Cross-Domain Few-Shot Learning via Multi-View Collaborative Optimization with Vision-Language Models

Dexia Chen[1,2], Wentao Zhang[1,2], Qianjie Zhu[3], Ping Hu[4], Weibing Li[1], Tong Zhang[5], Ruixuan Wang[1,2,5]*

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China [3]School of Computer, Electronics and Information, Guangxi University, Nanning, China
[4]School of Information Science and Engineering, Xinjiang University, Urumqi, China
[5]Peng Cheng Laboratory, Shenzhen, China
chendx27@mail2.sysu.edu.cn, wangruix5@mail.sysu.edu.cn

Abstract

Vision-language models (VLMs) pre-trained on natural image and language data, such as CLIP, have exhibited significant potential in few-shot image recognition tasks, leading to development of various efficient transfer learning methods. These methods exploit inherent pre-learned knowledge in VLMs and have achieved strong performance on standard image datasets. However, their effectiveness is often limited when confronted with cross-domain tasks where imaging domains differ from natural images. To address this limitation, we propose **Co**nsistency-guided **Mu**lti-view **Co**llaborative Optimization (**CoMuCo**), a novel fine-tuning strategy for VLMs. This strategy employs two functionally complementary expert modules to extract multi-view features, while incorporating prior knowledge-based consistency constraints and information geometry-based consensus mechanisms to enhance the robustness of feature learning. Additionally, a new cross-domain few-shot benchmark is established to help comprehensively evaluate methods on imaging domains distinct from natural images. Extensive empirical evaluations on both existing and newly proposed benchmarks suggest CoMuCo consistently outperforms current methods.

Introduction

Current deep learning methods often require vast amounts of labeled data which may be prohibitively expensive and difficult to obtain in domains such as rare disease diagnosis and industrial defect detection. To address this challenge, various few-shot learning techniques (Gharoun et al. 2024; Vettoruzzo et al. 2024) have been developed to enable models to learn effectively from limited data. While previous methods are effective in some scenarios, they generally suffer from limited generalization ability.

In recent years, the emergence of pre-trained vision-language models (VLMs) (Gao et al. 2024b; Huang et al. 2024b; Jia et al. 2021; Wei, Pan, and Owens 2024; Li et al. 2023), especially CLIP (Radford et al. 2021), offer new solutions for few-shot learning. An image encoder and a text encoder are commonly included in these models to align image features and text embeddings. The alignment is facilitated by enhancing the cosine similarity of the corresponding image-text pairs. After being pre-trained on large amounts of data, these models acquire strong semantic understanding and effective zero-shot image recognition abilities. The powerful feature representation ability and open-vocabulary recognition ability effectively alleviate the problems faced by few-shot learning. To enable efficient transfer learning of pre-trained VLMs in few-shot scenarios, a series of fine-tuning techniques have been proposed, *e.g.*, methods based on prompt tuning (Zhou et al. 2022b,a; Chen et al. 2023; Khattak et al. 2023; Zhu et al. 2023) or adapter tuning (Gao et al. 2024a; Zhang et al. 2022; Huang et al. 2024a).

As shown in Fig. 1, since these methods are initially designed to leverage intrinsic knowledge in VLMs, their performance is inherently dependent on the alignment between pre-learned knowledge in VLMs and the to-be-learned

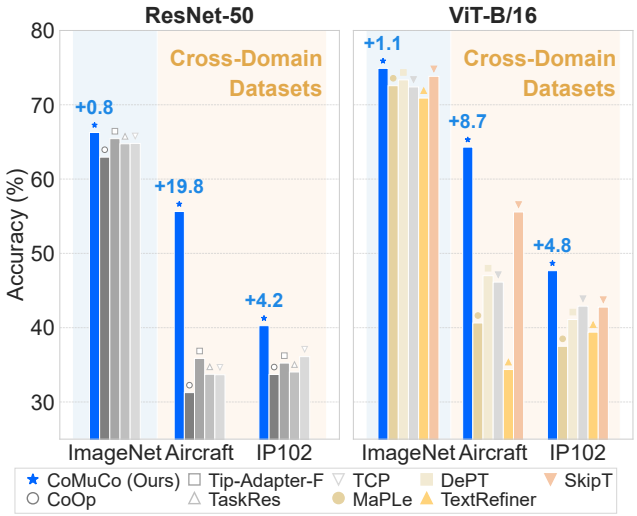

Figure 1: Accuracy comparison on in-domain (ImageNet) and cross-domain (Aircraft & IP102) datasets under 16-shot setting using ResNet-50 (left) and ViT-B/16 (right). The '+' marks improvement of our method over the best baseline.

*Corresponding author

knowledge in the downstream task. Strong alignment typically results in better performance, whereas in cross-domain settings with substantial domain shifts, the reduced alignment significantly limits their effectiveness. Furthermore, simple fine-tuning strategies may only assimilate a subset of discriminative features present in the training dataset, while comprehensive discriminative characteristics remain inadequately captured (Allen-Zhu and Li 2023), thus constraining model performance especially on cross-domain datasets.

To address the challenges of applying VLMs to few-shot learning in cross-domain scenarios and enhance the extraction of discriminative features with VLMs, we propose CoMuCo, a consistency-guided multi-view collaborative optimization framework. By establishing diverse learning preferences, this framework effectively facilitates the acquisition of multi-view features. Specifically, our framework consists of two functionally complementary expert modules, i.e., a Feature Integrator, which extracts and refines knowledge relevant to cross-domain classification from pre-trained models, and a Feature Refiner, which actively learns task-specific features from cross-domain data. Both modules are governed by a consensus constraint based on information geometry theory, which promotes the learning of mutually compatible and robust feature representations. Additionally, a prior consistency constraint is implemented to preserve logits consistency across the fine-tuning process by constraining logits deviations to follow a Laplacian prior distribution, mitigating catastrophic forgetting of general knowledge.

