
DOTA: DistributiOnal Test-time Adaptation of Vision-Language Models

**Zongbo Han^{1,2*}, Jialong Yang³, Guangyu Wang^{1,2}, Junfan Li⁴, Qianli Xu⁵,
Mike Zheng Shou^{6*}, Changqing Zhang^{3*}**

School of Information and Communication Engineering, Beijing University of Posts
and Telecommunications¹

State Key Laboratory of Networking and Switching Technology, Beijing University of
Posts and Telecommunications²

College of Intelligence and Computing, Tianjin University³

School of Computer Science and Technology, Harbin Institute of Technology Shenzhen⁴

Institute for Infocomm Research (I²R), A*STAR Research Entities (ARES)⁵

Show Lab, National University of Singapore⁶

Abstract

Vision-language foundation models (VLMs), such as CLIP, exhibit remarkable performance across a wide range of tasks. However, deploying these models can be unreliable when significant distribution gaps exist between training and test data, while fine-tuning for diverse scenarios is often costly. This creates a need for methods that can efficiently adapt to new data at test time without expensive retraining. Cache-based test-time adapters serve this purpose by storing representative test samples to guide subsequent classifications. Yet, these methods typically employ naive cache management with limited capacity, leading to severe catastrophic forgetting when samples are inevitably dropped during updates. In this paper, we propose DOTA (DistributiOnal Test-time Adaptation), a simple yet effective method addressing this limitation. Crucially, instead of merely memorizing individual test samples, DOTA continuously estimates the underlying distribution of the test data stream. Test-time posterior probabilities are then computed using these dynamically estimated distributions via Bayes' theorem for adaptation. This distribution-centric approach enables the model to continually learn and adapt to the deployment environment. Extensive experiments validate that DOTA significantly mitigates forgetting and achieves state-of-the-art performance compared to existing methods. Code is available at <https://github.com/skylineeeeeen/DOTA>.

1 Introduction

Recent advances in vision-language foundation models have shown remarkable vision understanding capabilities across a broad range of tasks by training on web-scale image-text pairs [39, 30, 51]. Taking CLIP as an example, it can conduct zero-shot classification without the need for additional training data using predefined prompts [39]. However, CLIP may still face challenges when handling various specific applications during test time, especially when there is a significant distribution gap between the training and test data [41, 25, 13]. To adapt a foundational model to diverse deployment environments and personalized application requirements, fine-tuning is often necessary, which can be resource-intensive in terms of time, computational effort, and training data requirements.

Test-time adaptation (TTA) methods provide an efficient solution for addressing distributional shifts between training and testing domains [5, 6, 45]. TTA allows for the dynamic adjustment of a pre-

*Corresponding author: zongbo@bupt.edu.cn, mike.zheng.shou@gmail.com, zhangchangqing@tju.edu.cn

trained model during the inference phase by leveraging incoming test data to refine the model's parameters. This adaptation process optimizes the model's performance for the specific test distribution, eliminating the need for resource-intensive retraining. As such, TTA aligns seamlessly with real-world applications, where models must rapidly adapt to diverse and changing environments. There are two primary lines to achieve TTA on VLMs. Early works advocate learning prompts during test time with the test data [41, 13]. However, these methods require significant computational resources to optimize the learnable prompts via backpropagation and gradient descent, rendering them unsuitable for applications demanding fast test-time inference. Therefore, more efficient methods, cache-based methods, have been proposed recently [25, 55]. Typically, to avoid the need for training with backpropagation, Training-free Dynamic Adapter (TDA) maintains a lightweight cache during testing to store representative test samples, which helps guide the classification of subsequent test samples.

