Advancing Prompt Learning through an External Layer

Fangming Cui
Shanghai Jiaotong University
Shanghai, China
cuifangming@sjtu.edu.cn

Xun Yang
MoE Key Laboratory of
Brain-inspired Intelligent Perception
and Cognition, University of Science
and Technology of China
Hefei, China
xyang21@ustc.edu.cn

Chao Wu
Zhejiang University
Hangzhou, China
chao.wu@zju.edu.cn

Liang Xiao*
Defense Innovation Institute
Beijing, China
Intelligent Game and Decision
Laboratory
Beijing, China
xiaoliang@nudt.edu.cn

Xinmei Tian
MoE Key Laboratory of
Brain-inspired Intelligent Perception
and Cognition, University of Science
and Technology of China
Hefei, China
xinmei@ustc.edu.cn

ABSTRACT

Prompt learning represents a promising method for adapting pre-trained vision-language models (VLMs) to various downstream tasks by learning a set of text embeddings. One challenge inherent to these methods is the poor generalization performance due to the invalidity of the learned text embeddings for unseen tasks. A straightforward approach to bridge this gap is to freeze the text embeddings in prompts, which results in a lack of capacity to adapt VLMs for downstream tasks. To address this dilemma, we propose a paradigm called *EnPrompt* with a novel ***E***xter***n***al ***La***yer (EnLa). Specifically, we propose a textual external layer and learnable visual embeddings for adapting VLMs to downstream tasks. The learnable external layer is built upon valid embeddings of pre-trained CLIP. This design considers the balance of learning capabilities between the two branches. To align the textual and visual features, we propose a novel two-pronged approach: i) we introduce the optimal transport as the discrepancy metric to align the vision and text modalities, and ii) we introduce a novel strengthening feature to enhance the interaction between these two modalities. **Four** representative experiments (i.e., base-to-novel generalization, few-shot learning, cross-dataset generalization, domain shifts generalization) across **15** datasets demonstrate that our method outperforms the existing prompt learning method.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

*Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680953

KEYWORDS

Vision-language Model, Prompt Learning, Optimal Transport

ACM Reference Format:
Fangming Cui, Xun Yang, Chao Wu, Liang Xiao, and Xinmei Tian. 2024. Advancing Prompt Learning through an External Layer. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680953

1 INTRODUCTION

Vision-language models (VLMs), such as CLIP [52] and ALIGN [30], have demonstrated remarkable generalization performance for various downstream tasks [63–65]. VLMs are typically trained to align textual and visual modalities using large-scale datasets, which allows them to encode open-vocabulary [17, 18] concepts in a shared embedding space. Thanks to the modality matching ability, CLIP has achieved remarkable success across various downstream tasks [24, 25, 45–47, 78], such as action recognition [60], image generation [55], and image segmentation [43, 77]. One of the attractive features of CLIP is the ability to perform zero-shot inference, where some pre-defined text inputs (a.k.a. prompts) are used to generate classification weights for predicting image features during inference.

Seminal explorations involve hand-crafted text prompts, e.g., "a photo of a [class]" as the prompt for text encoder. Advanced works introduce a set of learnable parameters to adapt VLMs to downstream tasks. For instance, CoOp [80] keeps the weights of CLIP frozen while learning the text embeddings of prompts for efficient task-specific adaptation. These prompt-learning approaches achieve significant performance improvements over manually tuned prompts on various typical scenarios [6, 8, 14]. However, the learned text embeddings typically produce worse performance compared with hand-crafted prompts [52] for unseen tasks. Specifically, applying these learned text prompts to unseen classes leads to degenerated model generalization performance. Recent works mitigate this problem through the overfitting view. To overcome the challenge, an image-conditional prompt (CoCoOp) learning

approach [79] is proposed to improve the generalization performance on unseen classes. Unfortunately, CoCoOp exhibits lower generalization performance than hand-crafted zero-shot CLIP when applied to novel classes, as depicted in Table 4. This could be attributed to the fact that learned embeddings are usually invalid on novel classes.

To address this dilemma, we propose a paradigm called EnPrompt with a novel External Layer (EnLa). Specifically, we propose adding a textual external layer on top of frozen text embeddings and using learnable visual embeddings for adapting VLMs to downstream tasks. To align the textual and visual features, we introduce the optimal transport [58] as a discrepancy metric that measures the difference between the visual and textual features, where the optimal transport can calculate the distance between two distributions under the form of multiple sampling. Further, we connect the text and image encoders through a strengthening feature instead of ordinary coupling for the dimension transformation of learnable hidden vectors. As shown in Figure 1, EnPrompt outperforms the existing state-of-the-art method on the base-to-novel generalization task. Our main contributions can be summarized as follows:

- We address the performance-degeneration issue of VLMs on unseen tasks by introducing a novel External Layer (EnLa) and freezing the text embeddings.
- We align the visual and textual features of the two modalities by introducing a novel two-pronged approach. Specifically, we introduce the optimal transport as the discrepancy metric to measure and mitigate the difference between the visual and textual features. Meanwhile, we introduce a novel interaction strategy between modalities to strengthen the modality fusion.
- **Four** highly representative experiments demonstrate that our method can consistently outperform the existing methods across **15** datasets, achieving state-of-the-art performance.

2 METHODOLOGY

2.1 Preliminaries

We provide a brief introduction to vision-language pre-training, with a specific focus on CLIP, which is a zero-shot learning approach without fine-tuning. We build our approach based on CLIP. CLIP has a text encoder $\mathcal{F}_t(\cdot)$ and an image encoder $\mathcal{F}_v(\cdot)$, which separately map a textual input (a.k.a. prompt) $\mathbf{p}$ and a visual input, i.e., an image, $\boldsymbol{x}$ into a shared feature space through many transformer blocks. The outputs of two encoders are denoted as $\boldsymbol{t} = \mathcal{F}_t(\mathbf{p})$ and $\boldsymbol{v} = \mathcal{F}_v(\mathbf{x})$. The image encoder aims to transform the input images into feature embeddings, while the text encoder generates representations for word embedding sequences.

