Dude: Dual Distribution-Aware Context Prompt Learning For Large Vision-Language Model

Duy M. H. Nguyen[*1,2,3], **An T. Le**[*4], **Trung Q. Nguyen**[2,5], **Nghiem T. Diep**[2], **Tai Nguyen**[2], **Duy Duong-Tran**[6,8], **Jan Peters**[2,4,9], **Li Shen**[8], **Mathias Niepert**[1,3], **Daniel Sonntag**[2,7]

[1] *University of Stuttgart,* [2] *German Research Center for Artificial Intelligence (DFKI),*
[3] *Max Planck Research School for Intelligent Systems,* [4] *Technical University of Darmstadt*
[5] *Technical University of Munich,* [6] *United States Naval Academy,* [7] *Oldenburg University*
[8] *University of Pennsylvania,* [9] *Hessian.AI.* [*] *Co-equal contribution.*

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Prompt learning methods are gaining increasing attention due to their ability to customize large vision-language models to new domains using pre-trained contextual knowledge and minimal training data. However, existing works typically rely on optimizing unified prompt inputs, often struggling with fine-grained classification tasks due to insufficient discriminative attributes. To tackle this, we consider a new framework based on a dual context of both *domain-shared* and *class-specific contexts*, where the latter is generated by Large Language Models (LLMs) such as `GPT`s. Such dual prompt methods enhance the model's feature representation by joining implicit and explicit factors encoded in LLM knowledge. Moreover, we formulate the Unbalanced Optimal Transport (UOT) theory to quantify the relationships between constructed prompts and visual tokens. Through partial matching, UOT can properly align discrete sets of visual tokens and prompt embeddings under different mass distributions, which is particularly valuable for handling irrelevant or noisy elements, ensuring that the preservation of mass does not restrict transport solutions. Furthermore, UOT's characteristics integrate seamlessly with image augmentation, expanding the training sample pool while maintaining a reasonable distance between perturbed images and prompt inputs. Extensive experiments across few-shot classification and adapter settings substantiate the superiority of our model over current state-of-the-art baselines.

Keywords: prompt learning, adapter learning, unbalanced optimal transport, large vision-language models.

1. Introduction

Recent advancements in vision-language models (VLMs), exemplified by `CLIP` (Radford et al., 2021), `ALIGN` (Jia et al., 2021), or `Flava` (Singh et al., 2022), have demonstrated remarkable capabilities in learning comprehensive visual and textual concepts in classification, generation, or recognition. During pre-training, these models leverage web-scale image-text pairs to establish aligned representations of images and text through contrastive loss. For instance, through prompts like "`A picture of a {label}`", VLMs seamlessly transfer their knowledge into downstream applications, employing zero-shot learning by comparing task-specific descriptions with encoded images and texts (Figure 1 (a)). Such approaches eliminate the need for extensive fine-tuning, underscoring their adaptability and efficiency in various practical scenarios.

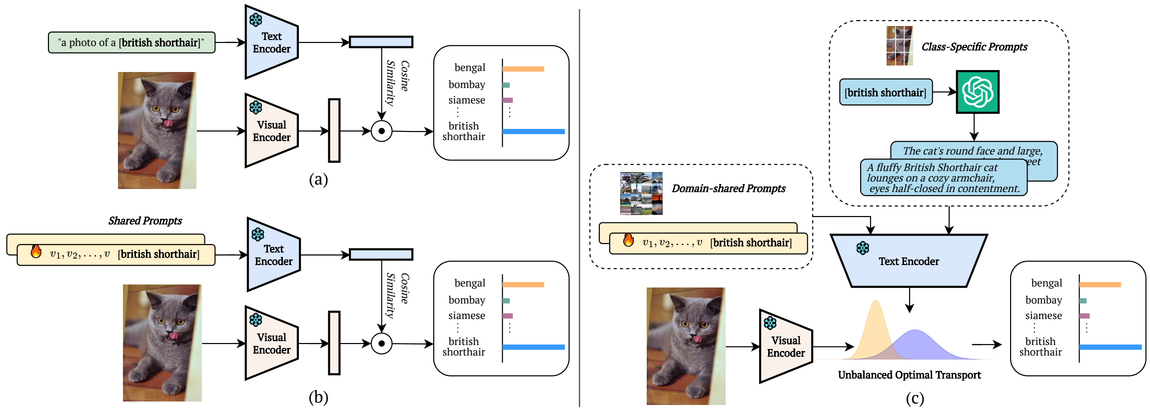

Figure 1: (a) Zero-shot learning; (b) Shared classes prompt learning; (c) Our method with dual prompts and Unbalanced Optimal Transport (UOT) as the distance between visual tokens and prompt sets.

However, the effectiveness of these zero-shot capabilities is highly dependent on the quality of the information embedded in the manually created prompts (Zhou et al., 2022a). While significant improvement can be achieved through prompt engineering (Gu et al., 2023), it is time-consuming, requires domain expertise, and has unpredictable performance under domain shifts. As a direct consequence, data-driven approaches (e.g., prompt learning) are introduced to leverage the rich context of additional information for classification. Early efforts considered a single learnable prompt (Zhou et al., 2022a), image conditional information (Zhou et al., 2022b), or multiple prompts (Lu et al., 2022; Chen et al., 2023) to implicitly formulate the shared class context, instance-specific context, or context variance, respectively. Another line of work is built on adapter learning, which refines the textual classifier or visual features with a simple learnable feature modulation for each specific task (Gao et al., 2024; Zhang et al., 2022a; Li et al., 2024).

Despite achieving promising records in few-shot learning, these models face two significant limitations. First, they often employ unified, learnable context prompts shared across all classes, which can overlook the subtle and unique attributes necessary to differentiate closely related or merely indistinguishable categories. As a result, the model may struggle to accurately identify and classify fine details with many contexts (e.g., fur color, eye color, cat species, etc. as semantics contexts), such as those required in cat classification tasks (cf. Fig. 2). Although methods like `CoOP` (Zhou et al., 2022a) or those leveraging `GPT` models (Naeem et al., 2023) adopt class-specific prompts, they are generally limited to zero-shot learning or require substantial labeled data to learn these class-specific prompts to avoid over-fitting due to the increasing of trainable parameters relative to the number of classes. Second, most prompt-based and adapter models utilize cosine distance between global visual and prompt features to measure affinity, potentially ignoring intricate relationships between image features and textual descriptions (Figure 1 (b)). This, in turn, falls short of reflecting the fact that different prompts may correspond to distinct image patches. Consequently, the model has difficulty capturing underlying structures and variability within the data that might distinguish closely related objects, resulting in degraded performance when handling fine-grained classification tasks.

