Inverse Optimal Transport for Efficient Adaptation of Vision-Language Models

Shupeng Qiu, Chuan-Xian Ren*

School of Mathematics, Sun Yat-Sen University, China
qiushp3@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

Vision–language models (VLMs) such as CLIP have unlocked powerful zero-shot transfer, yet efficient adaptation to downstream tasks remains challenging. Existing methods often depend on graph structures and dataset-specific tuning, making them sensitive to modality gaps and computationally costly at scale. In this paper, we propose IOTA (Inverse Optimal Transport Adaptation), a lightweight algorithm that reformulates VLMs inference from the perspective of inverse optimal transport (IOT), providing a unified view of training and inference. Under the IOT framework, IOTA enhances zero-shot alignment via a theory-guided unbalanced OT strategy and refines textual prototypes using OT-based pseudo-labels with a marginal-aware adaptive threshold, enabling reliable supervision without gradient updates. The framework naturally extends to few-shot scenarios through a label-guided masking mechanism. By decoupling image–text interactions from other inter-modal dependencies, IOTA avoids task-specific tuning and expensive affinity construction. Extensive experiments on standard benchmarks show that IOTA consistently improves zero-shot and few-shot performance while reducing memory and computation overhead, validating both its theoretical insight and plug-and-play practicality.

Introduction

With the availability of large-scale image–text data, vision–language models (VLMs) such as CLIP (Radford et al. 2021) have become a key paradigm for multimodal representation learning. By leveraging contrastive learning on paired image–text data, VLMs achieve effective cross-modal alignment and strong zero-shot transfer capabilities.

To efficiently adapt generic representations to domain-specific tasks, recent work explores parameter-efficient tuning, especially in zero- and few-shot settings. Popular approaches include prompt learning (Shu et al. 2022; Pratt et al. 2023) and adapter-based methods (Zhang et al. 2022; Silva-Rodriguez et al. 2024), which are mostly *inductive*, treating each test sample independently. In contrast, *transductive* methods exploit structural patterns across test samples and are gaining increasing attention. For example, TransCLIP (Zanella, Gérin, and Ayed 2024) formulates prediction as GMM-based clustering and applies a Laplacian regularizer on image features to encourage nearby samples to share similar assignments, effectively performing label propagation (Zhu and Ghahramani 2002). Similarly, ZLaP (Kalantidis, Tolias et al. 2024) and ECALP (Li et al. 2025) propagate labels over a graph connecting test samples with prompts and optional few-shot exemplars. OT-CLIP (Shi, Fan, and Yan 2024) instead frames CLIP training as a bilevel inverse optimal transport (IOT) problem, replacing softmax inference with fused Gromov–Wasserstein OT for structure-aware graph matching. Overall, these transductive approaches leverage graph structures to capture both intra- and cross-modal relationships.

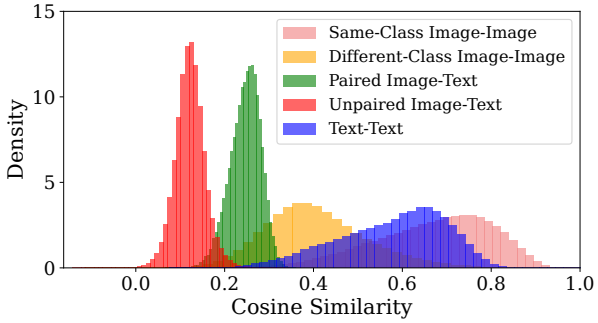


Figure 1: Cosine similarity distributions on the ImageNet (Deng et al. 2009) validation set, illustrating both intra-modal and cross-modal relationships.

While the above adaptation methods have shown promise, two critical issues remain. First, graph-based transductive approaches typically construct sample-wise affinity matrices to enforce label consistency but are highly sensitive to modality discrepancies (Liang et al. 2022), which can degrade cross-modal alignment. This sensitivity originates from the representation space itself: CLIP's contrastive pretraining explicitly aligns image-text pairs, while leaving image-image and text-text relations implicit. As shown in Figure 1, paired image–text similarities can even be lower than unpaired image–image ones, revealing a scale mismatch between intra- and cross-modal scores. Existing methods attempt to address this via balancing hyperparameters tuned on validation data, but such dataset-specific calibration is infeasible in test-time adaptation without labeled

*Corresponding Author.

target samples. Second, constructing such dense affinity matrices in intra-modal spaces, particularly for large datasets like ImageNet, incurs prohibitive memory and computational costs. These challenges highlight an adaptation framework that avoids dataset-specific tuning and expensive computation while maintaining robust cross-modal alignment.

In this paper, we introduce *IOTA (Inverse Optimal Transport Adaptation)*, as shown in Figure 2, a lightweight algorithm designed for adapting vision-language models within the inverse optimal transport framework. We demonstrate that standard CLIP transductive prediction via cosine similarity can be viewed as a special case of IOT, where the transport plan encodes image-text interactions. However, modality gaps often hinder text embeddings from capturing fine-grained visual details, leading to suboptimal transport plans. To address this, IOTA employs two complementary strategies: (i) *theory-guided alignment via unbalanced OT*, which applies a soft regularization on text-side marginals to capture the intrinsic structure in unlabeled test data. This refinement enhances the transport plan and boosts zero-shot inference, which benefits from clear theoretical interpretation through Sinkhorn iterations. (ii) *prototype refinement via OT-based pseudo-labels*, which utilizes a marginal-aware adaptive threshold to calibrate class-wise confidence. This strategy adapts to the OT plan's structure by tailoring the refinement process to different marginal priors, yielding high-quality pseudo-labels. Furthermore, IOTA naturally extends to few-shot scenarios through a transport-aware masking mechanism that anchors labeled samples. By decoupling image–text interactions from other inter-modal dependencies, IOTA eliminates dataset-specific tuning and expensive affinity construction, and remains modular for seamless integration into existing inductive adaptation pipelines. Overall, our contributions are summarized as follows:

- Standard CLIP prediction is reformulated as a special case of IOT framework, unifying training and inference and laying the foundation for structure-aware adaptation.
- An efficient algorithm is developed within the IOT framework to enhance zero-shot performance through theory-guided alignment via unbalanced OT and refine textual prototypes with OT-based pseudo-labels. The framework naturally extends to few-shot settings via a transport-aware masking mechanism.
- Extensive experiments on standard zero-shot and few-shot benchmarks validate the theoretical insights of IOTA, demonstrating competitive performance with significantly lower memory and computational overhead, and confirming its practical plug-and-play utility.