During the pre-training phase of CLIP, these two encoders are simultaneously trained on large-scale datasets of text-image pairs, where a contrastive loss is employed to maximize the cosine similarity of text-image pairs and minimize the cosine similarity between unmatched pairs in the feature space. The final prediction probability of alignment is computed by the matching score as follows:

$$p(y \mid \boldsymbol{x}, \mathcal{P}) = \frac{\exp\left\{sim\left(\mathcal{F}_t\left(\mathbf{p}_y\right), \mathcal{F}_v(\mathbf{x})\right)/\tau\right\}}{\sum_{\mathbf{p}_i \in \mathcal{P}} \exp\left\{sim\left(\mathcal{F}_t\left(\mathbf{p}_i\right), \mathcal{F}_v(\mathbf{x})\right)/\tau\right\}}, \tag{1}$$

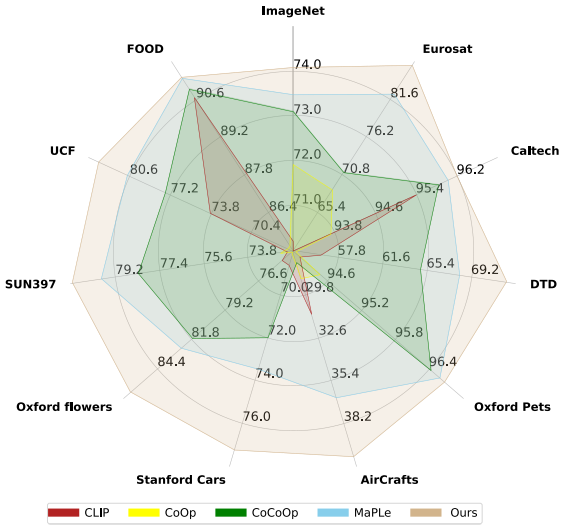

Figure 1: Performance comparison on base-to-novel generalization. EnPrompt (Ours) outperforms previous state-of-the-art methods on 11 datasets.

where $y \in \mathcal{Y}$ is the label of $\mathbf{x} \in \mathcal{X}$, $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{C}$ denotes the set of C pre-defined prompts, $sim(\cdot,\cdot)$ stands for cosine similarity between two vectors, and $\tau > 0$ represents a temperature parameter. Here, the classifier consists of C textual features derived from pre-defined prompts $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{C}$, where the prompt $\mathbf{p}_i$ for the i-th class may have the form of "a photo of a [class]".

Advanced works take a further step to investigate the possibility of aligning images and prompts. The insight of these works is that the prompts tuning for given images could be superior to hand-crafted prompts. Specifically, the class name is retained as prior knowledge to ensure the learned prompts can form a classifier, while the word (a.k.a. context) embeddings of prompts are modeled as learnable parameters. Here, the learnable words in the above prompt are typically initialized using "a photo of a []". The embeddings optimization approach can be formalized as follows,

$$\min_{\mathbf{w}} \ell(\mathbf{w}) = -\log p(y \mid \mathbf{x}, \mathcal{P}(\mathbf{w})), \tag{2}$$

where $\mathbf{w}$ stands for learnable embeddings used to model the context in prompts. The prompt $\mathbf{p}_i \in \mathcal{P}(\mathbf{w})$ learned by the context optimization approach may have the form of,

$$\mathbf{p}_i := [\mathbf{w}] \copyright [ClassName_i], \tag{3}$$

where © is the concatenating operation. Following CoOp [80], the embeddings $\mathbf{w}$ can be shared across classes. In the inference phase, the prompts with the learned embeddings can produce textual features for classification.

2.2 EnLa with Frozen Text Embeddings

As depicted in Table 4, the existing methods CoOp [80] and CoCoOp [79] work well on the base classes observed during training and beat the hand-crafted prompts method employed by CLIP [52] by a significant margin. However, these methods typically perform worse than hand-crafted prompts for novel classes (unseen tasks).

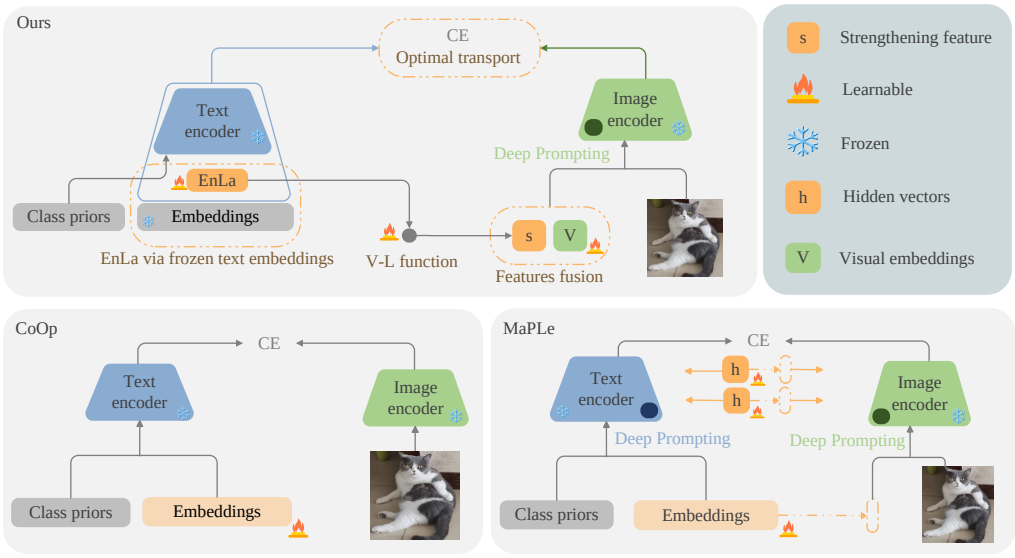

Figure 2: Comparison of our method with CoOp and MaPLe. Our method freezes the textual embeddings and learns an External Layer (EnLa) on top of it. Optimal transport is introduced to align visual and textual modalities apart from the conventional cross entropy (CE) loss. The textual and visual encoders are connected through a strengthening feature. This feature and the learnable visual prompt are fused and utilized for deep prompting of the visual encoder.

This may be attributed to the non-adaptive of the learned text embeddings on unseen tasks. Learning task-specific text embeddings can result in the loss of the essential general textual knowledge having a strong generalization ability.

Frozen Text Embeddings. Built upon the observation, we propose to freeze the text embeddings of prompts for unseen tasks. Specifically, the text embeddings are set to be frozen and non-learnable, which aims to release the potential generalization ability of CLIP and ensure pre-trained ability for unseen tasks. Here, the input words in the prompt are typically initialized using "a photo of a", which can be vectorized into embeddings. In our upcoming ablation experiments, we plan to incorporate other manual template inputs for validation. However, freezing text embeddings of prompts will cause limited model capacity, which results in a lack of capacity to adapt VLMs for downstream tasks.

External Layer (EnLa). To address this dilemma, we propose to introduce an External Layer (EnLa) of the text branch and learnable visual embeddings of the visual branch to learn with VLMs for adapting downstream tasks. This design considers the balance of learning capabilities between the two branches. As depicted in Figure 2, EnLa can be considered as an extension of the text encoder of VLMs, which is an auxiliary layer introduced to the architecture. Introducing the EnLa will make VLMs learnable while keeping the text embeddings frozen. In contrast to CLIP [52], our method is being explored in the field of few-shot learning.

Let $\mathcal{E}_{\boldsymbol{\theta}}(\cdot)$ denote the EnLa parameterized by $\boldsymbol{\theta}$, which is updated for adapting VLMs to downstream tasks. Moreover, let $\mathcal{P}$ denote the frozen text embeddings. The $\mathcal{E}_{\boldsymbol{\theta}}(\cdot)$ is also realized as an incomplete embedding transformer, showing the same spirit with neural networks. Specifically, let $\boldsymbol{e}$ denote the output of the EnLa,

$$\boldsymbol{e} = \mathcal{E}_{\boldsymbol{\theta}}(\mathcal{P}), \tag{4}$$

where $\boldsymbol{e}$ is further transferred as input to two directions: text encoder $\mathcal{F}_t(\cdot)$ and vision-language interaction, particularly towards the image encoder $\mathcal{F}_v(\cdot)$.

