

Modality-Consistent Prompt Tuning With Optimal Transport

Hairui Ren^{ID}, Fan Tang^{ID}, Huangjie Zheng^{ID}, He Zhao, Dandan Guo^{ID}, and Yi Chang^{ID}

Abstract—Prompt tuning has been successfully used in leveraging the knowledge of Large-scale Vision-Language Pre-trained (VLP) models on downstream tasks. Most existing prompt tuning approaches learn prompts by maximizing the pairwise similarity. Although samples in different modalities might be relatively aligned pairwise, such alignment does not fully utilize the information between samples, which can be less consistent on the modality level. In this paper, we propose a novel prompt tuning strategy by distributionally matching different modalities. Specifically, we minimize the distribution-wise distance between the image and text modalities with optimal transport (OT) theory. Simultaneously, we add a constraint on the learned transport plan during the modality matching to enhance the learning of vision and text prompts. Our proposed one can be applied to improve existing uni-modal and multi-modal prompt learning methods for being a plug-and-play method, which can generate modality-consistent representations. Experiments on eleven public datasets demonstrate that our proposed method has excellent performance, achieving substantial improvements on both uni-modal and multi-modal prompt tuning methods.

Index Terms—Prompt tuning, modality-consistent, optimal transport, distribution matching.

I. INTRODUCTION

LARGE-SCALE Vision-Language Pre-trained (VLP) models such as CLIP [1], ALIGN [2], and BLIP [3] have made significant advancements in open-world visual concept learning. Pretrained on vast text-image pairs (e.g., 400 million pairs for CLIP), VLPs build the connections between language and vision modalities. Despite the remarkable process of learning universal knowledge, fine-tuning the pre-trained models on

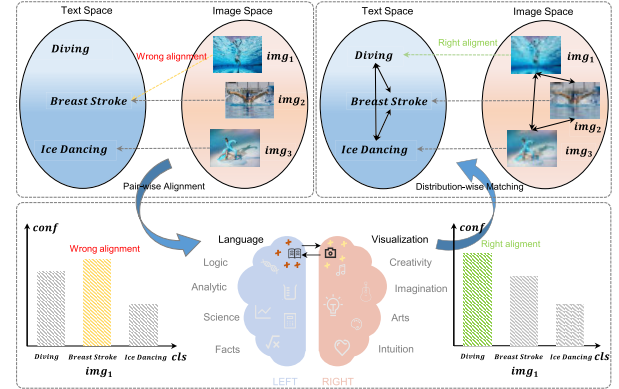


Fig. 1. ‘img1’ and ‘img2’ represent ‘Diving’ and ‘Breaststroke’ images. When only pairwise alignment is performed, the two image types are highly similar both semantically and visually. This similarity leads the model to confidently misclassify ‘Diving’ as ‘Breaststroke’. Our brains encode multiple modalities to create a modality-consistent representation [13], [14], [15], [17]. In contrast to previous pair-wise alignment models, we emulate the way the human brain integrates different modal features and design a new prompt learning method from the perspective of distribution matching. Distribution-wise matching enables different categories of text and images to exchange information, helping the model to learn the differences between easily confused samples and correctly classify the images.

downstream tasks with limited data is particularly challenging due to the substantial size of VLPs.

The predominant approach for adapting VLP models to downstream tasks under data constraints is prompt tuning. In particular, prompt tuning aims to identify the most suitable textual or visual prompts for a given task, thereby enabling the effective transfer of knowledge from pre-train to specific downstream applications. Recently, Zhou et al. [4] proposed CoOp, the first VLP prompt tuning approach, by replacing discrete manual prompts with continuous vectors and subsequently optimizing these vectors on downstream tasks. MaPle [5] introduces multi-modal prompt tuning within CLIP to better align its vision-language representations with pairwise image-text embeddings. In line with these methods, researchers have focused on proposing various formulations of textual/visual prompts [6], [7], [8], [9], [10], [11].

However, there are two limitations to most prompt learning methods: First, [4], [5], [6], [7], [8], [9], [10], [11] only model pair-wise knowledge from a single sample to another as shown in Fig. 1. The basic prompt-tuning paradigm is optimized by maximizing the cosine similarity [4], [5], [6] or minimizing the distance [7], [8] of paired image-text instances. Although image and text samples might be relatively aligned

Received 26 June 2024; revised 22 September 2024; accepted 25 October 2024. Date of publication 31 October 2024; date of current version 7 March 2025. This work was supported by NSFC under Grant 62306125, Grant 2023YFF0905400, and Grant U2341229. This article was recommended by Associate Editor Z. R. Lei. (Corresponding authors: Dandan Guo; Yi Chang.)

Hairui Ren and Dandan Guo are with the School of Artificial Intelligence, Jilin University, Changchun, Jilin 130012, China (e-mail: renhr22@mails.jlu.edu.cn; guodandan@jlu.edu.cn).

Fan Tang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tangfan@ict.ac.cn).

Huangjie Zheng is with the Department of Statistics and Data Science, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: huangjie.zheng@utexas.edu).

He Zhao is with CSIRO’s Data61, Eveleigh, NSW 2015, Australia (e-mail: he.zhao@ieee.org).

Yi Chang is with the School of Artificial Intelligence and the International Center of Future Science, Jilin University, Changchun, Jilin 130012, China, and also with the Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education (MOE), Changchun 130000, China (e-mail: yichang@jlu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3489024

pairwisely, the samples' relationship within each modality has not been fully explored, which may hinder alignment between different modality features of the same class [12]. Second, [4], [5], [6], [7], [8], [9], [10], [11] overlook the explicit exploitation of the modality-consistent knowledge for downstream tasks. Substantial evidence in cognitive science and neuroscience literature supports the idea that our brains encode multiple views to create one powerful view-invariant representation [13], [14], [15]. As shown in Fig. 1, The images 'img1' and 'img2' represent the actions 'Diving' and 'Breaststroke', respectively. Due to the visual and semantic similarity between these two actions, the model mistakenly classified 'Diving' as 'Breaststroke' with relatively high confidence when only pair-wise alignment was applied. However, if our model can behave more like a brain—treating samples within each modality as distributions and matching across modalities in a distribution-wise manner, the model would better identify the true 'Breaststroke' image based on the context of the original 'Diving' image, correcting the classification with higher confidence. Accordingly, an effective prompt tuning framework should produce a modality-consistent representation [16], [17] as the human brain does.

To mitigate the above limitations, this paper develops **Modality-Consistent Prompt Tuning (McPt)**, a novel prompt tuning framework from the perspective of cross-modality distribution-to-distribution matching, which aims to model modality-consistent knowledge for downstream tasks with the fused textual and visual knowledge. McPt tackles two crucial challenges: how to effectively leverage knowledge within each modal distribution and perform matching between distributions, and how to fully adapt the matched modality-consistent knowledge to downstream tasks. Recently, Optimal Transport (OT) [18] has demonstrated the prevailing performance in modeling distribution knowledge. PLOT [7] and ALIGN [8] have applied OT to the field of prompt learning, both of which replace the cosine similarity between the visual features and textual features in classification loss with the OT distance. However, they only consider a pairwise OT distance between the textual features of each category and the visual features of each image. The potential of OT in improving the efficient transfer learning of VLP models has not yet been fully explored. Inspired by this, for the first challenge, McPt aims to exploit OT to model the distribution-wise matching to fully utilize the inter-samples' relationship for downstream tasks. Consequently, we represent the batch set of two modalities as two discrete empirical distributions over all samples. Therefore, modality-consistent learning can be reformulated as the process of learning the image distribution to be as close to the text distribution as possible, an unsupervised process facilitated by leveraging the OT. Interestingly, the learned transport plan based on the above-mentioned OT can reflect the similarity between textual and visual embeddings from the view of distribution in an adaptive way. Inspired by this, for the second challenge, we further introduce a cross-entropy constraint on the transport plan by viewing it as the prediction probability of each sample in the distribution, assisting the learning of prompts in the form of supervised learning for downstream tasks.

Without losing generality, the progressive matching is conducted in a distribution-to-distribution manner that could be seamlessly integrated as a plug-and-play component into prompt tuning models with arbitrary structures. Beyond that, to evaluate our approach, we conducted extensive experiments across eleven datasets, following the standard setup of CLIP [1] and CoOp [4]. These experiments encompass various visual classification tasks involving generic objects, scenes, actions, fine-grained categories, and more. The desired improvements demonstrate that the proposed McPt effectively fosters modality consistency and in turn, enhances model performance. Our contributions can be included:

- We highlight the significance of effectively leveraging knowledge between samples within each modal by modality matching at the distribution level.
- We introduce a novel approach McPt: **Modality-Consistent Prompt Tuning**, which utilizes optimal transport to facilitate the distribution matching and cross-entropy constraint on the transport plan from both modalities, to generate modality-consistent representations akin to the human dose.
- McPt shows substantial improvements across eleven datasets and can be seamlessly integrated as a plug-and-play component into any prompt tuning model, thereby contributing to enhanced model performance.

II. RELATED WORK

A. Parameter-Efficient Fine-Tuning (PEFT)

Numerous efforts have been dedicated to developing various PEFT approaches. Adapter [19] introduces adapter-based parameter-efficient transfer learning techniques to reduce computational costs. Prefix-tuning [20] keeps the model parameters frozen while optimizing a small continuous task-specific vector, referred to as the prefix. Lora [21] freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the model.

CoOp [4] optimizes the continuous set of prompt vectors instead of discrete vectors for downstream tasks. CoCoOp [6] addresses the generalization challenge of CoOp by incorporating Meta-Net, achieving a balance between performance on base classes and novel classes. Compared with CoOp and CoCoOp, PLOT [7] further improves prompt learning by focusing on different attributes and learning multiple prompts. In addition to the uni-modal prompt tuning approaches mentioned above, some work investigates the effectiveness of multi-modal prompt tuning. MaPle [5] introduces multi-modal prompt tuning within CLIP to better align its vision-language representations. ALIGN [8] introduces a multi-mode token-level tuning framework that utilizes OT to learn and align a set of prompt tokens across different modalities. However, most prior work focused on improving the structure or number of prompts. Moving beyond most of them that learn the prompts only with the classification loss, we further improve the prompt tuning from the perspective of modality-consistent, which is orthogonal to the above methods. Therefore, ours can be flexibly combined with these uni-modal and multi-modal prompt tuning methods to enhance their performance.

