Semantic Anchor Transport: Robust Test-Time Adaptation for Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large pre-trained vision-language models (VLMs), such as CLIP, have shown unprecedented zero-shot performance across a wide range of tasks. Nevertheless, these models may be unreliable under distributional shifts, as their performance is significantly degraded. In this work, we investigate how to efficiently utilize class text information to mitigate distribution drifts encountered by VLMs during inference. In particular, we propose generating pseudo-labels for the noisy test-time samples by aligning visual embeddings with reliable, text-based semantic anchors. Specifically, to maintain the regular structure of the dataset properly, we formulate the problem as a batch-wise label assignment, which is efficiently solved using Optimal Transport. Our method, Semantic Anchor Transport (SAT), utilizes such pseudo-labels as supervisory signals for test-time adaptation, yielding a principled cross-modal alignment solution. Moreover, SAT further leverages heterogeneous textual clues, with a multi-template distillation approach that replicates multi-view contrastive learning strategies in unsupervised representation learning without incurring additional computational complexity. Extensive experiments on multiple popular test-time adaptation benchmarks presenting diverse complexity empirically show that SAT achieves consistent performance gains over recent state-of-the-art methods while being computationally efficient.

1 Introduction

Large pre-trained vision-language models (VLMs), such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021), have emerged as a new paradigm shift in machine learning, revealing promising zero-shot transferability. Nevertheless, if the model is exposed to domain drifts at test time, its performance can be largely degraded Yu et al. (2023); Silva-Rodriguez et al. (2024). While a straightforward solution to bridge this gap involves fine-tuning the trained model using domain-specific labeled data Lai et al. (2023); Goyal et al. (2023), this strategy presents several limitations in real-world scenarios, which may hinder its scalability and deployment. First, adapting to new domains requires collecting labeled samples drawn from each distinct distribution. This might be impractical for specific domains and further hinder the real-time adaptation of the trained model for each input test sample. Furthermore, fine-tuning the model may undermine its desirable zero-shot capabilities Wortsman et al. (2022).

Test-Time Adaptation (TTA) presents a realistic and practical scenario for unsupervised domain adaptation, where a pre-trained model requires real-time adaptation to new data to address unknown distribution drifts without access to supervisory signals Wang et al. (2021); Iwasawa & Matsuo (2021); Yuan et al. (2023). Nevertheless, despite the recent rise of VLMs, and the popularity of TTA in more traditional deep models, such as CNNs and ViTs, the study of TTA in large pre-trained vision-language models remains less explored. Standard approaches, e.g., TENT Wang et al. (2021), which minimizes a Shannon entropy objective, have been adopted in CLIP test-time adaptation Shu et al. (2022). On the other hand, more recent strategies utilize pseudo-labels within the inference batch to guide model adaptation Osowiechi et al. (2024); Hakim et al. (2025); Maharana et al. (2025). In particular, the core idea of these later methods is to minimize a classification loss, typically the standard cross-entropy, between the generated pseudo-labels and model predictions, which guides the network updates over multiple iterations. Nonetheless, while this strategy is common in semi-supervised learning, where additional labeled instances are available for all categories,

applying it naively in unsupervised scenarios, such as test-time adaptation, can lead to degenerate solutions over multiple updates Iwasawa & Matsuo (2021), a problem we refer to as *error accumulation*.

This problem is particularly prevalent in the literature on test-time adaptation in VLMs. To illustrate this phenomenon, we conducted a simple experiment, the findings of which are depicted in Figure 1. Considering all corruptions in CIFAR10C, we identify samples misclassified by zero-shot CLIP and track their behavior across consecutive TTA steps. Specifically, we compute the average cosine similarity between the adapted visual embeddings and the corresponding class text embeddings of both the *wrong* (i.e., predicted) and *true* (i.e., expected) category. Ideally, a robust TTA method should reduce the similarity to the *wrong* class prototype, gradually correcting the initial misalignment. Nevertheless, as shown in Figure 1, existing methods tend to reinforce these mistakes, thereby increasing the similarity between the visual and text prototypes of the incorrect class, even when the prediction is wrong.

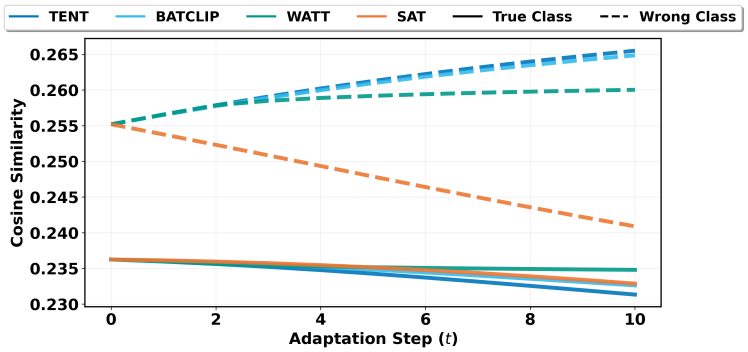

Figure 1: **Error Accumulation.** We track examples from all corruptions from CIFAR10C dataset with initial zero-shot CLIP predictions (at $t = 0$) that are misclassified. Baselines Wang et al. (2021); Maharana et al. (2025); Osowiechi et al. (2024) catastrophically reinforce this error: similarity to the 'Wrong Class' (dashed lines) increases while similarity to the 'True Class' (solid lines) decreases. In contrast, our method, SAT (orange), is the only one that provides a corrective signal, actively reducing similarity to the wrong class and breaking the cycle of error accumulation.

In light of the above limitation, we aim to provide a more robust TTA framework that can correct incorrect initial supervisory signals. To do so, we reframe TTA as a cross-modal alignment problem, where the goal is *to align the embeddings of test images to their corresponding fixed semantic anchors provided by the language modality*. This alignment can be viewed through the lens of clustering, where the unlabeled samples are grouped around class text prototypes. Hence, we can take into account the structure of the joint test batch to better correct sample-specific inconsistencies. Nevertheless, instead of jointly learning the cluster centroids during training, we use the fixed class prototypes derived from the text representations, which provide stable semantic guidance without requiring explicit supervision. A key challenge is the distribution mismatch between text prototypes and test image embeddings, since they originate from different modalities. Traditional clustering methods, which typically assume unimodal distributions, often struggle to bridge this gap, resulting in suboptimal performance. In contrast, we formulate label assignment as an Optimal Transport (OT) problem, yielding a global, cost-aware correspondence between visual embeddings and text prototypes that naturally model multi-modal distributions Lee et al. (2019). This perspective is more robust to outliers or noisy predictions, directly mitigating the error accumulation that undermines existing self-training in TTA.

