 | 69.67±0.80 | 79.68±0.74 | 22.64±0.41 | 51.24 |
| | IM-DCL [47] | RN10 | - | Y | 38.13±0.57 | 77.14±0.71 | 84.37±0.99 | 23.98±0.79 | 55.91 |
| | StepSTP [48] | ViT/CLIP | - | Y | 32.97±0.27 | 70.01±0.21 | 84.84±0.72 | **22.84±0.95** | 52.68 |
| | CoOp [62] | ViT/CLIP | - | Y | 32.86±0.47 | 72.08±0.66 | 80.50±0.74 | 21.65±0.32 | 51.77 |
| | Tip-Adapter [56] | ViT/CLIP | - | Y | 32.68±0.37 | 75.44±0.51 | 77.15±0.66 | 22.24±0.26 | 51.87 |
| | PromptSRC [19] | ViT/CLIP | - | Y | 31.86±0.57 | 73.44±0.71 | 76.15±0.89 | 21.16±0.36 | 50.65 |
| | PDA [1] | ViT/CLIP | - | Y | 31.45±0.44 | 69.68±0.71 | 79.20±0.83 | 20.66±0.28 | 50.25 |
| | AMU-Tuning [36] | ViT/CLIP | - | Y | 32.29±0.67 | 72.24±0.71 | 80.20±0.86 | 21.56±0.36 | 51.57 |
| | LP++ [17] | ViT/CLIP | - | Y | 33.63±0.41 | 73.05±0.55 | 81.84±0.66 | 21.72±0.42 | 52.56 |
| | LDC [25] | ViT/CLIP | - | Y | 33.72±0.46 | 74.19±0.52 | 83.77±0.81 | 22.12±0.36 | 53.45 |
| | Maple [18] | ViT/CLIP | - | Y | 33.38±0.49 | 76.05±0.63 | 81.78±0.72 | 21.09±0.31 | 53.07 |
| | **Maple + OURS** | ViT/CLIP | - | Y | 34.06±0.53 | 82.25±0.75 | 82.66±0.72 | 21.64±0.34 | 55.15 |
| | CLIP-LoRA-Vision [53] | ViT/CLIP | - | Y | 36.40±0.42 | 81.72±0.52 | 84.22±0.62 | 21.86±0.32 | 55.97 |
| | **CLIP-LoRA-Vision + OURS** | ViT/CLIP | - | Y | **38.20±0.45** | **85.01±0.41** | **87.00±0.53** | 22.70±0.33 | **58.23** |
| | SigLIP2-LoRA [39] | ViT/SigLip2 | - | Y | 33.47 | 74.16 | 87.50 | 21.44 | 54.14 |
| | **SigLIP2-LoRA + OURS** | ViT/SigLip2 | - | Y | **35.34** | **76.10** | **89.72** | **22.00** | **55.79** |
| | PE-Core-LoRA [3] | ViT/PE-Core | - | Y | 40.89 | 84.49 | 91.75 | 22.02 | 59.78 |
| | **PE-Core-LoRA + OURS** | ViT/PE-Core | - | Y | **42.20** | **86.16** | **92.61** | **23.04** | **61.00** |
| 5-way 5-shot | ATA [43] | RN10 | Y | - | 44.91±0.40 | 83.75±0.40 | 90.59±0.30 | 24.32±0.40 | 60.89 |
| | AFA [16] | RN10 | Y | - | 46.01±0.40 | 85.58±0.40 | 88.06±0.30 | 25.02±0.20 | 61.17 |
| | wave-SAN [8] | RN10 | Y | - | 44.93±0.67 | 85.22±0.71 | 89.70±0.64 | 25.63±0.49 | 61.37 |
| | StyleAdv [9] | RN10 | Y | - | 45.77±0.51 | 86.58±0.54 | 93.65±0.39 | 26.07±0.37 | 63.02 |
| | ATA-FT [43] | RN10 | Y | Y | 49.79±0.40 | 89.64±0.30 | 95.44±0.20 | 25.08±0.20 | 64.99 |
| | DARA [58] | RN10 | Y | Y | 56.28±0.66 | 85.84±0.54 | 95.32±0.34 | 27.54±0.42 | 66.25 |