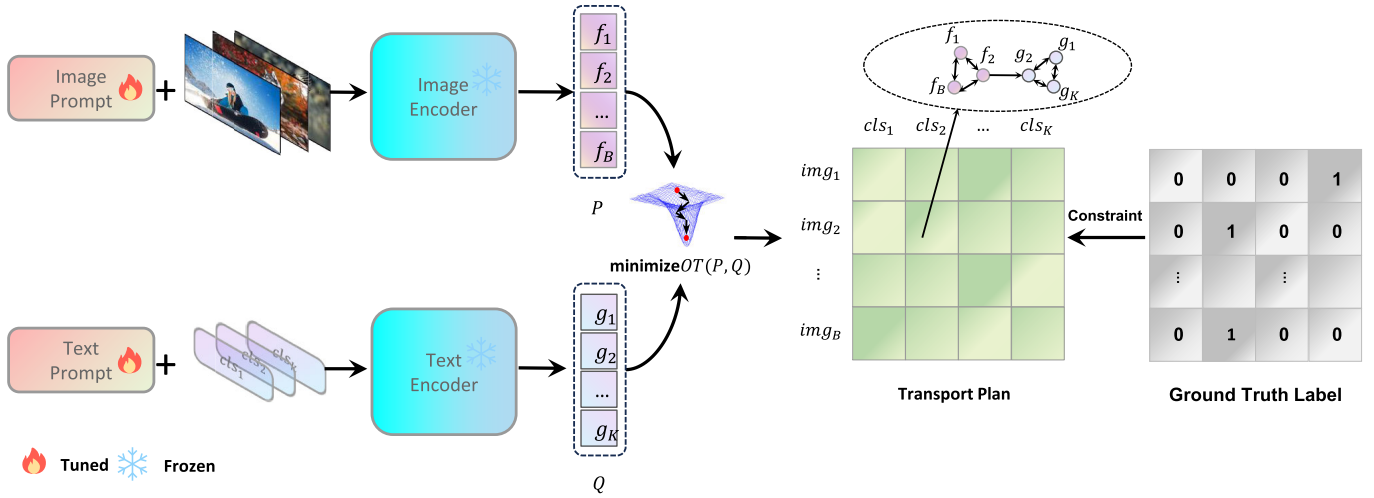


Fig. 2. Image embeddings and text embeddings first form two empirical distributions P and Q , by minimizing the optimal transport distance between the two distributions, the text and image modalities are aligned in an unsupervised distribution-wise manner, allowing the matching of text and image to consider the feature information between samples within each modality fully. Simultaneously Ground Truth labels assist in guiding the learned transport plan in the matching process, enhancing the adaptability of the learned text and image prompts to downstream tasks.

B. Optimal Transport

OT has found wide-ranging applications in various machine learning domains, including generative models [22], [23], text analysis [24], [25], and structural matching [26], [27]. Recently, PLOT [7] and ALIGN [8] apply OT into the field of prompt learning, both of which replace the cosine similarity between the visual features and textual features in classification loss with the OT distance between them. That is to say, they only consider a pairwise OT distance between the textual features of each category and the visual features of each image. However, we not only consider the pairwise alignment but also minimize the distribution-wise OT for matching the image and text modalities. Besides, we further enhance the learning of vision and text prompts by constraining the learned transport plan with cross-entropy, using ground truth labels. Therefore, ours does not conflict with PLOT or ALIGN, which instead can be integrated with them as a plug-and-play method.

C. Cross-Modal Alignment

A significant portion of work has achieved outstanding performance in cross-modal alignment. GOT [28] leverages Optimal Transport for graph matching to tackle cross-domain alignment, while NAAF [29] employs a two-branch matching module that utilizes both mismatched and matched textual fragments to achieve accurate image-text matching. UARDA [30] introduces a unified learning framework with joint optimization to produce more discriminative fragment features and determine optimal relevance thresholds. ESL [31] extends static, single-dimensional correspondence into a dynamic mechanism that enhances combinatorial expressive power, and BOOM [32] focuses on bidirectional semantic alignment consistency, addressing both ‘word-to-region’ and ‘region-to-word’ alignments in image-text matching. However, these approaches do not emphasize matching different modalities through distributional matching, and this aspect

has not been thoroughly explored in the context of prompt learning.

III. METHOD

In this section, we first revisit the prompt tuning and optimal transport (Sec. III-A) and then introduce the details of our proposed distribution matching (Sec. III-B) and cross-entropy constraint (Sec. III-C). Finally, our overall training procedure is shown in Sec. III-D. The pipeline of our proposed McPt is illustrated in Fig. 2.

A. Preliminaries

1) *Revisiting Prompt Tuning:* We provide a brief introduction to two forms of prompt tuning: uni-modal prompt tuning (UPT) and multi-modal prompt tuning (MPT). For UPT, the only parameter to learn is the text prompt $\mathbf{t} = \{t_i\}_{i=1}^N$, where N is the number of text prompts, e.g. $N=1$ for CoOp [4] and $N=4$ for PLOT [7]. Each class name $\{c_k\}_{k=1}^K$ is concatenated with the text prompt to formulate a text input $\mathbf{w}_{ki} = \{t_i, c_k\}$ with $i \in [1, N]$, which can be fed into the text encoder G in CLIP to yield a text embedding $\mathbf{g}_k = \frac{1}{N} \sum_{i=1}^N G(\mathbf{w}_{ki})$. Concurrently, for the input image $\mathbf{x}$ from the vision dataset, we can embed it with the image encoder F of CLIP to produce a visual embedding $\mathbf{f} = F(\mathbf{x})$. The prediction probabilities over classes are then given as:

$$p(y|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{f}, \mathbf{g}_k)/\tau)}{\sum_{k'=1}^K \exp(\text{sim}(\mathbf{f}, \mathbf{g}_{k'})/\tau)}, \quad (1)$$

where τ represents the temperature coefficient. Besides, $\text{sim}(\cdot)$ refers to a metric function for measuring the similarity between the input image and the text, which is realized with cosine similarity in CoOp and CoCoOp and with the OT distance (between the local visual features and multiple textual features) in PLOT. For these methods, the text prompts $\{t_i\}_{i=1}^N$

can be optimized via back-propagating the following loss:

$$\mathcal{L}_{ce} = \mathbb{E}_{(\mathbf{x}_i, y_i)} [-\log p(y_i | \mathbf{x}_i)]. \quad (2)$$

After learning, the text prompts can be used to classify test samples. Different from UPT, MPT considers learning both vision and text prompts to improve prompt learning. For example, in MaPle [5], each image $\mathbf{x}$ is associated with $M = 4$ learnable vision prompts $\mathbf{v} = \{\mathbf{v}_i\}_{i=1}^M$, which can be aggregated to obtain $\mathbf{x}$'s visual embedding: $\mathbf{f} = \frac{1}{M} \sum_{i=1}^M F(\mathbf{v}_i, \mathbf{x})$. Text embedding $\mathbf{g}$ is obtained from the text prompts in a similar way as UPT. Accordingly, the vision prompts $\{\mathbf{v}_i\}_{i=1}^M$ and text prompts $\{\mathbf{t}_i\}_{i=1}^N$ can be learned with Eqs. 1 and 2.

2) *Optimal Transport Theory*: Within the rich field of OT theory, this work provides a concise introduction to OT for discrete probability distributions; see [33] for more details. Consider two discrete distributions, $P = \sum_{i=1}^n a_i \delta_{x_i}$ and $Q = \sum_{j=1}^m b_j \delta_{y_j}$, where x_i and y_j exist in an arbitrary common space, and δ represents the Dirac function. Both $\mathbf{a} \in \sum^n$ and $\mathbf{b} \in \sum^m$ are discrete probability vectors summing to 1. The OT distance between P and Q can be expressed as follows:

$$\text{OT}(P, Q) = \min_{\mathbf{T} \in \Pi(P, Q)} \langle \mathbf{T}, \mathbf{C} \rangle, \quad (3)$$

where $\langle, \rangle$ denotes the Frobenius dot-product, and $\mathbf{C} \in \mathbb{R}_{\geq 0}^{n \times m}$ is the transport cost matrix constructed by $C_{ij} = C(x_i, y_j)$. The transport probability matrix $\mathbf{T} \in \mathbb{R}_{\geq 0}^{n \times m}$, which satisfies $\Pi(P, Q) := \{\mathbf{T} | \sum_{i=1}^n T_{ij} = b_j, \sum_{j=1}^m T_{ij} = a_i\}$, is learned by minimizing $\text{OT}(P, Q)$. Directly optimizing Eq. 3 often incurs significant computational demands. To address this challenge, the optimization problem is often adapted to include a popular entropic regularization term $h(\mathbf{T}) = -\sum_{ij} T_{ij} \ln T_{ij}$ for reducing the computational cost [34].