The main contributions of this paper can be summarized as:

- We propose Semantic Anchor Transport (SAT). This novel framework casts the pseudo-labeling strategy in CLIP test-time adaptation as a batch-wise Optimal Transport assignment, which leverages the class text information available in vision-language models in the form of fixed robust cluster centroids without requiring further annotations.
- To solve the label assignment task, we resort to the Sinkhorn algorithm, as it can handle multi-modal distributions and compute label assignments efficiently.
- We introduce a multi-template knowledge distillation approach that leverages richer information derived from different text prompts to better guide adaptation without incurring significant computational overhead.
- Comprehensive experiments across multiple visual corruptions and domain shifts benchmarks demonstrate the superiority of SAT over recent state-of-the-art methods.

2 Related Work

Test-Time Adaptation (TTA) aims at adapting a pre-trained model to a stream of incoming unlabeled target domain data, processed in batches during testing Wang et al. (2021); Zhang et al. (2022); Choi et al. (2022); Niu et al. (2022b). Existing approaches in traditional unimodal models can be roughly categorized into: *i)* normalization-based methods, which leverage the statistics of the test data to adjust the BatchNorm statistics of the model Mirza et al. (2022); Schneider et al. (2020); *ii)* entropy-based approaches Wang et al. (2021); Niu et al. (2022a); Goyal et al. (2022), where the model is adapted optimizing the Shannon entropy of the predictions; and *iii)* pseudo-label strategies Liang et al. (2020), which employ the test-time generated labels for supervising the model. With the advent of VLMs, several works have attempted to accommodate some of these techniques to adapt pre-trained foundation models, notably CLIP Shu et al. (2022); Ma et al. (2023); Osowiechi et al. (2024); Hakim et al. (2025), which mainly differ on the different parameter groups updated during adaptation, i.e., prompt tuning Shu et al. (2022); Ma et al. (2023); Döbler et al. (2024) or layer-norm Hakim et al. (2025); Osowiechi et al. (2024) strategies. For example, Test-Time Prompt Tuning (TPT) Shu et al. (2022) optimizes input text prompts by minimizing the model's prediction entropy. In contrast, Vision-Text-Space Ensemble (VTE) Döbler et al. (2024) uses an ensemble of prompts as input to the text encoder. Nevertheless, keeping both the text and vision encoder frozen makes it difficult for the model to adapt to images with severe noise effectively. Recent VLM adaptation methods, such as WATT Osowiechi et al. (2024) and BATCLIP Maharana et al. (2025), which employ parameter-efficient updates of the encoders, have developed sophisticated bimodal heuristics to improve pseudo-label quality. WATT Osowiechi et al. (2024) generates them by combining both image-to-image and text-to-text similarity matrices. Taking this a step further, BATCLIP Maharana et al. (2025) first generates pseudo-labels via a standard argmax prediction, which are refined via additional losses. Nevertheless, both approaches rely on locally derived pseudo-labels, failing to explicitly optimize the sample-to-class assignment group cost, making them more sensitive to outliers or ambiguous samples. **Clustering for unsupervised representation learning.** Jointly adapting the parameters of a deep network while inferring the class assignments can be viewed as *clustering* and *unsupervised representation learning*. Thus, our work is closely related to recent literature on deep clustering-based approaches Asano et al. (2020); Caron et al. (2018; 2019); Huang et al. (2019); Yan et al. (2020); Yang et al. (2016); Jabi et al. (2019); Caron et al. (2020). In Caron et al. (2018), k-means assignments are leveraged as pseudo-labels to learn visual representations, a strategy later employed in Caron et al. (2019) to pre-train standard supervised deep models. Nevertheless, applying naive k-means risks collapsing to only a few imbalanced clusters, making it ineffective. A more principled approach consists in framing the label assignment task as an instance of the Optimal Transport problem Asano et al. (2020), whose global, batch-aware optimization and balancing constraints naturally prevent degenerate solutions. Furthermore, Caron *et al.* Caron et al. (2020) enforce consistency between cluster assignments across different image views or augmentations, avoiding expensive pairwise comparisons typically performed in contrastive learning. While our formulation shares similarities, key differences exist. In particular, all these methods, i.e., Caron et al. (2019); Asano et al. (2020); Caron et al. (2018) operate in unsupervised representation learning, where prototypes are learned from the data distribution. In contrast, we leverage the text representations as fixed semantic anchors. Note that text embeddings are accessible *for free* in VLMs, not incurring additional supervision, as the test image label remains unknown. A natural benefit from this strategy is that we do not need to resort to additional augmentation or "multi-views" strategies Caron et al. (2020), which introduce a computational burden. Instead, we treat individual text templates as "augmented views", enhancing the representation learning of the adapted model without incurring extra cost. In fact, these embeddings are computed only once *offline*, whereas image augmentation or "multi-views" strategies must create the additional images for each incoming batch, performing one forward pass per new augmentation at test time.

3 Background

3.1 Vision-language models

CLIP Radford et al. (2021) is a foundation vision-language model, trained via contrastive learning to produce visual representations from images $\boldsymbol{x}$ paired with their associated text descriptions T. To do so, CLIP consists of an image encoder $\boldsymbol{\theta}$ and a text encoder $\boldsymbol{\phi}$. This generates the corresponding vision

$\mathbf{z} \in \mathbb{R}^d$ and class text $\mathbf{t}_k \in \mathbb{R}^d$ embeddings (column vectors), which are typically projected into an ℓ_2-normalized shared embedding space. At inference, this learning paradigm enables zero-shot prediction. More concretely, for a given set of K classes, and an ensemble of M different templates, we can generate the set of available prompts as $\mathcal{T} = \{\{T_{mk}\}_{m=1}^{M}\}_{k=1}^{K}$, whose embedding for template m and class k is obtained as $\mathbf{t}_{mk} = \boldsymbol{\phi}(\text{“A PHOTO OF A } [\text{CLASS}_k]\text{”})$. Then, a popular strategy Radford et al. (2021); Gao et al. (2023); Wortsman et al. (2022) consists in obtaining a class zero-shot prototype, which is computed as $\mathbf{t}_k = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\phi}(T_{mk})$. Then, for a given test image $\boldsymbol{x}_i$, its zero-shot prediction, $\mathbf{p}_i = (p_{ik})_{1\leq k\leq K}$, can be obtained as:

$$p(y = k|\boldsymbol{x}_i) = \frac{\exp\left(\mathbf{z}_i^\top \mathbf{t}_k/\tau\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{z}_i^\top \mathbf{t}_j/\tau\right)}, \tag{1}$$

where τ is a temperature parameter learned during pre-training Radford et al. (2021), and $\mathbf{z}^\top \mathbf{t}$ indicates dot product.