Furthermore, recent efficient transfer learning methods commonly adopt the CLIP Benchmark[1] for performance evaluation. However, many of its datasets have substantial domain overlap with CLIP's pretraining corpus. While datasets such as DTD (Cimpoi et al. 2014) and EuroSAT (Helber et al. 2019) provide some cross-domain variation, their diversity remains limited. To enable a more comprehensive evaluation across distinct visual domains, we curated a set of datasets that differ significantly from natural images and proposed a new cross-domain few-shot benchmark. Our method was evaluated on both the CLIP Benchmark and our proposed benchmark, consistently achieving state-of-the-art performance. The main contributions of this study are summarized below.

- A novel **Co**nsistency-guided **Mu**lti-view **Co**llaborative Optimization (**CoMuCo**) is proposed to effectively learn knowledge from downstream task data in few-shot scenarios, especially on cross-domain tasks.
- A prior consistency constraint is proposed, achieving the preservation of prior knowledge by constraining logits drift to satisfy the Laplace distribution.
- A novel multi-view geodesic consensus mechanism is proposed to facilitate the learning of more robust discriminative representations.
- Extensive empirical evaluations were performed on both the existing benchmark and the cross-domain benchmark, with SOTA performance achieved by CoMuCo.

[1]The CLIP Benchmark (Zhou et al. 2022b) consists of 11 widely used datasets for evaluating few-shot learning.

Related Work

Vision-Language Models The recently developed pretrained VLMs (Gao et al. 2024b; Huang et al. 2024b; Jia et al. 2021; Wei, Pan, and Owens 2024; Li et al. 2023; Zhai et al. 2023; Tschannen et al. 2025; Sun et al. 2023; Cherti et al. 2023; Xu et al. 2024) have been widely applied to few-shot learning. Among these VLMs, CLIP (Radford et al. 2021) has garnered significant attention for its generalization capability for downstream tasks. CLIP is pre-trained on a vast number of image-text pairs, learning the semantic relationships between images and text, which enables it to extract visual features rich in semantic information. This powerful pretraining capability renders CLIP a promising base model for transfer learning.

Efficient Transfer Learning To fully leverage the pretrained knowledge of VLMs in few-shot learning scenarios, a series of efficient transfer learning methods have been developed (Zhou et al. 2022b; Chen et al. 2023; Guo et al. 2023; Huang et al. 2024a; Khattak et al. 2023; Yao, Zhang, and Xu 2024; Zhang et al. 2024; Zhou et al. 2022a; Zhang et al. 2022; Yu et al. 2023). These methods can be primarily divided into two groups, prompt-tuning (Chen et al. 2023; Khattak et al. 2023; Yao, Zhang, and Xu 2024; Zhou et al. 2022b,a; Zhang et al. 2024) and adapter-tuning (Huang et al. 2024a; Zhang et al. 2022; Gao et al. 2024a; Yu et al. 2023). Prompt-tuning methods like CoOp (Zhou et al. 2022b) use learnable prompts in the text encoder, while adapter-tuning methods such as CLIP-Adapter (Gao et al. 2024a) add lightweight modules to encoders. Although these methods demonstrate robust performance, their model adaptations are predominantly constrained to input tokens and output features, thereby limiting their effectiveness in cross-domain scenarios. By introducing two complementary expert modules alongside prior constraints and information-geometric consensus constraints, our method markedly mitigates this issue and improves the model's generalization ability.

Method

Preliminary

Our method is built upon CLIP, which consists of an image and a text encoder. To perform zero-shot classification for an image into one of C classes , with each class associated with a textual sentence, CLIP firstly extracts the image feature $\mathbf{z} \in \mathbb{R}^d$ and text embeddings $\mathbf{t} \in \mathbb{R}^{C \times d}$ via the image encoder and the text encoder, respectively, where d represents the feature dimension. Then, the similarity between the image feature and each text embedding is computed, resulting in a similarity vector $\mathbf{s} = sim(\mathbf{z}, \mathbf{t}) \in \mathbb{R}^C$, where $sim(\cdot, \cdot)$ represents the cosine similarity measurement. Consequently, the probability output is given by $\mathbf{p} = softmax(\mathbf{s}/\tau)$, where τ is the temperature coefficient. The class with the highest similarity score is selected as the prediction.