B. Distribution Matching

As discussed above, most of the existing methods learn the prompts based on the pairwise alignment between a text and an image embedding, which is optimized by the classification loss in Eq. 2. Despite their effectiveness, they ignore the correlations between the text and image embeddings at the distributional level when learning the prompts. Moving beyond them, we further consider matching the text and image embeddings from a distributional point of view. Specifically, we take the MPT as an example, which aims to learn both the visual and textual prompts. Now, we denote $\{\mathbf{f}_b(\mathbf{v})\}_{b=1}^B$ as the mini-batch of visual embeddings with the mini-batch size B and $\{\mathbf{g}_k(\mathbf{t})\}_{k=1}^K$ as the textual embeddings with the number of training classes K , where $\mathbf{f}_b(\mathbf{v})$ indicates the dependence of visual embedding on vision prompt and $\mathbf{g}_k(\mathbf{t})$ the dependence of textual embedding on text prompt. Each visual and textual embedding is independently and identically distributed, the probability of each embedding is $1/B$ and $1/K$, following a uniform distribution. Different visual and textual embeddings, along with their occurrence probabilities, describe how the corresponding data is distributed in the feature space. Therefore, the visual embeddings and textual

embeddings can be represented as two empirical distributions:

$$P(\mathbf{v}) = \sum_{b=1}^B \frac{1}{B} \delta_{\mathbf{f}_b(\mathbf{v})}, \quad Q(\mathbf{t}) = \sum_{k=1}^K \frac{1}{K} \delta_{\mathbf{g}_k(\mathbf{t})}, \quad (4)$$

where $\delta_{\mathbf{f}_b(\mathbf{v})}$ and $\delta_{\mathbf{g}_k(\mathbf{t})}$ are Dirac delta functions placed at support points in the embedding space and for each sample we use the uniformly distributed probability. Considering distribution $P(\mathbf{v})$ and distribution $Q(\mathbf{t})$ describe the mini-batch samples from the image modality and text modality respectively, it is natural to learn the prompt by minimizing the OT distance between them:

$$\min_{\mathbf{t}, \mathbf{v}} \text{OT}(P(\mathbf{v}), Q(\mathbf{t})) = \min_{\mathbf{t}, \mathbf{v}} \min_{\mathbf{T} \in \Pi(P(\mathbf{v}), Q(\mathbf{t}))} \langle \mathbf{T}, \mathbf{C} \rangle, \quad (5)$$

where $\mathbf{C} \in \mathbb{R}_{\geq 0}^{B \times K}$ represents a cost function with element C_{bk} measuring the distance between sample b and class k , which can be flexibly defined in different ways. And we empirically find that $C_{bk} = 1 - \cos(\mathbf{f}_b(\mathbf{v}), \mathbf{g}_k(\mathbf{t}))$ is a good choice. The transport plan matrix $\mathbf{T} \in \mathbb{R}_{\geq 0}^{B \times K}$ should satisfy $\{\mathbf{T} | \sum_{i=1}^n T_{ij} = 1/K, \sum_{j=1}^m T_{ij} = 1/B\}$. For Eq.5, we utilize the entropy regularized OT loss [34] with an entropic constraint and rewrite the Eq.5 as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{t}, \mathbf{v}} \mathcal{L}_m &= \langle \mathbf{C}, \mathbf{T}^* \rangle, \\ \text{s.t., } \mathbf{T}^* &= \arg \min_{\mathbf{T} \in \Pi(P(\mathbf{v}), Q(\mathbf{t}))} \langle \mathbf{T}, \mathbf{C} \rangle - \lambda h(\mathbf{T}), \end{aligned} \quad (6)$$

where $\lambda \geq 0$ serves as a hyper-parameter of the entropic regularization. Specifically, we follow the typical setting of OT problem to first minimize the entropy regularized OT loss to obtain the transport plan between two distributions, where $\mathbf{T}^*$ relies on the to-be-learned vision and text prompts. Then, we can use $\mathbf{T}^*$ to compute the entropic regularized OT loss $\mathcal{L}_m$ for learning the vision and text prompts.

C. Cross Entropy Constraint on Transport Plan

Minimizing the OT distance between $P(\mathbf{v})$ and $Q(\mathbf{t})$ can encourage the visual embeddings and its corresponding textual embeddings to be aligned in an unsupervised way according to their similarity (i.e., cost function). Now the learned transport plan $\mathbf{T}^*$ in Eq.6 is the probability matrix of the B visual embeddings being transported to the K textual embeddings during the distribution matching process. That is to say, $\mathbf{T}^*$ indicates that how the B sample be classified into the K classes in the mini-batch. Considering the special property of the learned transport plan $\mathbf{T}^*$ and availability of the label information during the training stage, we further propose a supervised constraint on $\mathbf{T}^*$ to enhance the learning of vision and text prompts. Specifically, we can utilize the learned transport plan to predict the probability of sample b in the mini-batch over K classes:

$$p(y|x_b, \mathbf{T}^*) = \frac{\exp(\mathbf{T}_{bk}^*/\tau)}{\sum_{k'=1}^K \exp(\mathbf{T}_{bk'}^*/\tau)}, \quad (7)$$

Notably, $p(y|x_b, \mathbf{T}^*)$ is different from the pairwise prediction probability in Eq.1, where the similarity $\mathbf{T}_{bk}^*$ in the former is optimized from the view of distribution matching

Algorithm 1 Workflow About Our McPt Method for Optimizing $\mathbf{t}$, $\mathbf{v}$

Require: initial visual prompts $\mathbf{v}$, initial textual prompts $\mathbf{t}$;
for $iter = 1, 2, 3, \dots$ **do**
 Sample a batch of B images ;
 Use image embeddings $\mathbf{f}_b$ to build P in Eq.4 and text embeddings $\mathbf{g}_k$ to build Q in Eq.4 ;
 Calculate the transport plan $\mathbf{T}^*$ between image modality and text modality in Eq.6 ;
 Calculate the OT distance between image modality and text modality $\mathcal{L}_m$ in Eq.6 ;
 Calculate the cross entropy constraint loss between image modality and text modality $\mathcal{L}_c$ in Eq.8 ;
 Update $\mathbf{t}^{(iter+1)}, \mathbf{v}^{(iter+1)} \leftarrow \mathbf{t}^{(iter)}, \mathbf{v}^{(iter)} - \nabla \mathcal{L}_{match}$ where $\mathcal{L}_{match} = \mathcal{L}_{ce} + \beta_m * \mathcal{L}_m + \beta_c \mathcal{L}_c$;
end

and the $\text{sim}(\mathbf{f}, \mathbf{g}_k)$ in the latter is computed in a point-to-point fashion. In other words, since $\mathbf{T}$ should satisfy the two marginal distributions when solving the optimal transport problem, the element $\mathbf{T}_{bk}^*$ in the transport plan matrix is affected by all K textual embeddings and B samples in the current mini-batch rather than the k -th textual embedding and b -th sample. After that, we use the prediction probability $p(y|x_b, \mathbf{T}^*)$ to compute the cross entropy loss and optimize the vision and text prompts:

$$\min_{\mathbf{t}, \mathbf{v}} \mathcal{L}_c = \mathbb{E}_{(x_b, y_b, \mathbf{T}^*)} [-\log p(y_b|x_b, \mathbf{T}^*)], \quad (8)$$

where y_b denotes the ground truth label of x_b . Designed in this way, the prompts are learned to enforce the image sample and its corresponding class close to each other from the view of distribution matching. The probability based on the optimized transport plan is only used for training. At the test stage, we still use the commonly-used probability in Eq.1 as the predicted probability of the test sample.

D. Overall Training Procedure

To summarize, the overall matching loss for optimizing the text and vision prompts ($\mathbf{t}$ and $\mathbf{v}$) in our McPt can be expressed as:

$$\mathcal{L}_{match} = \mathcal{L}_{ce} + \beta_m * \mathcal{L}_m + \beta_c \mathcal{L}_c, \quad (9)$$

where β_m and β_c represent the weights of distribution matching and cross-entropy constraint loss, respectively.

Notably, $\mathcal{L}_{ce}$ can be replaced with the loss used by any existing uni-modal and multi-modal prompt tuning method. Our two loss functions, $\mathcal{L}_m$ and $\mathcal{L}_c$, are designed to enhance the attention to modality matching of text and image from the view of distribution matching. Therefore, our method can be seamlessly combined as a plug-and-play component with any model without adding any additional parameters. We describe our proposed framework in Algorithm 1.

IV. EXPERIMENTS

A. Benchmark Setting

1) *Datasets*: The experiments are conducted on 11 public datasets including Caltech101 [35], DTD [36], EuroSAT [37], FGVC [38], Flowers102 [39], Food101 [40], ImageNet [41], OxfordPets [42], StanfordCars [43], SUN397 [44], and UCF101 [45]. These datasets cover topics such as the classification of generic objects, scenes, actions, and fine-grained categories, as well as specialized tasks like recognizing textures and satellite imagery.

2) *Baselines*: Uni-modal: CoOp [4] optimizes the continuous prompt vectors instead of discrete vectors. CoCoOp [6] extends CoOp by further learning a lightweight network to generate for each image an input-conditional token. PLOT [7] learn multiple comprehensive prompts to describe diverse characteristics of categories. Multi-modal: MaPle [5] promotes strong coupling between the vision-language prompts. ALIGN [8] proposes a multi-mode token-level tuning framework.

3) *Few-Shot Learning*: We follow the few-shot evaluation protocol as outlined in CoOp [4] and CoCoOp [6], employing training sets with 1, 2, 4, 8, and 16 shots while evaluating the original full test set. The reported results are averaged across three runs for comparison.

4) *Generalization From Base-to-Novel Classes*: We assess the generalizability of McPt and adopt a zero-shot setting in which the datasets are partitioned into base and novel classes. The model is exclusively trained on the base classes within a 16-shot framework and subsequently evaluated on both base and novel classes.

5) *Domain Generalization*: We assess the robustness of our method on out-of-distribution datasets. We directly assess the performance of our ImageNet-trained model on four other ImageNet datasets, including ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R, each characterized by different types of domain shifts.

6) *Implementation Details*: As our method is intended to be employed as a plug-and-play component in conjunction with other methods, we follow the settings of the original models, including the choice of backbone and optimizer. For CoOp [4] and PLOT [7], we employ ResNet50 [46] as the image encoder's backbone, with the number of context tokens set to 16. we train our model with a batch size of 16, SGD optimizer with an initial learning rate of 0.002 and train for 200 epochs. For CoCoOp [6], MaPle [5], and Align [8], we adopt the settings as detailed in their respective papers. We use ViT-B/16 as the image encoder backbone and set the number of textual and visual prompts to 4. We employ the batch size of 4, the SGD optimizer with an initial learning rate of 0.0035, and optimize the number of epochs for each task. The weights of distribution matching and cross-entropy constraint loss are set to 0.5 and 0.3, respectively. The weight of the regularization term in Sinkhorn loss is set to 0.1.