Related Work

Vision-Language Model Adaptation. Efficient adaptation of pre-trained VLMs to downstream tasks has garnered significant interest. Existing approaches can be broadly divided into two categories, i.e. prompt tuning and adapter tuning. Prompt tuning like CoOp (Zhou et al. 2022), PLOT (Chen et al. 2022), TaskRes (Yu et al. 2023), learn continuous vectors prepended to input texts, effectively modifying the text embedding space by the few labeled samples. TPT (Shu et al. 2022), UPL (Huang, Chu, and Wei 2022) extend this concept to zero-shot settings by using the predictive confidence. While these approaches are parameter-efficient by design, they still rely on relatively costly optimization. Adapter tuning methods, such as CLIP-Adapter (Gao et al. 2024), introduce lightweight modules after frozen encoders and learn residual adapters. More relevant to efficient adaptation are training-free, cache-based methods such as Tip-Adapter (Zhang et al. 2022), APE (Zhu et al. 2023), SuS-X (Udandarao, Gupta, and Albanie 2023), DMN (Zhang et al. 2024), and TDA (Karmanov et al. 2024). These approaches leverage CLIP-extracted features from few-shot examples to construct a non-parametric cache model for nearest-neighbor retrieval, thereby explicitly incorporating CLIP's visual priors. In the transductive setting, InMaP (Qian, Xu, and Hu 2023) updates class proxies using pseudo-labels with unlabeled test set in the zero-shot context. Graph-based methods, including ZLaP (Kalantidis, Tolias et al. 2024), TransCLIP (Zanella, Gérin, and Ayed 2024), ECALP (Li et al. 2025), exploit the data manifold structure via label propagation to implicitly capture relations between image features and textual prototypes, enabling effective adaptation in both zero-shot and few-shot scenarios.

Optimal Transport (OT). As a mathematical tool for distribution alignment, OT has been widely used in the machine learning community. Our work is particularly related to the IOT problem (Cui et al. 2019), which learns a cost function to explain an observed semantic alignment. (Shi et al. 2023) interpret contrastive learning through IOT, showing that the InfoNCE objective implicitly solves a bi-level IOT problem. OT-CLIP (Shi, Fan, and Yan 2024) formulates CLIP pretraining as bi-level IOT optimization. GCA (Chen et al. 2024) further connects InfoNCE losses with entropic OT, providing new theoretical insight into contrastive representation learning. Several recent works integrate OT into unsupervised and semi-supervised settings by leveraging label-aware guidance. KPG-RL (Gu et al. 2022) introduces a keypoint-masked OT to preserve the matching of keypoint pairs, while MOT (Luo and Ren 2023) imposes mask for domain adaptation under label shift. Similarly, we propose a label-guided masking strategy to enhance transport accuracy in few-shot scenarios. Other works employ OT for pseudo-label refinement. P^2OT (Zhang, Ren, and He 2024) formulates the pseudo label generation for imbalanced clustering as an unbalanced OT problem. Similar to our work, InMap (Qian, Xu, and Hu 2023) shows the modality gap in CLIP is insufficiently reduced during pretraining, and improves zero-shot performance by refining pseudo labels via OT, while its use of OT is indirect and relies on heuristics.

Methodology

In this section, we show that standard VLM inference is a special case of IOT, enabling natural improvements in adaptation and supporting zero-shot and few-shot classification.

Preliminaries

Contrastive Vision-Language Pretraining. Let $\mathcal{X}$ and $\mathcal{Z}$ denote the image and text input spaces. Contrastive vision-

language models such as CLIP (Radford et al. 2021) employ dual encoders $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and $g_\phi : \mathcal{Z} \to \mathbb{R}^d$ that map inputs into a shared ℓ_2-normalized embedding space. Given a batch of aligned image-text pairs $\{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^n$, the model is trained to maximize similarity for matched pairs while suppressing mismatches using a symmetric InfoNCE loss:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2n}\sum_{i=1}^{n}\left[\log\frac{\exp(f_\theta(\mathbf{x}_i)^\top g_\phi(\mathbf{z}_i)/\tau)}{\sum_{j=1}^{n}\exp(f_\theta(\mathbf{x}_i)^\top g_\phi(\mathbf{z}_j)/\tau)} + \log\frac{\exp(g_\phi(\mathbf{z}_i)^\top f_\theta(\mathbf{x}_i)/\tau)}{\sum_{j=1}^{n}\exp(g_\phi(\mathbf{z}_i)^\top f_\theta(\mathbf{x}_j)/\tau)}\right], \tag{1}$$

where $\tau > 0$ is a temperature parameter controlling the sharpness of the similarity distribution.

Optimal Transport Reformulation. OT provides a distributional perspective on contrastive learning. Let $\mu \in \Delta^n$ and $\nu \in \Delta^m$ be discrete probability distributions over n and m samples, where Δ^k is the k-dimensional probability simplex. Given a ground cost matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$, with $\mathbf{C}_{ij}$ denoting the cost of moving mass from the i-th support of μ to the j-th support of ν, OT seeks a coupling matrix $\mathbf{P} \in \mathbb{R}_+^{n\times m}$ that minimizes the total transport cost:

$$\min_{\mathbf{P}\in\Pi(\mu,\nu)} \langle \mathbf{C}, \mathbf{P}\rangle,$$

where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product and $\Pi(\mu,\nu) = \{\mathbf{P} \geq 0 \mid \mathbf{P}\mathbf{1}_m = \mu,\ \mathbf{P}^\top\mathbf{1}_n = \nu\}$.