EnLa Design Analysis. We compare the performance of a single layer with a two-layer, referring to the bottleneck structure (Linear-ReLU-Linear), on 11 datasets with 16 shots for base-to-novel tasks. HM refers to the harmonic mean [61]. The results in Table 1 indicate that the single 512x512 structure (row-5) provides better performance. It achieves the highest HM of 80.64%, which surpasses all reduction factors of the hidden layers that have 32x, 16x, 8x, and 4x. Thus, we used a single layer of EnLa in all of our experiments.

Table 1: Ablation study on the different designs of EnLa. Results are averaged over 11 datasets.

Layer Design	Reduction factor	Base Acc	Novel Acc.	HM
1: (512 x 16) (16 x 512)	32x	83.9	73.40	78.3
2: (512 x 32) (32 x 512)	16x	84.6	76.8	80.51
3: (512 x 64) (64 x 512)	8x	84.45	76.70	80.41
4: (512 x 128) (128 x 512)	4x	84.40	**77.0**	80.54
5: (512 x 512)	/	**84.71**	76.90	**80.64**

2.3 Alignment with Optimal Transport

To align two modalities for generalization, we introduce the optimal transport as a discrepancy metric and formulate the feature sets as discrete probability distributions for generalization performance.

In contrast to conventional distance metrics such as Euclidean distance, optimal transport [58] learns an adaptive transport plan to calculate the cross-modality distance, facilitating fine-grained matching across the two modalities. It enables us to align visual features in a more precise and adaptive manner, capturing subtle nuances and enhancing the matching accuracy between modalities.

Optimal Transport. The Optimal Transport (OT) distance is a commonly employed metric for comparing distributions. In the context of our framework, we specifically concentrate on the discrete situation as it aligns more closely with our approach. In this scenario, we consider two sets of points or features.

Given two sets that contain N and M points respectively, the discrete distributions can be formulated as follows:

$$\mathbf{Z}_\theta = \sum_{n=1}^{N} \theta_n \delta_{\boldsymbol{e}_n} \quad \text{and} \quad \mathbf{Q}_\beta = \sum_{m=1}^{M} \beta_m \delta_{\boldsymbol{l}_m} \tag{5}$$

where $\boldsymbol{\theta} \in \Delta^N$ and $\boldsymbol{\beta} \in \Delta^M$ are discrete probability vectors that sum to 1 , and δ_e refers to a point mass located at point e in the embedding space. The OT distance between $\mathbf{Z}_\theta$ and $\mathbf{Q}_\beta$ is defined as:

$$d_{\mathrm{OT}}(\boldsymbol{\theta}, \boldsymbol{\beta}) \coloneqq \min_{\mathbf{T}} < \mathbf{T}, \mathbf{C} >, \tag{6}$$

$$\text{s.t. } \mathbf{T} \cdot \mathbf{1}_M = \boldsymbol{\theta},\ \mathbf{T}^\intercal \cdot \mathbf{1}_N = \boldsymbol{\beta}, \tag{7}$$

where $< \cdot, \cdot >$ denotes the Frobenius dot-product and $\mathbf{1}_N$ is the N dimensional vector of ones. $\mathbf{C} \in \mathbb{R}_{>0}^{N \times M}$ is the cost matrix of the transport, and C_{nm} denotes the transport cost between points $\boldsymbol{e}_n$ and $\boldsymbol{l}_m$, such as the cosine distance $C_{nm} = 1 - \cos(\boldsymbol{e}_n, \boldsymbol{l}_m)$.$\mathbf{T} \in \mathbb{R}_{>0}^{N \times M}$ denotes the transport plan to be learned. OT distance is then minimized over all the joint probabilities of $N \times M$ space with two marginal constraints. As computing the above OT distance has the cubic time complexity, we apply the Sinkhorn distance [10] that regularizes with an entropic constraint:

$$d_{\mathrm{OT},\lambda}(\boldsymbol{\theta}, \boldsymbol{\beta}) = d_{\mathrm{OT}}(\boldsymbol{\theta}, \boldsymbol{\beta}) - \lambda h(\mathbf{T}), \tag{8}$$

$$\text{s.t. } \mathbf{T} \cdot \mathbf{1}_M = \boldsymbol{\theta},\ \mathbf{T}^\intercal \cdot \mathbf{1}_N = \boldsymbol{\beta}, \tag{9}$$

where $h(\mathbf{T})$ is the entropy of transport plan $\mathbf{T}$ and $\lambda \geq 0$ is a hyper-parameter. It can be optimized within a few iterations by the Sinkhorn algorithm with the Lagrange multiplier of the entropy constraint.

Modalities Alignment with Optimal Transport. The transport plan is efficiently computed through a limited number of matrix multiplications as a forward module. These matrix multiplications are crucial for determining the gradients that are then preserved for back-propagation.

Specifically, let $\boldsymbol{v}$ denote the visual feature and $\boldsymbol{t}$ denote the textual feature. The output of the image encoder is a tensor with the shape, where H and W are the height and width. Therefore, we can obtain $M = H \times W$ local visual features. Further, let $\boldsymbol{V} = \left\{ \boldsymbol{v}_m |_{m=1}^{M} \right\}$ denote the local visual features as the fixed set, let $\boldsymbol{t}_k$ denote the text feature as the fixed set for class k. Our method learns the transport plan $\mathbf{T}$ by minimizing the following OT distance to push $\boldsymbol{t}_k$ to $\boldsymbol{V}$ for fine-grained alignment:

$$d_{\mathrm{OT}}(k) = d_{\mathrm{OT}}\left(\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{1} - \boldsymbol{V}^\intercal \boldsymbol{t}_k\right), \tag{10}$$

where $\boldsymbol{C} = \mathbf{1} - \boldsymbol{V}^\intercal \boldsymbol{t}_k$ denotes that we use the cosine distance between $\boldsymbol{V}$ and $\boldsymbol{t}_k$ as the cost matrix. Then, we can obtain the solution of transport plan $\mathbf{T}^*$ and the final OT distance $d_{\mathrm{OT}}(k)$. Given the OT distance between $\boldsymbol{t}_k$ and $\boldsymbol{V}$, and image $\boldsymbol{x}$, we reformulate the final prediction probability of V-L alignment as follows:

$$p_{\mathrm{OT}}(y = k \mid \boldsymbol{x}) = \frac{\exp\left(\left(1 - d_{\mathrm{OT}}(k)\right)/\tau\right)}{\sum_{k'=1}^{K} \exp\left(\left(1 - d_{\mathrm{OT}}\left(k'\right)\right)/\tau\right)}. \tag{11}$$

In our method, we first fix the visual and textual features to optimize the optimal transport problem to calculate the cross-modality distance, obtaining the transport plan $\mathbf{T}^*$. Further, we back-propagate the gradient with cross-entropy loss to learn the learnable parameters of our method by fixing $\mathbf{T}^*$. This process is more robust to variations in visual domain shift tasks and tolerant to generalization.