B. Evaluation With the Standard Setting

1) *Comparison on Few-Shot Learning*: To assess the efficacy of ours in few-shot learning, we integrate it with three

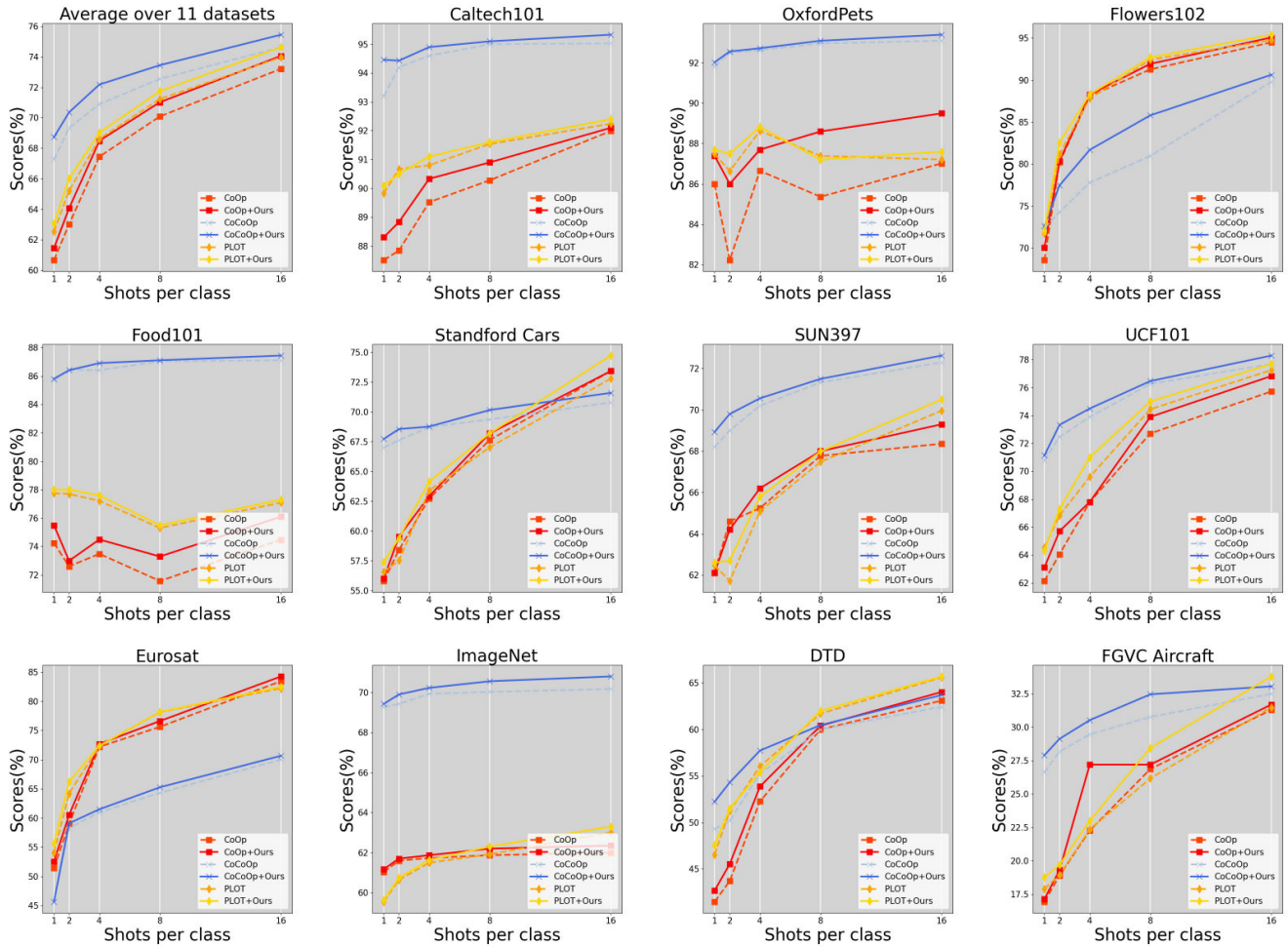


Fig. 3. The few-shot learning results on eleven datasets. We compare our McPt with CoOp, CoCoOp, and PLOT methods and observe the consistent and significant performance improvement on most datasets. (The average accuracy on all datasets is shown on the left top.)

prevalent prompt tuning methods: CoOp [4], CoCoOp [6], and PLOT [7]. Note that our evaluation of CoCoOp follows their original setting by utilizing ViT-B/16 as the backbone, which differs from the ResNet50 in CoOp [4] and PLOT [7]. We conduct the experiments using training sets with 1, 2, 4, 8, and 16 shots while assessing the original full test set across 11 datasets. Fig 3 showcases the accuracy result for each prompt learning model across different few-shot settings. Overall, ours results in an average improvement of over 1% point across the 11 datasets. This demonstrates our capacity by introducing distribution matching and cross entropy constraints between text and image modalities.

2) *Base-to-Novel Generalization*: To verify our method's generalization capabilities, we adopt the protocol established in ALIGN [8], where the training set is divided into non-overlapping base and novel classes. Training is exclusively performed on the base classes using 16 shots while testing is conducted across all novel classes. We use CoCoOp [6] as our UPT baseline and ALIGN [8] as our MPT baseline. The results provided in Table I demonstrate the performance improvements for both CoCoOp+Ours and ALIGN+Ours across base and novel classes. CoCoOp+Ours displays enhanced performance, improving the base class by

an average of **0.3%** points and the novel class by **1.5%** points compared to the baseline CoCoOp. For ALIGN+Ours, our method demonstrates a stronger improvement in the base class, achieving an increase of **0.59%** points and **0.25%** points in novel classes. These findings suggest that learning modality-consistent representation through distribution-wise matching not only enhances the model's performance on base classes of downstream tasks but also boosts its generalization ability for new classes. Ultimately, this approach improves the model's overall recognition performance across various downstream tasks.

3) *Domain Generalization*: In this setting, we train our model on ImageNet and test it across other 4 domain shift datasets, demonstrating the model's generalizability. As listed in Table II, our method not only shows improvements within the source domain but also outperforms the baseline methods on **3/4** datasets on domain generalization. We achieved the highest improvement of 0.67% on UPT+Ours on ImageNet-A and 0.85% on MPT+Ours on ImageNet-R. This suggests that ours for considering modality match enhances the generality of modal features across diverse domains, subsequently bolstering the model's robustness in recognizing and processing information.

TABLE I

BASE-TO-NEW ON ELEVEN DATASETS. THE PROMPTS ARE LEARNED FROM THE 16-SHOT BASE SET. WE REPORT THE CLASSIFICATION ACCURACY ON THE BASE SET (BASE), NEW SET (NEW), AND THEIR HARMONIC MEAN (H). UPT STANDS FOR UNI-MODAL PROMPT TUNING, MPT STANDS FOR MULTI-MODAL PROMPT TUNING. WE RUN THE RESULTS OF ALIGN FOLLOWING THE OFFICIALLY RELEASED CODE. THE BEST RESULTS IN MPT ARE IN **BOLD** WHILE THE SECOND BEST RESULTS ARE MARKED WITH UNDERLINE

		Year	Publish	Average			ImageNet			Caltech101			Oxford Pets		
				Base	New	H	Base	New	H	Base	New	H	Base	New	H
UPT	CLIP	2021	ICML	69.34	74.22	71.69 ^{+5.27}	72.34	68.14	70.21 ^{+2.95}	96.84	94.00	95.39 ^{+0.52}	91.17	97.26	94.11 ^{+2.44}
	CoOp	2022	IJCV	82.66	63.22	71.65 ^{+5.31}	76.14	67.88	71.77 ^{+1.39}	98.00	89.81	93.72 ^{+2.19}	93.67	95.29	94.47 ^{+1.88}
	CoCoOp	2022	CVPR	80.47	71.69	75.83 ^{+1.13}	75.98	70.43	73.10 ^{+0.06}	97.96	93.81	95.84 ^{+0.07}	95.20	97.69	96.43 ^{+0.12}
	PLOT	2023	ICLR	77.20	60.38	67.76 ^{+9.20}	75.97	69.23	72.44 ^{+0.72}	96.53	82.86	89.17 ^{+6.74}	93.45	79.76	86.06 ^{+10.49}
	CoCoOp+Ours			80.75	73.17	76.96	75.98	70.73	73.16	98.13	93.90	95.91	95.40	97.71	96.55
MPT	MaPLe	2023	CVPR	82.28	75.14	78.55 ^{+1.06}	76.66	70.54	73.47 ^{+0.96}	97.71	94.36	96.02 ^{+0.74}	95.43	97.76	96.58 ^{+0.41}
	ALIGN	2023	NIPS	83.33	75.28	79.17 ^{+0.44}	76.89	72.15	74.45 ^{-0.02}	98.37	94.70	96.50 ^{+0.26}	95.67	97.93	96.79 ^{+0.20}
	ALIGN+Ours			83.92	75.53	79.61	76.98	<u>72.05</u>	<u>74.43</u>	98.67	<u>94.63</u>	96.76	95.76	98.23	96.99

		Year	Publish	Stanford Cars			Flowers 102			Food101			FGVC Aircraft		
				Base	New	H	Base	New	H	Base	New	H	Base	New	H
UPT	CLIP	2021	ICML	63.37	74.89	68.65 ^{+3.47}	72.08	77.80	74.83 ^{+7.10}	90.10	91.22	90.65 ^{+0.45}	27.19	36.29	31.08 ^{+1.22}
	CoOp	2022	IJCV	78.12	60.40	68.12 ^{+4.00}	97.60	59.67	74.06 ^{+7.87}	88.33	82.26	85.18 ^{+5.92}	40.44	22.30	28.74 ^{+3.56}
	CoCoOp	2022	CVPR	70.49	73.59	72.10 ^{+0.02}	94.87	71.75	81.71 ^{+0.22}	90.70	91.29	90.99 ^{+0.01}	33.41	23.71	27.71 ^{+4.59}
	PLOT	2023	ICLR	61.41	42.69	50.37 ^{+21.75}	95.26	56.03	70.56 ^{+11.37}	88.45	85.28	86.84 ^{+4.26}	29.63	16.17	20.92 ^{+11.38}
	CoCoOp+Ours			70.50	73.75	72.12	95.50	71.80	81.93	90.63	91.60	91.10	34.50	30.10	32.30
MPT	MaPLe	2023	CVPR	72.94	74.00	73.47 ^{+2.26}	95.92	72.46	82.56 ^{+0.81}	90.71	92.05	91.38 ^{+0.08}	37.44	35.61	36.50 ^{+1.33}
	ALIGN	2023	NIPS	77.24	74.03	75.63 ^{+0.10}	97.43	72.25	83.17 ^{+0.20}	90.23	92.07	91.15 ^{+0.28}	37.90	36.25	37.07 ^{+0.76}
	ALIGN+Ours			<u>77.19</u>	74.20	75.73	<u>97.20</u>	72.87	83.37	90.75	92.10	91.46	39.10	36.50	37.83