To improve scalability, entropic OT (Cuturi 2013) introduces entropy regularization:

$$\min_{\mathbf{P}\in\Pi(\mu,\nu)} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}), \tag{2}$$

where $H(\mathbf{P}) = -\sum_{i,j}\mathbf{P}_{ij}(\log\mathbf{P}_{ij} - 1)$ and $\varepsilon > 0$ controls plan smoothness. The solution can be efficiently approximated by the Sinkhorn algorithm. When the row marginals are uniform, the optimal plan admits a closed form:

$$\mathbf{P}_{ij} = \frac{\exp(-\mathbf{C}_{ij}/\varepsilon)}{n\sum_{k=1}^{m}\exp(-\mathbf{C}_{ik}/\varepsilon)}. \tag{3}$$

Notably, the temperature τ in Eq. (1) is mathematically equivalent to the OT regularization coefficient ε.

Building on this, recent work (Shi, Fan, and Yan 2024) reformulates CLIP training as an IOT problem. The goal is to learn encoder parameters $\Theta = (\theta, \phi)$ for the image and text encoders so that the OT plan matches the ideal diagonal coupling $\tilde{\mathbf{P}} \in \mathbb{R}^{n\times n}$, where $\tilde{\mathbf{P}}_{ij} = \delta_{ij}$ and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Consequently, the InfoNCE objective in Eq. (1) can be written as the following bilevel optimization:

$$\begin{aligned} &\min_{\Theta} \mathrm{KL}(\tilde{\mathbf{P}} \,\|\, \mathbf{P}_I) + \mathrm{KL}(\tilde{\mathbf{P}} \,\|\, \mathbf{P}_T) \\ &\text{s.t.} \quad \mathbf{P}_I = \arg\min_{\mathbf{P}\in\Pi_I}\langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}), \\ &\qquad \mathbf{P}_T = \arg\min_{\mathbf{P}\in\Pi_T}\langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}), \end{aligned} \tag{4}$$

where $\mathbf{C}_{ij} = 1 - f_\theta(\mathbf{x}_i)^\top g_\phi(\mathbf{z}_j)$ is the cost matrix, and Π_I, Π_T are the row- and column-normalized marginal constraints. This formulation aligns the OT plan with the ideal alignment. Follow up, we will provide a unified view of pretraining and inference through the IOT framework.

Transductive Inference via IOT Framework

Let $\mathcal{D} = \{1, \ldots, N\} = \mathcal{Q} \cup \mathcal{S}$ denote the test set indices, partitioned into an unlabeled query set $\mathcal{Q}$ and an optional labeled support set $\mathcal{S}$ providing reference labels in the few-shot setting. For each query index $i \in \mathcal{Q}$, let $\mathbf{x}_i \in \mathcal{X}$ be the corresponding test image. Class prompts $\{\mathbf{c}_k\}_{k=1}^C \subset \mathcal{Z}$ are constructed using predefined templates such as `"a photo of a [class name]"`. Using the frozen pretrained encoders, we can obtain image features and textual prototypes:

$$\mathbf{f}_i = f_\theta(\mathbf{x}_i), \quad \mathbf{t}_k = g_\phi(\mathbf{c}_k).$$

In the zero-shot setting, $\mathcal{S} = \emptyset$. The standard inductive zero-shot classification yields prediction via cosine similarity:

$$y_i = \arg\max_{k\in[C]} \mathbf{f}_i^\top \mathbf{t}_k. \tag{5}$$

While effective, DN (Zhou et al. 2023) reveals that this dot product is a zeroth-order approximation of the InfoNCE objective, as it neglects negative sample distributions. Meanwhile, OT-CLIP (Shi, Fan, and Yan 2024) reframes CLIP training as an IOT problem, as described in (4). However, its inference phase focuses solely on the lower-level OT solver, while neglecting the upper-level KL alignment. Building on these insights, we investigate whether inference can be viewed as a counterpart to the full IOT process. From the IOT perspective, the goal of test-time prediction is to recover the most likely label assignment $\tilde{\mathbf{P}} \in \{0,1\}^{N\times C}$. The following proposition formalizes this relationship:

Proposition 1 (KL-Optimal Hard Label)**.** *Let $\mathbf{p}' \in \Delta^C$ be a soft label distribution over C classes, and let $\mathbf{e}_k$ denote the one-hot basis vector. The one-hot label $\mathbf{p}^* \in \{0,1\}^C$ minimizing the KL divergence $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{p}')$ is*

$$\mathbf{p}^* = \mathbf{e}_{k^*}, \quad k^* = \arg\max_j \mathbf{p}'_j. \tag{6}$$

For a batch of N samples with soft label matrix $\mathbf{P}' \in \mathbb{R}_+^{N\times C}$ where each row $\mathbf{P}'_{i,:} \in \Delta^C$, the optimal hard label matrix $\mathbf{P}^* \in \{0,1\}^{N\times C}$ is obtained by row-wise projection:

$$\mathbf{P}^*_{i,:} = \mathbf{e}_{k^*(i)}, \quad k^*(i) = \arg\max_j \mathbf{P}'_{ij}, \ \ \forall i \in [N]. \tag{7}$$

Remark. Proposition 1 reveals that the $\arg\max$ operation is not an ad-hoc choice but the KL-optimal completion of the IOT bilevel optimization. In other words, the entropy-regularized OT solves for the soft OT plan $\mathbf{P}'$ in the lower-level optimization, while the final label prediction, made by the $\arg\max$ decision rule, implicitly completes the upper-level IOT process. This shows that both training and inference follow the same IOT framework, providing a unified view of the two phases. Inspired by this mathematical equivalence, we now develop a refined inference algorithm, optimizing the IOT structure during test-time prediction.

Zero-Shot Label Refinement with IOTA

In the IOT framework, aligning image and text features through the OT plan matrix is crucial for label prediction. However, the modality gap between images and text (Liang et al. 2022) limits the ability of text representations to capture fine-grained visual details, resulting in suboptimal cost matrix accuracy and degrading label prediction. To address this, we introduce two strategies to enhance OT plan.