| | StyleAdv-FT [9] | RN10 | Y | Y | 53.05±0.54 | 91.64±0.43 | 96.51±0.28 | 26.24±0.35 | 66.86 |
| | PMF [15] | ViT/DINO | Y | Y | 50.12 | 85.98 | 92.96 | 27.27 | 64.08 |
| | StyleAdv-FT [9] | ViT/DINO | Y | Y | 51.23±0.51 | 90.12±0.33 | 95.99±0.27 | 26.97±0.33 | 66.08 |
| | FLoR [67] | ViT/DINO | Y | Y | 53.06 | 90.75 | 96.47 | 27.02 | 66.83 |
| | CD-CLS [68] | ViT/DINO | Y | Y | 54.69 | 91.53 | 96.27 | 27.66 | 67.54 |
| | AttnTemp [66] | ViT/DINO | Y | Y | 54.91 | 90.82 | 96.66 | 28.03 | 67.61 |
| | FN+VDB [52] | RN18 | - | Y | 47.48±0.59 | 87.31±0.50 | 94.63±0.37 | 25.55±0.43 | 64.74 |
| | IM-DCL [47] | RN10 | - | Y | 52.74±0.69 | 89.47±0.42 | 95.73±0.38 | 28.93±0.41 | 66.72 |
| | StepSTP [48] | ViT/CLIP | - | Y | 52.12±0.36 | 89.40±1.05 | 96.01±0.88 | 26.36±0.97 | 65.97 |
| | CoOp [62] | ViT/CLIP | - | Y | 45.78±0.75 | 85.88±0.49 | 93.31±0.57 | 23.35±0.50 | 62.08 |
| | Tip-Adapter [56] | ViT/CLIP | - | Y | 46.96±0.59 | 87.24±0.33 | 94.19±0.39 | 24.07±0.44 | 63.12 |
| | PromptSRC [19] | ViT/CLIP | - | Y | 46.09±0.48 | 86.54±0.49 | 89.97±0.41 | 23.51±0.47 | 61.52 |
| | PDA [1] | ViT/CLIP | - | Y | 45.19±0.62 | 86.21±0.44 | 92.67±0.39 | 21.87±0.33 | 61.48 |
| | AMU-Tuning [36] | ViT/CLIP | - | Y | 44.60±0.62 | 88.47±0.39 | 94.26±0.52 | 23.34±0.41 | 62.66 |
| | LP++ [17] | ViT/CLIP | - | Y | 48.49±0.44 | 87.48±0.42 | 94.47±0.38 | 23.89±0.29 | 63.58 |
| | LDC [25] | ViT/CLIP | - | Y | 49.70±0.33 | 90.82±0.22 | 96.71±0.34 | 25.89±0.21 | 65.78 |
| | Maple [18] | ViT/CLIP | - | Y | 48.35±0.75 | 89.04±0.52 | 93.50±0.54 | 22.96±0.50 | 63.46 |
| | **Maple + OURS** | ViT/CLIP | - | Y | 49.81±0.78 | 92.24±0.42 | 94.62±0.53 | 24.04±0.50 | 65.18 |
| | CLIP-LoRA-Vision [53] | ViT/CLIP | - | Y | 52.22±0.71 | 93.31±0.47 | 95.88±0.42 | 24.61±0.47 | 66.50 |
| | **CLIP-LoRA-Vision + OURS** | ViT/CLIP | - | Y | **56.20±0.41** | **94.58±0.31** | **97.21±0.35** | **26.42±0.31** | **68.57** |
| | SigLIP2-LoRA [39] | ViT/SigLip2 | - | Y | 51.79 | 91.39 | 96.43 | 24.24 | 65.96 |
| | **SigLIP2-LoRA + OURS** | ViT/SigLip2 | - | Y | **55.11** | **92.70** | **97.63** | **25.54** | **67.75** |
| | PE-Core-LoRA [3] | ViT/PE-Core | - | Y | 58.81 | 94.07 | 97.25 | 24.44 | 68.64 |
| | **PE-Core-LoRA + OURS** | ViT/PE-Core | - | Y | **60.03** | **94.67** | **98.36** | **27.05** | **70.05** |