Overview

To facilitate the comprehensive learning of discriminative features from downstream task data, CoMuCo, a consistency-guided multi-view collaborative optimization

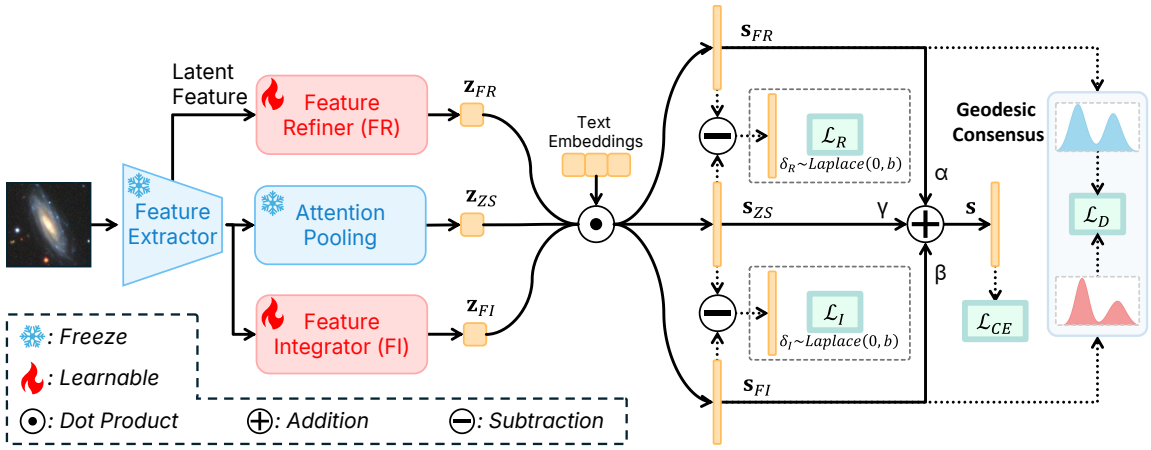

Figure 2: Overview of CoMuCo. The proposed framework is constructed around two core modules: the Feature Integrator (FI) and the Feature Refiner (FR). A consensus constraint aligns FI and FR for enhanced feature learning, while a prior consistency constraint regulates logit deviation to preserve zero-shot knowledge. Feature vectors $\mathbf{z}_{FI}$, $\mathbf{z}_{FR}$, and $\mathbf{z}_{ZS}$ are extracted from FI, FR, and frozen CLIP modules, with corresponding logits $\mathbf{s}_{FI}$, $\mathbf{s}_{FR}$, and $\mathbf{s}_{ZS}$ obtained through class text embedding alignment. "Attention Pooling" is configured at the final transformer block in the ViT architecture.

framework, is introduced. As illustrated in Fig. 2, this framework facilitates the learning of features from different perspectives by incorporating two functionally complementary modules: the Feature Integrator (FI) and the Feature Refiner (FR). Specifically, FI is designed to extract and refine knowledge pre-learned by the VLM that remains relevant to the downstream classification task, whereas FR actively learns novel task-specific knowledge from downstream data. To prevent excessive forgetting of the pre-trained model's general knowledge within each module, a prior consistency constraint is enforced in the logit space. Specifically, the deviation between each module's logits and those of zero-shot CLIP is encouraged to follow a zero-mean Laplacian distribution, thereby promoting minimal modifications to the logits and ensuring that pre-learned knowledge is preserved throughout the training process of each module. Furthermore, to enhance the robustness of feature learning, we propose a multi-view consensus mechanism grounded in information geometry theory, which approximately minimizes the squared geodesic distance between probability distributions from different perspectives on a statistical manifold, thereby fostering compatibility across views and promoting more robust feature learning.

Dual-Expert Framework

To comprehensively learn discriminative features for downstream tasks, two structurally decoupled and functionally complementary expert modules are introduced. As shown in Fig. 3, these modules are designed to implicitly capture features from different perspectives of the downstream task data through distinct fine-tuning strategies:

- **Feature Integrator** (Invariant Expert) is designed to preserve existing knowledge from the VLM through conservative parameter modifications, focusing updates solely on the last module.

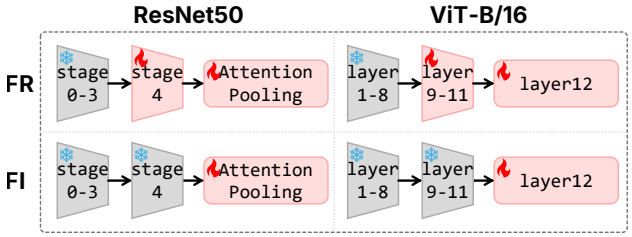

Figure 3: Illustrations of FI and FR under different architectures. They are initialized with the same architecture and weights as the pre-trained model, with intermediate results from frozen CLIP forward propagation being reused as FI and FR inputs to reduce computation.

- **Feature Refiner** (Adaptive Expert) captures novel patterns induced by downstream task data, employing fine-tuning of deeper network layers to achieve superior adaptation to domain-specific data distributions.

By learning features from different perspectives, this dual-expert framework enables effective information extraction.

Prior Consistency Constraint

Since substantial prior knowledge is embedded in pre-trained models during their pre-training phase, its complete erasure during fine-tuning is considered harmful. To address catastrophic forgetting of prior knowledge triggered by fine-tuning, a prior consistency constraint is implemented to reduce the number of elements modified in the fine-tuning branch's logits when compared to original CLIP logits, thereby necessitating sparsity in the logits offset. This sparsity is enforced by requiring that the logits offset follows a sparse prior distribution, modeled using zero-mean Laplacian distribution.

We consider the deviation between the expected logits $\mathbf{s}$ produced by each expert and the zero-shot logits $\mathbf{s}_{\text{ZS}}$ from frozen CLIP. Let $\boldsymbol{\delta} = \mathbf{s} - \mathbf{s}_{ZS}$ denote the deviation vector in the logit space.

Definition 1 (Laplace Prior on Logits Deviation) *Each component δ_c of the deviation vector $\boldsymbol{\delta}$ is independently drawn from a zero-mean Laplace distribution:*

$$p(\delta_c) = \frac{1}{2b} \exp\left(-\frac{|\delta_c|}{b}\right), \tag{1}$$

where b is a scale hyper-parameter.