In this paper, we propose bridging both *domain-shared* and *class-specific prompts* initialized from GPT, aiming to enrich class-wise descriptions. Learnable domain-shared prompts serve to establish foundational understanding across various categories, ensuring broad applicability and robust generalization capabilities. Concurrently, trainable class-specific prompts derived from GPT facilitate specificity by capturing the diverse attributes unique to each object, thereby favoring discriminative abilities in fine-grained distinction tasks. In particular, we learn a shared self-attention mechanism to mitigate the increase in trainable parameters linked with class-specific prompts. This module takes GPT prompts as inputs and generates textual vectors tailored to various categories, which is parameter efficiency while maintaining discriminative power.

Given dual-composed prompts, we compute their textual embedding by feeding into the frozen text encoder (e.g., CLIP text encoder). Then, we express the distance between visual features and prompt embedding as a distance between discrete probability distributions using the unbalanced optimal transport theory (Liero et al., 2018). Specifically, we extract all local visual maps for each image rather than a single global representation. This corresponds to a 7×7 spatial dimension in the case of ResNet-50 or outputs taken from the multi-head self-attention layer with the Vision Transformer (Dosovitskiy et al., 2021). The local visual tokens are subsequently aligned to each prompt feature using transport plans computed by solving the UOT and then averaging two distance values to form a final correlation score. Compared with other distances, such as Euclidean or cosine distances, the UOT can properly align diverse visual features to local prompts and be resilient against misalignment or feature shift, benefiting from its partial matching flexibility. This is particularly advantageous when some visual tokens do not have corresponding matches in the prompt sets. Furthermore, these properties make UOT particularly suitable for data augmentation, where input images are augmented with random transformations before alignment with contextual prompts, aiming to enrich training data and enhance the model's generalization capabilities (Figure 1 (c)). It is worth noting that while a few current works also employ optimal transport between visual and prompt sets, they typically enforce balanced mass preservation constraints between two sets (Chen et al., 2023; Kim et al., 2023), resulting in sub-optimal mappings in the essence of misalignment or noise outliers.

In summary, we make the following contributions:

- We propose a dual prompt learning approach that captures both unified domain-shared and class-specific contexts, enriched by descriptions generated by GPT.
- The Unbalanced Optimal Transport (UOT) is formulated to capture underlying relationships between local visual tokens and multi-prompt features while being robust to noise and misalignment.
- We assess our performance on fine-grained classification using both few-shot and adapter-based settings and attain state-of-the-art results compared to other leading benchmarks.

2. Related Works

Vision-Language Pre-training Algorithms. Several approaches are used to *pre-train vision-language models* with large-scale data. They can be divided into reconstruction (Hong et al., 2021; Kim et al., 2021), contrastive learning (Jia et al., 2021; Yuan et al., 2021), graph matching (MH Nguyen et al., 2024; Ektefaie et al., 2023), or fusing several objective losses (Kamath et al., 2021; Bao et al., 2022). In this work, we implement data augmentation on input images similarly to contrastive learning but apply it within the context of prompt learning. Here, perturbed images are aligned with prompt embeddings, with features extracted from frozen text encoders. The distance between the augmented visual features and the prompt visual features is then estimated using the Unbalanced Optimal Transport (UOT).

Efficient Transfer Learning. Prompt tuning and adapter-based methods are two prominent directions for transferring task-specific knowledge to downstream tasks by tuning minimal parameters. In prompt tuning, early efforts focus on prompt engineering to seek optimal template inputs, aiming at maximum performance of a non-trainable scheme such as a zero-shot `CLIP` (Radford et al., 2021). Afterward, `CoOP` (Zhou et al., 2022a) as the pioneer work extends to learnable prompts in few-shot tasks. Following this trend, several works (Zhou et al., 2022b; Zhang et al., 2022b; Lu et al., 2022; Chen et al., 2023) further improve prompt tuning from multiple aspects, such as image-conditional generalization or multiple prompts for diversity. In contrast, adapter-style approaches customize vision-language models for particular tasks by incorporating lightweight learnable modules on top of the textual and visual feature outputs. For example, `CLIP-Adapter` (Gao et al., 2024) introduces a trainable bottleneck layer to produce adapted features, which are then merged with the original `CLIP` outputs via a residual connection. Other advanced adapter-based techniques have also been exploited, such as those employing task-independent strategies (Yu et al., 2023) or leveraging the structural knowledge of data (Li et al., 2024).

In contrast to the aforementioned ones, our formulation bridges both domain-shared and specific-class contextual prompts, leveraging `GPT`-generated descriptions for enhanced model capacities when dealing with fine-grained tasks. We also implement a distance metric between visual tokens and multiple prompts using UOT, which is effective for both *prompt learning* (Section 4.2) and *adapter learning* (Section 4.3).

Representation Learning with Optimal Transport. Optimal Transport (OT) has been widely adopted in machine learning as an objective comparing distributions. Most of the recent successful OT stories define auxiliary training objectives or transformation components (Montesuma et al., 2023) with applications in domain adaptation (Courty et al., 2014; Alvarez-Melis and Fusi, 2020), Wasserstein GAN (Arjovsky et al., 2017), molecular representation learning (Nguyen et al., 2024), and robotics planning (Le et al., 2023a). To address real-world scenarios where the mass preservation constraint is too strict, variants such as Unbalanced Optimal Transport (UOT) (Liero et al., 2018) and entropic regularization (Cuturi, 2013) have been introduced. These extensions have led to the development of entropic UOT (Chizat et al., 2018), which combines the flexibility of UOT with the computational advantages of entropic regularization. Till now, UOT has been used in domain adaptation with minibatch training on large datasets (Fatras et al., 2021), and recently on unsupervised action segmentation (Xu and Gould, 2024), or reactive policy blending in robotics (Le et al., 2023b). In this work, to the best of our knowledge, we first introduce UOT to prompt learning for large-vision language models.