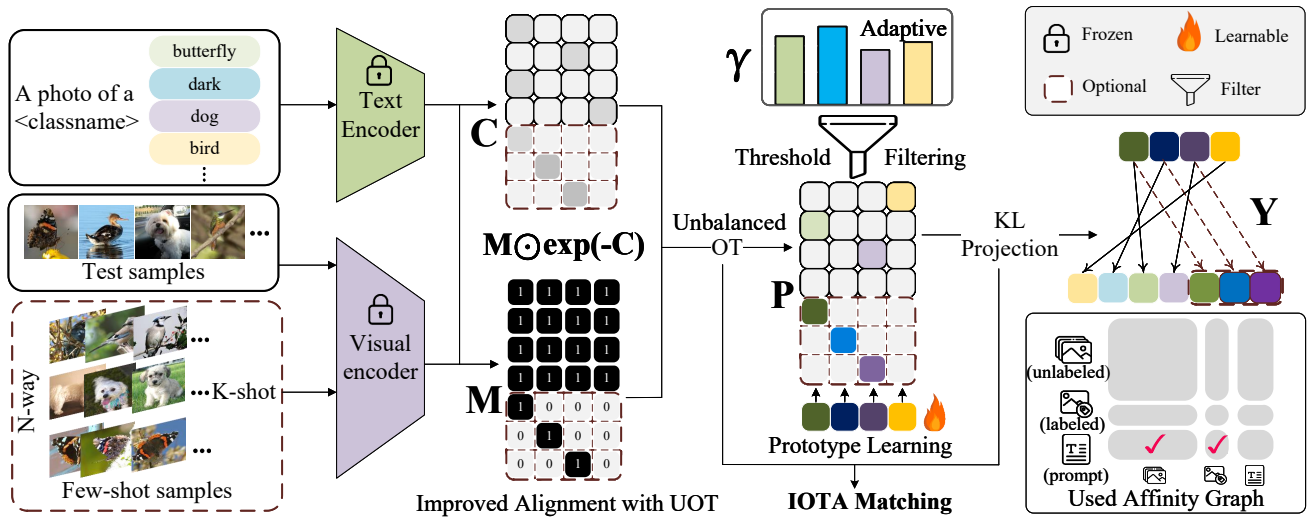


Figure 2: Overview of IOTA: Image–text alignment via unbalanced OT, robust pseudo-labeling via adaptive thresholds, and OT-based prototype learning support zero-/few-shot adaptation without large affinity graphs.

OT-based Label Refinement. While CLIP training imposes a single constraint on the transport plan, inference typically reuses this partial alignment without refinement and ignores the inherent distributional structure of unlabeled test samples. To complete the distributional alignment in the IOT framework, we propose introducing a soft constraint on the column (text-side) marginal during inference, thereby encouraging better alignment between image features and class prompts. This yields the following unbalanced OT objective:

$$\min_{\mathbf{P}\in\Pi(\mu,\nu)} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) + \lambda \,\mathrm{KL}(\mathbf{P}^\top \mathbf{1}_N \,\|\, \nu), \quad (8)$$

where $\mu = \frac{1}{N}\mathbf{1}_N$ denotes the uniform distribution over N query samples, ν is is either the empirical distribution or a uniform class prior, and λ is the parameter for marginal penalty, with the cost matrix defined as $\mathbf{C}_{ik} = 1 - \mathbf{f}_i^\top \mathbf{t}_k$.

The KL term serves as a soft regularizer that promotes class-wise assignments without relying on ground-truth labels, making the method compatible with the zero-shot setting. The resulting transport problem is a single-sided unbalanced OT (UOT) with entropy regularization, which can be efficiently solved by a generalized Sinkhorn algorithm (Chizat et al. 2018). Specifically, defining the kernel matrix $\mathcal{K} = \exp(-\mathbf{C}/\varepsilon) \in \mathbb{R}_+^{N\times C}$, the Sinkhorn iterations

$$\mathbf{u}^{(k)} = \frac{\mu}{\mathcal{K}\mathbf{v}^{(k-1)}}, \quad \mathbf{v}^{(k)} = \left(\frac{\nu}{\mathcal{K}^\top \mathbf{u}^{(k)}}\right)^{\frac{\lambda}{\lambda+\varepsilon}}. \quad (9)$$

Then the sequence $\mathrm{diag}(\mathbf{u}^{(k)})\mathcal{K}\mathrm{diag}(\mathbf{v}^{(k)})$ will converge to the solution of Eq. (8). The OT plan is then projected to hard labels via the KL-optimal rule described in Proposition 1.

Critically, contrastive learning theory (Piran et al. 2024) establishes that InfoNCE is a single-sided projection, equivalent to a half-step Sinkhorn iteration. By introducing soft regularization on text-side marginals, our method completes this partial projection and refines the transport plan at test time. This yields provable theoretical improvements:

Proposition 2 (Improved Alignment with UOT)**.** *Assume the prescribed marginals μ and ν match the empirical marginals of the ground-truth label matrix $\mathbf{Y}$. Let $\mathbf{P}^{(k)}$ be the transport plan at the k-th Sinkhorn iteration of solving Eq.* (8)*. Then the upper-level KL loss with the converged plan $\mathbf{P}^{(\infty)}$ satisfies:*

$$\mathrm{KL}(\mathbf{Y} \,\|\, \mathbf{P}^{(\infty)}) \le \mathrm{KL}(\mathbf{Y} \,\|\, \mathbf{P}^{(k)}), \quad \forall k,$$

where $\mathbf{Y} \in \{0,1\}^{N\times C}$ is the ideal label assignment.

Remark. Proposition 2 assumes the prescribed marginals μ, ν match the ground-truth distribution. In practice, an uniform or empirical distribution often suffices. Even imperfectly, unbalanced constraints promote structured alignment and boost zero-shot predictions in a training-free manner.