3.2 Test-Time Adaptation using CLIP

Problem setting. We address the problem of adapting a pre-trained VLM at test time. In particular, given a model trained on the source domain $\mathcal{D}_S$, our goal is to adjust this model online to the new target domain $\mathcal{D}_T$, where only unlabeled test data $\{\boldsymbol{x}_i\}_{i=1}^{N}$ is available, and which is received as a stream of batches, from which predictions must be provided.

Vanilla entropy minimization. Inspired by the semi-supervised learning literature, a straightforward solution to leverage the predicted probability in eq:clip in TTA would be to adapt the model parameters based on an entropy minimization objective, similar to TENT Wang et al. (2021):

$$\mathcal{L}(p) = -\frac{1}{B_T}\sum_{i=1}^{B_T}\sum_{k=1}^{K} p(y = k|\boldsymbol{x}_i) \log p(y = k|\boldsymbol{x}_i), \tag{2}$$

with B_T denoting the size of the test batch. However, relying on the Shannon entropy Wang et al. (2021) in the fully unsupervised case poses a significant risk, as it may potentially result in a degenerate solution, i.e., Eq. (2) might be trivially minimized by assigning all data points to a single, arbitrary label.

4 SAT: TTA as Cross-Modal Alignment

The proposed approach aims to overcome the limitations in eq:tent objective by finding reliable pseudo-labels in a batch of test-time samples without supervision. To achieve this, we propose **SAT**. We formulate TTA as a cross-modal alignment problem, where the core task is to generate a robust supervisory signal for each unlabeled test batch. This novel TTA method leverages the text embeddings generated by the frozen CLIP text encoder more effectively, serving as strong class descriptors to yield label assignments for test samples, which we propose propagating through Optimal Transport. Note that our motivation drastically differs from the existing self-supervised representation learning literature that also resorts to pseudolabels Caron et al. (2018), where cluster centroids are iteratively updated based on the prior label assignments. Furthermore, by solving a globally optimal matching problem, SAT is more robust to noisy predictions than existing TTA methods Osowiechi et al. (2024); Maharana et al. (2025). Below, we elaborate on the details of our method, which is illustrated in Figure 2.

To overcome the limitation of naively minimizing eq:tent in the fully unlabeled scenario, TTA literature typically encodes the model predictions as posterior distributions $q(y|\boldsymbol{x}_i)$, which results in:

$$\mathcal{L}(p, q) = -\frac{1}{B_T}\sum_{i=1}^{B_T}\sum_{k=1}^{K} q(y = k|\boldsymbol{x}_i) \log p(y = k|\boldsymbol{x}_i). \tag{3}$$

Thus, the adaptation is driven by the pseudo cross-entropy in Eq. (3), where a significant challenge lies in constructing a high-quality target distribution, or pseudo-label $q_{ik} = q(y = k|\boldsymbol{x}_i)$, without access to ground

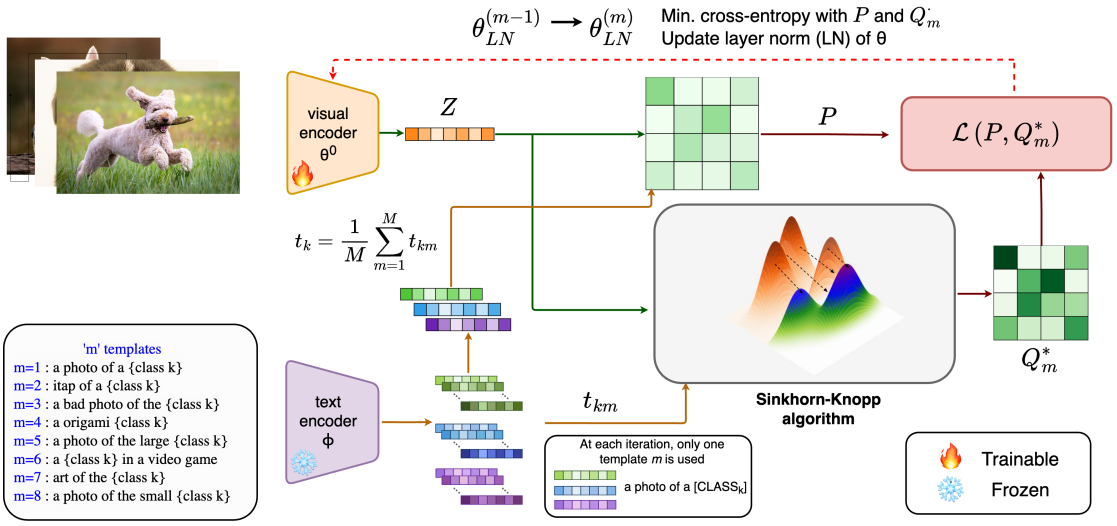

Figure 2: **SAT** leverages Optimal Transport (with the Sinkhorn algorithm) to yield soft-codes Q_m^* (eq:individual-codes). Then, it minimizes the pseudo cross-entropy between Q_m^* and the CLIP predictions P as unsupervised loss during test-time adaptation (eq:posterior). Note that, at each test batch, our model runs for m iterations, each using a different text template m to obtain the pseudo-codes Q_m^*.

truth labels. Indeed, as illustrated in Figure 1, prior TTA methods tend to amplify prediction errors, leading to a state of *error accumulation*. Our central thesis is that we can break this loop by exploiting the natural alignment between visual features and their text-based semantic anchors in the batch.