Table 6. Detailed results for meta-dataset [38] with the ViT-B/16 as visual backbone. Highest value is high lighted inbold.

| Shots | Method | Omniglot | Traffic Signs | MSCOCO | Textures | CUB | Quickdraw | Aircraft | VGG Flower | Fungi | Mini-test | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-shot | WiSE-FT | 83.56 | 60.84 | 67.28 | 63.55 | 81.16 | 62.54 | 62.64 | 73.14 | 59.10 | 93.55 | 70.73 |
| | FD-Align | 83.81 | 57.32 | 65.91 | 66.05 | 82.88 | 64.49 | 62.90 | 79.87 | 57.05 | 95.04 | 71.53 |
| | MaPLe | 77.82 | 56.45 | 68.55 | 66.05 | 94.08 | 72.54 | 79.76 | 98.24 | 55.85 | 97.17 | 76.65 |
| | PromptMargin | 87.01 | 67.24 | 72.86 | 69.25 | **96.97** | 74.84 | 83.94 | **98.43** | 61.16 | **99.17** | 81.08 |
| | CLIP-LoRA | 90.29 | 78.45 | 83.45 | 84.75 | 93.90 | 76.08 | 80.83 | 97.57 | 64.59 | 98.97 | 84.88 |
| | CLIP-LoRA + OURS | **95.01** | **85.40** | **83.89** | **86.27** | 95.97 | **82.37** | **84.02** | 97.94 | **66.68** | 98.99 | **87.64** |
| 5-shot | WiSE-FT | 95.26 | 78.11 | 81.08 | 83.31 | 93.41 | 82.78 | 77.66 | 99.06 | 73.28 | 98.44 | 86.24 |
| | FD-Align | 94.81 | 73.39 | 81.37 | 83.60 | 93.87 | 82.78 | 78.21 | 98.95 | 73.69 | 98.52 | 85.91 |
| | Maple | 96.23 | 85.21 | 75.13 | 88.45 | 97.65 | 85.08 | 87.56 | 99.23 | 79.69 | **99.39** | 89.06 |
| | PromptMargin | 96.37 | 87.55 | 76.68 | 88.71 | 97.12 | 85.21 | 86.53 | 99.27 | 80.91 | 99.19 | 89.75 |
| | CLIP-LoRA | 98.78 | 94.50 | 86.61 | 90.66 | 96.60 | 89.33 | 87.44 | 99.29 | 82.68 | 99.10 | 92.49 |
| | CLIP-LoRA + OURS | **99.40** | **96.73** | **86.67** | **91.06** | **97.83** | **90.11** | **87.63** | **99.51** | **84.18** | 99.20 | **93.22** |

# J. Detailed Related Work

## J.1. Source-Free Cross-domain Few shot Learning

Cross-Domain Few-Shot Learning (CDFSL) aims to train a model on a source domain that can generalize effectively to a target domain with limited examples. Existing methods are typically categorized into two types: meta-learning-based approaches [8, 12, 16, 43] and transfer learning-based approaches [12, 28, 60, 66–68]. Source-Free Cross-Domain Few-Shot Learning (SF-CDFSL) introduces a stronger constraint by making source domain data inaccessible. Current SF-CDFSL methods [48, 52, 64] primarily rely on large models, such as CLIP [32], leveraging their prior knowledge for classification in the target domain. However, no prior work has identified or analyzed the issue of the lost layer when using CLIP for CDFSL tasks.