		Year	Publish	SUN 397			DTD			EuroSAT			UCF 101		
				Base	New	H	Base	New	H	Base	New	H	Base	New	H
UPT	CLIP	2021	ICML	69.36	75.35	72.23 ^{+6.32}	53.24	59.90	56.37 ^{+10.66}	56.48	64.05	60.02	70.53	77.50	73.85 ^{+4.13}
	CoOp	2022	IJCV	80.60	65.89	72.50 ^{+6.05}	79.44	41.18	54.24 ^{+12.79}	92.19	54.74	68.69 ^{+9.89}	84.69	56.05	67.45 ^{+10.53}
	CoCoOp	2022	CVPR	79.74	76.86	78.27 ^{+0.31}	77.01	56.00	64.85 ^{+2.18}	87.49	60.04	71.21 ^{+7.37}	82.33	73.45	77.64 ^{+0.34}
	PLOT	2023	ICLR	78.56	72.34	75.32 ^{+3.23}	69.87	53.63	60.68 ^{+6.35}	87.39	67.63	74.30 ^{+4.28}	72.71	41.51	52.84 ^{+25.14}
	CoCoOp+Ours			79.90	77.20	78.55	77.57	56.50	67.03	88.46	68.70	78.58 ^{+18.56}	82.34	73.90	77.98
MPT	MaPLe	2023	CVPR	80.82	78.70	79.75 ^{+0.52}	80.36	59.18	68.16 ^{+3.70}	94.07	73.23	82.35 ^{+0.74}	83.00	78.66	80.77 ^{+2.04}
	ALIGN	2023	NIPS	81.96	78.33	80.14 ^{+0.13}	82.13	59.46	70.79 ^{+1.07}	94.43	73.43	83.93 ^{+0.16}	84.43	78.33	81.27 ^{+1.54}
	ALIGN+Ours			<u>81.69</u>	78.80	80.27	84.10	59.56	71.86	95.43	73.77	84.09	86.83	78.73	82.81

TABLE II

CROSS-DOMAIN GENERALIZATION ACCURACY RESULTS.
THE BEST RESULTS ARE IN **BOLD**

Method	Source ImageNet	-V2	Target -Sketch	-A	-R
CLIP [1]	66.73	60.83	46.15	47.77	73.96
CoOp [4]	71.51	64.20	47.99	49.71	75.21
CoCoOp [6]	71.02	64.07	48.75	50.63	76.18
CoCoOp+Ours	71.20	64.10	48.83	51.30	76.50
MaPLe [5]	70.72	64.07	49.15	50.90	76.98
ALIGN [8]	72.03	64.64	49.96	50.94	76.16
ALIGN+Ours	72.13	64.73	49.87	51.03	77.01

C. Ablation Study

In this subsection, we conduct ablation studies to investigate the effectiveness of different losses. We use the SOTA MPT method ALIGN [8] as a baseline and observe the model's accuracy on the four datasets encompassing fine-grained objects: FGVC Aircraft, action recognition: UCF101, texture recognition: DTD and satellite object: EuroSAT. Due to the fact that the transport plan is based on the calculation of distribution matching loss, we demonstrate the effect of the cross entropy constraint on the transport plan loss through overlapping use. As shown in Table III, both components are helpful for improving the prompt learning for different shot settings. Notably, using distribution matching loss alone has effectively improved the model under different few-shot settings: 1.96% on EuroSat with 16 shots and 1.92% on UCF101 with 8 shots. With the assistance of cross entropy constraint on the transport plan loss, the accuracy of the model will be further improved by about 0.2% points. This

TABLE III

ACCURACY SCORES (%) ON FGVC AIRCRAFT, UCF101, DTD, AND EUROSAT DATASETS ARE REPORTED TO VERIFY THE EFFECTS OF DIFFERENT LOSSES. $\mathcal{M}$ AND $\mathcal{C}$ REPRESENT THE DISTRIBUTION-WISE AND CROSS ENTROPY CONSTRAINT ON TRANSPORT PLAN LOSS FUNCTIONS, RESPECTIVELY

	$\mathcal{M}$	$\mathcal{C}$	1 shot	2 shot	4 shot	8 shot	16 shot
FGVC			29.57	31.63	34.03	40.95	49.99
	✓		29.94	32.13	34.23	42.15	50.61
	✓	✓	30.23	32.34	34.41	42.83	50.82
UCF			74.42	75.87	80.18	81.99	85.69
	✓		74.88	76.83	80.64	83.43	86.12
	✓	✓	75.01	77.21	80.73	83.91	86.37
DTD			54.07	56.53	63.30	67.60	71.40
	✓		54.94	57.14	63.94	68.02	72.02
	✓	✓	55.13	57.41	64.13	68.25	72.28
EURO			53.23	71.43	80.93	85.97	90.77
	✓		54.65	71.95	81.21	87.16	92.73
	✓	✓	55.21	72.13	81.41	87.67	93.03

underscores that both components contribute to improving the accuracy and validating the efficacy of matching modalities in a progressive fashion.

To further explore the relationship between pair-wise alignment and distribution-wise matching, we conduct experiments on the DTD dataset, evaluating model performance under different few-shot settings by removing either the distribution-wise loss or the pairwise cross-entropy loss. As shown in Table IV, when only the pairwise cross-entropy loss is used, the model degenerates to the original baseline, ALIGN [8], where no information interaction occurs between

TABLE IV

ACCURACY SCORES (%) ON DTD DATASET IS REPORTED TO VERIFY THE EFFECTS OF DIFFERENT LOSSES. $\mathcal{MC}$ AND $\mathcal{CE}$ REPRESENT THE DISTRIBUTION-WISE, CROSS ENTROPY CONSTRAINT ON TRANSPORT PLAN LOSS AND CROSS ENTROPY LOSS FUNCTIONS, RESPECTIVELY. THE BEST RESULTS ARE IN **BOLD**

$\mathcal{CE}$	$\mathcal{MC}$	1 shot	2 shot	4 shot	8 shot	16 shot
✓		54.07	56.53	63.30	67.60	71.40
	✓	50.97	51.26	53.37	55.43	58.04
✓	✓	55.13	57.41	64.13	68.25	72.28

visual embeddings, and the model can only achieve its initial performance. When only the distribution-wise loss is used to optimize the model, the absence of pairwise loss guidance causes the model to lose focus on aligning specific visual and text embeddings. Although information is exchanged between visual embeddings, this confuses the optimization direction, leading to a significant drop in performance. This is because the pairwise alignment loss is a supervised task loss, while our distribution-wise loss serves as an unsupervised regularization. As a result, the supervised loss is mandatory. However, when both pairwise and distribution-wise losses are used simultaneously, the model benefits from the interaction of information across multiple visual samples, which enhances pairwise matching and improves recognition performance. This demonstrates a complementary relationship between pairwise and distribution-wise losses.

D. Analysis of the Batchsize

In examining the impact of varying batch sizes on model performance, ALIGN [8] and CoOp [4] serve as the UPT and MPT baselines, respectively. Employing the base-to-novel (16-shots) setting across FGVC Aircraft, DTD, and EuroSat datasets, we compare model accuracy with and without our McPt method at different batch sizes. The average harmonic accuracy of the three datasets is reported. As depicted in Fig. 4, an unexpected trend emerges where model accuracy steadily decreases as the batch size increases. We attribute this phenomenon to the large batch size inducing overfitting of the downstream task's training set, thereby diminishing the model's generalization capacity on the test set and novel classes. Despite these challenges, McPt consistently enhances the performance of both baselines under various overfitting scenarios. Remarkably, when the batch size is set to 4, McPt exhibits the highest accuracy improvement of 0.63% for ALIGN, while at a batch size of 32, it achieves a notable 2.42% improvement for CoOp. This underscores the efficacy of McPt in enhancing model classification by leveraging sample relationships and modality distribution matching. However, it also underscores the nuanced balance between classification ability and overfitting concerns in UPT and MPT models across downstream tasks, suggesting that higher sample sizes do not necessarily equate to superior performance.

E. Analysis of the Weights of Loss Functions

To determine the optimal values for β_m and β_c in Eq.9, we use ALIGN [8] as a baseline and conduct experiments

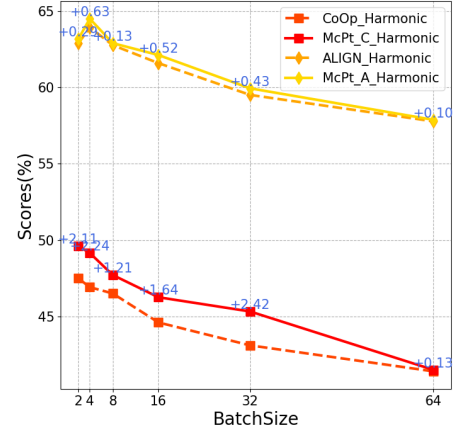


Fig. 4. UPT (CoOp) and MPT (ALIGN) are integrated with our McPt to assess performance variations across different batch sizes. We utilize the base-to-novel setting and present the harmonic accuracy of the model output. The highlighted values in blue denote the degree to which our method surpasses the respective baseline.