4.1 Batch-Aware Cross-modal Alignment

In the following, we detail how to encode robust posteriors into eq:posterior by exploiting multi-modal capabilities of CLIP and the test-time batch data distribution. Considering the probability is obtained as in eq:clip, the objective in eq:posterior can be expressed as:

$$\mathcal{L}(p, q) = -\frac{1}{B_T}\sum_{i=1}^{B_T}\left[\frac{1}{\tau}\mathbf{z}_i^\top \mathbf{T}\mathbf{q}_i - \log\sum_{k=1}^{K}\exp(\frac{\mathbf{z}_i^\top \mathbf{t}_k}{\tau})\right], \tag{4}$$

where $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_K]$ represents the matrix containing the class text prototypes, and the logarithmic term in the right side does not depend on $\mathbf{q}$. If we now express eq:midstep in its matrix form over all the test images in the batch (with $\mathbf{Z} \in \mathbb{R}^{d \times B_T}$ the corresponding embeddings), and let $\mathbf{Q}$ be a matrix with columns $\mathbf{q}_i$, we have the following objective:

$$\max_{\mathbf{Q} \in \mathcal{Q}} tr(\mathbf{Q}^\top \mathbf{T}^\top \mathbf{Z}). \tag{5}$$

which seeks the optimal alignment between the batch's visual features $\mathbf{Z} \in \mathbb{R}^{d \times B_T}$ and a set of text-based *"semantic anchors"* $\mathbf{T} \in \mathbb{R}^{d \times K}$. Following Cuturi (2013), we enforce the matrix $\mathbf{Q}$ to be an element of the *transportation polytope*:

$$\mathcal{Q} := \{\mathbf{Q} \in \mathbb{R}_+^{K \times B_T} \mid \mathbf{Q}\mathbf{1}_{B_T} = \frac{1}{K}\mathbf{1}_K, \mathbf{Q}^\top \mathbf{1}_K = \frac{1}{B_T}\mathbf{1}_{B_T}\}, \tag{6}$$

with $\mathbf{1}_K$ and $\mathbf{1}_N$ denoting the vectors of ones in dimension K and N, respectively. The constraints on $\mathcal{Q}$ enforce that on average each prototype is selected at least $\frac{N}{K}$ times, encouraging $\mathbf{Q}$ to be a matrix with uniform marginals in both rows and columns. *Note that such constraints 'break' the direct dependency between the original prediction p in eq:tent and the posterior q in eq:midstep, since the latter considers the whole batch distribution and its marginal properties for the assignment.* This results in more robust supervisory signals, later demonstrated empirically (Figure 3).

Finding optimal Q. As the objective function in Eq. (5) is linear, and the constraints defining $\mathcal{Q}$ are also linear, this is a linear program. Nevertheless, directly optimizing the above learning objective might be time-consuming, particularly as the number of data points and classes increases. To address this issue and facilitate faster optimization, we apply the Sinkhorn algorithm Cuturi (2013), which incorporates an entropic constraint that enforces a simple structure on the optimal regularized transport. Hence, the optimization problem becomes:

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{T}^\top \mathbf{Z}) + \varepsilon \mathcal{H}(\mathbf{Q}), \tag{7}$$

where $\mathcal{H}(\mathbf{Q})$ denotes the Shannon entropy function $\mathcal{H}(\mathbf{Q}) = -\sum_{ij} \mathbf{Q}_{ij} \log \mathbf{Q}_{ij}$ and ε controls its weight. Furthermore, to align with TTA assumptions, we work on batches by restricting the transportation polytope to the current batch Caron et al. (2020), in contrast to other works that employ the full dataset Asano et al. (2020). Thus, the dimensionality of the $\mathbf{Q}$ matrix becomes $K \times B_T$, where B_T denotes the batch size. Now, the soft assignment matrix $\mathbf{Q}^*$ is the solution of the problem in (7) over the set $\mathcal{Q}$, which can be efficiently optimized with a few iterations:

$$\mathbf{Q}^* = \text{Diag}(\mathbf{u}^{(t)}) \exp\left(\frac{\mathbf{T}^\top \mathbf{Z}}{\lambda}\right) \text{Diag}(\mathbf{v}^{(t)}), \tag{8}$$

with $\mathbf{u}$ and $\mathbf{v}$ representing renormalization vectors in $\mathbb{R}^K$ and $\mathbb{R}^{B_T}$ respectively, and t the iteration.

Benefits of the proposed strategy. The proposed Optimal Transport (OT) based formulation provides: **i) Global consistency**, since OT is *batch-aware*. The assignment for any single image is dependent on the features of all other images in the batch. The solver finds the best overall configuration, making it inherently robust to individual outliers. **ii) Intrinsic regularization** given by the polytope constraints (Eq. 6), which enforce that, on average, all classes are represented in the pseudo-labels. This structurally prevents the model from collapsing to a single class, yielding robustness.

4.2 Semantic alignment with multiple templates

Recent literature Osowiechi et al. (2024) has pointed out the limitations of using the average text prototypes, and has instead suggested to resort to multiple category embeddings, each obtained through the different text templates in $\mathcal{T}$. Thus, to create a richer and more robust supervisory signal, we leverage the ensemble of M diverse text templates typically used in TTA, each providing a unique "semantic view", $\mathbf{T}_m = [\mathbf{t}_{1m}, ..., \mathbf{t}_{Km}]$, where $\mathbf{T}_m$ denotes the matrix containing the class text prototypes obtained from the m-th template. Compared to existing TTA approaches, we introduce a more sophisticated knowledge distillation strategy that disentangles the goals of assignment and generalization.

① **For Assignment, we need specificity.** We compute a separate, "clean" transport plan $\mathbf{Q}^*_m$ for each specific semantic view $\mathbf{T}_m$. This provides an unambiguous matching based on a single, clear context (e.g., aligning to "a sketch of a class" vs. "a photo of a class"). Concretely, we modify the solution presented in eq:codes to be specialized for specific class template[1], where the codes for each template m are:

$$\mathbf{Q}^*_m = \text{Diag}(\mathbf{u}^{(t)}_m) \exp\left(\frac{\mathbf{T}^\top_m \mathbf{Z}}{\lambda}\right) \text{Diag}(\mathbf{v}^{(t)}_m). \tag{9}$$

The renormalization vectors in our setting are computed trough a small number of matrix multiplications using the iterative Sinkhorn-Knopp algorithm Knight (2008), where in each iteration $\mathbf{u}^{(t)}_m = \mathbf{u}_m / ((\exp(\mathbf{T}^\top_m \mathbf{Z}/\lambda)\mathbf{v}^{t-1}_m)$ and $\mathbf{v}^{(t)}_m = \mathbf{v}_m / ((\exp(\mathbf{T}^\top_m \mathbf{Z}/\lambda)\mathbf{u}^{t}_m)$, with $\mathbf{v}^0_m = \mathbf{1}$.