2.4 Alignment with Strengthening Feature

In order to align two modalities for unseen tasks, we propose to connect the visual and textual modalities through a novel strengthening feature with a vision-language (V-L) function instead of coupling learnable hidden vectors for V-L alignment.

We introduce learnable visual embeddings for enhancing learnability of VLMs for downstream tasks. In contrast to VPT [31], our method employs a multi-modal prompt design. Specifically, let $\mathcal{T}_\epsilon(\cdot)$ denote the V-L function parameterized by ϵ to transfer features from textual branch to visual branch. The $\mathcal{T}_\epsilon(\cdot)$ is a function net with dimensionality of [512, 768] for dimension alignment of the visual branch and text branch. The input feature of the V-L function is the output produced by the EnLa. Further, we append L learnable visual input embeddings $\mathcal{P}_v$, which is a vector with [L, 768], given as follows,

$$\mathcal{P}_v = \left\{ \boldsymbol{p}_{\boldsymbol{v}}^1, \boldsymbol{p}_{\boldsymbol{v}}^2, \cdots, \boldsymbol{p}_{\boldsymbol{v}}^L \right\}. \tag{12}$$

To analyze generalization capabilities of alignment, we are conducting an evaluation to assess the performance of different connection positions of the image encoder, as depicted in Table 2.

Table 2: Comparisons of different positions of strengthening feature in base-to-novel generalization. Results are averaged over 11 datasets.

Position of Strengthening Feature	Base Acc.	Novel Acc.	HM
1: Deep layer of image encoder	83.85	76.31	79.90
2: Input layer of image encoder	**84.71**	**76.90**	**80.64**

The deep layer (row-1) of the image encoder is the hidden layer of the image encoder, without visual input (first) layer. The visual input layer (row-2) of the image encoder method is focused on the initialization of the visual embeddings. These designs combine prompting benefits in both branches by enforcing the image encoder representation ability through text knowledge transformation. The position (row-2) is more effective than the position (row-1) in unseen tasks. It can be attributed to that position (row-1) lacks visual embedding initialization, leading to a limited ability to adapt the image encoder with the unbalance of V-L alignment. As a consequence, in our subsequent comprehensive experiments, we adopted the connection position with the input layer of the image encoder (row-2) in our approach. Let $\boldsymbol{p}_{\boldsymbol{v}}(\boldsymbol{e})$ denote the fusion

embeddings, it can be expressed as:

$$\boldsymbol{p}_{\boldsymbol{v}}(\boldsymbol{e}) = \boldsymbol{p}_{\boldsymbol{v}} + \mathcal{T}_{\boldsymbol{\epsilon}}(\boldsymbol{e}), \tag{13}$$

the $\boldsymbol{e}$ is the output of EnLa, and the fusion embeddings are further introduced in the transformer of the image encoder for deep prompting.

In contrast to EnPrompt (Ours), CoCoOp transfers the output features of the image encoder to the text embeddings, which is more inference time consumption. MaPLe initializes the multi-learnable hidden vectors for V-L alignment and couples the text embeddings to the visual encoder. The coupling function maps these initialized hidden vectors of deep layer and text embeddings with the image encoder and text encoder for more parameters. However, our approach introduces the learnable visual initialization embeddings of the image encoder with the strengthening feature fusion. However, MaPLe learns to visual branch with the coupled visual embeddings without the visual embeddings initialization step, which limits the ability of the visual branch to adapt downstream data distributions.

3 EXPERIMENTS

3.1 Benchmark Setting

We compare the performance with CLIP [52], CoOp [80], Clip-Adapter [23], CoCoOp [79], PLOT [7], and MaPLe [32]. Please note that CLIP was originally a zero-shot method. However, in this context, we linearly process it using a few-shot method. The Clip-Adapter [23] is a frozen prompt method with the adapter component behind the encoder, which is different from EnPrompt. We compared it with PLOT [7], which is an optimal transport method. PLOT optimizes four sentences simultaneously ("a photo of a dog", "a picture of a dog", "a drawing of a dog", "a good drawing of a dog"). The output text feature of EnPrompt is global feature, the input sentence is one text prompt such as "a photo of a", and the text prompt is non-learnable.

Few-shot Learning. We employ this setup to evaluate EnPrompt in conditions of highly restricted supervision. We evaluate the model on 11 datasets by conducting tests with varying K-shots per class, where K takes the values of 1, 2, 4, 8, and 16.

Base-to-Novel Generalization. We evaluate the ability to generalize following the setting where the datasets are split into base and novel classes. The model is trained only on the base classes in 16 shots setting and evaluated on base and novel classes.

Cross-dataset Evaluation. To validate the potential of our approach in cross-dataset transfer, we evaluate our ImageNet-trained model in 16 shots directly on other 10 datasets. This experiment aims to verify whether our model can successfully complete the unseen generalization.

Domain Generalization. We evaluate the robustness of our approach to domain shift datasets. Similar to cross-dataset evaluation, we train our model using the ImageNet in 16 shots and evaluate its performance directly on 4 different variants of the ImageNet.

Datasets. For base to novel class generalization, few-shot settings, and cross-dataset evaluation, we use 11 image recognition datasets. The datasets cover multiple recognition tasks including ImageNet [11] and Caltech101 [22] which consists of generic objects, OxfordPets [51], StanfordCars [33], Flowers102 [49], Food101 [3], and FGVCAircraft [48] for fine-grained classification, SUN397 [62] for scene recognition, UCF101 [56] for action recognition, DTD [9] for texture classification, and EuroSAT [27] which consists of satellite images. For domain generalization benchmark, we use ImageNetA [29], ImageNet-R [28], ImageNet-Sketch [59] and ImageNetV2 [53] as domain shift datasets.

Implementation Details. We use a ViT-B/16-based CLIP model, and report results averaged over 3 runs. We set the learnable visual embedding length to 4. We use deep prompting [32] for the image encoder. Training for 50 epochs for few-shot setting, 20 epochs for domain generalization setting and cross-dataset evaluation setting. In the base-to-novel setting, we apply 30 epochs. We use an SGD optimizer with a learning rate of 0.0025. All experiments are run on a single Nvidia A6000 GPU. For domain generalization and cross-dataset evaluation, we train the ImageNet source model on all classes with K = 16 shots in the first 3 transformer layers of the image encoder. For the few-shot learning and base-to-novel tasks, we set visual embeddings learning depth to 9. We initialize the text embeddings using the hand-crafted words of "a photo of a []".

3.2 Effect of Components

In Table 3, we disentangle the components and show the individual contributions to the base-to-novel generalization task. Results are averaged over 11 datasets with 16-shot. HM refers to harmonic mean. Integrating EnLa (row-2) outperforms baseline methods (row-1) on novel classes while maintaining base class gains. By enforcing visual embeddings (row-3), HM performance significantly increases due to the ability to adapt the image encoder for better V-L alignment. Integrating optimal transport (row-4) outperforms component (row-3) with an improvement of 1.08% in HM. This suggests that the regularization of OT is more effective than the traditional metric function in calculating the cross-modality distance. Finally, combined with strengthening features and connecting to overcome the mismatch between the text and visual branch (row-5), our method achieves improvements on both base and novel classes.