Then, we demonstrate the negative log-prior can be interpreted as an L_1 regularization:

Theorem 1 (Laplace Prior Equivalence) *Imposing an L_1 regularization on the deviation vector $\boldsymbol{\delta}$ is equivalent to assuming an independent Laplace prior:*

$$-\log p(\boldsymbol{\delta}) \propto \frac{1}{b}\|\boldsymbol{\delta}\|_1. \tag{2}$$

Proof is provided in Appendix A.1

Based on this theory, the resulting prior consistency losses are defined as:

$$\mathcal{L}_R = \|\mathbf{s}_{FR} - \mathbf{s}_{ZS}\|_1, \quad \mathcal{L}_I = \|\mathbf{s}_{FI} - \mathbf{s}_{ZS}\|_1. \tag{3}$$

This Laplace-based regularization enforces sparse adaptation during fine-tuning. The model is encouraged to modify its predictions only for a small set of classes while maintaining consistency with the powerful prior for most categories. This mechanism enables local adaptation without erasing general visual knowledge.

Multi-view Consensus Constraint

As the features derived from different perspectives are intended to collaboratively address the same classification task, a consensus between their predictions is expected rather than mutual contradiction. To facilitate this, the expected predictive distributions generated by the two expert branches are aligned by minimizing Jeffreys divergence on the statistical manifold $\mathcal{M}_P$ of output probabilities.

Definition 2 (Jeffreys Divergence) *The Jeffreys divergence is defined as the symmetric form of KL divergence:*

$$D_J(\mathbf{p}\|\mathbf{q}) = D_{KL}(\mathbf{p}\|\mathbf{q}) + D_{KL}(\mathbf{q}\|\mathbf{p}). \tag{4}$$

The consensus loss is defined via the Jeffreys divergence:

$$\mathcal{L}_D = \frac{1}{2} D_J(\mathbf{p}_{FR}, \mathbf{p}_{FI}). \tag{5}$$

This formulation is grounded in the intrinsic geometry of statistical manifolds. Specifically, we demonstrate that the Jeffreys divergence admits a higher-order approximation to the squared geodesic distance between two probability distributions on such a manifold:

Theorem 2 (Geodesic Divergence Approximation) *Let $\mathcal{M}$ be a statistical manifold, where points on $\mathcal{M}$ are parameterized by a local coordinate system π. For any two points P and Q on $\mathcal{M}$, with coordinates π_P and π_Q respectively, the squared geodesic distance $d^2(P, Q)$ connecting them can be approximated to fourth order through the Jeffreys divergence:*

$$D_J(P, Q) = d^2(P, Q) + O(\|\pi_Q - \pi_P\|^4), \tag{6}$$

where $\|\pi_Q - \pi_P\|$ represents the norm of the parametric coordinate difference between the two points. (Proof is provided in Appendix A.2)

This geometric perspective suggests that minimizing $\mathcal{L}_D$ effectively reduces the geodesic distance between the prediction distributions of the two experts on the statistical manifold, thereby encouraging predictive consensus and enhances the robustness of each expert.

Training and Inference

For each image $\mathbf{x}_i$, the fused logits, used for both training and inference, are obtained by aggregating the logits from FR, FI, and the frozen CLIP, which are denoted as $\mathbf{s}_{FR}(\mathbf{x}_i)$, $\mathbf{s}_{FI}(\mathbf{x}_i)$, and $\mathbf{s}_{ZS}(\mathbf{x}_i)$, respectively:

$$\mathbf{s}_i = \alpha \cdot \mathbf{s}_{FR}(\mathbf{x}_i) + \beta \cdot \mathbf{s}_{FI}(\mathbf{x}_i) + \gamma \cdot \mathbf{s}_{ZS}(\mathbf{x}_i), \tag{7}$$

The weights α, β, and γ are expert coefficients, with $\gamma = 1 - \alpha - \beta$. The cross-entropy loss over the training set serves as the expected likelihood objective:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{y}_i|\mathbf{s}_i). \tag{8}$$

where $p(y_i|\mathbf{s}_i)$ represents the predicted probability for the ground-truth label y_i given the fused logits $\mathbf{s}_i$.

The complete training objective integrates multiple regularization terms:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}}}_{\text{Joint Likelihood}} + \underbrace{\lambda_1\mathcal{L}_R + \lambda_2\mathcal{L}_I}_{\text{Prior Regularization}} + \underbrace{\lambda_3\mathcal{L}_D}_{\text{Consensus Regularization}}. \tag{9}$$

Experiments

Experimental Settings

Datasets To address the CLIP Benchmark's limitations in evaluating model performance across various visual domains, a novel benchmark is introduced, which incorporates seven diverse image datasets, *i.e.*, Skin40 (Yang et al. 2023) with 40 classes of skin disease, TCGA12 (Chen et al. 2022) with 12 classes of tissue pathology, RFMiD12 (Panchal et al. 2023) with 12 classes of fundus, NWPU-RESISC45 (Cheng, Han, and Lu 2017) with 45 classes of remote sensing, NEU-CLS (Song and Yan 2013) with 6 classes of hot-rolled steel defect, IP102 (Wu et al. 2019) with 102 classes of crop pest and disease, and Galaxy10 DECaLS (Leung and Bovy 2019) with 10 classes of galaxy. Collectively, these datasets cover a wide range of fields, enabling more comprehensive assessment of model adaptability.