Prototype Learning via Pseudo-Label Alignment. To bridge the modality gap between text prototypes and test-time visual features, a standard approach for adapting frozen VLMs is *linear probing*, which learns a set of class-wise prototypes $\mathbf{W} = \{\mathbf{w}_k\}_{k=1}^C$ via supervised softmax regression:

$$\min_{\mathbf{W}} -\sum_{i\in S} y_i \log \frac{\exp(\mathbf{f}_i^\top \mathbf{w}_k/\tau)}{\sum_{l=1}^C \exp(\mathbf{f}_i^\top \mathbf{w}_l/\tau)}. \quad (10)$$

This requires labeled data, failing in zero-shot scenarios. To address this, we propose an IOT-guided formulation for prototype learning that replaces ground-truth supervision with soft pseudo-labels $\hat{\mathbf{P}}$ derived from Eq. (8). Specifically, we reformulate Eq. (10) as:

$$\begin{aligned} &\min_{\mathbf{W}} \ \mathrm{KL}(\hat{\mathbf{P}} \,\|\, \mathbf{P}_I), \\ &\text{s.t.} \quad \mathbf{P}_I = \arg\min_{\mathbf{P}\in\Pi_I} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon_I H(\mathbf{P}), \end{aligned} \quad (11)$$

where $\mathbf{C}_{ik} = 1 - \mathbf{f}_i^\top \mathbf{w}_k$, and the regularization coefficient ε_I acts as the softmax temperature. While enforcing both row- and column-wise constraints on Π_I offers theoretical

elegance, it entails significant computational overhead due to OT solver iterations. Since the OT-based pseudo-labels $\hat{\mathbf{P}}$ already capture inter-class relationships, retaining only the image-side marginal constraint is sufficient for effective zero-shot prototype learning, as by the results in Table 4.

To enhance prototype learning with more reliable OT-based pseudo-labels, *an adaptive thresholding strategy* is adopted, inspired by recent advances in semi-supervised learning (Wang et al. 2022; Tan, Zheng, and Huang 2023). Unlike fixed thresholds that ignore class-level confidence imbalance across simple and fine-grained categories, the class-wise threshold combines two components:

$$\gamma(k) = \underbrace{\frac{\mathbf{p}(k)}{\max_{k'} \mathbf{p}(k')}}_{\text{class-wise calibration}} \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} \max(\hat{\mathbf{P}}_{i:})}_{\text{global baseline}} \tag{12}$$

where $\mathbf{p}(k) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{P}}_{ik}$ is the mean confidence for class k. The global term keeps overall top-1 confidence to capture dataset-level variations, while class-wise calibration ensures thresholds are globally consistent and locally adaptive, reducing overconfidence and improving rare-class recall.

Furthermore, the threshold $\gamma(k)$ is applied differently based on the marginal constraints of Eq (8). Under *uniform marginals*, the OT plan $\hat{\mathbf{P}}$ is constrained to match uniform class priors, producing inherently smoother predictions. To exploit this soft structure without forcing premature one-hot assignments, a partial hardening strategy is adopted:

$$\hat{\mathbf{P}}_{i,:} = \begin{cases} \mathbf{e}_k & \text{if } \hat{\mathbf{P}}_{ik} > \gamma(k) \text{ and } k = \arg\max \hat{\mathbf{P}}_{ij} \\ \hat{\mathbf{P}}_{i,:} & \text{otherwise.} \end{cases} \tag{13}$$

Under *empirical marginals*, the OT plan better reflects the initial data manifold, making confidence scores more meaningful. In this case, we adopt a binary masking strategy:

$$\hat{\mathbf{P}}_{ij} = \begin{cases} 1 & \text{if } \hat{\mathbf{P}}_{ij} > \gamma(j) \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where entries above the class-wise threshold are treated as reliable (possibly multi-label) positives, while low-confidence entries are suppressed to 0. More details of this marginal-aware mechanism are provided in the Appendix.

Extension to the Few-Shot Setting

Our zero-shot formulation naturally extends to the few-shot setting ($\mathcal{S} \neq \emptyset$) via *label-guided masking* with theoretical guarantees for feasible OT solutions (Luo and Ren 2023). Specifically, a binary mask $\mathbf{M} \in \{0,1\}^{N \times C}$ is defined as

$$\mathbf{M}_{i,:} = \begin{cases} \mathbf{e}_k & \text{if } i \in \mathcal{S} \text{ with label } k, \\ \mathbf{1}_C & \text{if } i \in \mathcal{Q}, \end{cases} \tag{15}$$

to anchor labeled samples in Eq. (8). It can be incorporated into the Sinkhorn updates via a modified Gibbs kernel:

$$\mathcal{K} = \mathbf{M} \odot \exp(-\mathbf{C}/\varepsilon), \tag{16}$$

where $\odot$ denotes element-wise multiplication. From the cost matrix perspective, this is equivalent to setting $\mathbf{C}_{ik} = +\infty$ where $\mathbf{M}_{ik} = 0$, explicitly suppresses negative class assignments. Notably, this mechanism integrates seamlessly with our adaptive threshold strategy.

Algorithm 1: IOTA

Input: Test indices $\mathcal{D} = \{1, \ldots, N\} = \mathcal{Q} \cup \mathcal{S}$, image embeddings $\{\mathbf{f}_i\}_{i \in \mathcal{D}}$, class prototypes $\{\mathbf{t}_k\}_{k=1}^{C}$, temperature of pretrained CLIP model τ

Parameters: UOT regularization weight λ, temperature ε_I

Output: Refine predictions $\hat{\mathbf{P}}$

1: Initialize adaptive prototypes $\mathbf{w}_k \leftarrow \mathbf{t}_k, k = 1, ..., C$
2: Compute binary mask matrix $\mathbf{M}$ using Eq. (15) and kernel matrix $\mathcal{K}$ using Eq. (16)
3: Compute OT-based refined label $\hat{\mathbf{P}}$ using Eq. (8) via Sinkhorn iterations Eq. (9)
4: Apply adaptive threshold γ from Eq. (12) and harden pseudo-labels via Eq. (13) or Eq. (14)
5: Update adaptive prototypes $\{\mathbf{w}_k\}_{k=1}^{C}$ by minimizing KL divergence in Eq. (11)
6: **return** $\{\mathbf{w}_k\}_{k=1}^{C}$

Overall Flow of the Algorithm

We summarize our method, **IOTA** (Inverse Optimal Transport Adaptation), in Algorithm 1, which naturally supports both *zero-shot* and *few-shot* learning. Unlike prior graph-based methods such as TransCLIP (Zanella, Gérin, and Ayed 2024), that construct dense image-image affinity graphs with $\mathcal{O}(N^2)$ memory cost, IOTA directly operates on the image-text similarity matrix, reducing the complexity to $\mathcal{O}(NC)$.