## J.2. Parameter-Efficient Fine-Tuning

A crucial area of research is the efficient application of Vision-Language Models (VLMs) to downstream tasks. A widely used approach to achieve this is parameter-efficient fine-tuning (PEFT), which involves utilizing only a small number of samples from the target task. PEFT focuses on adjusting a limited subset of the VLM's parameters, enabling the model to adapt to diverse applications without altering all pre-trained parameters. There are three main categories of PEFT methods: prompt learning, adapters, and LoRA (along with its variants). Prompt learning involves converting fixed templates into learnable parameters, as demonstrated in works like CoOp [62], CoCoOp [61], MaPLe [18], PLOT [6], ProGrad [63], PromptSRC [20], KgCoOp [50], PCB [2], DynaPrompt [46], TCP [51], and ATPrompt [27]. Moreover, Customized Ensemble [29] improves performance by combining outputs from multiple models, while PromptKD [26] explores knowledge distillation within prompt learning. Adapter-based methods, such as CLIP-Adapter [10], Tip-Adapter [57], LP++ [17], AMU-Tuning [36], LatHAdapter [59], MMA [49] and LDC [25]. Low-Rank Adaptation (LoRA) [14, 53] fine-tunes the model by adding learnable low-rank matrices while keeping the original parameters fixed. The new weights can be merged with the original ones, and LoRA does not add extra inference time. Various studies have extended LoRA by adapting the rank for each matrix [40, 55], improving its performance [5, 21, 65], or reducing memory usage through quantization [7, 33]

## J.3. Layer Redundancy

Research related to layer redundancy, such as [11, 23] in LLMs, differs from ours. We focus on VLMs, such as CLIP [32], and examine this phenomenon in the SF-CDFSL task setting. Unlike previous studies [24, 31, 37, 44] which focus on in-domain scenarios, consider these layers as redundant, and employ a removal strategy, we study the SF-CDFSL problem on VLM and find the information in these layers actually beneficial for SF-CDFSL tasks, instead of removing layers [23, 31, 37], we reclaim the lost layer and demonstrate that this reclamation is a superior strategy compared to removal. Our work provides a new perspective for analyzing similar issue.

# K. Broader Impact

In this paper, we observe the Lost Layer in CLIP under cross-domain scenarios and reveal that the information within the Lost Layer is actually beneficial for SF-CDFSL tasks. However, changes in the visual domain lead to the under-utilization of this information. We then propose the VtT model to reutilize this information. Extensive experiments validate our rationale and effectiveness. Our research is crucial for future studies on fine-tuning VLM models in cross-domain scenarios. From the perspective of shortcut learning, it highlights the impact of the Lost Layer and discusses how to effectively utilize it. While our method has been evaluated across four distinct target domains, offering a promising initial assessment of its cross-domain applicability, the diversity of these domains may not fully capture all potential real-world scenarios. Future work will focus on expanding our evaluations to include a broader range of target domains to better understand the method's performance in diverse real-world contexts.

# References

[1] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 729–737, 2024. 6

[2] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27004–27014, 2024. 7

[3] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 5, 6

[4] Debarshi Brahma, Anuska Roy, and Soma Biswas. Prompt tuning vision language models with margin regularizer for few-shot learning under distribution shifts. *arXiv preprint arXiv:2505.15506*, 2025. 5

[5] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023. 7

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 7

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[8] Yuqian Fu, Yu Xie, Yanwei Fu, Jingjing Chen, and Yu-Gang Jiang. Wave-san: Wavelet based style augmentation network for cross-domain few-shot learning. *arXiv preprint arXiv:2203.07656*, 2022. 5, 6, 7

[9] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24575–24584, 2023. 5, 6

[10] 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. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 7

[11] Ramón Calvo González, Daniele Paliotta, Matteo Pagliardini, Martin Jaggi, and François Fleuret. Leveraging the true depth of llms. *arXiv preprint arXiv:2502.02790*, 2025. 7

[12] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*, pages 124–141. Springer, 2020. 7

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 4

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5, 7

[15] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 5, 6

[16] Yanxu Hu and Andy J Ma. Adversarial feature augmentation for cross-domain few-shot classification. In *European conference on computer vision*, pages 20–37. Springer, 2022. 5, 6, 7

[17] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23773–23782, 2024. 5, 6, 7

[18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 5, 6, 7

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15190–15200, 2023. 6

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 7

[21] Sanghyeon Kim, Hyunmo Yang, Yunghyun Kim, Youngjoon Hong, and Eunbyung Park. Hydra: Multi-head low-rank adaptation for parameter efficient fine-tuning. *Neural Networks*, page 106414, 2024. 7

[22] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019. 2

[23] Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference?, 2024. *URL https://arxiv. org/abs/2406.19384*. 7