TABLE V

COMPARISON ACCURACY SCORES (%) WITH DIFFERENT WEIGHTS VALUE. WE USE ALIGN [8] AS A BASELINE AND CONDUCT EXPERIMENTS UNDER THE BASE-TO-NOVEL (16-SHOTS) SETTING ACROSS THE FGVC AIRCRAFT, DTD, AND EUROSAT DATASETS. WE MAINLY COMPARE THE HARMONIC ACCURACY OF THE MODELS. THE BEST RESULTS ARE IN **BOLD**

β_m	Base	New	H	β_c	Base	New	H
0.3	72.70	55.71	64.21	0.1	72.88	55.42	64.15
0.5	72.78	56.26	64.52	0.3	72.78	56.26	64.52
1.0	72.87	55.91	64.39	0.5	72.69	55.68	64.19

under the base-to-novel (16-shots) setting across the FGVC Aircraft, DTD, and EuroSat datasets. We test different values for β_m and β_c and compare the final model's accuracy results. As shown in Table V, we increase β_m and β_c incrementally from 0.3 to 1 and 0.1 to 0.5, respectively. Adjusting β_m resulted in the model's harmonic accuracy fluctuating between a minimum of 64.21% and a maximum of 64.52%, a variation of only 0.31%. Adjusting β_c led to the model's harmonic accuracy ranging from a minimum of 64.15% to a maximum of 64.52%, a difference of 0.37%. These findings indicate that our model's performance is not highly sensitive to the values of the two weights and consistently maintains high accuracy. Notably, the model performed best with an accuracy of 64.52% when β_m was set to 0.5 and β_c was set to 0.3. Therefore, when applying our McPt method, we recommend setting β_m to 0.5 and β_c to 0.3.

F. Analysis of the λ of the Optimal Transport Theory

In Eq. 6, the parameter λ represents the introduction of an entropy regularization term during the optimal transport process. Its main function is to smooth the optimization problem, ensure numerical stability, and speed up the solution. As λ increases, the optimization process becomes smoother. To investigate the effect of different values of λ on model performance, we conduct experiments using a 16-shot setting on the EuroSat dataset, adjusting λ values to compare the

TABLE VI

COMPARISON ACCURACY SCORES (%) WITH DIFFERENT λ VALUE. WE USE ALIGN [8] AS A BASELINE AND CONDUCT EXPERIMENTS UNDER FEW SHOT (16-SHOTS) SETTING ON THE EUROSAT DATASET. THE BEST RESULTS ARE IN **BOLD**

λ	0.5	0.1	0.01	0.001
McPt	92.96	93.03	92.74	92.73

TABLE VII

COMPARISON TO OTHER METHODS WITH COMPUTATION COST

Method	Parameters (M)	Trainable Parameters (M)	fps (frames/s)
CoOp [4]	76.71	0.01	619.79
CoOp+Ours	76.71	0.01	607.27
CoCoOp [6]	124.35	0.04	49.87
CoCoOp+Ours	124.35	0.04	46.30
ALIGN [8]	127.90	3.57	52.39
ALIGN+Ours	127.90	3.57	51.25

results. As shown in Table VI, different values of λ resulted in minimal fluctuations in model performance, within 0.3%, demonstrating that our McPt method is robust to parameter changes. A smaller λ value indicates a steeper distribution when matching text and images, which may prevent the image embedding from fully interacting with all text embedding information. On the other hand, a larger λ weakens the matching relationship between the image embedding and the corresponding text embedding. However, overall, the effect of varying λ values on the results is not significant. Thus, we select a common λ value of 0.1, which provided relatively strong results, for use in our experiments.

G. Analysis of Computation Cost

To demonstrate the computational cost, we provide a comparison of CoOp [4], CoCoOp [6], and ALIGN [8] with and without our McPt. We report the number of overall parameters, trainable parameters, and frames per second (fps) during training. As depicted in Table VII, the number of trainable parameters in multi-modal prompt tuning models (MPT) surpasses that in uni-modal prompt tuning models (UPT), and the training speed of MPT is slower compared to UPT. However, our proposed McPt does not require additional training parameters for both UPT and MPT. Additionally, by incorporating our McPt, the training speed of the model does not significantly decrease. The most substantial decrease is observed in CoOp, with only a 12fps difference. For CoCoOp and ALIGN, the differences are merely 3fps and 1fps, respectively, which can be considered negligible. Overall, our method enhances model performance without substantially increasing computational complexity.

H. Analysis of Different Backbone

Our McPt has been integrated with different backbones (ResNet50 for CoOp and ViT-B/16 for CoCoOp) to demonstrate its effectiveness. To further explore the impact of our method across different backbones, we sample three different datasets: DTD, EuroSat, and FGVC-AirCRAFT, and compare the classification accuracy results with CoOp under varying

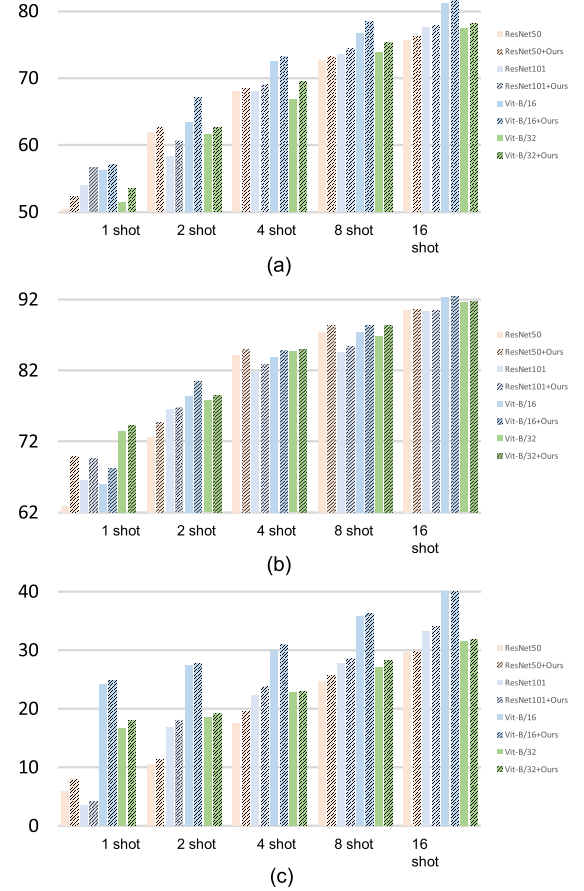


Fig. 5. Accuracy scores of different backbones: ResNet50, ResNet101, ViT-B/16, and ViT-B/32 on (a) DTD, (b) EuroSat, and (c) FGVC-AirCRAFT dataset.

backbones. As presented in Fig. 5, generally, the classification results with ViT [47] as the backbone tends to outperform those with ResNet, and the model's performance gradually improves with an increase in the number of shots. Importantly, the addition of our McPt consistently enhances the model's performance, regardless of whether ViT or ResNet is used as the backbone. This indicates that matching the distributions of different modalities can contribute to the generation of more robust features, regardless of the specific backbone employed.

I. Analysis of Different Distance Metrics

We use Cosine Distance as the cost function in Optimal Transport (OT). To evaluate the impact of different distance metrics on the results, we conduct 16-shot experiments across four datasets (FGVC Aircraft, UCF101, DTD, and EuroSat) and compare the outcomes. As shown in Table IX, we evaluate Manhattan distance, Euclidean distance, and Cosine distance. The results indicate that the choice of distance metric has minimal effect on model accuracy, with fluctuations within 0.2%. However, Cosine distance performs slightly better. We attribute this to two reasons: First, during the pre-training phase of the CLIP model, cosine similarity is used as the self-supervised loss for optimization, so using the same distance metric during model transfer to downstream tasks is advantageous. Second,

text-image matching is more concerned with similarity rather than absolute distance in a coordinate system, making cosine similarity more suitable. For these reasons, we use Cosine Distance as the cost measure in McPt's OT calculations.

J. Analysis of Different Distribution Matching Methods

To investigate the impact of different distribution matching methods on model performance, we apply Maximum Mean Discrepancy (MMD) and Kullback-Leibler Divergence (KL Divergence) to the FGVC Aircraft, UCF101, DTD, and EuroSat datasets under a 16-shot setting and compared the experimental results. As shown in Table X, the model's performance decreases by approximately 1.2% when using MMD and KL compared to Optimal Transport (OT). We attribute this to the distinct focus of each distribution matching method. MMD primarily measures the distance between two distributions, $MMD(P, Q) = \|\mathbb{E}_{x \sim P}[\phi(x)] - \mathbb{E}_{y \sim Q}[\phi(y)]\|_{\mathcal{H}}$, where $\phi(x)$ is the kernel function that maps samples to an RKHS. While using it as a loss function brings the text and image modalities closer together, it is primarily concerned with matching the shape of the overall distribution and does not necessarily lead to more accurate alignment between the text and image embeddings, resulting in limited improvement. KL divergence, $D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$ on the other hand, measures the information loss between two distributions but does not facilitate interaction between the current visual embedding and other samples, limiting its impact on model enhancement. In contrast, $OT(P, Q) = \min_{T \in \Pi(P, Q)} \langle T, C \rangle$, provides a more geometric and robust comparison between distributions and ensures interaction between different samples, leading to more substantial performance improvements.

K. Comparison on Tip-Adapter

Our McPt method is initially devised to address the distribution matching challenge in prompt learning. However, the Adapter method is also a prevalent solution in multimodal learning. Consequently, we integrated our method into the classic adapter model Tip-Adapter [48] and conducted training and testing in a 16-shot scenario on three different types of datasets. As depicted in Table VIII, although the improvement is modest, our model still imparts benefits to the performance of the adapter on diverse datasets. The relatively subdued enhancement can be attributed to the fact that the adapter adds a linear layer for feature mapping after the backbone. In comparison to prompt learning, which adjusts from the input end, the adapter's adjustment is less pronounced, resulting in a slightly milder performance improvement. Nevertheless, our method of distribution-wise modality matching proves equally effective in the adapter framework.