3. Methodology

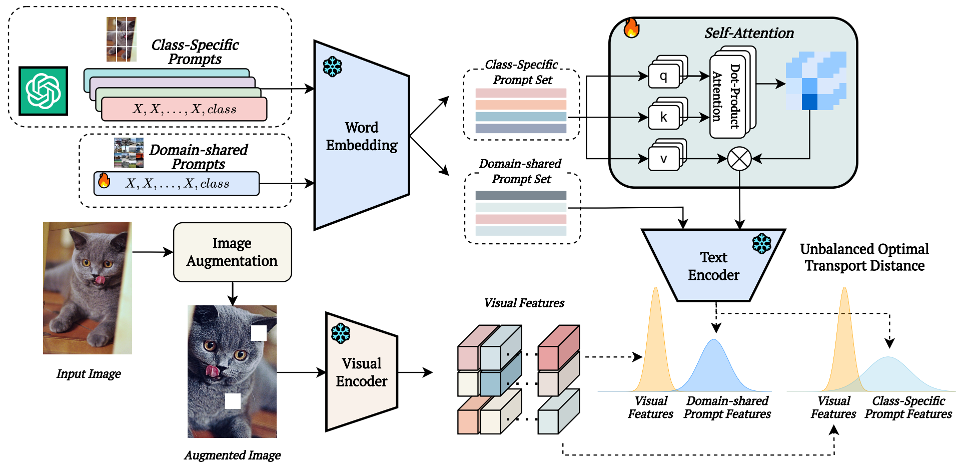

Figure 2: Overview of the proposed framework. `CLIP`'s vision and text encoders are frozen, training only domain-shared prompt embeddings and self-attention model.

3.1. Revisit Zero-shot Learning to Single Prompt Learning

Zero-shot Learning. Pre-training `CLIP` (Radford et al., 2021) involves learning to match images with their textual descriptions, allowing zero-shot inference on a downstream recognition task by manually designing the prompt template. Let $\boldsymbol{f}$ be the feature vector representing an image $\boldsymbol{x} \in \mathcal{X}$, and $\{\boldsymbol{t}_i\}_{i=1}^{K}$ be the prompt tokens generated from an encoder, assuming $\boldsymbol{f}, \boldsymbol{t}_i$ having the same dimension. K is the total number of classes, and $\boldsymbol{t}_i$ is generated from a prompt such as "`an image of {label}`". The classification likelihood of a class i can be defined as a softmax

$$\mathbb{P}(c = i \mid \boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{t}_i, \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{t}_j, \boldsymbol{f})/\tau)}, \tag{1}$$

where τ is fixed temperature scalar from `CLIP`, and $\cos(\cdot, \cdot)$ denotes cosine similarity (Figure 1 (a)). This differs from traditional classification learning from pre-defined categories in the sense that `CLIP` leverages natural language descriptions, enabling it to explore a wider range of visual concepts and produce more transferable representations for various tasks.

Prompt Learning. Zero-shot prediction with fixed prompt features can suffer from domain shift problems. To alleviate, Zhou et al. (2022c,b) has demonstrated prompt learning to outperform zero-shot adaptions using manual prompts or linear probe models (Tian et al., 2020). Specifically, let $\boldsymbol{w}$ be the learnable context vector. The learnable prompt is denoted as the concatenation $\boldsymbol{t}_i = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{N-1}, \boldsymbol{c}^i]$, with $\boldsymbol{c}^i$ is the word token corresponding to class i, having the same dimension as $\boldsymbol{w}$. Either same context $\{\boldsymbol{w}_k\}_{k=1}^{N-1}$ with all classes or different context $\{\boldsymbol{w}_k^i\}_{k=1}^{N-1}$ per class i can be optimized w.r.t. a cross-entropy loss between the labeled target and the prediction

$$\mathbb{P}(c = i \mid \boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{f})/\tau)}, \tag{2}$$

where $g(\cdot)$ is a text encoder. Recently, Chen et al. (2023); Kim et al. (2023) generalize the prompt learning to multi-prompt and multi-visual-features alignment with a distribution-aware OT metric, e.g., the scenarios where many prompts can describe an image and many image regions can be related to a prompt (i.e., many-to-many alignment). In this work, we look deeper into the matching problem of visual-language alignment, especially the unbalanced problem of visual-language embedding matching described in the next section.

3.2. Aligning Prompts and Visual Token via Unbalanced Optimal Transport

Formulating alignment between (multi-)prompt and visual tokens as an OT objective (Chen et al., 2023; Kim et al., 2023) with entropic is efficient and scalable. However, we observe that the marginal constraints of OT are restrictive in some settings, e.g., there are many irrelevant image embeddings that are far from true embeddings and hence introduce noises, which should be discouraged entirely (Figure 5 (Left)). Below, we formally describe the OT and its entropic relaxation, then introduce further relaxation on the marginal constraints, addressing the mentioned problem.

Algorithm 1 Solving UOT_λ in dual form.

Input: $k = 0$ and $\boldsymbol{u}^0 = \boldsymbol{v}^0 = \mathbf{0}$.
while *not all batch instances converged* **do**

$\boldsymbol{n}^k = \boldsymbol{W}_{k,k}\mathbf{1}_m$

$\boldsymbol{u}^{k+1} = \left[\dfrac{\boldsymbol{u}^k}{\lambda} + \log(\boldsymbol{n}) - \log(\boldsymbol{n}^k)\right]\dfrac{\lambda\rho_1}{\lambda+\rho_1}$

$\boldsymbol{m}^k = \boldsymbol{W}^\intercal_{k+1,k}\mathbf{1}_n$

$\boldsymbol{v}^{k+1} = \left[\dfrac{\boldsymbol{v}^k}{\lambda} + \log(\boldsymbol{m}) - \log(\boldsymbol{m}^k)\right]\dfrac{\lambda\rho_2}{\lambda+\rho_2}$

$k \leftarrow k + 1$

end

Output: $\boldsymbol{W}^*$.