In addition, the CLIP Benchmark was still used to thoroughly evaluate the model's performance, which includes 11 datasets, *i.e.*, ImageNet (Deng et al. 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2007), Food101 (Bossard, Guillaumin, and Gool 2014), DTD (Cimpoi et al. 2014), EuroSAT (Helber et al. 2019), FGVCAircraft (Maji et al.

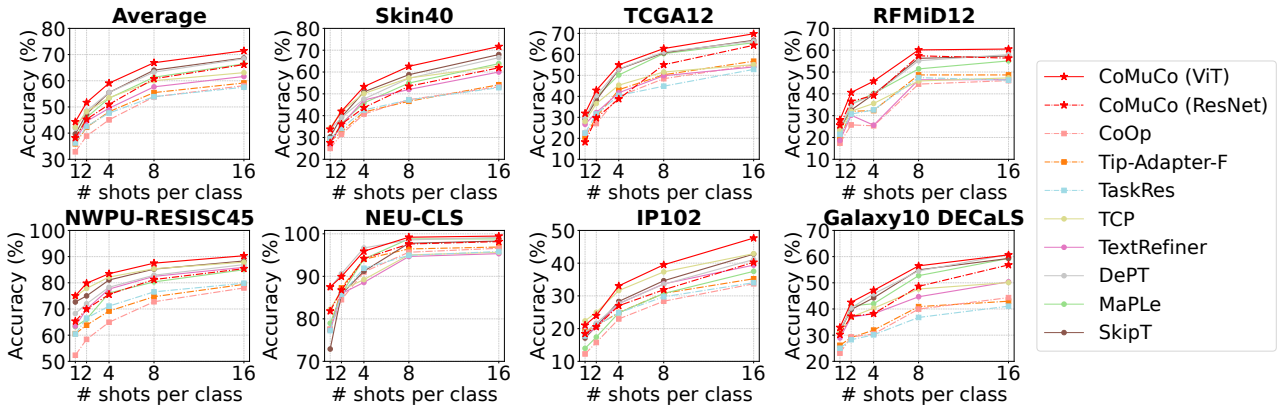

Figure 4: Performance comparison on the cross-domain benchmark. Dashed lines for ResNet50 and solid lines for ViT-B/16.

2013), Flowers102 (Nilsback and Zisserman 2008), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), SUN397 (Xiao et al. 2010), and UCF101 (Soomro, Zamir, and Shah 2012). Moreover, ImageNet-Sketch (Wang et al. 2019) and ImageNet-V2 (Recht et al. 2019) are incorporated to assess the model's domain generalization capability.

Implementation Following previous studies (Yu et al. 2023; Zhou et al. 2022b), we trained models under K-shot settings ($K = 1, 2, 4, 8, 16$) with K images per class and evaluated on the full test set. Unless otherwise stated, ResNet-50 was used as the visual backbone, and pre-defined text templates (Yu et al. 2023) were used for text encoding. Training was conducted using SGD with cosine learning rate decay for 50 epochs (300 for cross-domain settings), starting with a warm-up from $1e{-5}$ to 0.002 in the first epoch. The default batch size was 32. Data augmentations from CoOp (Zhou et al. 2022b) (random crop and flip) were applied. Hyperparameters were fixed as $\alpha = \beta = 0.2$, and $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$. All results were averaged over three runs with different seeds.

Baselines To validate the effectiveness of our method, comparisons were made with SOTA efficient transfer learning methods, including CoOp (Zhou et al. 2022b), Tip-Adapter-F (Zhang et al. 2022), TaskRes (Yu et al. 2023), MaPLe (Khattak et al. 2023), TCP (Yao, Zhang, and Xu 2024), DePT (Zhang et al. 2024), TextRefiner (Xie et al. 2024) and SkipT (Wu et al. 2025).

Efficacy of the Proposed Method

Results on Cross-Domain Few-Shot Benchmark Our method was first assessed on the cross-domain few-shot benchmark. To confirm the challenges for cross-domain few-shot recognition, zero-shot CLIP as the naive baseline was evaluated on the benchamrk, which revealed that the frozen pre-trained CLIP model is unable to perform effective classification in cross-domain tasks (Appendix C.2).

As shown in Fig. 4, our method CoMuCo consistently outperforms all baselines. With ResNet50 as the visual encoder, it achieves superior results, particularly with a slightly larger number of training images, surpassing the strongest baseline by 5.27% and 7.03% under the [8, 16]-shot settings.

Components						Datasets		
$\mathcal{L}_{CE}$	FI	FR	$\mathcal{L}_I$	$\mathcal{L}_R$	$\mathcal{L}_D$	ImageNet	Stanford Cars	Galaxy
						58.18	55.61	13.90
✓	✓					65.40	79.27	51.23
✓	✓		✓			65.50	79.53	50.10
✓		✓				64.33	81.27	53.30
✓		✓		✓		65.10	83.53	56.23
✓	✓	✓				64.47	81.23	53.26
✓	✓	✓			✓	65.73	80.23	54.43
✓	✓	✓	✓	✓		65.63	83.87	55.67
✓	✓	✓	✓	✓	✓	**66.27**	**85.07**	**56.83**

Table 1: Ablation study of our method on 3 representative datasets under the 16-shot setting.

When ViT-B/16 is used, CoMuCo exhibits superior performance across all competing approaches, yielding improvements of 2.03%, 3.23%, 3.59%, 2.83%, and 2.78% over the best baseline under the [1, 2, 4, 8, 16]-shot settings. Notably, even with ResNet50, the performance of CoMuCo remains comparable to certain ViT-B/16-based baselines, exceeding TextRefiner and TCP while matching MaPLe under [8, 16]-shot settings. These results support that our method can more effectively learn knowledge from the limited training data when the imaging modality of the downstream task is significantly different from those used in CLIP pre-training.