[24] Tim Lawson and Laurence Aitchison. Learning to skip the middle layers of transformers. *arXiv preprint arXiv:2506.21103*, 2025. 7

[25] Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25411–25421, 2025. 5, 6, 7

[26] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 7

[27] Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3618–3627, 2025. 7

[28] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9424–9434, 2021. 7

[29] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023. 7

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 4

[31] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024. 7

[32] 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. 5, 7

[33] Hossein Rajabzadeh, Mojtaba Valipour, Tianshu Zhu, Marzieh Tahaei, Hyock Ju Kwon, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. Qdylora: Quantized dynamic low-rank adaptation for efficient large language model tuning. *arXiv preprint arXiv:2402.10462*, 2024. 7

[34] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 3

[35] Kun Song, Huimin Ma, Bochao Zou, Huishuai Zhang, and Weiran Huang. Fd-align: Feature discrimination alignment for fine-tuning pre-trained models in few-shot learning. *Advances in Neural Information Processing Systems*, 36:43579–43592, 2023. 5

[36] Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Amu-tuning: Effective logit bias for clip-based few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23323–23333, 2024. 5, 6, 7

[37] Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. *arXiv preprint arXiv:2505.19536*, 2025. 7

[38] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019. 5, 7

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5, 6

[40] Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*, 2022. 7

[41] 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

[42] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 2

[43] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. *arXiv preprint arXiv:2104.14385*, 2021. 5, 6, 7

[44] Yizhou Wang, Song Mao, Yang Chen, Yufan Shen, Yinqiao Yan, Pinlong Cai, Ding Wang, Guohang Yan, Zhi Yu, Xuming Hu, et al. Investigating redundancy in multimodal large language models with multiple vision encoders. *arXiv preprint arXiv:2507.03262*, 2025. 7

[45] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 5

[46] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 7

[47] Huali Xu, Li Liu, Shuaifeng Zhi, Shaojing Fu, Zhuo Su, Ming-Ming Cheng, and Yongxiang Liu. Enhancing information maximization with distance-aware contrastive learning for source-free cross-domain few-shot learning. *IEEE Transactions on Image Processing*, 2024. 5, 6

[48] Huali Xu, Yongxiang Liu, Li Liu, Shuaifeng Zhi, Shuzhou Sun, Tianpeng Liu, and MingMing Cheng. Step-wise distribution alignment guided style prompt tuning for source-free cross-domain few-shot learning. *arXiv preprint arXiv:2411.10070*, 2024. 5, 6, 7

[49] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on*

*Computer Vision and Pattern Recognition (CVPR)*, pages 23826–23837, 2024. 7

[50] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 7

[51] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 7

[52] Moslem Yazdanpanah and Parham Moradi. Visual domain bridge: A source-free domain adaptation for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2868–2877, 2022. 5, 6, 7

[53] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 5, 6, 7

[54] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 3

[55] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023. 7

[56] 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. 5, 6

[57] 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 *European conference on computer vision*, pages 493–510. Springer, 2022. 7

[58] Yifan Zhao, Tong Zhang, Jia Li, and Yonghong Tian. Dual adaptive representation alignment for cross-domain few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11720–11732, 2023. 5, 6

[59] Yumiao Zhao, Bo Jiang, Yuhe Ding, Xiao Wang, Jin Tang, and Bin Luo. Fine-grained vlm fine-tuning via latent hierarchical adapter learning. *arXiv preprint arXiv:2508.11176*, 2025. 7

[60] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20061–20070, 2023. 7

[61] 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. 7

[62] 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. 5, 6, 7

[63] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 7

[64] Linhai Zhuo, Zheng Wang, Yuqian Fu, and Tianwen Qian. Prompt as free lunch: Enhancing diversity in source-free cross-domain few-shot learning through semantic-guided prompting. *arXiv preprint arXiv:2412.00767*, 2024. 7

[65] Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023. 7

[66] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 5, 6, 7

[67] Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23575–23584, 2024. 5, 6

[68] Yixiong Zou, Shuai Yi, Yuhua Li, and Ruixuan Li. A closer look at the cls token for cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37: 85523–85545, 2025. 5, 6, 7