While cache-based classifiers like TDA offer significant efficiency, they are fundamentally constrained by their finite cache capacity. These methods typically store a limited set of 'typical' test samples, updating the cache by replacing older entries with newer, high-confidence ones. However, this reliance on instance-level memorization and forced sample discarding inevitably leads to catastrophic forgetting. As the model adapts to new data patterns, it loses information about previously seen variations, hindering the formation of a stable and comprehensive understanding of the evolving test distribution. Consequently, classifiers depending solely on cached samples can be suboptimal, and ultimately performance-limited by the cache size [55]. As illustrated in Fig. 1, to overcome these limitations of instance-level caching, we introduce DistributiOnal Test-time Adaptation (Dota). This represents a paradigm shift in test-time adaptation. Instead of passively memorizing discrete samples in the cache with limited size, the core motivation of Dota lies in actively and continuously estimating the underlying statistical distribution of the incoming test data, thereby enabling the full utilization of all available test samples.

Specifically, Dota continually estimates the distribution of test samples to adapt the test environment. Under the mild assumption that the embedding distribution of each class follows a Gaussian distribution [19], we propose an efficient method to continually estimate the distribution of different classes. Once the distributions of different classes are estimated, we can easily calculate the posterior probabilities of subsequent test samples based on Bayes' theorem and obtain a test-time classifier for test-time adaptation. Similar to cache-based methods, this process does not require gradient backpropagation, avoiding the complex computational overhead during testing, leading to more than 20 times faster inference speed. Moreover, unlike cache-based methods memorizing only representative test samples, Dota can continually adapt to the test environment by estimating the distribution of test samples. The contributions of this paper are:

- We propose a novel continual test-time learning framework which improves the performance of pretrained foundation models in downstream tasks by learning the statistical distribution of test data, rather than caching a limited set of individual samples.
- Within this framework, we propose a simple and effective method to enhance foundation models by estimating distribution of different categories at test time and using Bayes' theorem to create an adaptive classifier.
- Extensive experiments on diverse datasets validate the effectiveness of the proposed method, demonstrating a significant improvement. The code has been released at here.

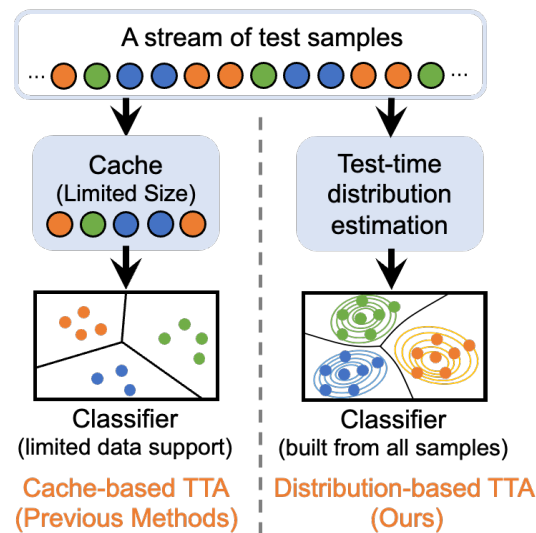


Figure 1: Cache-based TTA methods store individual test samples within a limited cache, which often leads to underutilization of the available test data. In contrast, the proposed method continuously estimates the underlying distribution of the test data, enabling the full exploitation of all available test samples.

2 Related work

Test-time adaptation (TTA) for classical classification neural networks focuses on addressing the distribution shift between training and test data by learning from the test data. Early efforts to improve TTA performance primarily involve adjusting batch normalization layers and designing unsupervised objective functions [35, 44, 28, 32]. For example, TENT [44] optimizes the affine parameters in batch normalization layers by minimizing the entropy of the prediction probability. MEMO [53] applies variant augmentation methods to a single test sample and optimizes model parameters by minimizing the entropy of the prediction probability. A recent advancement in test-time adaptation with distribution shift, which introduces the concept of universal TTA to address domain non-stationarity and temporal correlation, ensuring robust model performance across diverse scenarios [34]. The most relevant of these traditional TTA methods to us is T3A [24], which achieves test-time adaptation by adjusting the trained linear classifier using prototypes. Compared to T3A, which naively stores typical test samples, we achieve continuous adaptation by estimating the distribution of test samples.