Experiments and Analysis

Datasets and Implementation Details. We evaluate on 11 *fine-grained classification* benchmarks: ImageNet (Deng et al. 2009), Flowers102 (Nilsback and Zisserman 2008), DTD (Cimpoi et al. 2014), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), UCF101 (Soomro, Zamir, and Shah 2012), Caltech101 (Fei-Fei, Fergus, and Perona 2004), Food101 (Bossard, Guillaumin, and Van Gool 2014), SUN397 (Xiao et al. 2010), Aircraft (Maji et al. 2013), and EuroSAT (Helber et al. 2019). In addition, four variants of ImageNet: ImageNet-V2 (Recht et al. 2019), ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), and ImageNet-Sketch (Wang et al. 2019) are employed to test the *domain generalization* performance. We use the pretrained CLIP model (Radford et al. 2021) with ResNet-50 (He et al. 2016) (default) and ViT-B/16 (Dosovitskiy et al. 2020) as backbone. IOTA† constructs OT-based pseudo-labels with empirical distributions while IOTA adopts an uniform prior. *All settings share unified hyperparameters. More details are provided in the Appendix.*

Comparison with SOTA Methods. We employ some state-of-the-art methods for comparison, which is mostly capable in zero-shot and few-shot setting: CoOp (Zhou et al. 2022), TPT (Shu et al. 2022), Tip-Adapter, DiffTPT (Feng et al. 2023), SuS-X, DMN, TDA (Karmanov et al. 2024), APE, ZLaP, TransCLIP, ECALP (Li et al. 2025).

Zero-shot evaluation. The results on *fine-grained categorization tasks* are shown in Table 1. IOTA† achieves the best average accuracy of **64.64%** with ResNet-50 and **71.73%** with ViT-B/16 backbones when using empirical

Method	ImageNet	Flower	DTD	Pets	Cars	UCF	Caltech	Food	SUN	Aircraft	EuroSAT	Avg.
CLIP-RN50	58.16	61.75	40.37	83.57	55.70	58.84	85.88	73.97	58.80	15.66	23.69	56.04
TPT	60.74	62.69	40.84	84.49	58.46	60.82	87.02	74.88	61.46	17.58	28.33	57.94
DiffTPT	60.80	63.53	40.72	83.40	60.71	62.67	86.89	**79.21**	62.72	17.60	41.04	59.94
SuS-X	61.84	67.72	50.59	85.34	57.27	61.54	89.53	77.58	62.95	19.47	45.57	61.76
DMN	62.02	68.33	50.53	86.29	58.36	64.02	89.09	74.69	63.70	20.22	44.94	62.02
TDA	61.35	68.74	43.74	86.18	57.78	64.18	**89.70**	77.75	62.53	17.61	42.11	61.06
ZLaP	62.20	69.27	42.79	80.32	56.42	62.81	86.90	77.87	61.83	17.37	31.85	59.06
TransCLIP	60.81	**72.15**	47.78	89.30	57.89	68.84	88.60	78.01	64.22	16.60	**59.58**	63.98
ECALP	62.64	69.39	54.49	88.20	60.56	66.67	89.94	76.97	64.97	21.12	49.09	64.00
IOTA†	**63.26**	72.07	54.26	**89.34**	**61.77**	**67.25**	87.38	77.10	**66.37**	20.04	52.17	**64.64**
IOTA	**64.25**	69.51	**55.85**	**89.72**	**64.02**	**68.68**	76.84	77.30	**67.47**	**22.26**	51.85	**64.32**
CLIP-ViTB/16	66.73	64.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.87
TPT	68.98	68.98	47.75	87.79	66.87	68.04	94.16	84.67	65.50	24.78	42.44	65.45
DiffTPT	70.30	70.10	47.00	88.20	67.01	68.22	92.49	87.23	65.74	25.60	43.13	65.90
DMN	70.51	75.32	54.85	91.22	67.01	71.95	93.63	84.05	69.14	28.29	56.22	69.29
TDA	69.51	71.42	47.40	88.63	67.28	70.66	94.24	86.14	67.62	23.91	58.00	67.71
ZLaP	70.17	73.49	48.58	87.14	65.63	71.45	93.06	86.92	67.44	25.44	55.62	67.72
TransCLIP	70.31	76.72	49.54	92.63	68.89	74.38	92.67	**87.13**	68.80	26.91	**65.10**	70.22
ECALP	71.27	75.52	54.67	92.04	68.03	75.92	**94.24**	85.65	70.61	29.13	55.95	70.28
IOTA†	**72.78**	**76.86**	**57.09**	**93.08**	**72,60**	**76.42**	92.62	86.03	**72.58**	**31.68**	57.32	**71.73**
IOTA	**72.89**	75.27	**57.86**	**93.05**	**72.62**	**77.00**	77.97	86.04	**73.03**	**31.17**	58.05	**70.45**

Table 1: Evaluation of zero-shot adaptation on fine-grained categorization tasks.

distributions, while IOTA obtains competitive accuracies of **64.32%** and **70.45%** respectively. Compared with TransCLIP and ECALP, which rely on graph-based to leverage intra-modal relationships, IOTA directly explores the image-text interactions, offering a more efficient and principled solution. We truthfully note that IOTA underperforms on Flower102 and Caltech due to highly imbalanced, with the latter even falling below the baseline. While more relaxed regularization (even reduced to IOTA†) could mitigate this issue, real-world test scenarios typically require a unified, dataset-agnostic configuration. It cannot be denied that under such conditions, our method still consistently yields improvements, with an average gain of approximately **8%** than baseline. We further report results on the *domain generalization tasks* in Table 2. IOTA is robust to distributional shifts and performs competitively against SOTA adaptation methods, highlighting its strong generalization capability.