② **For Generalization, we need consistency.** To ensure generalization, model predictions must remain consistent across semantic views, preventing the visual encoder from overfitting to any single template, or "view". In the proposed adaptation strategy, the model's prediction $p(y = k|\boldsymbol{x}_i)$ in Eq. 1 is always computed using a single classification head, defined by the *averaged* text class prototypes $\mathbf{T}$. This design enforces stability across templates, ensuring that the pseudo-supervision $q(y = k|\boldsymbol{x}_i)$ is always applied against a

[1]The text templates are the same as in WATT Osowiechi et al. (2024) – see Appendix A.

consistent prediction function, preventing the encoder from overfitting to template-specific fluctuations and promoting robust generalization.

Coupling both together. First, to optimize the objective in eq:posterior, we employ the soft label assignments from (9), as they yield superior performance compared to their hard counterpart in other problems Caron et al. (2020). Now, to distill the knowledge of the more diverse, and richer representations derived from multiple text predictions, the cross-entropy in eq:posterior is optimized by iterating through each of the M views, which can be formally defined for each image i and template m as $\ell(\mathbf{p}_i, \mathbf{q}_{im}^*) = -\mathbf{q}_{im}^* \log \mathbf{p}_i$, where $\mathbf{p}_i$ is obtained with the average text template. Thus, for a given test image i in a batch, after each update produced by the m-th text template, the visual encoder produces novel visual embeddings $\mathbf{z}_i$, which are then used to generate new predictions $\mathbf{p}_i$ for i.

4.3 Whole learning strategy

The complete Semantic Anchor Transport (SAT) algorithm (Alg. 1) consists in learning a label assignment matrix $\mathbf{Q}_m$ per text template by solving the optimization problem presented in Eq. (7), and then updating the affine parameters of the visual encoder's Layer Normalization layers, such as Wang et al. (2021); Osowiechi et al. (2024); Hakim et al. (2025). This is done iteratively by alternating two steps across the batches and text templates: *i)* with the learnable parameters fixed, we compute the visual features $\mathbf{Z} \in \mathbb{R}^{d \times B_s}$ of the test batch samples, and find $\mathbf{Q}_m^*$ through Eq. (9) by iterating the updates on $\mathbf{u}_m$ and $\mathbf{v}_m$; and *ii)* given label assignments $\mathbf{Q}_m^*$, and the softmax predictions for the test samples obtained with each average class prototype $\mathbf{t}_k$, the set of learnable parameters is updated by minimizing Eq. (3) w.r.t. the layer norm parameters, where stochastic gradient descent is applied over the whole batch. This process is repeated M times. Finally, the class predictions on the batch are inferred with the updated model.

Algorithm 1 Semantic Anchor Transport for one test batch.

Input: Test batch $\{\boldsymbol{x}_i\}_{i=1}^{B_T}$, set of M semantic anchor matrices $\{\mathbf{T}_m\}_{m=1}^{M}$, visual encoder $\boldsymbol{\theta}$.
// $\mathbf{T}_m$ (and thus $\mathbf{T}$) are pre-computed offline from a systematically generated set of text templates.
$\mathbf{T} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \mathbf{T}_m$
// — **Adaptation Phase** —
for each template m in a random permutation of $\{1, \ldots, M\}$ **do**
    // Step 1: Align - Compute soft assignments that maximize cross-modal alignment for the m-*th* "view"
    $\mathbf{Z} \leftarrow [\boldsymbol{\theta}(\boldsymbol{x}_1), \ldots, \boldsymbol{\theta}(\boldsymbol{x}_{B_T})]$
    $\mathbf{Q}_m^* \leftarrow \text{SolveOT}(\mathbf{Z}, \mathbf{T}_m)$ Eq. 9
    // Step 2: Adapt - Distill knowledge into a general representation
    $\mathbf{P} \leftarrow \text{Predict}(\mathbf{Z}, \mathbf{T})$ Eq. 1
    $\mathcal{L} \leftarrow \text{CrossEntropy}(\mathbf{P}, \mathbf{Q}_m^*)$
    Update LayerNorm parameters of $\boldsymbol{\theta}$ using $\nabla_{\boldsymbol{\theta}} \mathcal{L}$.
end for
// — **Inference Phase** —
$\mathbf{Z}_{\text{adapted}} \leftarrow [\boldsymbol{\theta}_{\text{adapted}}(\boldsymbol{x}_1), \ldots, \boldsymbol{\theta}_{\text{adapted}}(\boldsymbol{x}_{B_s})]$
$\mathbf{P}_{\text{final}} \leftarrow \text{Predict}(\mathbf{Z}_{\text{adapted}}, \mathbf{T})$ Eq. 1
Return $\mathbf{P}_{\text{final}}$

5 Experiments

5.1 Setting

Datasets. We evaluate on standard TTA benchmarks Maharana et al. (2025); Osowiechi et al. (2024); Hakim et al. (2025), covering two primary types of domain shift: ***i)* Visual corruptions**: CIFAR-10C, CIFAR-100C, ImageNet-C, and Tiny-ImageNet-C Hendrycks & Dietterich (2019); and ***ii)* Domain generalization**: PACS Li et al. (2017), VLCS Fang et al. (2013), OfficeHome Venkateswara et al. (2017), and VisDA-C Peng et al.

(2018). We also use the original clean benchmarks (e.g., CIFAR-10 Krizhevsky (2012), CIFAR-10.1 Recht et al. (2018), and Tiny-ImageNet Wu et al. (2017)) for ablation. Full dataset details are presented in Appendix A.

Baselines: We resort to relevant CLIP-based TTA parameter-efficient adaptation approaches, i.e., modify a set of learnable parameters, including TENT Wang et al. (2021), SAR Niu et al. (2022b), VTE Döbler et al. (2024), TPT Shu et al. (2022), CLIPArTT Hakim et al. (2025), WATT Osowiechi et al. (2024), and BATCLIP Maharana et al. (2025), where Hakim et al. (2025); Osowiechi et al. (2024); Maharana et al. (2025) represent the state-of-the-art.

Architectures: Following prior work, we use Vision Transformer (ViT) backbones from CLIP Radford et al. (2021). Our main experiments use ViT-B/32 as in Hakim et al. (2025); Osowiechi et al. (2024), and we demonstrate generalization with additional evaluations on larger ViT-B/16 and L/14 backbones, particularly CLIP's and SigLIP's Zhai et al. (2023).