TTA for VLMs. To enhance the performance of VLMs during testing, TPT [41] introduces adaptive text prompts and optimizes the prompts through entropy minimization. Building on this, DiffTPT [13] leverages pre-trained stable diffusion models to generate diverse augmented data for use in test-time prompt tuning. However, TPT and DiffTPT rely heavily on gradient backpropagation to optimize the prompts, making them computationally expensive and resource-intensive. TDA [25] proposes a lightweight test-time adaption method by storing representative test samples. Building upon TDA, Boostadapter [55] enhances the cache sample selection strategy through Regional Bootstrapping. Compared to TDA and Boostadapter, which naively stores typical test samples, we achieve continuous adaptation by estimating the distribution of test samples, leading to a more adaptive solution.

Distribution estimation for recognition. Distribution estimation plays a crucial role in adapting recognition models by leveraging data's statistical properties for dynamic updates. This approach is particularly effective when encountering new classes or shifting data distributions [19, 48]. For example, [3] developed an open-world recognition system that continuously learns new object categories by evolving the nearest class mean algorithm into a nearest non-outlier variant. Similarly, Prototypical Networks [42] utilize distribution estimation by defining class prototypes as the mean of embedded support examples; classification then relies on metric distances, yielding strong performance in few-shot and zero-shot settings. Further addressing dynamic data, previous method [9] proposed a system for continual prototype evolution, facilitating online learning from non-stationary streams via efficient memory management and a novel objective function. Recently, a training-free CLIP adaptation method has been proposed by introducing Gaussian Discriminant Analysis [48], which estimates class means and shared covariance from data to build an ensemble classifier integrating zero-shot CLIP predictions. Building upon these principles, particularly from the continual learning literature, this paper introduces Dota to enhance the test-time performance of vision-language models.

Vision-language models have demonstrated strong vision understanding capabilities benefiting from training on large-scale datasets [39, 51, 30]. Among them, CLIP [39] is the most representative method by maximizing the similarity between image and their corresponding text embeddings. To further enhance performance of CLIP on downstream tasks, prompt learning-based methods have been proposed by optimizing the prompts of the text encoder [57, 58, 2, 27, 2]. Moreover, to reduce the computational cost associated with gradient calculations in prompt learning, efficient CLIP adaptation methods have been introduced [16, 54, 48, 31, 50]. These methods enable downstream task adaptation using only a small number of training samples in the embedding space. Orthogonal to above methods, this paper focuses on continuously adapting to environments during testing by leveraging test samples.

3 Method

3.1 Zero-Shot Classification with Prompt

Zero-shot classification. During the pre-training stage, CLIP² trains its image and text encoders using large-scale image-text pairs. This is achieved by maximizing the cosine similarity between the image

²Dota is also applicable to other similar models.

and text embeddings through contrastive loss. Unlike traditional classifiers trained on closed-set labels, CLIP leverages open-set semantic information in the image-text pairs to learn a broader range of visual concepts. Consequently, during the test stage, CLIP can perform zero-shot classification without additional training. Specifically, given a test sample $\mathbf{x}$ for K -class classification, where $\mathbf{x}$ represents the image embedding obtained from the image encoder, the corresponding zero-shot prediction probability P_k^{zs} for class k is calculated as:

$$P_k^{\text{zs}}(y = k|\mathbf{x}) = \frac{\exp(\cos(\mathbf{x}, \mathbf{w}_k)/\tau)}{\sum_{k=1}^K \exp(\cos(\mathbf{x}, \mathbf{w}_k)/\tau)}, \quad (1)$$

where zs refers to zero-shot. $\mathbf{w}_k$ is the classification weight for class k , obtained by encoding the corresponding prompt, e.g., “a photo of {class}”, with the class token replaced by the specific category name. τ is the learned temperature parameter in CLIP, and $\cos(\cdot, \cdot)$ denotes the cosine similarity. The above classification process can be understood as comparing the obtained image embedding with the text prompt and selecting the most similar category as the final decision.