Table 3: Effect of our components in base-to-novel generalization. Results are averaged over 11 datasets.

Components	Base Acc.	Novel Acc.	HM
1: Frozen text embeddings	69.34	74.22	71.70
2: + EnLa	80.24	74.43	77.25
3: + Visual embedding learning	82.25	75.94	78.98
4: + Optimal transport	84.50	76.05	80.06
5: + Strengthening feature	**84.71**	**76.90**	**80.64**

3.3 Few-shot Learning Experiments

EnPrompt (Ours) consistently provides improvements on all few-shot settings compared to existing approaches. We note that it is effective for EnPrompt to enhance the few-shot image recognition task through our novel designs. Furthermore, we note that EnPrompt achieves relatively larger gains in minimal data cases, such as for $K = 2$ for almost all datasets. This finding suggests that EnPrompt achieves better alignment between the visual and language branches in seen class tasks, even with limited training resources. Moreover, EnPrompt demonstrates significant improvements compared to linear probe CLIP.

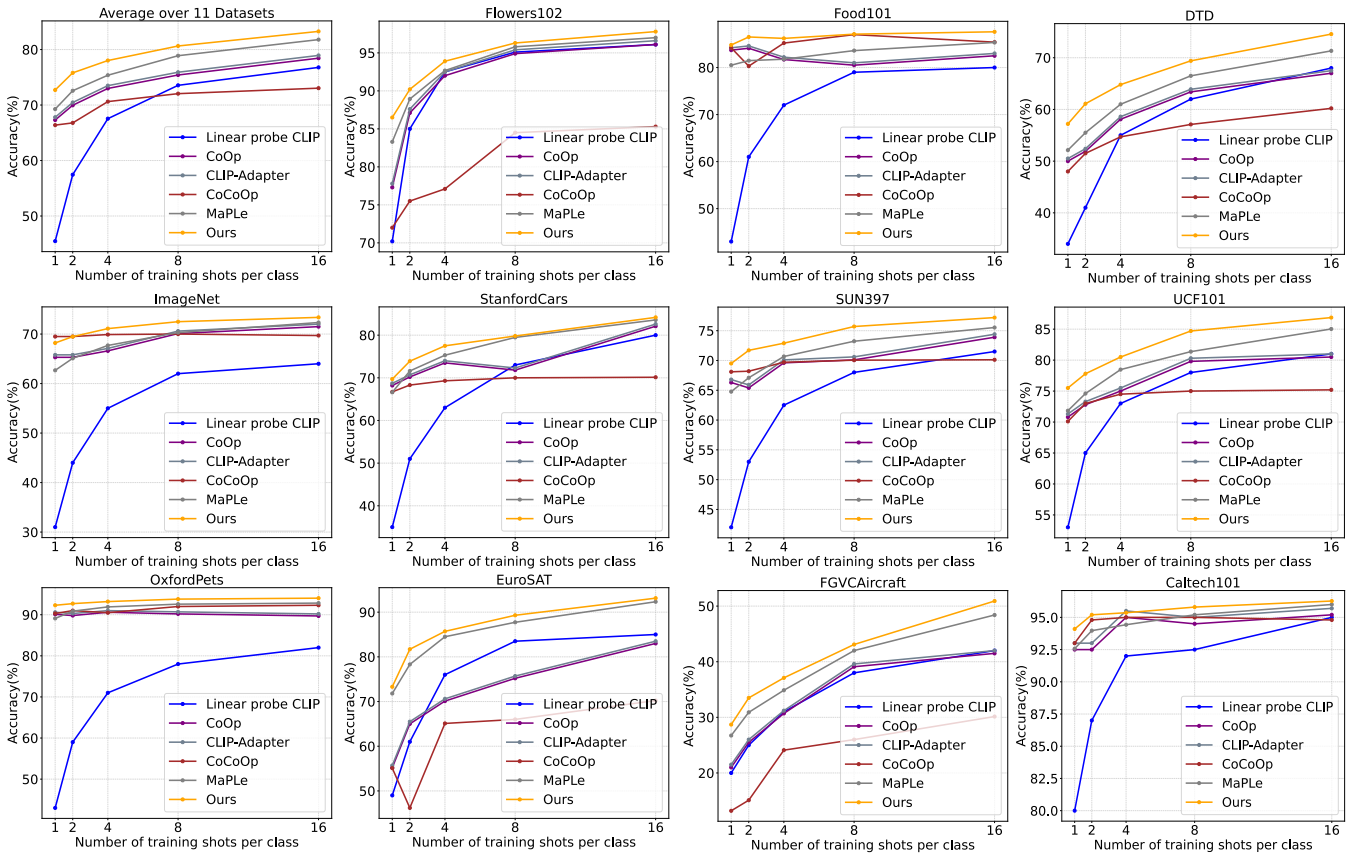

Figure 3: Few-shot Learning Experiments. All methods are trained on the ViT-B/16 CLIP backbone. EnPrompt (Ours) demonstrates consistent improvements over existing methods, specifically for minimal training data such as K=2. On average, EnPrompt provides the highest performance gains for all shots (K = 1,2,4,8,16).

3.4 Base-to-Novel Generalization Experiments

In Table 4, EnPrompt (Ours) demonstrates significant improvements on all 11 datasets. In general, existing approaches CoOp and CoCoOp demonstrate better performance than zero-shot CLIP on base classes. However, when it comes to novel classes, these approaches tend to exhibit normal performance. However, EnPrompt significantly enhances the performance on base classes while also improving the accuracy of zero-shot CLIP on novel classes by 2.69%. Overall, EnPrompt provides the best-averaged results of 84.71%, 76.90%, and 80.64% on the base classes, novel classes, and harmonic mean, respectively.

3.5 Cross Dataset Evaluation

To verify the cross-dataset generalization ability, we train EnPrompt (Ours) on the ImageNet dataset with 1,000 classes, and test it on the remaining 10 datasets. As shown in Table 5, EnPrompt shows competitive performance and achieves better generalization in 8/10 over the CoCoOp. Compared with MaPLe, EnPrompt achieves better performance. This indicates that EnPrompt favors better generalization for a wide range of datasets.

3.6 Domain Generalization Experiments

In Table 6, EnPrompt (Ours) consistently outperforms all existing methods on target datasets, with an overall highest average accuracy of 60.74%. EnPrompt achieves average gains of +1.46% over the CoOp. Compared with MaPLe, EnPrompt shows improved performance in all ImageNet variants datasets. This suggests that EnPrompt is focused on improving generalization capability for domain shifts. Furthermore, EnPrompt provides the highest accuracy of 51.45% on ImageNetA [29].

4 ABLATIVE ANALYSIS

4.1 Comparison of Different Templates

We conducted an experiment using various templates as inputs to the model. Remarkably, we observed that the HM values generated by different templates were very close to each other (see Table 7). The prompt templates can be relatively diversified because of the relatively stronger learning ability of EnLa on unseen tasks, due to the black box [74] nature of EnLa models.