Notation. $\mathbf{1}_d$ is the vector of ones in $\mathbb{R}^d$. The scalar product for vectors and matrices is $x, y \in \mathbb{R}^d$, $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^d \boldsymbol{x}_i\boldsymbol{y}_i$; and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \sum_{i,j=1}^d \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}$, respectively. For two histograms $\boldsymbol{n} \in \Sigma_n$ and $\boldsymbol{m} \in \Sigma_m$ in the simplex $\Sigma_d := \{\boldsymbol{x} \in \mathbb{R}^d_+ : \boldsymbol{x}^\intercal\mathbf{1}_d = 1\}$, we define the set $U(\boldsymbol{n}, \boldsymbol{m}) := \{\boldsymbol{W} \in \mathbb{R}^{n\times m}_+ \mid \boldsymbol{W}\mathbf{1}_m = \boldsymbol{n}, \boldsymbol{W}^\intercal\mathbf{1}_n = \boldsymbol{m}\}$ containing $n\times m$ matrices with row and column sums $\boldsymbol{n}$ and $\boldsymbol{m}$ respectively. The entropy for $\boldsymbol{A} \in U(\boldsymbol{n}, \boldsymbol{m})$ is defined as $H(\boldsymbol{A}) = -\sum_{i,j=1}^{n,m} a_{ij}\log a_{ij}$. $\widetilde{\mathrm{KL}}(\boldsymbol{w}||\boldsymbol{z}) = \boldsymbol{w}^\intercal\log(\boldsymbol{w}\oslash\boldsymbol{z}) - \mathbf{1}^\intercal\boldsymbol{w} + \mathbf{1}^\intercal\boldsymbol{z}$ is the generalized Kullback-Leibler (KL) divergence between two positive vectors $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^d_+$ ($\oslash$ is the element-wise division), with the convention $0\log 0 = 0$.

Let $\boldsymbol{C} \in \mathbb{R}^{n\times m}_+$ be the positive cost matrix, the OT between $\boldsymbol{n}$ and $\boldsymbol{m}$ given cost $\boldsymbol{C}$ is $\mathrm{OT}(\boldsymbol{C}) := \min_{\boldsymbol{W}\in U(\boldsymbol{n},\boldsymbol{m})}\langle \boldsymbol{W}, \boldsymbol{C}\rangle$. Traditionally, this Kantorovich formulation does not scale well with high dimensions. To address this, Cuturi (2013) proposes to regularize its objective with an entropy term, resulting in the entropic OT

$$\mathrm{OT}_\lambda(\boldsymbol{C}) := \min_{\boldsymbol{W}\in U(\boldsymbol{n},\boldsymbol{m})}\langle \boldsymbol{W}, \boldsymbol{C}\rangle - \lambda H(\boldsymbol{W}), \tag{3}$$

with entropic scalar $\lambda > 0$, which can be solved with the Sinkhorn algorithm (Sinkhorn and Knopp, 1967) with complexity of $\tilde{\mathcal{O}}(n^2/\epsilon^3)$ (Altschuler et al., 2017), where ϵ is the approximation error w.r.t. the original $\mathrm{OT}(\boldsymbol{C})$. Small λ produces fast and biased solutions, or vice versa.

Further relaxing marginal constraints leading to entropic UOT (Chizat et al., 2018)

$$\mathrm{UOT}_\lambda(\boldsymbol{C}) := \min_{\boldsymbol{W}\in\mathbb{R}^{n\times m}_+}\langle \boldsymbol{W}, \boldsymbol{C}\rangle - \lambda H(\boldsymbol{W}) + \rho_1\widetilde{\mathrm{KL}}(\boldsymbol{W}\mathbf{1}_m \,\|\, \boldsymbol{n}) + \rho_2\widetilde{\mathrm{KL}}(\boldsymbol{W}^\intercal\mathbf{1}_n \,\|\, \boldsymbol{m}) \tag{4}$$

where now $\boldsymbol{n} \in \mathbb{R}^n_+, \boldsymbol{m} \in \mathbb{R}^m_+$ are arbitrary positive vectors, $\rho_{1,2}$ are the marginal regularization scalars. Equation 4 is well-known as Wasserstein-Fischer-Rao distance on the set of positive Radon measures with entropic regularization (Liero et al., 2018; Séjourné et al., 2023), which is desirable as a metric quantifying the alignments of unbalanced embedding distributions on a common latent space. Pham et al. (2020) shows that the generalized matrix scaling Algorithm 1 (Chizat et al., 2018) solves the dual of Equation 4

$$\min_{\boldsymbol{u}\in\mathbb{R}^n,\boldsymbol{v}\in\mathbb{R}^m} \lambda \sum_{i,j=1}^{n} \exp\left(\frac{\boldsymbol{u}_i + \boldsymbol{v}_j - \boldsymbol{C}_{ij}}{\lambda}\right) + \rho_1 \left\langle e^{-\boldsymbol{u}/\rho_1}, \boldsymbol{n}\right\rangle + \rho_2 \left\langle e^{-\boldsymbol{v}/\rho_2}, \boldsymbol{m}\right\rangle, \tag{5}$$

with the complexity of $\tilde{\mathcal{O}}(n^2/\epsilon)$. Denoting the dual vectors $(\boldsymbol{u}^k, \boldsymbol{v}^k)$ at iteration k, the optimal coupling is computed as $\boldsymbol{W}_{i,j} = \mathrm{diag}(e^{\boldsymbol{u}^i/\lambda})\, e^{-\frac{\boldsymbol{C}}{\lambda}}\, \mathrm{diag}(e^{\boldsymbol{v}^j/\lambda})$. Iterating the Sinkhorn projections (Algorithm 1) is guaranteed to converge to a fixed point $\boldsymbol{W}^*$ (Theorem 4.1 in Chizat et al. (2018)). Note that Algorithm 1 is vectorizable, which is desirable for scaling training with multi-prompt alignments with augmented image patches. We implement the entropic UOT as the alignment distance between a set of image embeddings and a set of word embeddings for each class, with vectorization for minibatch training (i.e., a minibatch of matching sets), described in the next section.

3.3. Dual Context Prompt Learning

Despite the scalability and simplicity of using shared multi-prompts (Lu et al., 2022; Chen et al., 2023), we observe that such a sharing prompt limits its effectiveness in many prompt learning scenarios, such as failing to capture the diverse contexts associated with fine-grained classes.