L. Comparison With PEFT and Cross-Modal Learning Methods

To further validate the effectiveness of our McPt method, a Parameter-Efficient Fine-Tuning (PEFT) approach for cross-modal learning, we compare McPt with other PEFT and

TABLE VIII
COMPARISON TO TIP-ADAPTER WITH FEW-SHOT SETTING

Method	DTD	FGVC Aircraft	EuroSat
Tip-Adapter	67.08	35.55	84.79
Tip-Adapter+Ours	67.20	35.67	84.93

TABLE IX
COMPARISON OF DISTANCE METRICS WITH ACCURACY SCORES (%) ON FGVC AIRCRAFT, UCF101, DTD, AND EUROSAT DATASETS

Method	FGVC	UCF101	DTD	EuroSat	Average
Manhattan Distance	50.60	86.12	72.03	92.89	75.41
Euclidean Distance	50.75	86.24	72.13	92.98	75.52
Cosine Distance	50.82	86.37	72.28	93.03	75.62

TABLE X
COMPARISON OF DISTRIBUTION MATCHING METHODS WITH ACCURACY SCORES (%) ON FGVC AIRCRAFT, UCF101, DTD, AND EUROSAT DATASETS. THE BEST RESULTS ARE IN **BOLD**

Method	FGVC	UCF101	DTD	EuroSat	Average
Maximum Mean Discrepancy	49.85	85.57	71.29	91.03	74.43
Kullback-Leibler	50.03	85.80	71.52	90.92	74.56
Optimal Transport	50.82	86.37	72.28	93.03	75.62

cross-modal methods on the FGVC Aircraft, UCF101, DTD, and EuroSat datasets using a 16-shot setting, focusing on recognition accuracy. As shown in Table XI, Adapter [19], Prefix-tuning [20], and Lora [21] are PEFT methods. We combine these methods with CLIP to optimize continuous prompt parameters. For Lora [21], we fix the input prompt as 'A photo of a' and optimize the model's linear layer using a low-rank approach. Additionally, we compare McPt with other cross-modal methods: GOT [28] uses Optimal Transport for graph matching to address cross-domain alignment, NAAF [29] employs a two-branch matching module that utilizes both mismatched and matched textual fragments for accurate image-text matching, UARDA [30] provides a unified learning framework with joint optimization for more discriminative fragment features and optimal relevance thresholds, ESL [31] extends static single-dimensional correspondence to a dynamic mechanism for enhancing combinatorial expressive power, and BOOM [32] considers bidirectional semantic alignment consistency for both 'word-to-region' and 'region-to-word' alignments in image-text matching. We use CLIP's image/text encoder in conjunction with these cross-modal methods.

As shown in Table XI, our McPt method outperforms most PEFT methods, indicating that fine-tuning parameters in a distribution-wise manner is highly effective for adapting models to downstream tasks within the PEFT framework. This effectiveness is also reflected in the strong performance of GOT. However, unlike McPt, GOT requires the construction of graphs with node and edge information, which introduces additional computational complexity compared to McPt's direct matching of text and image data, therefore McPt appears to be a bit more cost-effective. McPt also maintains a state-of-the-art (SOTA) level compared to other cross-modal methods. It's worth noting that methods like BOOM, which performs

TABLE XI

COMPARISON WITH OTHER PEFT AND CROSS-MODAL LEARNING METHODS WITH ACCURACY SCORES (%) ON FGVC AIRCRAFT, UCF101, DTD, AND EUROSAT DATASETS

Method	FGVC	UCF101	DTD	EuroSat	Average
Adapter [19]	35.55	83.92	67.08	84.79	64.83
Prefix-tuning [20]	43.28	83.16	69.73	74.92	67.90
Lora [21]	50.22	86.71	72.03	92.17	75.28
GOT [28]	50.67	86.13	72.02	92.99	75.45
NAAF [29]	50.01	85.63	71.40	91.89	74.73
UARDa [30]	50.05	85.73	71.33	92.46	74.89
ESL [31]	50.37	86.03	71.82	92.73	75.23
BOOM [32]	50.43	86.12	71.91	92.82	75.32
ALIGN+Ours	50.82	86.37	72.28	93.03	75.62

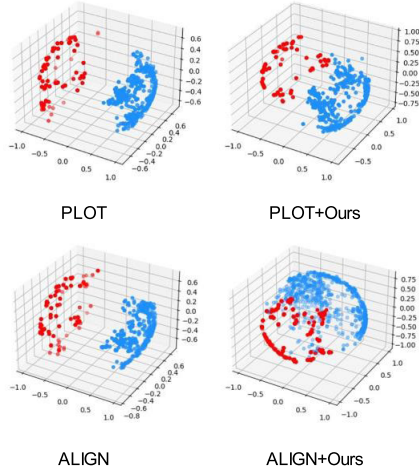


Fig. 6. DOSNES embedding visualization of PLOT, PLOT+Ours, ALIGN and ALIGN+Ours. Blue denotes the image embedding, and red denotes the text embedding.

bidirectional matching between text and images, also improve upon the CLIP baseline. However, BOOM's approach is primarily designed for image-text retrieval tasks, where images contain multiple regions and texts contain multiple nouns. This makes it less suitable for scenarios where each image contains only one instance and each text corresponds to a single category, as is the case in the current image classification task. Despite these improvements, McPt remains superior in maintaining a higher performance standard. Overall, by comparing various PEFT and cross-modal approaches, we verify that our McPt method, which matches the textual and image modalities in a distribution-wise manner, enables better model adaptation to downstream tasks.

M. Qualitative Analysis

1) *Analysis of Modality Distribution*: To explore the distribution of the model's embeddings in the feature space before and after applying our McPt, we visualize the text embeddings and image embeddings of the model's output on the UCF101 dataset with DOSNES [49]. Furthermore, to highlight the distinctions between our method and PLOT and ALIGN, which also utilize the optimal transport theory, we use PLOT and ALIGN as baselines. Additionally, we report the distribution of the distance between positive pairs in each case, employing the cosine distance. As shown in Fig. 6, before utilizing our

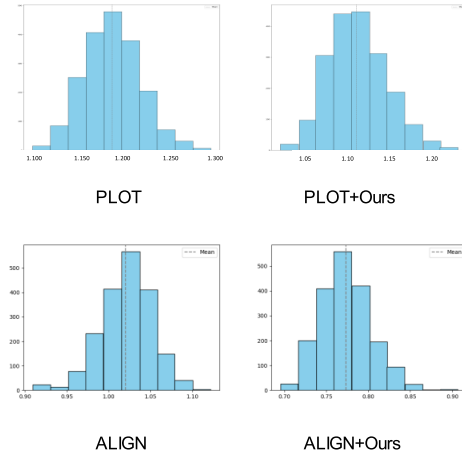


Fig. 7. Histogram of the distribution of Cosine distances between positive pairs of PLOT, PLOT+Ours, ALIGN, and ALIGN+Ours. Our methods induce a more improved matched distribution and well-aligned distance.

McPt, the text embeddings and image embeddings from the original PLOT are ALIGN distributed at two ends of the space with an accuracy of **77.9%** and **86.6%**. Corresponding to this, in Fig. 7, the mean matching distance between positive pairs in PLOT is **1.18**, and in ALIGN, it is **1.00**. This suggests that there is indeed a discernible gap between the representations of different modalities in the model trained using the pairwise method. However, after incorporating our method, the embeddings of the two modalities matched with an accuracy of **78.3%** and **87.2%**, and the distance between positive pairs contracts to a mean value of **1.10** and **0.75**. This suggests that our distribution-wise matching method McPt is indeed distinct from the pair-wise method PLOT and ALIGN. Through matching between modalities, we have effectively enhanced the similarity between positive pairs and consequently improved the overall model accuracy.

2) *Analysis of Transport Plan*: To illustrate the effectiveness of our distribution-wise McPt, let's consider a toy example. We use ALIGN + ours and highlight the distinctions in pair-wise cost matrix and distribution-wise transport plans. As shown in Fig. 8, we randomly select 10 categories from UCF101 as text modalities and pick 3 samples for each category as image modalities. The horizontal axis represents the 10 classes, and the vertical axis represents the corresponding 30 image samples. It is evident that pair-wise alignment concentrates on instance-level matching. Some samples are inaccurately learned and misclassified into other categories in the cost matrix obtained through pair-wise alignment. However, our McPt, focusing on distribution-level matching, considers relationships among all samples. By leveraging the relationships between different visual embeddings, our McPt model enables them to interact with one another during the matching process, thereby enhancing and refining the pairwise visual-text matching. Therefore, the learned transport plan can align the visual embeddings and their corresponding class textual embeddings in a better way, where some misaligned samples in the cost matrix can be corrected. It thus indicates the difference between pair-wise alignment

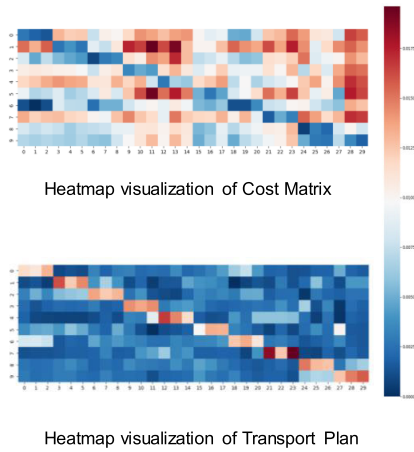


Fig. 8. Heatmap visualization of the pair-wise cost matrix and distribution-wise transport plan. Our McPt corrected the samples that were originally transmitted incorrectly through distribution-wise matching.

and distribution-wise alignment and explains the effectiveness of supervised constraint on the transport plan.

3) *Analysis of t-SNE*: The previous visualizations illustrate the changes between modalities before and after applying our method. The following visualizations will showcase changes within modalities. Using ALIGN as a baseline and Caltech101 dataset, given that text embeddings are just a bunch of discrete points, we conduct t-SNE visualization of the image embeddings. As depicted in Fig. 9, different colors represent the different classes. Even though we matched the two modalities closely, the relative relationship between the classes is more distinct within a single modality. This is attributed to the fact that after modality matching, aligning between positive pairs becomes easier, improving the feature distribution of the image embedding. This observation supports the notion that matching multiple modalities enhances the robustness of representations within a single modality.

4) *Analysis of Calibrated Samples*: To further investigate the effectiveness of our McPt method, we compare the misclassified samples from the original ALIGN model with the McPt-corrected logits on the UCF101 dataset to observe how McPt influences the model's predictions. As shown in Fig. 10, the 'Eye Makeup' and 'Lipstick' images are visually quite similar, leading the baseline model to misclassify 'Eye Makeup' samples as 'Lipstick' by a narrow margin. However, after optimization with our modality-consistent learning model McPt, the previously misclassified 'Eye Makeup' samples outperform the 'Lipstick' category, enabling the model to correctly differentiate between these easily confused classes. A similar improvement is observed for the 'Cricket Bowling' and 'Cricket Shot' categories, where McPt helps maintain consistency between these closely related actions. This suggests that distribution-wise matching significantly enhances the model's classification performance by allowing intra-modality samples to interact and share information, thereby guiding the model's learning process more effectively.