Results on CLIP Benchmark As shown in Fig. 5 (top-left subfigure), on the widely used CLIP Benchmark for few-shot learning, our method consistently achieves superior average performance compared to SOTA methods across all the few-shot settings. As the number of training samples increases, the performance gap progressively widens. Under the [1, 4, 16]-shot settings, our method outperforms the best baseline by 1.48%, 2.77%, and 4.65% on ResNet50, and by 0.25%, 0.54%, and 1.70% on ViT-B/16, respectively. Notably, our method excels on the two fine-grained datasets StanfordCars and FGVCAircraft, and on the texture dataset DTD. Classification of fine-grained classes often requires more specialized knowledge (Gao et al. 2024a), as does clas-

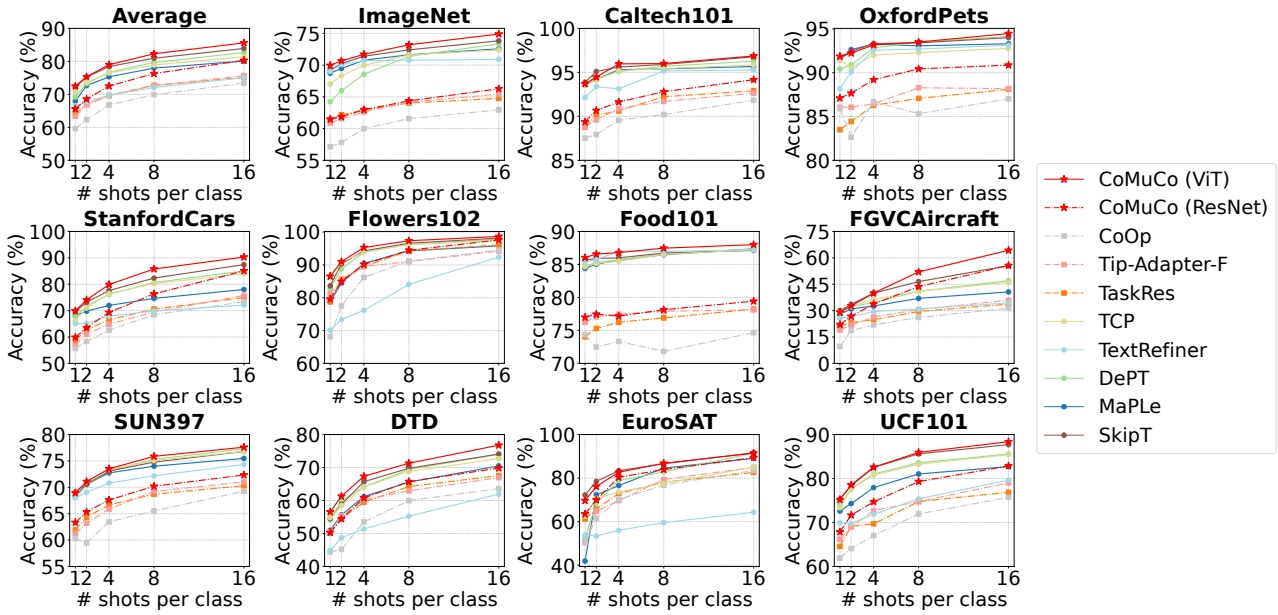

Figure 5: Performance comparison on the CLIP Benchmark. Dashed lines for ResNet50 and solid lines for ViT-B/16.

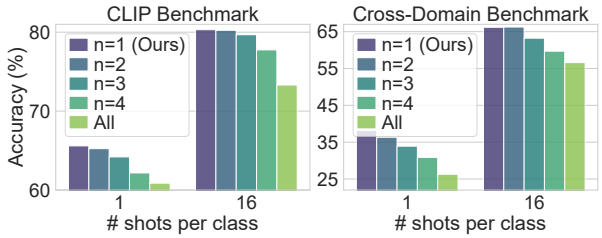

Figure 6: Average results on CLIP and cross-domain benchmarks for different FR fine-tuning strategies. Here, 'n' indicates the count of fine-tuned layers closest to output.

sification in the texture classes, which is less frequently encountered during CLIP's pre-training phase. In this case, our method enables the model to effectively learn from new data, resulting in superior performance. Conversely, on the ImageNet, Flowers102, and Food101 datasets, our method performs on par with competing methods, as CLIP's pre-trained knowledge is sufficient to acquire substantial task-specific knowledge with minimal additional learning. To conclude, our method exhibits marked superiority in standard few-shot learning tasks, particularly in the domains of fine-grained classification and cross-domain data recognition.

Ablation Study

Ablation Study on Model Components Ablation studies were performed on ImageNet, Stanford Cars, and Galaxy10 DECaLS ('Galaxy') under the 16-shot setting to evaluate the impact of key components in the proposed framework. These datasets respectively represent natural image classification, fine-grained classification, and cross-domain classification tasks, allowing for a comprehensive evaluation. As shown in Tab. 1, FR is more advantageous for fine-grained and cross-domain classification, whereas FI excels in natural image classification (row 2 vs. 4). This discrepancy arises because FI retains most pre-learned knowledge and efficient learning from natural image data is achieved through the correction of attention pooling. However, in tasks involving fine-grained distinctions or large domain shifts, FR demonstrates superior results by enabling more comprehensive feature refinement and enhanced knowledge adaptation. Combining FR and FI (row 6) yields intermediate performance.