Diversifying prompts using LLM. Due to being pre-trained on extensive corpora, LLMs have acquired substantial common knowledge on a wide range of topics and can serve as external knowledge bases for downstream tasks (Jang et al., 2021; Ke et al., 2023; Razdaibiedina et al., 2023). For each class, we construct a system prompt shown in Figure 3 to query the LLM. This aims to obtain image descriptions, providing varied local context information for class i as a class-specific prompt set $H_i = \mathrm{LLM}(Q_i)$, where Q_i is the question for class i.

Prompting the LLM to generate image descriptions

```
System Prompt: Given the input text indicating the category name of a certain object, your task
involves the following steps:

  1. Imagine a scene containing the input object.
  2. Generate 4 descriptions about different key appearance features of the input object from
     the imagined scene, with each description having a maximum of 16 words.
  3. Output a JSON object containing the following key: {"description": <list of 4
     descriptions>}
```

Figure 3: Prompt supplied for the class-specific prompt generation

For example, in Figure 1, with the given certain class of "british shorthair" in `OxfordPets` dataset, the response can be *"A fluffy British Shorthair cat lounges on a cozy armchair, eyes half-closed in contentment."* or *"The cat's round face and large, expressive eyes give it a sweet and gentle appearance."*. Then, the prompts are tokenized using a frozen word embedding into token set $\{\boldsymbol{w}_j^i\}_{j=1}^{N-1} \cup \boldsymbol{c}^i$ from H_i.

Self-attention adapter. Since class-specific prompts can induce exponential increases in token size with an increasing number of classes, we adopt a shared trainable self-attention adapter (Vaswani et al., 2017) before forwarding the transformed tokens to the frozen text encoder $g(\cdot)$. The self-attention module shown in Figure 2 is trained on all prompt sets associated with all classes to cope with the exponentially large token number. Intuitively, this module compresses the diverse class contexts with $\text{Attention}(\boldsymbol{T}) = \text{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^\intercal/\sqrt{d_k}\right)\boldsymbol{V}$, where $\boldsymbol{Q} = \boldsymbol{T}\boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{T}\boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{T}\boldsymbol{W}^V \in \mathbb{R}^{N\times d_k}$ are the products of the class-specific token matrix $\boldsymbol{T}^i = [\boldsymbol{w}_1^i, \dots, \boldsymbol{w}_{N-1}^i, \boldsymbol{c}^i]^\intercal$ of class i with their associate query, key, value weighting matrices $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d\times d_k}$. The trained self-attention adapter allows the model to focus on relevant token contexts represented by the latent vectors, thereby compressing input information.

Image augmentation for diverse visual embeddings. Data augmentation is a standard technique for combating overfitting (Shorten and Khoshgoftaar, 2019). In the prompt learning setting, we observe that image augmentation generates diverse visual embeddings, preventing overfitting to a subset of local features. Additionally, this technique enhances robustness against common image transformations that happen frequently in practice. Furthermore, the UOT formulation synergizes with data augmentation techniques, as the optimal coupling solution in a balanced problem can constrain the matching to heavily deformed data. For instance, in Figure 2, we apply `random flip`, `colorjitter`, and `cutout` transformations on the input images and feed the perturbed outputs to the vision encoder.

Distribution-aware distance between visual and prompt embedding. Let $\boldsymbol{F} \in \mathbb{R}^{M\times d}$ be the image embedding matrix representing M local features from the augmented image $\boldsymbol{x}$, $\boldsymbol{G}_{\text{ds}}^i = g(\boldsymbol{T}_{\text{ds}}^i) \in \mathbb{R}^{N\times d}$ be the prompt embedding matrix representing learnable domain-shared prompts, $\boldsymbol{G}_{\text{cs}}^i = g(\text{Attention}(\boldsymbol{T}_{\text{cs}}^i)) \in \mathbb{R}^{N\times d}$ are class-specific prompt embeddings. We also assume both image embeddings and prompt embeddings lie in the same space $\mathbb{R}^d$, and are represented by discrete distributions

$$\alpha = \sum_{i=1}^{M} m_i \delta_{\boldsymbol{f}_i},\ \boldsymbol{f}_i \in \boldsymbol{F} \quad \beta = \sum_{i=1}^{N} n_i \delta_{\boldsymbol{g}_i},\ \boldsymbol{g}_i \in \boldsymbol{G}^i, \tag{6}$$

where the weights are elements of the marginals $\boldsymbol{m} = [m_i]_{i=1}^M, \boldsymbol{n} = [n_i]_{i=1}^N$ and can be selected as uniform weights. The cost between two domains now is defined as $\boldsymbol{C} = \mathbf{1}_{n\times m} - \cos(\boldsymbol{F}, \boldsymbol{G}^i)$, where $\cos(\cdot,\cdot)$ denotes the pairwise cosine similarity between embeddings. Then, the embedding matching objective for class i can be defined as two UOT distances (Eq. (4)), including: (i) class-specific prompts $\text{UOT}_\lambda^i(\boldsymbol{C}_{\text{cs}})$ and (ii) domain-shared prompts alignments $\text{UOT}_\lambda^i(\boldsymbol{C}_{\text{ds}})$, respectively. Given this, the final alignment objective is the weighted sum $d^i = \gamma_{\text{cs}}\text{UOT}_\lambda^i(\boldsymbol{C}_{\text{cs}}) + \gamma_{\text{ds}}\text{UOT}_\lambda^i(\boldsymbol{C}_{\text{ds}})$ with the type weighting scalars $\gamma_{\text{cs}}, \gamma_{\text{ds}} > 0$ (Figure 2). The classification likelihood can be written as

$$\mathbb{P}(c = i \mid \boldsymbol{x}) = \frac{\exp((1 - d^i)/\tau)}{\sum_{j=1}^{K} \exp((1 - d^j)/\tau)}. \tag{7}$$

For each inner iteration, we optimize UOT objectives in batches of K classes and fix $\{\boldsymbol{W}_i^*\}_{i=1}^K$, then, using Danskin theorem (Danskin, 1966), we can optimize the prompts the cross-entropy objective (Chen et al., 2023; Lu et al., 2022; Zhou et al., 2022a)

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^{K} y_{i,\boldsymbol{x}} \mathbb{P}(c = i \mid \boldsymbol{x}), \tag{8}$$

where $\boldsymbol{y}_{\boldsymbol{x}}$ is the one-hot label for image $\boldsymbol{x}$.