Table 4: Base-to-novel generalization experiments. EnPrompt (Ours) demonstrates strong generalization results over existing methods on 11 recognition datasets. Here, the CLIP refers to the linear probe CLIP.

Dataset		CLIP	CoOp	CLIP-Adapter	CoCoOp	PLOT	MaPLe	Ours	Δ
Average	Base	69.34	82.69	82.91	80.47	81.3	82.28	**84.71**	+2.4
	Novel	74.22	63.22	63.98	71.69	72.2	75.14	**76.90**	+1.8
	HM	71.70	71.66	72.23	75.83	76.48	78.55	**80.64**	+2.1
ImageNet	Base	72.43	76.47	76.88	75.98	75.33	76.66	**77.70**	+1.1
	Novel	68.14	67.88	68.1	70.43	70.48	70.54	**70.65**	+0.1
	HM	70.22	71.92	72.23	73.10	72.83	73.47	**74.07**	+0.6
Caltech101	Base	96.84	98.00	98.1	97.96	97.86	97.74	**98.40**	+0.7
	Novel	94.00	89.81	90.00	93.81	93.99	**94.36**	94.07	-0.3
	HM	95.40	93.73	93.89	95.84	95.92	96.02	**96.2**	+0.2
OxfordPets	Base	91.17	93.67	93.88	95.20	95.7	95.43	**95.67**	+0.2
	Novel	97.26	95.29	95.55	97.69	98.1	**97.76**	97.63	-0.1
	HM	94.12	94.47	94.74	96.43	76.80	96.58	**96.67**	+0.1
StanfordCars	Base	63.37	78.12	78.35	70.49	71.5	72.94	**78.70**	+5.8
	Novel	74.89	60.40	60.55	73.59	73.77	74.00	**75.67**	+1.6
	HM	68.65	68.13	68.33	72.01	72.62	73.47	**77.22**	+3.8
Flowers102	Base	72.08	97.60	97.61	94.87	95.1	95.92	**98.47**	+2.5
	Novel	77.80	59.67	59.98	71.75	72.2	72.46	**77.00**	+4.4
	HM	74.83	74.06	74.32	81.71	82.10	82.56	**86.43**	+4.0
Food101	Base	90.10	88.33	88.55	90.70	90.98	90.71	**91.00**	+0.3
	Novel	91.22	82.26	82.35	91.29	91.54	**92.05**	91.80	-0.9
	HM	90.66	85.19	85.36	90.99	91.28	91.38	**91.41**	+0.1
FGVCAircraft	Base	27.19	40.44	40.66	33.41	35.6	37.44	**43.27**	+5.8
	Novel	36.29	22.30	23.1	23.71	28.5	35.61	**37.77**	+2.0
	HM	31.09	28.75	29.46	27.74	31.66	36.50	**40.34**	+3.7
SUN397	Base	69.36	80.60	80.85	79.74	79.96	80.82	**82.77**	+2.0
	Novel	75.35	65.89	65.91	76.86	77.33	78.70	**79.07**	+0.3
	HM	72.23	72.51	72.62	78.27	78.64	79.75	**80.91**	+1.2
DTD	Base	53.24	79.44	80.56	77.01	78.9	80.36	**83.87**	+3.5
	Novel	59.90	41.18	45.30	56.00	57.9	59.18	**63.67**	+3.5
	HM	56.37	54.24	58	64.85	66.8	68.16	**72.20**	+4.0
EuroSAT	Base	56.48	92.19	92.5	87.49	90.2	94.07	**94.50**	+0.5
	Novel	64.05	54.74	55.65	60.04	63.5	73.23	**79.60**	+6.4
	HM	60.03	68.69	69.49	71.21	74.54	82.35	**86.43**	+4
UCF101	Base	70.53	84.69	84.10	82.33	82.56	83.00	**87.47**	+4.5
	Novel	77.50	56.05	57.35	73.45	75.56	78.66	**79.17**	+0.5
	HM	73.85	67.46	68.21	77.64	78.92	80.77	**83.13**	+2.5

Table 5: Cross-dataset benchmark evaluation. EnPrompt (Ours) achieves overall favorable performance.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
Linear probe CLIP	66.73	92.94	89.07	65.29	71.30	86.11	24.87	62.62	44.56	47.69	66.77	65.12
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CLIP-Adapter	**71.40**	93.85	89.57	64.66	68.85	85.54	18.53	64.35	41.86	46.43	66.77	64.04
CoCoOp	71.02	**94.43**	90.14	65.32	**71.88**	86.06	22.94	67.36	45.73	45.37	68.21	65.74
PLOT	70.15	94.60	90.23	65.41	71.97	86.32	22.87	67.22	44.99	46.57	68.32	65.85
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	**46.49**	48.06	68.69	66.30
Ours	71.03	93.93	**91.20**	**65.63**	71.73	**86.40**	**25.13**	**67.67**	46.47	**48.96**	**69.73**	66.69

Table 6: Domain generalization. These approaches are trained on imageNet and tested on datasets with domain shifts.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
Linear probe CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	**71.51**	64.2	47.99	49.71	75.21	59.28
CLIP-Adapter	**71.40**	64.5	47.72	49.75	75.55	59.38
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
PLOT	70.15	64.17	49.15	50.83	76.5	60.16
MaPLe	70.72	64.07	49.15	50.9	76.98	60.27
Ours	71.03	**64.3**	**49.5**	**51.45**	**77.83**	60.77

Table 7: Comparison of different templates on Flowers102.

Templates	Base Acc	Novel Acc.	HM	GAP
"a drawing of a"	98.3	77.10	86.42	±0.05
"a painting of the"	97.2	**77.7**	86.39	±0.05
"a photo of a"	**98.47**	77	**86.43**	/

4.2 Training Strategy

In Table 8, we evaluate the performance of different EnLa training strategies as an ablation. In our approach, the EnLa is implemented as a neural network, which is essential to learn all potentially valuable features during the training stage in order to achieve effective generalization [36, 75, 76]. To this end, we focus on training the model for more epochs to learn richer features (row-2), resulting in improved generalization performance [15, 73].

Table 8: Comparison of EnLa train strategies. Results are averaged over 11 datasets.

Training epoch	Base Acc.	Novel Acc.	HM
1: EnLa (2 epoch)	83.22	75.91	79.68
2: EnLa (30 epoch)	**84.71**	**76.90**	**80.64**

4.3 Visual Embeddings Length

Figure 4 (left) shows the effect of visual embedding length on the harmonic mean. The visual embeddings are learnable, and the text embeddings are non-learning. We observed a significant decrease in HM when the visual embedding length exceeds 4. It suggests that too many learnable visual parameters inherently decrease the ability of V-L alignment. Overall, we have determined that using 4 visual embeddings is the most suitable method.