Few-shot evaluation. We present few-shot results on fine-grained datasets under {1, 2, 4, 8, 16}-shot settings using ResNet-50 backbones in Figure 3. IOTA demonstrates outstanding performance across all shot levels on commonly challenging large datasets. Furthermore, as the number of shots increases, IOTA *consistently improves*, whereas other methods suffer under unified hyperparameter settings, underscoring its adaptability in low-data scenarios.

Plug-and-Play Compatibility. IOTA is orthogonal to existing adaptation methods, as it operates purely on extracted features without architectural modifications. We validate its plug-and-play compatibility by applying it to prompt tuning (CoOp) and adapter-based method (CLAP (Silva-Rodriguez et al. 2024)), as shown in Figure 4a and 4b. IOTA consistently improves performance across all shot settings, with the largest gains up to **4.77**% for CoOp and **2.63**% for CLAP.

Method	-A	-V2	-R	-Sketch	Avg.
CLIP-RN50	21.83	51.41	56.15	33.37	40.69
CoOp	23.06	55.40	56.60	34.67	42.43
TPT	26.67	54.70	59.11	35.09	43.89
DiffTPT	31.06	55.80	58.80	37.10	45.69
CALIP	23.96	53.70	60.81	35.61	43.52
TDA	**30.29**	55.54	62.58	38.12	46.63
DMN	28.57	56.12	61.44	39.84	46.49
ECALP	28.80	56.92	**63.68**	41.51	47.73
IOTA†	28.96	56.79	63.50	39.35	47.15
IOTA	27.71	**58.36**	63.16	**44.51**	**48.43**

Table 2: Domain generalization performance (%) of IOTA.

Ablation Study Compared to vanilla CLIP (Fig.4c), IOTA (Fig.4d) shows a clearer diagonal structure and achieves a **+20.39**% absolute gain in intra-class transport mass, confirming the theoretical insight of Proposition 2. Table 3 further shows that augmenting UOT with prototype adaptation yields an additional **+1.68%**, while incorporating threshold filtering brings another **+1.0%**. Additional thresholding results are in the Appendix. In the few-shot setting, IOTA-FS gains **+5.63%** over the variant without the label-guided mask, showing effective use of limited supervision.

Effect of Marginal Constraints. Table 4 compares the double-marginal constraint in Eq. (11) with simplified image-side-only constraint. The latter achieves competitive accuracy while reducing runtime by up to **6**×, showing that

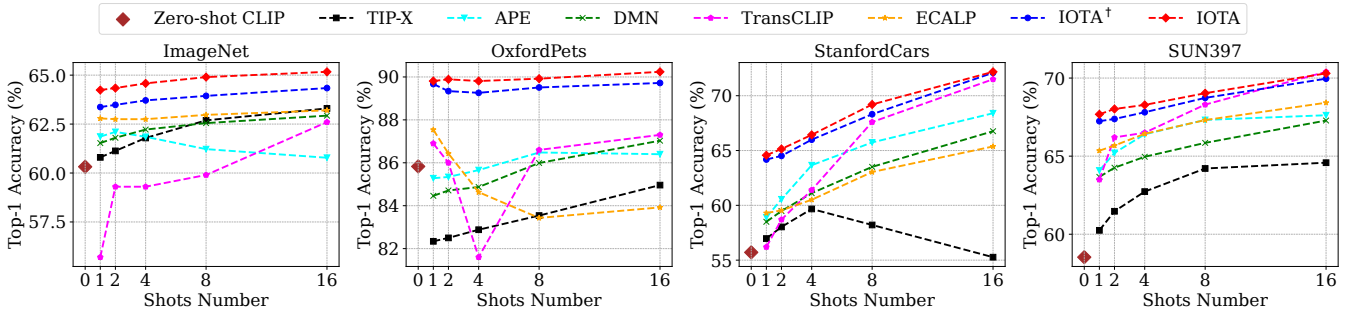
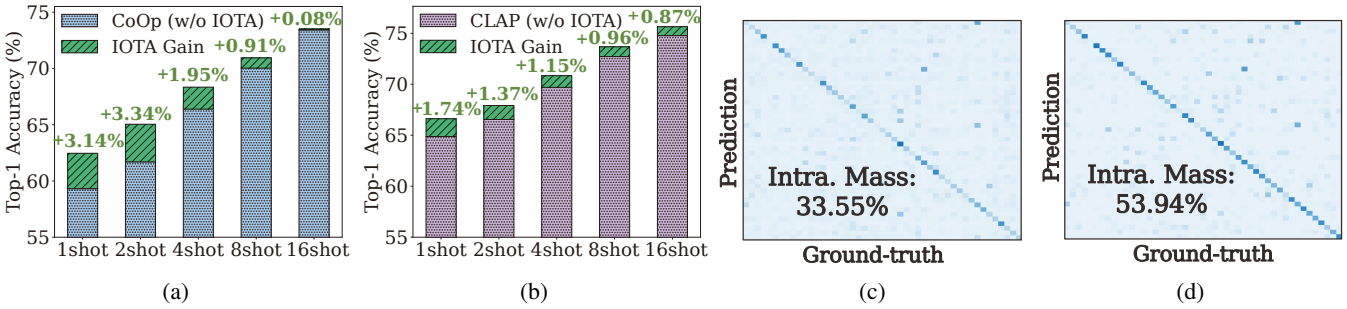


Figure 3: Evaluation of few-shot adaptation on fine-grained categorization tasks.



(a) (b) (c) (d)

Figure 4: Analysis on IOTA. (a) and (b) respectively show IOTA implemented atop the advanced prompt-tuning method CoOp and the adapter-based method CLAP, reporting average performance on 11 fine-grained datasets. while (c) and (d) represent heat-maps of vanilla CLIP and IOTA on DTD, respectively, where darker colors represent larger probability coupling values.