4. Experiment Results

4.1. Datasets and Implementation Details

Datasets. We conduct *few-shot learning* on five fine-grained datasets, including Flowers102 (Nilsback and Zisserman, 2008), FGVCAircraft (Maji et al., 2013), StanfordCars (Krause et al., 2013), OxfordPets (Parkhi et al., 2012), and Food101 (Bossard et al., 2014).

Implementation Details Our implementation builds on the CoOp codebase (Zhou et al., 2022a). We conducted all experiments using CLIP with ViT-B/16 and ResNet-50 backbones. The number of domain-shared and class-specific prompts is chosen to be either 2 or 4, depending on the dataset. We use ChatGPT APIs to generate prompts for each class, the system prompt is shown in Figure 3. The final results were averaged over three random seeds (1/2/3) for a fair comparison. We used the Adam optimizer with a learning rate of $2e^{-3}$ and a batch size of 32, running for 50 epochs. We configured self-attention with a single-head output for data efficiency. The UOT problem is solved using the Sinkhorn algorithm, as described in Algorithm 1. We tuned the hyperparameters of ρ_1, ρ_2 in range of $\{\infty, 0.001 \to 0.023\}$ based on validation performance. All experiments were performed on A100 GPUs.

4.2. Few-shot Learning with Prompt-based Methods

Baselines. We compare with ten *prompt-based methods* including CoOp (Zhou et al., 2022a), CoCoOp (Zhou et al., 2022b), DAPT (Cho et al., 2023), ProGrad (Zhu et al., 2023), ProDA (Lu et al., 2022), KgCoOp (Yao et al., 2023), RPO (Lee et al., 2023), Plot (Chen et al., 2023), MaPLe (Khattak et al., 2023a), and PromptSRC (Khattak et al., 2023b). Results for baseline are summarized from the literature. Among these, DAPT, PLOT and ProDA relate to prompt distribution learning, and ProDA or PLOT also adapt multi-prompt mechanisms.

Few-shot learning with K-shot labeled images. We conduct the few-shot classification on five datasets using K-shot labeled images and evaluate trained performance on the testing domain *within the same class space as training ones*. It is worth noting that we freeze both CLIP's vision and text encoders during training. We only train our prompt embeddings and self-attention model. Table 1 summarizes our results using 4-shot per class with ViT-B/16. We observe that DUDE achieves the best performance in three out of five settings and has a higher average performance than state-of-the-art methods, reaching 76.84%. Notably, on some datasets like StanfordCars, DUDE significantly outperforms zero-shot CLIP and single shared-prompt methods like CoOp, with substantial margins of 10.65% and 3.62%, respectively.

Table 1: Few-shot learning compared with **prompt-based methods**.

	CLIP	CoOp	CoCoOp	ProGrad	KgCoOp	MaPLe	DAPT	PromptSRC	PLOT	DUDE
OxfordPets	89.10	91.30	**93.01**	93.21	93.20	92.05	92.17	93.23	92.55	92.01
StandfordCars	65.70	72.73	69.10	71.75	71.98	68.70	74.40	71.83	74.93	**76.35**
Flowers	70.70	91.14	82.56	89.98	90.69	80.80	92.37	91.31	92.93	**94.50**
Food101	85.90	82.58	86.64	85.77	86.59	**86.90**	83.60	86.06	86.46	84.90
FGVCAircraft	24.90	33.18	30.87	32.93	32.47	29.03	32.47	32.80	35.29	**36.45**
Average	67.26	74.19	72.44	74.73	74.99	71.50	75.00	75.05	76.43	**76.84**

Table 2: Comparison on the **base-to-new generalization** setting with 16-shot samples.

		CoOp	CoCoOp	DAPT	ProGrad	ProDA	KgCoOp	RPO	PLOT	MaPLe	DUDE
OxfordPets	Base	94.47	95.20	95.00	95.07	95.43	94.65	94.63	94.50	**95.43**	94.87
	New	96.00	97.69	95.83	97.63	**97.83**	97.76	97.50	96.83	97.76	97.16
StanfordCars	Base	75.67	70.49	75.80	77.68	74.70	71.76	73.87	79.07	72.94	**80.75**
	New	67.53	73.59	63.93	68.63	71.20	75.04	**75.53**	74.80	74.00	74.23
Flowers	Base	97.27	94.87	96.97	95.54	97.70	95.00	94.13	**97.93**	95.92	97.53
	New	67.13	71.75	60.90	71.87	68.68	74.73	76.67	73.53	72.46	**76.73**
Food101	Base	89.37	**90.70**	90.37	90.37	90.30	90.5	90.33	89.80	90.71	90.37
	New	88.77	91.29	91.30	89.59	88.57	91.7	90.83	91.37	**92.05**	91.37
FGVC-Aircraft	Base	39.67	33.41	39.97	40.54	36.90	36.21	37.33	**42.13**	37.44	42.02
	New	31.23	23.71	29.80	27.57	34.13	33.55	34.20	33.73	**35.61**	34.53
Average	Base	79.29	76.93	79.62	79.84	79.01	77.62	78.06	80.69	78.49	**81.12**
	New	70.13	71.61	68.35	71.06	72.12	74.56	74.95	74.05	74.38	**75.08**