Implementation details: Following prior work Hakim et al. (2025); Osowiechi et al. (2024), we use a batch size of 128 for all experiments (see Appendix B.4 for a sensitivity analysis on batch size). All reported results are the average of three runs using different random seeds. We follow the officially recommended hyperparameters for all baseline methods. Specifically, TENT Wang et al. (2021) and CLIPArTT Hakim et al. (2025) are run for 10 adaptation steps per batch. For WATT Osowiechi et al. (2024), we use their suggested $L = 2$, $M = 5$, and 8 templates, totaling 80 iterations per batch. For BATCLIP Maharana et al. (2025), we adapt both the text and visual encoders. For all other baselines, we use the learning rates specified in their respective papers for each dataset. We set the learning rate for our method to 10^{-4} for all datasets, except for ImageNet-C, where we use 5×10^{-5} as in Maharana et al. (2025). Our formulation introduces only two new hyperparameters: the entropic constraint weight, ϵ from eq:main, and the number of Sinkhorn iterations, T. These are fixed across all datasets and experiments. We set $\epsilon = 0.7$ (see Appendix B.3) and $T = 3$. We found $T = 3$ to be sufficient for convergence, as in prior work Caron et al. (2018).

5.2 Main results

Performance under common visual corruptions. We present a unified evaluation across four standard benchmarks in Table 1, which shows the superiority of SAT in adapting CLIP in the presence of common corruptions against a comprehensive suite of recent TTA methods. Compared to vanilla CLIP, our model brings performance gains of 17.8% (CIFAR-10C) and 17.9% (CIFAR-100C) without requiring additional supervision. These performance gains are similar when compared to other popular TTA methods, e.g., TENT or TPT, with differences ranging from 10% to 21%. While this gap is reduced compared to recent approaches, such as WATT, CLIPArTT, and BATCLIP, the differences remain significant. SAT outperforms the second-best competitor, i.e., WATT, by up to 3.2% in CIFAR-10C (e.g., 5.1% in *pixelate*) and 1.8% in the more challenging scenario of CIFAR-100C (which contains $\times 10$ classes). Additionally, it achieves significant gains over recent baselines in specific corruptions, e.g., 19.8% compared to BATCLIP in *pixelate* (CIFAR-10C) or 16.3% in *Glass Blur* (CIFAR-100C). Also, compared to CLIPArTT, SAT shows gains of nearly 10.6% on *pixelate* (CIFAR-100C) and 8.1% on *Gaussian Noise* (CIFAR-100C).

What if the number of classes increases? The gap is even more pronounced on the more difficult Tiny-ImageNet-C testbed, where our 34.9% mean accuracy is 5.6% higher than that of the strongest baseline, BATCLIP. Finally, on the highly challenging 1000-class ImageNet-C benchmark, where many methods show limited gains, SAT again proves its robustness. Specifically, achieves the highest mean accuracy, 26.0%, a promising +5.5% gain over the CLIP baseline, outperforming all other TTA baselines.

Performance under additional shifts: Table 2 (VisDA-C) reports the results of *simulated* (3D) and *video* (YT) shifts, which showcase the superiority of SAT. More concretely, the differences compared to the second-competitor SoTA method (WATT) in these two datasets are 5.4% and 0.8%. Regarding other popular baselines, per-dataset differences can go up to 11%, e.g., 11.4% compared to TPT in the 3D domain. It is worth noting also the significant performance drop of the very recent BATCLIP on this benchmark, which underscores the instability of methods that rely too heavily on their own initial predictions in the face of large domain gaps. Furthermore, Table 2 results on *texture and style shifts* show that the trend of superior performance is observed in other datasets (i.e., PACS and OfficeHome), whereas performance is degraded in only one out of the five datasets, i.e., VLCS.

Table 1: **Performance in visual corruptions benchmarks.** Results using CLIP ViT-B/32. Best method in **bold**, second best underlined.