Strategy	UOT	Adapting	Threshold	Mask	Avg.
-	✓	-	-	-	61.64
-	✓	✓	-	-	63.32
-	-	✓	✓	-	62.85
IOTA-ZS	✓	✓	✓	-	**64.32**
-	✓	✓	✓	-	65.25
IOTA-FS	✓	✓	✓	✓	**70.88**

Table 3: Ablation of IOTA components. (ZS: zero-shot; FS: 16-shot; Adapting: Eq. (11) via vanilla CLIP pseudo-labels)

Constraint	ImageNet		UCF	
	Accuracy	Runtime	Accuracy	Runtime
double	**64.26**	56.76s	**68.86**	0.71s
single	64.25	**8.93s**	68.68	**0.10s**

Table 4: Effect of marginal constraint on IOTA performance.

λ	ImageNet	DTD	UCF	SUN397	Avg.
0.05	64.22	55.79	**68.83**	67.48	**64.37**
0.1	**64.26**	55.73	68.75	67.47	64.25
0.5	64.25	**55.85**	68.68	67.47	64.32
5	64.25	55.61	68.81	**67.53**	64.30

Table 5: Sensitivity analysis of the hyperparameter.

$\hat{\mathbf{P}}$ sufficiently encodes inter-class structure without the need for additional column-wise constraints. Empirically, IOTA processes 50K ImageNet samples in 8.93 s, faster than TransCLIP (14.4 s) and other methods.

Hyper-parameter Sensitivity. Table 5 shows IOTA's robust stability across a wide range of UOT regularization weights λ (from 0.05 to 5) on fine-grained datasets. The average accuracy fluctuates within a narrow band of **0.12%** (64.25% to 64.37%), indicating that our method is highly robust to the choice of λ. This stability also stems from our adaptive threshold compensating suboptimal transport plans.

Conclusion

In this work, we recast standard CLIP inference under the IOT framework. Building on this, we introduce IOTA, a lightweight adaptation algorithm that decouples image–text interactions from other inter-modal dependencies for efficient adaptation of VLMs. IOTA enhances zero-shot alignment via a theory-guided UOT strategy and further bridges the modality gap by refining textual prototypes with OT-guided pseudo-labels. Besides, a marginal-aware threshold mechanism tailors refinement to different OT marginal priors. The framework naturally extends to few-shot scenarios through a label-guided masking strategy. Benefiting from its principled IOT foundation, IOTA supports hyperparameter-free deployment across diverse datasets, achieving robust and efficient zero-/few-shot adaptation with seamless plug-and-play compatibility on a wide range of downstream tasks.

Acknowledgments

This work is supported in part by National Key R&D Program of China (2024YFA1011900), National Natural Science Foundation of China (Grant No. 62376291), Guangdong Basic and Applied Basic Research Foundation (2023B1515020004), Science and Technology Program of Guangzhou (2024A04J6413), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (24xkjc013).

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, 446–461. Springer.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*.

Chen, Z.; Lin, C.-H.; Liu, R.; Xiao, J.; and Dyer, E. 2024. Your contrastive learning problem is secretly a distribution alignment problem. *Advances in Neural Information Processing Systems*, 37: 91597–91617.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314): 2563–2609.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Cui, L.; Qi, X.; Wen, C.; Lei, N.; Li, X.; Zhang, M.; and Gu, X. 2019. Spherical optimal transportation. *Computer-Aided Design*, 115: 181–193.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Feng, C.-M.; Yu, K.; Liu, Y.; Khan, S.; and Zuo, W. 2023. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2704–2714.

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

Gu, X.; Yang, Y.; Zeng, W.; Sun, J.; and Xu, Z. 2022. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. *Advances in Neural Information Processing Systems*, 35: 14972–14985.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Huang, T.; Chu, J.; and Wei, F. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Kalantidis, Y.; Tolias, G.; et al. 2024. Label propagation for zero-shot classification with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23209–23218.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14162–14171.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Li, Y.; Su, Y.; Goodge, A.; Jia, K.; and Xu, X. 2025. Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. *arXiv preprint arXiv:2412.18303*.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

Luo, Y.-W.; and Ren, C.-X. 2023. Mot: Masked optimal transport for partial domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3531–3540. IEEE.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth*

Indian conference on computer vision, graphics & image processing, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Piran, Z.; Klein, M.; Thornton, J.; and Cuturi, M. 2024. Contrasting multiple representations with the multi-marginal matching gap. *arXiv preprint arXiv:2405.19532*.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15691–15701.

Qian, Q.; Xu, Y.; and Hu, J. 2023. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*, 36: 25461–25474.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Shi, L.; Fan, J.; and Yan, J. 2024. Ot-clip: Understanding and generalizing clip via optimal transport. In *Forty-first International Conference on Machine Learning*.

Shi, L.; Zhang, G.; Zhen, H.; Fan, J.; and Yan, J. 2023. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning*, 31408–31421. PMLR.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289.

Silva-Rodriguez, J.; Hajimiri, S.; Ben Ayed, I.; and Dolz, J. 2024. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23681–23690.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tan, Z.; Zheng, K.; and Huang, W. 2023. Otmatch: Improving semi-supervised learning with optimal transport. *arXiv preprint arXiv:2310.17455*.

Udandarao, V.; Gupta, A.; and Albanie, S. 2023. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2725–2736.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. 2022. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10899–10909.

Zanella, M.; Gérin, B.; and Ayed, I. 2024. Boosting vision-language models with transduction. *Advances in Neural Information Processing Systems*, 37: 62223–62256.

Zhang, C.; Ren, H.; and He, X. 2024. P² OT: Progressive Partial Optimal Transport for Deep Imbalanced Clustering. *arXiv preprint arXiv:2401.09266*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, 493–510. Springer.

Zhang, Y.; Zhu, W.; Tang, H.; Ma, Z.; Zhou, K.; and Zhang, L. 2024. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 28718–28728.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Y.; Ren, J.; Li, F.; Zabih, R.; and Lim, S. N. 2023. Test-time distribution normalization for contrastively learned visual-language models. *Advances in Neural Information Processing Systems*, 36: 47105–47123.

Zhu, X.; and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. *ProQuest number: information to all users*.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2605–2615.