Adding Prior Consistency Constraint improves results by 1.99% for FR (row 4 vs. 5) and 2.07% for the dual-expert integration (row 6 vs. 8), which supports that the prior constraint alleviates the overfitting issue by preventing excessive forgetting of pre-learned knowledge in CLIP. Additionally, inclusion of the multi-view consensus constraint enhances the performance of dual-expert integration (row 6 vs. row7). When prior constraint was employed, the consensus mechanism demonstrated enhanced performance gains (rows 7 & 8 vs. row 9), indicating that prior constraints assist the consensus mechanism in improving the model's generalization capability. These ablation results confirm the effectiveness of all CoMuCo components.

Impact of Fine-tuning Layer Configurations in FR The impact of fine-tuning depths in the FR was evaluated. Fig. 6 shows that, performance generally declines with more fine-tuned layers, especially with fewer or cross-domain samples. Under the 16-shot setting, fine-tuning 3 layers or the entire visual encoder reduced average performance by 0.64% and 7.01% on the CLIP benchmark and by 3.08% and 9.18% on the cross-domain benchmark, relative to tuning the last layer only. In the 1-shot cross-domain case, declines extended to 3.67% and 11.73%. These findings indicate that early layers are more prone to overfitting when data is scarce, and that

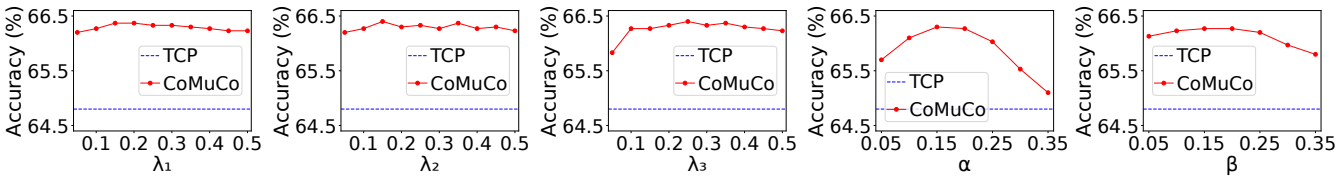

Figure 7: Sensitivity study on ImageNet. TCP is a representative strong baseline.

Method	Visual Backbone	Source	Target	
		ImageNet	-V2	-Sketch
Zero-Shot CLIP	ResNet-50	58.18	51.34	33.32
Linear Probe CLIP		55.87	45.97	19.07
CoOp		62.95	55.11	32.74
TCP		64.8	56.27	34.23
CoMuCo		**66.27**	**57.13**	**35.03**
Zero-Shot CLIP	ResNet-101	61.62	54.81	38.71
Linear Probe CLIP		59.75	50.05	26.80
CoOp		66.60	58.66	39.08
TCP		67.53	59.10	40.37
CoMuCo		**69.30**	**60.60**	**41.93**
Zero-Shot CLIP	ViT-B/32	62.05	54.79	40.82
Linear Probe CLIP		59.58	49.73	28.06
CoOp		66.85	58.08	40.44
TCP		67.73	58.50	41.50
CoMuCo		**69.60**	**60.17**	**42.57**
Zero-Shot CLIP	ViT-B/16	66.73	60.83	46.15
Linear Probe CLIP		65.85	56.26	34.77
CoOp		71.92	64.18	46.71
TCP		72.40	64.83	48.17
CoMuCo		**74.90**	**66.80**	**49.47**

Table 2: Performance in domain adaption and with different CLIP visual backbones.

limiting fine-tuning to the final layer helps preserve generalizable representations learned during pre-training.

Sensitivity Study

Our method contains five hyperparameters, including the logit weights α and β, the consistency constraint weights λ_1 and λ_2, and the consensus constraint weight λ_3. The sensitivity study (Fig. 7) of these parameters on ImageNet under the 16-shot setting shows that our method's performance remains stable when each hyperprameter varies within certain range, demonstrating its robustness to hyperparameters.

Generalization Study

Domain Adaption A domain adaptation study was performed to assess the adaptability of CoMuCo to new domains during inference. The model was trained on ImageNet with 16-shot samples and evaluated on ImageNet-V2 and ImageNet-Sketch. As presented in Tab. 2, CoMuCo achieves up to 1.97% and 1.56% higher accuracy than the best baseline on the two target datasets, demonstrating solid generalization across domains.

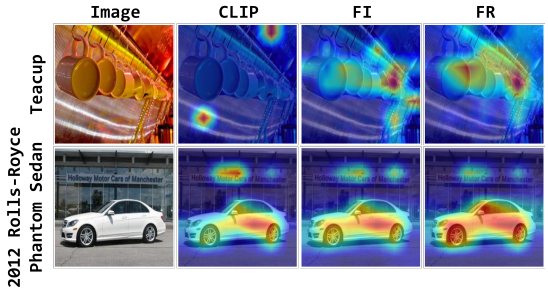

Figure 8: GradCAM visualization on exemplar images. Columns show: (left) original images, (center-left to right) GradCAM heatmaps from CLIP visual encoder, FI, and FR respectively. Warmer colors indicate higher attention.

Backbone Generalization We evaluate CoMuCo on various visual backbones, including ResNet-50, ResNet-101, ViT/B-32, and ViT/B-16. As shown in Tab. 2, CoMuCo consistently outperforms all baselines, with an average gain of 1.53% across all three datasets. These results confirm its robustness across different architectures.