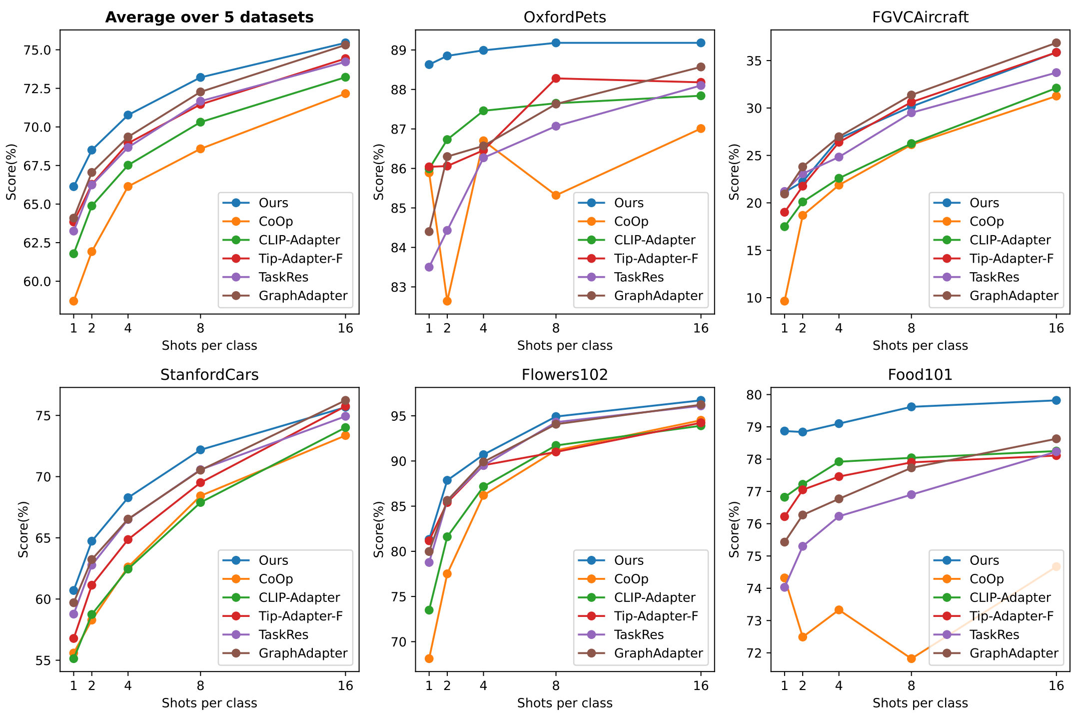

Figure 4: Few-shot learning results on five datasets with adapter learning. Curves are drawn from $1, 2, 4, 8, 16$ shots.

Base-to-New Class Generalization within Same Domain. We investigate the generalization of prompt tuning by splitting each dataset into two disjoint subsets: *Base* and *New* classes where *Base* categories are utilized for training learnable prompts and *New* categories are used to evaluate performance (Lee et al., 2023). In this setting, we use the ViT-16 CLIP as the base model and train models with 16-shot samples. Table 2 shows that with increased training examples, all methods improve their performance compared to using only 4-shot, as seen in Table 1. Overall, DUDE achieves the highest performance in both the *Base* and *New* settings, showcasing its ability to generalize to unseen classes. This capability is attributed to initializing class-specific prompts with external GPT knowledge. Other top-performing baselines, such as PLOT and RPO, also excel by learning multiple prompts and applying regularization to internal feature representations.

4.3. Few-shot learning with Adapter-based Methods

Settings. We validate our DUDE approach using adapter-based techniques. Rather than optimizing prompt embedding inputs, we focus on training small module networks on the outputs of frozen VML models to adapt to new domains quickly. Our base model uses `Tip-Adapter` (Zhang et al., 2022a) with ResNet-50. Specifically, we enhance the original `Tip-Adapter` by extending from a single learnable linear model to learnable multi-linear models. Additionally, we replace the global embedding in `Tip-Adapter` with local visual features. Our UOT then reformulates the global embedding-based distance in `Tip-Adapter` to measure the distance between two distributions.

Baseline. We benchmark Adapter-based DUDE with the advanced adapter-based baselines involving: `Clip-Adapter` (Gao et al., 2024), `Tip-Adapter` (Zhang et al., 2022a), `TaskRes` (Yu et al., 2023), and `GraphAdapter` (Li et al., 2024). All methods are based on the `CLIP` ResNet-50.

Results. The experimental results are presented in Figure 4, proving evidence that our DUDE consistently performs better than previous adapter-based methods across $1, 2, 4, 8$, and 16 shots on four datasets, as well as in average performance. Notably, for datasets such as `OxfordPets` and `Food101`, our curves surpass competitive ones by significant margins across all shots. These records, therefore, validate the effectiveness of DUDE, validating its advantage in both prompt learning and adapter cases.

4.4. Ablation Study

We implement the following variations to understand the effects of critical components in DUDE. (i) Without learning class-specific context prompts for each class; (ii) without learning domain-shared prompts; (iii) without using `GPT` to initialize parameters for class-specific prompts, i.e., initialization randomly; (iv) without using unbalanced optimal transport and using standard optimal transport distance; (v) without using shared self-attention to learn class-specific prompt embedding, i.e., each class will initialize separate parameters to train.

Table 3 summarizes the performance of DUDE utilizing `CLIP` ResNet-50 on the `Food101` and `OxfordPets` datasets. The results indicate that each component is crucial in achieving optimal performance. Among those, the most important factors include using class-specific context prompts, unbalanced optimal transport as the distance between domains, and the parameter efficiency of shared self-attention for learning per class prompt representations, which avoids amounts of number parameters scaled to the number of categories.

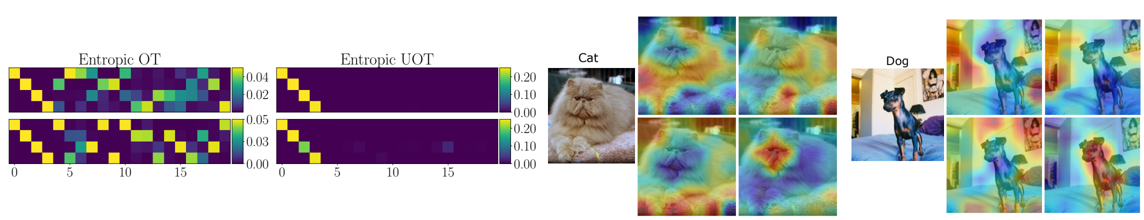

Figure 5: (**Left**) Comparison between Balanced OT and Unbalanced OT on the `Food101` (top) and the `OxfordPets` dataset (bottom); (**Right**) heatmaps of optimal transport plan related to each of class-specific context prompts learned from `GPT` on two examples of `Cat` and `Dog`.