Method	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixelate	JPEG	Mean ↑
CIFAR-10C																
CLIP [ICLR'21]	35.27	39.67	42.61	69.76	42.40	63.97	69.83	71.78	72.86	67.04	81.87	64.37	60.83	50.53	55.48	59.22
TENT [ICLR'21]	41.27	47.20	48.58	77.12	52.65	71.25	76.20	78.29	79.84	77.39	87.78	79.47	70.00	63.74	62.64	67.56
SAR [ICLR'22]	47.58	50.39	47.19	71.65	49.34	70.27	72.63	71.66	72.82	69.48	82.34	70.54	60.98	48.07	58.48	62.89
VTE [ECCVw'24]	44.40	47.70	42.90	64.90	45.00	66.70	67.00	67.40	64.50	65.30	74.90	53.60	61.20	42.60	50.80	57.26
TPT [NeurIPS'22]	33.90	38.20	37.66	67.83	38.81	63.39	68.95	70.16	72.39	64.31	81.30	62.26	56.43	42.80	53.67	56.80
WATT [NeurIPS'24]	63.84	65.28	58.64	78.94	65.12	77.81	79.32	79.79	80.54	78.53	87.11	81.20	72.66	71.11	67.36	73.82
CLIPArTT [WACV'25]	59.90	62.77	56.02	76.74	61.77	76.01	77.40	77.29	79.20	75.74	86.59	77.82	70.20	66.52	63.51	71.17
BATCLIP [ICCV'25]	50.89	56.01	54.35	76.17	56.11	74.71	76.01	77.74	79.33	75.87	86.46	78.65	68.76	56.41	61.79	68.62
SAT	**64.85**	**67.34**	**62.27**	**82.09**	**68.07**	**81.30**	**83.13**	**83.71**	**83.40**	**82.56**	**89.90**	**84.86**	**76.08**	**76.25**	**70.03**	**77.06**
CIFAR-100C																
CLIP [ICLR'21]	14.80	16.03	13.85	36.74	14.19	36.14	40.24	38.95	40.56	38.00	48.18	29.53	26.33	21.98	25.91	29.43
TENT [ICLR'21]	14.38	17.34	10.03	49.05	3.71	46.62	51.84	46.71	44.90	47.31	60.58	45.90	33.09	26.47	29.89	35.19
SAR [ICLR'22]	22.82	25.10	18.68	44.51	21.78	43.04	47.04	46.75	47.34	44.62	57.00	42.17	31.51	25.09	30.83	36.55
VTE [ECCVw'24]	10.00	10.30	13.30	36.10	20.40	37.90	39.80	42.20	40.80	36.60	45.50	29.20	30.80	17.00	20.70	28.71
TPT [NeurIPS'22]	14.03	15.25	13.01	37.60	16.41	37.52	42.99	42.35	43.31	38.81	50.23	28.09	28.12	20.43	28.82	30.46
WATT [NeurIPS'24]	32.07	34.36	30.33	52.99	32.15	50.53	55.30	52.77	53.79	51.49	63.57	52.76	40.90	40.97	39.59	45.57
CLIPArTT [WACV'25]	25.32	27.90	25.62	49.88	27.89	47.93	52.70	49.72	49.63	48.77	61.27	48.55	37.45	33.88	36.07	41.51
BATCLIP [ICCV'25]	17.25	19.76	18.98	42.20	19.00	40.81	46.59	41.34	40.14	41.56	53.85	34.07	31.38	25.51	28.96	33.43
SAT	**33.43**	**35.60**	**30.94**	**53.87**	**35.26**	**52.77**	**56.71**	**54.30**	**54.92**	**53.57**	**64.43**	**55.01**	**43.79**	**44.51**	**40.83**	**47.33**
Tiny-ImageNet-C																
CLIP [ICLR'21]	7.08	9.41	3.44	21.71	9.12	34.52	27.44	32.51	36.33	25.94	40.15	1.81	30.40	22.78	29.59	22.15
TENT [ICLR'21]	8.01	10.04	4.18	24.53	10.09	36.94	29.48	32.20	35.72	27.46	39.79	2.24	31.92	24.79	30.93	23.22
SAR [ICLR'22]	9.09	10.94	3.65	5.50	1.68	14.02	12.08	20.72	24.62	8.37	32.35	0.71	15.32	12.39	25.35	13.12
VTE [ECCVw'24]	18.63	20.34	4.71	9.62	2.21	30.37	21.68	38.84	40.27	17.41	41.22	0.63	31.64	25.33	37.79	22.71
TPT [NeurIPS'22]	9.29	11.70	4.85	27.56	11.03	38.97	34.29	34.45	37.13	28.89	43.31	3.15	33.88	27.70	33.60	25.32
WATT [NeurIPS'24]	13.02	15.94	6.90	29.91	14.01	41.26	33.96	37.76	39.65	32.13	46.93	3.53	35.01	31.55	36.46	27.87
CLIPArTT [WACV'25]	14.44	17.44	10.37	31.46	15.84	41.34	35.06	36.86	38.20	33.44	46.43	6.24	33.89	34.85	37.32	28.88
BATCLIP [ICCV'25]	11.96	15.48	10.05	31.89	14.76	43.31	39.07	39.02	39.05	31.91	49.06	5.65	32.79	36.63	39.12	29.32
SAT	**21.40**	**24.90**	**17.34**	**35.39**	**21.16**	**46.26**	**40.93**	**42.32**	**44.97**	**38.60**	**53.10**	**11.88**	**40.73**	**41.84**	**42.86**	**34.91**
ImageNet-C																
CLIP [ICLR'21]	11.30	11.58	12.28	20.88	8.92	19.78	17.62	19.92	23.48	25.90	47.34	15.48	17.02	28.00	27.60	20.47
TENT [ICLR'21]	8.00	7.20	9.20	23.04	10.84	22.86	19.04	21.24	23.86	26.54	48.54	18.32	17.90	30.32	29.66	21.10
SAR [ICLR'22]	13.07	15.69	13.92	22.74	14.53	23.41	19.49	22.65	24.89	29.47	48.39	18.88	19.61	31.68	29.07	23.17
VTE [ECCVw'24]	7.12	10.24	9.18	**27.31**	10.27	**26.42**	**27.36**	24.28	26.15	31.22	49.37	13.09	14.18	32.44	31.33	22.66
TPT [NeurIPS'22]	8.94	7.22	7.55	20.47	9.13	21.78	23.92	24.61	21.54	24.98	40.37	15.22	13.18	30.74	24.63	20.01
WATT [NeurIPS'24]	7.76	7.06	8.94	24.16	12.46	25.00	21.52	21.58	24.16	26.62	**49.74**	**21.14**	19.90	32.70	32.16	22.33
CLIPArTT [WACV'25]	14.74	15.10	15.30	10.82	9.02	13.82	12.30	16.96	22.52	19.90	41.78	0.26	12.84	22.80	31.94	18.15
BATCLIP [ICCV'25]	14.84	15.10	15.52	24.42	17.18	25.64	23.08	25.06	25.58	**31.08**	49.66	18.44	22.20	**33.42**	**33.02**	24.95
SAT	**18.98**	**24.54**	**25.04**	25.46	**21.98**	26.34	24.46	**27.78**	**28.74**	29.32	41.98	17.90	**25.70**	26.32	25.68	**26.01**

All considered, SAT emerges as a powerful and reliable TTA approach, where average improvement gains compared to recent methods are 2.2% (BATCLIP), 1.2% (WATT), and 1.9% (CLIPArTT), showcasing its versatility.

5.3 In-depth studies

Impact of each component. In Fig. 3, we empirically validate the benefits of each component in SAT (each configuration is detailed in Appendix B.1). The 'Training-Free OT' (using

Table 2: **Performance comparison under domain shifts**. Results using CLIP ViT-B/32. Best method in **bold**, second best underlined.

Method	PACS	O-Home	VLCS	VisDA-3D	VisDA-YT	Mean
	(texture / style)			*(simulated / video)*		
CLIP	93.65	77.53	80.16	84.43	84.45	84.04
TENT	93.81	77.68	80.27	84.86	84.68	84.26
TPT	93.23	77.20	74.57	79.35	83.57	81.58
CLIPArTT	93.95	77.56	80.06	85.09	84.40	84.21
WATT	94.80	78.83	**81.14**	85.36	84.69	84.96
BATCLIP	94.52	78.90	80.78	81.97	83.60	83.95
SAT	**95.94**	**80.15**	78.33	**90.73**	**85.44**	**86.12**

Table 3: **TTA results using larger CLIP ViT backbones.** SAT consistently outperforms baselines across different architectures.