N. Dissicusion

We propose the modality-consistent prompt tuning model McPt, it is still challenging for McPt to handle situations

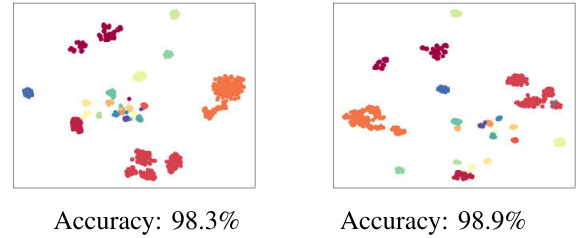


Fig. 9. t-SNE visualization of image embedding before and after using our McPt method. Accuracy 98.3% for ALIGN (left) and Accuracy 98.9% for ALIGN+ours (right). Different colors represent different image classes. Our McPt can improve the inter-class differentiation within the image modality.

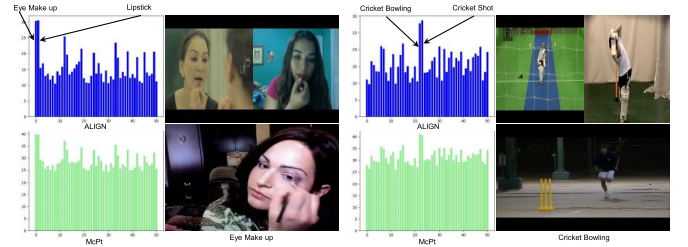


Fig. 10. Visualization of the distribution of logits for misclassified samples and the corrected logits for split pairs after being processed by the McPt model. The images of 'Eye Makeup' and 'Lipstick' are quite similar, leading the ALIGN model to incorrectly classify the 'Eye Makeup' image as 'Lipstick' by a small margin. McPt successfully corrects this misclassification. Similarly, McPt corrects the 'Cricket Bowling' image, originally misclassified as 'Cricket Shot'.

where multiple instances exist in one image. When an image contains multiple instances, the visual embeddings extracted by the model represent information from all instances, making it difficult to fully align with the textual embeddings corresponding to a specific instance. This mismatch can reduce the model's recognition accuracy. In future work, we will explore ways to better align the information between both models, focusing on improving the accuracy for individual instances within a single sample.

V. CONCLUSION

From the standpoint of how the human brain functions. This paper introduces a novel modality-consistent learning method, McPt, for prompt tuning based on OT, which matches multiple modalities distributionally and constrains the transport plan with cross-entropy, which is a complementary approach to existing methods that usually match samples pairwise. Our proposed one can be applied to improve existing uni-modal and multi-modal prompt learning methods for being a plug-and-play method, which can generate modality-consistent representations. Extensive experimental results on a variety of recognition datasets have validated the effectiveness and flexibility of our proposed method. We believe that ours comes from a different view to most of the existing prompt tuning methods and may provide new thoughts for future work.

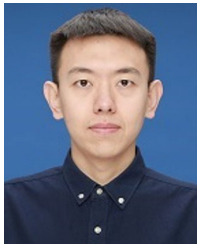
REFERENCES

- [1] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [2] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, pp. 4904–4916. [Online]. Available: <https://proceedings.mlr.press/v139/jia21b.html>

- [3] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, p. 12.
- [4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.
- [5] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19113–19122.
- [6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16795–16804.
- [7] G. Chen, W. Yao, X. Song, X. Li, Y. Rao, and K. Zhang, "Plot: Prompt learning with optimal transport for vision-language models," 2022, *arXiv:2210.01253*.
- [8] D. Wang, M. Li, X. Liu, M. Xu, B. Chen, and H. Zhang, "Tuning multi-mode token-level prompt alignment across modalities," 2023, *arXiv:2309.13847*.
- [9] X. Hao, X. Jiang, W. Ni, W. Tan, and B. Yan, "Prompt-guided semantic-aware distillation for weakly supervised incremental semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jun. 11, 2024, doi: [10.1109/TCSVT.2024.3412996](https://doi.org/10.1109/TCSVT.2024.3412996).
- [10] X. Nie et al., "Pro-tuning: Unified prompt tuning for vision tasks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4653–4667, Jun. 2024.
- [11] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4616–4629, Sep. 2023.
- [12] C. Oh et al., "Geodesic multi-modal mixup for robust fine-tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Red Hook, NY, USA: Curran Associates, 2023, pp. 52326–52341.
- [13] L. Smith and M. Gasser, "The development of embodied cognition: Six lessons from babies," *Artif. Life*, vol. 11, nos. 1–2, pp. 13–29, Jan. 2005.
- [14] H. E. M. den Ouden, P. Kok, and F. P. de Lange, "How prediction errors shape perception, attention, and motivation," *Frontiers Psychol.*, vol. 3, p. 548, Dec. 2012.
- [15] J. Hohwy, *The Predictive Mind*. Oxford, U.K.: OUP Oxford, 2013.
- [16] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 776–794.
- [17] R. Girdhar et al., "ImageBind one embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15180–15190.
- [18] C. Villani et al., *Optimal Transport: Old and New*, vol. 338. Cham, Switzerland: Springer, 2009.
- [19] Y.-L. Sung, J. Cho, and M. Bansal, "VL-ADAPTER: parameter-efficient transfer learning for Vision-and-Language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5217–5227.
- [20] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Long Papers*, vol. 1, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Association for Computational Linguistics, Aug. 2021, pp. 4582–4597.
- [21] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
- [22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
- [23] T. Salimans, H. Zhang, A. Radford, and D. Metaxas, "Improving GANs using optimal transport," 2018, *arXiv:1803.05573*.
- [24] H. Zhao, D. Phung, V. Huynh, T. Le, and W. Buntine, "Neural topic model via optimal transport," 2020, *arXiv:2008.13537*.
- [25] D. Wang et al., "Representing mixtures of word embeddings with mixtures of topic embeddings," 2022, *arXiv:2203.01570*.
- [26] L. Chen et al., "Improving sequence-to-sequence learning via optimal transport," 2019, *arXiv:1901.06283*.
- [27] J. Xu, H. Zhou, C. Gan, Z. Zheng, and L. Li, "Vocabulary learning via optimal transport for neural machine translation," 2020, *arXiv:2012.15671*.
- [28] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu, "Graph optimal transport for cross-domain alignment," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2020, pp. 1542–1553.
- [29] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15640–15649.
- [30] K. Zhang, Z. Mao, A. Liu, and Y. Zhang, "Unified adaptive relevance distinguishable attention network for image-text matching," *IEEE Trans. Multimedia*, vol. 25, pp. 1320–1332, 2023.
- [31] K. Zhang, B. Hu, H. Zhang, Z. Li, and Z. Mao, "Enhanced semantic similarity learning framework for image-text matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2973–2988, Apr. 2024.
- [32] Z. Li, L. Zhang, K. Zhang, Y. Zhang, and Z. Mao, "Improving image-text matching with bidirectional consistency of cross-modal alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6590–6607, Jul. 2024.
- [33] G. Peyré et al., "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.
- [34] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Red Hook, NY, USA: Curran Associates, 2013, pp. 2292–2300.
- [35] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2004, p. 178.
- [36] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.
- [37] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.
- [38] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.
- [39] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
- [40] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—mining discriminative components with random forests," in *Proc. Comput. Vis. ECCV 13th Eur. Conf.*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 446–461.
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.
- [42] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3498–3505.
- [43] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
- [44] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.
- [45] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Res. Comput. Vis.*, vol. 2, no. 11, 2012.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [47] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [48] R. Zhang et al., "Tip-adapter: Training-free CLIP-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.
- [49] Y. Lu, J. Corander, and Z. Yang, "Doubly stochastic neighbor embedding on spheres," *Pattern Recognit. Lett.*, vol. 128, pp. 100–106, Dec. 2019.



Hairui Ren received the B.Sc. degree in mechanical engineering from Jilin University in 2017, where he is currently pursuing the Ph.D. degree in computer science with the School of Artificial Intelligence. His research interests include image recognition and object detection.



Fan Tang received the B.Sc. degree in computer science from North China Electric Power University in 2013 and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2019. He is currently an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.



Dandan Guo received the B.S. degree from the North University of China, Taiyuan, China, in 2014, and the Ph.D. degree in electronic engineering from Xidian University, Xi'an, China, in 2020. After that, she was a Post-Doctoral Researcher with The Chinese University of Hong Kong, Shenzhen, China. She is currently a tenure-track Professor with Jilin University, Jilin, China. She has been working on machine learning algorithms and applications, including few-shot learning, long-tailed classification, and prompt tuning.



Huangjie Zheng received the joint B.S. and M.S. degree from Shanghai Jiao Tong University in 2019. He is currently pursuing the Ph.D. degree with the University of Texas at Austin. His research interests include machine learning and Bayesian deep learning, with specific interests in probabilistic modeling and deep generative models with application to computer vision, natural language processing, and multi-modal data.



He Zhao received the Ph.D. degree from Monash University, Australia. He is currently a Machine Learning Researcher with CSIRO's Data61, Australia. His research interests include probabilistic approaches that improve robustness, generalization, uncertainty estimation, and the interpretation of machine/deep learning.



Yi Chang is currently the Dean of the School of Artificial Intelligence, Jilin University, Changchun, China. He was elected as a Chinese National Distinguished Professor in 2017 and an ACM Distinguished Scientist in 2018. Before joining academia, he was the Technical Vice President with Huawei Research America and the Research Director of Yahoo Labs. He is the author of two books and more than 100 papers in top conferences or journals. His research interests include information retrieval, data mining, machine learning, natural language processing, and artificial intelligence. He won the Best Paper Award on KDD'2016 and WSDM'2016. He has served as the Conference General Chair for WSDM'2018, SIGIR'2020, and WWW'2025.