Visualization Analysis

To further elucidate CoMuCo, a visual analysis of its dual modules was performed. Specifically, GradCAM (Selvaraju et al. 2017) was employed to visualize the model's attention regions when presented with category text and query images. Fig. 8 reveals that while the original CLIP model fails to properly focus on the target object, the adapted FR and FI successfully identify it. Moreover, when the CLIP model successfully detected the targets, enhanced comprehensive attention to target objects is achieved through the adapted FR and FI. Refer to Appendix D for more results.

Conclusion

In this study, we propose CoMuCo, a Consistency-guided Multi-view Collaborative Optimization framework, for few-shot learning especially in cross-domain scenarios. The method employs dual expert modules with prior consistency constraint and multi-view consensus mechanism to enhance learning capacity. Additionally, we establish a novel cross-domain benchmark for thorough performance assessment across various imaging domains. Extensive experiments support that CoMuCo substantially boosts model performance. This study offers a new perspective on efficient transfer learning with vision-language models, and CoMuCo is expected to work well under more scenarios.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (grant No. 62571559 and No. 62206317), the Major Key Project of PCL (grant No. PCL2025AS209), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

References

Allen-Zhu, Z.; and Li, Y. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. In *ICLR*.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 - Mining Discriminative Components with Random Forests. In *ECCV*.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2023. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *ICLR*.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2818–2829.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1): 59–70.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024a. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Gao, Y.; Liu, J.; Xu, Z.; Wu, T.; Zhang, E.; Li, K.; Yang, J.; Liu, W.; and Sun, X. 2024b. Softclip: Softer cross-modal alignment makes clip stronger. In *AAAI*.

Gharoun, H.; Momenifar, F.; Chen, F.; and Gandomi, A. 2024. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12): 1–41.

Guo, Z.; Zhang, R.; Qiu, L.; Ma, X.; Miao, X.; He, X.; and Cui, B. 2023. CALIP: Zero-Shot Enhancement of CLIP with Parameter-Free Attention. In *AAAI*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7): 2217–2226.

Huang, Y.; Shakeri, F.; Dolz, J.; Boudiaf, M.; Bahig, H.; and Ben Ayed, I. 2024a. LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP. In *CVPR*.

Huang, Y.; Tang, J.; Chen, Z.; Zhang, R.; Zhang, X.; Chen, W.; Zhao, Z.; Zhao, Z.; Lv, T.; Hu, Z.; and Zhang, W. 2024b. Structure-CLIP: Towards Scene Graph Knowledge to Enhance Multi-modal Structured Representations. In *AAAI*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Khattak, M. U.; Rasheed, H. A.; Maaz, M.; Khan, S. H.; and Khan, F. S. 2023. MaPLe: Multi-modal Prompt Learning. In *CVPR*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV*.

Leung, H. W.; and Bovy, J. 2019. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, 483(3): 3255–3277.

Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; and He, K. 2023. Scaling language-image pre-training via masking. In *CVPR*.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M. B.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. *CoRR*, abs/1306.5151.

Nilsback, M.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, 722–729.

Panchal, S.; Naik, A.; Kokare, M.; Pachade, S.; Naigaonkar, R.; Phadnis, P.; and Bhange, A. 2023. Retinal Fundus Multi-Disease Image Dataset (RFMiD) 2.0: a dataset of frequently and rarely identified diseases. *Data*, 8(2): 29.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? In *ICML*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Song, K.; and Yan, Y. 2013. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285: 858–864.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0402.

Sun, Q.; Fang, Y.; Wu, L. Y.; Wang, X.; and Cao, Y. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *ArXiv*.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*.

Vettoruzzo, A.; Bouguelia, M.; Vanschoren, J.; Rögnvaldsson, T. S.; and Santosh, K. 2024. Advances and Challenges in Meta-Learning: A Technical Review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7): 4763–4779.

Wang, H.; Ge, S.; Lipton, Z. C.; and Xing, E. P. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *NeurIPS*.

Wei, Z.; Pan, Z.; and Owens, A. 2024. Efficient Vision-Language Pre-training by Cluster Masking. In *CVPR*.

Wu, S.; Zhang, J.; Zeng, P.; Gao, L.; Song, J.; and Shen, H. T. 2025. Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In *CVPR*.

Wu, X.; Zhan, C.; Lai, Y.; Cheng, M.-M.; and Yang, J. 2019. IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. In *CVPR*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Xie, J.; Zhang, Y.; Peng, J.; Huang, Z.; and Cao, L. 2024. TextRefiner: Internal Visual Feature as Efficient Refiner for Vision-Language Models Prompt Tuning. *arXiv preprint arXiv:2412.08176*.

Xu, H.; Xie, S.; Tan, X. E.; Huang, P.; Howes, R.; Sharma, V.; Li, S.; Ghosh, G.; Zettlemoyer, L.; and Feichtenhofer, C. 2024. Demystifying CLIP Data. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Yang, Y.; Cui, Z.; Xu, J.; Zhong, C.; Zheng, W.-S.; and Wang, R. 2023. Continual learning with Bayesian model based on a fixed pre-trained feature extractor. *Visual Intelligence*, 1(1): 5.

Yao, H.; Zhang, R.; and Xu, C. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *CVPR*.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task Residual for Tuning Vision-Language Models. In *CVPR*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *ICCV*.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In *ECCV*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *ICCV*.