4.5. Visualization

Transport Mapping from Balanced and Unbalanced OT. In Figure 5 (left), we provide an intuitive example of the output differences in optimal coupling of entropic OT and entropic UOT under outliers. In particular, we show multi-prompt alignment between 4 prompts and 20 images where only 4 images matched with prompts; others are negative samples. In the UOT setting, we set $\rho_1 \to \infty, \rho_2 = 0.04, \lambda = 0.01$ for conserving source marginal while relaxing target marginal. Clearly, the optimal coupling of entropic OT is blurry, thus introducing matching noises, while entropic UOT destroys noisy couplings and produces sharper matching. Intuitively, the total mass is conserved between the source and target distributions in entropic OT. However, this marginal constraint is restrictive in multi-prompt alignment problems where several word embeddings might not properly correspond to local visual ones, especially under data augmentation.

Table 3: **Ablation studies on few-shot recognition**: `CSC Prompt`: Class-specific context prompts for each class. `SC Prompt`: Domain-shared class prompts. `Self Att`: Shared Attention for all prompts

Dataset	Setting	1 shot	2 shot	4 shot	8 shot	16 shot
`Food101`	**Our (full)**	**77.8**	**77.8**	**77.9**	**78.5**	**78.7**
	w/o CSC Prompt	77.6	77.8	77.1	75.4	77.1
	w/o SC Prompt	75.4	77.1	77.3	77.8	78.4
	w/o GPT init	76.5	77.4	77.7	77.3	78.1
	w/o UOT	75.7	76.8	77.2	77.6	78.3
	w/o Self Att	61.8	68.0	70.8	74.1	75.7
`OxfordPets`	**Our (full)**	**87.5**	**87.5**	**88.1**	**88.9**	**88.4**
	w/o CSC Prompt	87.3	86.9	88.5	87.4	87.1
	w/o SC Prompt	84.3	87.0	87.5	87.7	88.1
	w/o GPT init	86.5	87.2	87.5	88.2	87.8
	w/o UOT	85.7	86.7	86.9	87.4	87.9
	w/o Self Att	82.5	83.1	85.5	85.8	87.6

Learnable Class-Specific Context Prompt. Figure 5 (right) presents the heatmap of four learnable prompts for each class. The UOT distance between each prompt embedding and visual local features is computed, illustrating correlations from transport plans. It is intuition to observe that each prompt targets distinct sub-regions of the image, covering object characteristics and relevant background. Such properties, therefore, may offer better guidance than a single shared class prompt, resulting in improved predictions.

5. Conclusion

This paper demonstrated that a large vision-language model like `CLIP` can be transformed into a data-efficient learner through prompt learning, utilizing a unified context and class-specific context initialized from the `GPT` model. Additionally, framing the distance between visual tokens and prompt features as an unbalanced optimal transport problem is essential for capturing misalignments and outliers between the two domains. This approach, combined with data augmentation to increase training samples, significantly enhances the model's few-shot learning abilities. Our results with prompt and adapter-based settings indicate substantial improvements over several competitive approaches. For future work, we propose to (i) test our framework on various types of adapter-based learning to validate its generalization capabilities and (ii) extend the method to vision-language model families trained with the autoregressive setting, such as LLAVA (Liu et al., 2024). This is particularly challenging since the learned embedding space structures in autoregressive models differ from those in `CLIP`, which is trained using a contrastive function.

Acknowledgement

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Duy M. H. Nguyen. Duy M. H. Nguyen and Daniel Sonntag are also supported by the XAINES project (BMBF, 01IW20005), No-IDLE project (BMBF, 01IW23002), and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University. An T. Le was supported by the German Research Foundation project METRIC4IMITATION (PE 2315/11-1). Mathias Niepert acknowledges funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC and support by the Stuttgart Center for Simulation Science (SimTech).

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *NeurIPS*, 2017.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *NeurIPS*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th ICML*, 2017.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *NeurIPS*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *ICLR*, 2023.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 2018.

Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. Distribution-aware prompt tuning for vision-language models. In *IEEE/CVF ICCV*, 2023.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML-PKDD*, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013.

John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 1966.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 2023.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, 2021.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2024.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *IEEE/CVF CVPR*, 2021.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *IEEE/CVF ICCV*, 2021.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*, 2023.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF CVPR*, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF ICCV*, 2023b.

Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE ICCV workshops*, 2013.

An T. Le, Georgia Chalvatzaki, Armin Biess, and Jan Peters. Accelerating motion planning via optimal transport. In *NeurIPS*, 2023a.

An T. Le, Kay Hansel, Jan Peters, and Georgia Chalvatzaki. Hierarchical policy blending as optimal transport. In *Learning for Dynamics and Control Conference*, 2023b.

Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *IEEE/CVF*

ICCV, 2023.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *NeurIPS*, 36, 2024.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 2018.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2024.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *IEEE/CVF CVPR*, 2022.

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

Duy MH Nguyen et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *NeurIPS*, 2024.

Eduardo Fernandes Montesuma, Fred Ngole Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *arXiv preprint arXiv:2306.16156*, 2023.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *IEEE/CVF CVPR*, 2023.

Duy MH Nguyen, Nina Lukashina, Tai Nguyen, An T. Le, TrungTin Nguyen, Nhat Ho, Jan Peters, Daniel Sonntag, Viktor Zaverkin, and Mathias Niepert. Structure-aware e (3)-invariant molecular conformer aggregation networks. *ICML*, 2024.

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

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE CVPR*, 2012.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *ICML*, 2020.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*, 2023.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 2023.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *IEEE/CVF CVPR*, 2022.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 1967.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Ming Xu and Stephen Gould. Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In *Proceedings of the IEEE/CVF CVPR*, 2024.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *IEEE/CVF CVPR*, 2023.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *IEEE/CVF CVPR*, 2023.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *IEEE/CVF CVPR*, 2021.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ECCV*, 2022a.

Yue Zhang, Hongliang Fei, Dingcheng Li, Tan Yu, and Ping Li. Prompting through prototype: A prototype-based prompt learning on pretrained vision-language models. *arXiv preprint arXiv:2210.10841*, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF CVPR*, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022c.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *IEEE/CVF ICCV*, 2023.