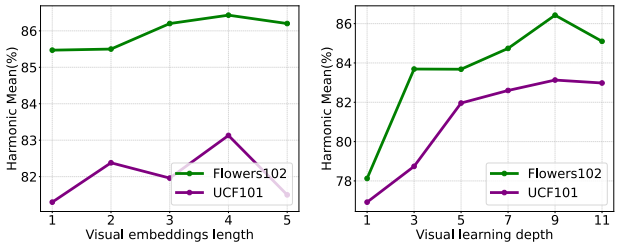

Figure 4: Ablation study for embedding length and learning depth of image encoder on Flowers102 and UCF101.

4.4 Visual Learning Depth

In Figure 4 (right), we ablate on the visual learning depth for HM. We apply deep prompting on the image encoder. We note that increasing the learning depth generally increases the performance. This suggests that using more layer depth for pre-trained features provides more rich supervision for the features in our approach. The HM value decreases as the number of visual layers increases to 10 or more. It indicates excessive fine-tuning of the model, causing it to lose CLIP generalization ability. Overall, EnPrompt achieves maximum performance at a depth of 9.

4.5 Analyzing of Multi-modal Methods

To validate the effect of EnPrompt (Ours), we conducted the base-to-novel task and non-generalization few-shot task using two-modal designs (row-1) and the method completely opposite to EnPrompt (row-2) in Table 9. The method (row-2) is a visual external layer based on frozen visual embeddings with learning textual embeddings. We observed that the method (row-2) has a lower performance in novel classes than the method (row-3), which is due to the learnable text embeddings of method (row-2) being weak in pre-trained generalization compared to the frozen text prompt. Finally, we observed that the method (row-3) and method (row-2) performed better in the novel classes than method (row-1) while still maintaining a non-generalization few-shot task. This is attributed to the fact that method (row-1) has neither learnable visual embeddings nor learnable text embeddings. It indicates that prompt learning is crucial for adapting pre-trained VLMs for generalization.

Table 9: We evaluate EnPrompt with different multi-modal designs. The V-L connection is the default setting.

Multi-modal Methods	Base	Novel	HM	Few-shot Task
1: *Visual + Text*	82.20	75.5	78.71	82.05
2: Opposite Ours (*Visual*)	84.15	75.20	79.43	82.69
3: Ours (*Text*)	**84.71**	**76.90**	**80.64**	**83.27**

4.6 Inference Stage Computational Cost

In Table 10, we show the compute cost analysis of EnPrompt (Ours) and compare it with text embedding learning approaches. Although EnPrompt uses the EnLa, its overall parameters exceed only by 0.52% over CLIP. Compared to MaPLe, EnPrompt has fewer parameters and lower coupling. In terms of inference speed, CoCoOp is significantly slower. In contrast, EnPrompt has no such overhead, obtaining a higher performance with less training time. Although CoOp has a small number of parameters, due to the mismatch of V-L alignment, the training time of 10 epochs for CoOp is similar to ours. EnPrompt is simpler than textual multi-prompt input method (PLOT) in textual design component. Further, EnPrompt provides better convergence as it gets better HM as compared to MaPLe in 10 epochs.

Table 10: The compute cost comparison using SUN397 dataset. Training time for all methods is calculated for 10 epochs on a single A6000 GPU.

Method	Params	Params % CLIP	Train time (min)	HM
CoOp	2048	0.002	10.88	71.65
CoCoOp	35360	0.03	39.53	75.83
CLIP-Adapter	0.52M	0.41	8.55	72.23
PLOT	8192	0.008	10.85	76.48
MaPLe	3.55 M	2.85	10.58	79.68
Ours	0.65M	0.52	10.21	**80.51**

5 RELATED WORK

5.1 Prompt Learning in Vision Language Models

Recently, the strong generalization capability of CLIP [52] has made it a foundation for many methods [4, 5, 19] in adapting pre-trained VLMs for downstream tasks [13, 35, 38, 44]. Prompt learning [34, 37, 39] is a widely used technique in NLP for learning downstream tasks. The use of text prompts, which are instructions provided to the language branch of a Vision-Language model (VLM), is a common practice to enhance task understanding [23, 69, 72]. CoOp [80] and CoCoOp [79] fine-tuning the CLIP model specifically for few-shot image recognition by optimizing a continuous set of token embeddings in the language branch. The image-conditional prompt of CoCoOp significantly contributes to enhancing generalization to unseen classes [16, 20, 21] and mitigating the risk of overfitting to the limited labeled data. Moreover, some approaches [41, 70, 71, 81] constrain the proposed learnable prompts to contain the essential general knowledge and prior distribution learning. By conditioning prompts on visual features, CoCoOp [79] ensures that the language model attends to relevant visual information when generating predictions. In addition to single-modal prompt tuning [31], some approaches [32, 40] introduce the multi-modal prompt-tuning designs in CLIP to effectively align V-L representations [12, 42, 50]. However, their methods have not fully released the potential generalization ability of CLIP.

5.2 Optimal Transport

Optimal Transport (OT) has recently gained significant attention in various theoretical and application tasks to measure the distance between two probability distributions over metric spaces [1, 2, 26, 57]. For instance, Redko et al.[54] tackle the target shift problem by aligning domain distributions using the OT framework. PLOT [7] introduces a local text feature with textual multi-prompt input. Ours obtain a global text feature for textual one prompt, instead of a local text feature with textual multi-prompt input.

6 CONCLUSION

Prompt learning is a promising method for adapting pre-trained vision-language models. However, these methods often struggle to tackle the challenge of generalization on unseen tasks effectively. In this paper, we propose a paradigm called EnPrompt with a novel External Layer (EnLa). Specifically, we propose a textual external layer and learnable visual embeddings for adapting VLMs to downstream tasks. To align the textual and visual features, we propose a novel two-pronged approach. (a) We introduce the optimal transport as the discrepancy metric to align the vision and text modalities. (b) We introduce a novel strengthening feature to enhance the interaction between these two modalities. Extensive experiments clearly demonstrated the effectiveness of our method. In the future, we will extend our effort to more challenging tasks, such as video moment retrieval [67, 68], and investigate the cross-domain generalization ability [66] comprehensively.

7 ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China (Grant U22A2094 and 62222117), National Key Research and Development Project of China (Grant 2021ZD0110505), and the Zhejiang Provincial Key Research and Development Project (Grant 2023C01043)

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[2] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. 2020. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems* 33 (2020), 12934–12944.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 446–461.

[4] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. 2024. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 936–944.

[5] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. 2024. Domain prompt learning with quaternion networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26637–26646.

[6] Yihong Cao, Hui Zhang, Xiao Lu, Zheng Xiao, Kailun Yang, and Yaonan Wang. 2024. Towards Source-free Domain Adaptive Semantic Segmentation via Importance-aware and Prototype-contrast Learning. *IEEE Transactions on Intelligent Vehicles* (2024).

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

[8] Zhuoxiao Chen, Yadan Luo, Zixin Wang, Zijian Wang, Xin Yu, and Zi Huang. 2023. Towards open world active learning for 3d object detection. *arXiv preprint arXiv:2310.10391* (2023).

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.