Method	a) ViT-B/16				b) ViT-L/14			
	C10-C	C100-C	TIN-C	IN-C	C10-C	C100-C	TIN-C	IN-C
CLIP	60.15	32.01	20.92	20.89	76.04	44.59	34.98	32.05
TENT	68.00	37.90	29.78	22.79	79.18	50.14	40.28	33.09
TPT	59.75	33.73	26.96	22.35	75.01	47.58	41.07	30.67
CLIPArTT	73.22	40.08	32.90	23.47	78.06	52.52	42.98	34.13
WATT	76.22	48.95	31.66	24.38	80.06	54.34	43.28	36.30
BATCLIP	73.52	38.85	29.30	27.16	83.79	48.84	35.72	37.07
SAT	**80.11**	**51.24**	**37.69**	**28.35**	**86.35**	**62.21**	**50.36**	**38.82**

OT for assignments but not updating the model) already improves over zero-shot CLIP, demonstrating the power of our batch-aware assignment. 'Average Template' (using an averaged text prototype) provides a further boost. Finally, our full method, which utilizes 'Multi-Templates' knowledge distillation, yields the largest gains, demonstrating that all components are necessary. The benefit of leveraging multiple textual semantic views is also supported in Appendix B.5 experiments.

Do observations hold across backbones? Table 3 reports the performance across different datasets when larger CLIP pre-trained models are employed.

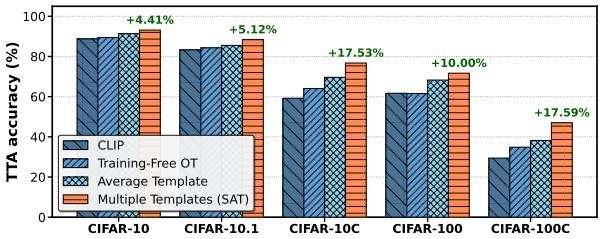

Figure 3: **Ablation on each component**. Results of adding each SAT element. **Green**: Gains vs Zero-shot CLIP.

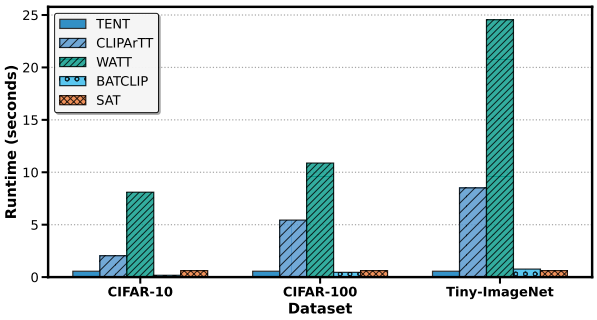

Figure 4: **Inference Runtime**. Seconds per batch ($N = 128$) on NVIDIA A6000 for ViT-B/32 methods.

Generalization across VLMs. SAT's principles are also model-agnostic beyond the CLIP pre-training framework. As shown in Table 4, SAT consistently delivers state-of-the-art performance when applied to a modern SigLIP backbone. These results demonstrate the model-agnostic nature of SAT, as its superiority remains consistent across different backbones.

Table 4: **Performance using SigLIP** Zhai et al. (2023).

Method	C-10C	C-100C	Tiny-IN	IN-C
SigLIP	59.04	34.76	22.00	26.49
BATCLIP	67.45	39.26	24.72	30.40
SAT	**76.20**	**49.72**	**26.81**	**31.92**

The gains are even more pronounced on this architecture, with a 15.0% improvement on CIFAR-100C and 5.4% on ImageNet-C over zero-shot SigLIP, proving the general applicability of our cross-modal alignment paradigm.

Computational analysis. Fig. 4 depicts the running time required for relevant baselines and the proposed SAT across several datasets. These results expose that different TTA methods, particularly SoTA, substantially differ in total runtimes. In particular, recent CLIPArTT and WATT constantly increase the required runtime with the number of classes, driven by their iterative nature. BATCLIP is also highly efficient, yet its runtime still shows a slight increase as the number of classes grows. In contrast, SAT avoids this overhead by distilling this information during the adaptation stage and computing multiple text embeddings *off-line* only once. It is worth noting that the Sinkhorn algorithm used for generating pseudo-codes is highly efficient, accounting for only nearly 1% of the total runtime.

6 Conclusions

We have presented Semantic Anchor Transport (SAT), a novel approach for the test-time adaptation of vision-language models. SAT reformulates TTA as a principled cross-modal alignment problem. It generates robust, batch-aware pseudo-labels by aligning visual embeddings to fixed text-based semantic anchors using Optimal Transport. This global assignment strategy fundamentally mitigates the error accumulation demonstrated by other TTA methods. Furthermore, SAT employs a sophisticated multi-template distillation strategy to harness diverse textual clues, enhancing robustness without incurring significant computational overhead. Extensive experiments demonstrate that SAT achieves state-of-the-art performance on TTA across multiple visual domain shift benchmarks and multi-modal backbones.

References

YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2959–2968, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems (NeurIPS)*, 33:9912–9924, 2020.

Sungha Choi, Seunghan Yang, Seokeon Choi, and Sungrack Yun. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *European Conference on Computer Vision (ECCV)*, pp. 440–458, 2022.

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

Mario Döbler, Robert A Marsden, Tobias Raichle, and Bin Yang. A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models. In *European Conference on Computer Vision (ECCV) Workshops*, 2024.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1657–1664, 2013.

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 (IJCV)*, 2023.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:6204–6218, 2022.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19338–19347, 2023.

Gustavo Adolfo Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. CLIPArTT: Light-weight adaptation of CLIP to new domains at test time. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and surface noise. In *International Conference on Learning Representations (ICLR)*, 2019.

Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning (ICML)*, pp. 2849–2858, 2019.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Mohammed Jabi, Marco Pedersoli, Amar Mitiche, and Ismail Ben Ayed. Deep clustering: On the link between discriminative models and k-means. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 1887–1896, 2019.

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 *International conference on machine learning (ICML)*, pp. 4904–4916, 2021.

Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Preprint*, 2012.

Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16155–16165, 2023.

John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. doi: 10.48550/arXiv.1906.11768.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5542–5550, 2017.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning (ICML)*, pp. 6028–6039, 2020.

Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Sarthak Kumar Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. BATCLIP: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14765–14775, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning (ICML)*, pp. 16888–16905, 2022a.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2022b.

David Osowiechi, Mehrdad Noori, Gustavo Adolfo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. WATT: Weight average test-time adaption of CLIP. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to cifar-10? In *Preprint*, volume abs/1806.00451, 2018.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems (NeurIPS)*, 33:11539–11551, 2020.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14274–14289, 2022.

Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. 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 (CVPR)*, pp. 23681–23690, 2024.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. In *Preprint*, 2017. URL `https://api.semanticscholar.org/CorpusID:212697711`.

Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6509–6518, 2020.

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5147–5156, 2016.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10899–10909, 2023.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems (NeurIPS)*, 35:38629–38642, 2022.