[10] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[12] Donglin Di, Jiahui Yang, Chaofan Luo, Zhou Xue, Wei Chen, Xun Yang, and Yue Gao. 2024. Hyper-3DG: Text-to-3D Gaussian Generation via Hypergraph. *arXiv preprint arXiv:2403.09236* (2024).

[13] Zheng Ding, Jieke Wang, and Zhuowen Tu. 2023. Open-Vocabulary Universal Image Segmentation with MaskCLIP. (2023).

[14] Djamahl Etchegaray, Zi Huang, Tatsuya Harada, and Yadan Luo. 2024. Find n'Propagate: Open-Vocabulary 3D Object Detection in Urban Environments. *arXiv preprint arXiv:2403.13556* (2024).

[15] Zhen Fang, Yixuan Li, Feng Liu, Bo Han, and Jie Lu. 2024. On the Learnability of Out-of-distribution Detection. *Journal of Machine Learning Research* 25 (2024).

[16] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. 2022. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems* 35 (2022), 37199–37213.

[17] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. 2021. Learning bounds for open-set learning. In *International conference on machine learning*. PMLR, 3122–3132.

[18] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. 2020. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems* 32, 10 (2020), 4309–4322.

[19] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. 2019. Unsupervised domain adaptation with sphere retracting transformation. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[20] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. 2022. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 1087–1105.

[21] Zhen Fang, Jie Lu, and Guangquan Zhang. 2023. An extremely simple algorithm for source domain reconstruction. *IEEE Transactions on Cybernetics* (2023).

[22] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*. IEEE, 178–178.

[23] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2023. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* (2023), 1–15.

[24] Ross Greer, Bjørk Antoniussen, Mathias V Andersen, Andreas Møgelmose, and Mohan M Trivedi. 2024. The Why, When, and How to Use Active Learning in Large-Data-Driven 3D Object Detection for Safe Autonomous Driving: An Empirical Exploration. *arXiv preprint arXiv:2401.16634* (2024).

[25] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking Micro-action Recognition: Dataset, Methods, and Applications. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 6238–6252. https://doi.org/10.1109/TCSVT.2024.3358415

[26] Dandan Guo, Zhuo Li, He Zhao, Mingyuan Zhou, Hongyuan Zha, et al. 2022. Learning to re-weight examples with optimal transport for imbalanced classification. *Advances in Neural Information Processing Systems* 35 (2022), 25517–25530.

[27] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. 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 (2019), 2217–2226.

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

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

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[31] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.

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

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*. 554–561.

[34] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[35] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. 2023. Transformer-based visual grounding with cross-modality interaction. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 6 (2023), 1–19.

[36] Wenxi Li, Zhuoqun Cao, Qian Wang, Songjian Chen, and Rui Feng. 2021. Learning error-driven curriculum for crowd counting. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 843–849.

[37] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[38] Yicong Li, Xun Yang, An Zhang, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023. Redundancy-aware transformer for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3172–3180.

[39] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602* (2021).

[40] Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023. Deeply Coupled Cross-Modal Prompt Learning. *arXiv preprint arXiv:2305.17903* (2023).

[41] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5206–5215.

[42] Chaofan Luo, Donglin Di, Yongjia Ma, Zhou Xue, Chen Wei, Xun Yang, and Yebin Liu. 2024. TrAME: Trajectory-Anchored Multi-View Editing for Text-Guided 3D Gaussian Splatting Manipulation. *arXiv preprint arXiv:2407.02034* (2024).

[43] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. 2023. SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic Segmentation. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 23033–23044. https://proceedings.mlr.press/v202/luo23a.html

[44] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*. PMLR, 23033–23044.

[45] Yadan Luo, Zhuoxiao Chen, Zhen Fang, Zheng Zhang, Mahsa Baktashmotlagh, and Zi Huang. 2023. Kecor: Kernel coding rate maximization for active 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18279–18290.

[46] Yadan Luo, Zhuoxiao Chen, Zijian Wang, Xin Yu, Zi Huang, and Mahsa Baktashmotlagh. 2023. Exploring active 3d object detection from a generalization perspective. *arXiv preprint arXiv:2301.09249* (2023).

[47] Yadan Luo, Ziwei Wang, Zi Huang, Yang Yang, and Cong Zhao. 2018. Coarse-to-fine annotation enrichment for semantic segmentation learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 237–246.

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

[49] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 722–729.

[50] Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multi-modal neurons in pre-trained transformer. In *ACL*.

[51] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3498–3505.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International conference on machine learning*. PMLR, 5389–5400.

[54] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. 2019. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on artificial intelligence and statistics*. PMLR, 849–858.

[55] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. 2021. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624* (2021).

[56] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[57] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. 2021. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems* 34 (2021), 17194–17208.

[58] Cédric Villani et al. 2009. *Optimal transport: old and new*. Vol. 338. Springer.

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

[60] Mengmeng Wang, Jiazheng Xing, and Yong Liu. [n. d.]. ActionCLIP: A New Paradigm for Video Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems* PP ([n. d.]).

[61] Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-Shot Learning - the Good, the Bad and the Ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[62] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.

[63] Fei Xie, Lei Chu, Jiahao Li, Yan Lu, and Chao Ma. 2023. Videotrack: Learning to track objects via video transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22826–22835.

[64] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. 2022. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8751–8760.

[65] Fei Xie, Zhongdao Wang, and Chao Ma. 2024. DiffusionTrack: Point Set Diffusion Model for Visual Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19113–19124.

[66] Xun Yang, Tianyu Chang, Tianzhu Zhang, Shanshan Wang, Richang Hong, and Meng Wang. 2024. Learning Hierarchical Visual Transformation for Domain Generalizable Visual Matching and Recognition. *International Journal of Computer Vision* (2024), 1–27.

[67] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[68] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.

[69] Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. 2024. Robust Video Question Answering via Contrastive Cross-Modality Representation Learning. *SCIENCE CHINA Information Sciences* (2024).

[70] Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6757–6767.

[71] Hantao Yao, Rui Zhang, and Changsheng Xu. 2023. Visual-Language Prompt Tuning With Knowledge-Guided Context Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6757–6767.

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

[73] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. 2021. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196* (2021).

[74] Yonggang Zhang, Ya Li, Tongliang Liu, and Xinmei Tian. 2020. Dual-path distillation: A unified framework to improve black-box attacks. In *International Conference on Machine Learning*. PMLR, 11163–11172.

[75] Yonggang Zhang, Xinmei Tian, Ya Li, Xinchao Wang, and Dacheng Tao. 2020. Principal component adversarial example. *IEEE Transactions on Image Processing* 29 (2020), 4804–4815.

[76] Yonggang Zhang, Zhiqin Yang, Xinmei Tian, Nannan Wang, Tongliang Liu, and Bo Han. 2024. Robust Training of Federated Models with Extremely Label Deficiency. *arXiv e-prints* (2024), arXiv–2402.

[77] Zhuowen Tu Zheng Ding, Jieke Wang. 2023. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In *International Conference on Machine Learning*.

[78] Chong Zhou, Chen Change Loy, and Bo Dai. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*. Springer, 696–712.

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

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

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