3.2 Distributional Test-time Adaptation

Key motivation. When CLIP is deployed across different environments, its performance often deteriorates due to changes in the data distribution, especially when the test data significantly deviates from the training data. TTA can effectively enable the foundational model to adapt quickly to new environments during the test phase. Current state-of-the-art methods typically maintain a cache of representative samples from various classes, which guide the classification of subsequent test samples [25, 55]. However, due to the limited cache size, cache-based TTA methods face several challenges:

- **Limited information storage:** These methods store embeddings for only a limited number of samples, which restricts the breadth and depth of information captured. As a result, they may fail to adequately represent the full complexity and nuances of the test data distribution.
- **Lack of deep association learning:** These methods primarily rely on cached samples for similarity matching or guidance, rather than learning or updating the model’s understanding of the deeper, intrinsic relationships between sample features and their corresponding semantic labels in the new test environment.
- **Significant risk of test-time forgetting:** Due to the above limitations, when the cache must be updated (e.g., by replacing older samples), the model struggles to retain the adaptive knowledge gained previously. This leads to a high risk of catastrophic forgetting, compromising both performance and stability as new data arrives.

To address these challenges, as shown in Fig. 1, we propose distributional test-time adaptation (Dota), which continuously estimates the evolving test sample distribution during testing. By leveraging Bayes’ theorem to infer the posterior distribution of different classes, Dota allows for dynamic adaptation based on an accurate understanding of the current data distribution. This distributional estimation ensures richer, more reliable information for classification, overcoming the limitations of cache-based methods and reducing the risk of forgetting, thus maintaining stable performance.

Classification with classical Gaussian discriminant analysis. Formally, inspired by classical Gaussian discriminant analysis [19], we assume that the embedding distribution of each class k follows a Gaussian distribution, i.e., $P(\mathbf{x}|y=k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of class k , respectively. Using Bayes’ theorem, the posterior probability $P(y=k|\mathbf{x})$ of class k can be given by

$$P(y=k|\mathbf{x}) = \frac{P(\mathbf{x}|y=k)P(y=k)}{P(\mathbf{x})},$$

where $P(\mathbf{x}) = \sum_{k=1}^K P(\mathbf{x}|y=k)P(y=k)$ and $P(y=k)$ is the prior probability. In practice, we set $P(y=k)$ to $1/K$ for simplicity. Then $P(y=k|\mathbf{x})$ can be obtained with

$$P(y=k|\mathbf{x}) = \frac{\exp(f_k(\mathbf{x}))}{\sum_{k=1}^K \exp(f_k(\mathbf{x}))}, \quad (2)$$

where $f_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k|$. The discriminant function $f_k(\mathbf{x})$ measures how well a sample $\mathbf{x}$ fits the distribution of class k . The detail can be found in the Appendix A.6. For

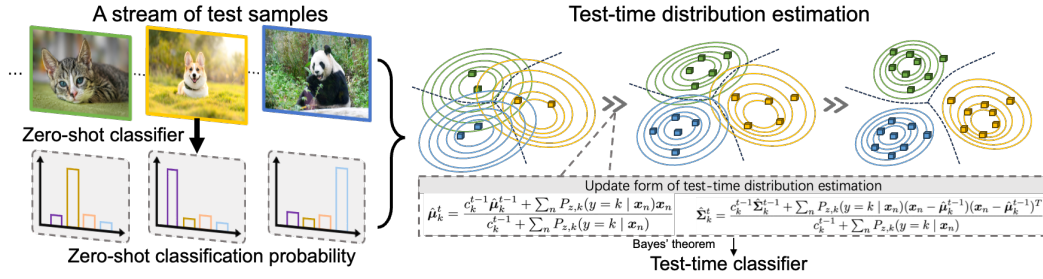


Figure 2: Pipeline of Dota. During test time, a stream of test samples is evaluated with original zero-shot classifier, and we estimate the distributions for the test samples during testing, enabling the model to continually learn from the test samples and the zero-shot classification probabilities. As the number of test samples increases, the estimated test sample data distribution will become more accurate. Finally the test-time classifier can be obtained with the estimated distributions according to Bayes' theorem for test-time adaptation.

the k -th class with N_k training samples, the mean and covariance matrix of different classes can be estimated as follows:

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N_k} \mathbf{x}_n}{N_k}, \hat{\Sigma}_k = \frac{\sum_{n=1}^{N_k} (\mathbf{x}_n - \hat{\mu}_k)(\mathbf{x}_n - \hat{\mu}_k)^T}{N_k}, \quad (3)$$

where $\hat{\mu}_k$ and $\hat{\Sigma}_k$ represent the estimated mean and covariance matrix of the k -th class, respectively. $\mathbf{x}_n$ denotes the input embeddings. However, there are two problems with the estimation of the mean and covariance in Eq. 3 that make it unsuitable for test-time distribution estimation. First, during testing, class labels are not available to compute the mean and covariance for class k . Second, during testing, it is impossible to access all test samples at once for distribution estimation, as the test samples are provided in the form of a data stream. We will address these two issues one by one below.

Parameter estimation with zero-shot predictive probability. When conducting test-time distribution estimation, one main challenge is that we cannot access to the ground-truth labels for the N test samples. Therefore, we try to use the zero-shot predictive probability to estimate the distribution [19]. To this end we introduce the Prop. 3.1. The proof can be found in the appendix A.1.

Proposition 3.1 (Parameter estimation with zero-shot predictive probability using the Expectation–Maximization (EM) algorithm). *Let $\{P_k^{zs}(y=k | \mathbf{x}_n)\}_{k=1}^K$ be the zero-shot prediction probabilities for n -th test sample. The means $\{\hat{\mu}_k\}_{k=1}^K$ and covariances $\{\hat{\Sigma}_k\}_{k=1}^K$ of the distribution can be estimated as a single iteration of the EM algorithm, where the expectation step computes the zero-shot prediction and the maximization step updates the parameters by maximizing the likelihood. The estimates are:*

$$\hat{\mu}_k = \frac{\sum_{n=1}^N P_k^{zs}(y=k | \mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^N P_k^{zs}(y=k | \mathbf{x}_n)}, \hat{\Sigma}_k = \frac{\sum_{n=1}^N P_k^{zs}(y=k | \mathbf{x}_n) (\mathbf{x}_n - \hat{\mu}_k)(\mathbf{x}_n - \hat{\mu}_k)^T}{\sum_{n=1}^N P_k^{zs}(y=k | \mathbf{x}_n)}. \quad (4)$$

Prop. 3.1 shows that the estimation process with zero-shot predictive probability. The estimation can also be intuitively understood as reweighting, where the zero-shot probabilities are used as weights to adjust the contributions of different samples, thereby mitigating the impact of the potential inaccuracies in the zero-shot prediction.

Online test-time distribution estimation. When estimating data distribution at test time, one another challenge is that we evaluate the test samples sequentially in a streaming manner instead of accessing all samples simultaneously. This necessitates a strategy to appropriately adjust the estimation method in Eq. 4 through effective initialization, and then allowing the parameters to be updated quickly as new test samples arrive. To achieve this goal, Dota maintains the distribution information of different classes (i.e., mean and covariance matrix) during testing, and updates its distribution information based on its representation information after obtaining new samples. **Initialization of $\{\hat{\mu}_k, \hat{\Sigma}_k\}_{k=1}^K$.** We can initialize the estimated mean of different classes in the following way:

$$\hat{\mu}_k^0 = \omega \mathbf{1} \quad \text{and} \quad \hat{\Sigma}_k^0 = \sigma^2 \mathbf{I}, \quad (5)$$

where ω and σ^2 is a hyperparameter that determines the initial mean and variance, $\mathbf{I}$ is the identity matrix. **Update of $\{\hat{\mu}_k, \hat{\Sigma}_k\}_{k=1}^K$.** We employ the update form described in [8], which is capable of estimating Gaussian distribution parameters in an online setting. Theoretically, for any sequence, the average regret of the update form converges to zero in the limit. Specifically, given a batch of test samples at step t , the updated $\hat{\mu}_k^t, \hat{\Sigma}_k^t$ can be computed based on the $\hat{\mu}_k^{t-1}, \hat{\Sigma}_k^{t-1}$ as follows:

$$\hat{\mu}_k^t = \frac{c_k^{t-1} \hat{\mu}_k^{t-1} + \sum P_k^{zs}(y = k | \mathbf{x}_n) \mathbf{x}_n}{c_k^{t-1} + \sum P_k^{zs}(y = k | \mathbf{x}_n)}, \hat{\Sigma}_k^t = \frac{c_k^{t-1} \hat{\Sigma}_k^{t-1} + \sum P_k^{zs}(y = k | \mathbf{x}_n) \mathbf{S}_k^{t-1}}{c_k^{t-1} + \sum P_k^{zs}(y = k | \mathbf{x}_n)}, \quad (6)$$

where $\mathbf{S}_k^{t-1} = (\mathbf{x}_n - \hat{\mu}_k^{t-1})(\mathbf{x}_n - \hat{\mu}_k^{t-1})^T$, c_k^{t-1} represents the effective sample size, defined by the cumulative confidences of the observed samples of class k at step $t - 1$. Specifically, we set $c_k^0 = 0$ and c_k^t is updated as $c_k^t = c_k^{t-1} + \sum P_k^{zs}(y=k|\mathbf{x}_n)$. When we obtain the estimated $\hat{\mu}, \hat{\Sigma}$, we can use Eq. 2 to calculate the test-time adapted posterior probability.

In practice, Eq. 6 is a generalized vector update version that works effectively with different test batch sizes. For consistency with comparison methods, we set the batch size to 1 in our experiments. To reduce computational complexity when inverting the covariance matrix $\hat{\Sigma}_k$, similar to the approach in [1, 14], we approximate the covariance by averaging across all classes, reducing the number of matrix inversions from K to 1, thereby improving efficiency and reducing the number of parameter estimates. Additionally, we apply shrinkage regularization to the precision matrix to enhance the stability of the inversion process as follows: $\hat{\Lambda} = [(1 - \epsilon)\hat{\Sigma} + \epsilon\mathbf{I}]^{-1}$, where $\epsilon = 10^{-4}$ is the shrinkage parameter. The term $\epsilon\mathbf{I}$ ensures that the eigenvalues of the covariance matrix are well-conditioned, maintaining the desired properties such as positive definiteness and rank stability.

3.3 Adaptive fusion of zero-shot and test-time classifier

As the number of test samples increases, the reliability of the estimated test sample distribution improves [8]. However, when the number of test samples is insufficient, the estimated distribution may be unreliable. To address this, we introduce a dynamic zero-shot classification and test-time result fusion approach, allowing the model to rely more on zero-shot classification during stages where the sample size for distribution estimation is insufficient. Formally, the final fusion probability is defined as follows:

$$P_k(y = k|x) = \frac{\exp(\cos(\mathbf{x}, \mathbf{w}_k)/\tau + \lambda f_k(\mathbf{x}))}{\sum_{k=1}^K [\exp(\cos(\mathbf{x}, \mathbf{w}_k)/\tau + \lambda f_k(\mathbf{x}))]}, \quad (7)$$

where $\lambda = \min(\rho c, \eta)$. Here, c represents the number of test samples, and ρ and η are hyperparameters that control the weight of the test-time classifier logits. The value of λ increases with the number of test samples when this number is insufficient, gradually approaching the maximum value η . This approach encourages the model to rely on the zero-shot classifier results when the test samples are insufficient to estimate the distribution, mitigating the potential negative impact of the test-time classifier. The whole pseudo code is shown in Alg. 1.

The additive fusion in Eq. 7 is a heuristic yet effective strategy designed to balance the robust prior knowledge from the pre-trained model with dynamic adaptation to test-time shifts. This logit-level combination is computationally efficient and has been widely adopted as a common and effective practice in the TTA literature for leveraging pre-trained knowledge while incorporating new test-time information[25, 52]. While this approach is largely heuristic, it dynamically ensures minimal interference in early stages while effectively refining predictions as Dota's per-class modeling becomes more accurate, proving highly effective in practice.

4 Experiments

Benchmarks. Consistent with prior works [41, 13, 25], we conduct our main experiments on cross-domain generalization and natural distribution shifts (NDS) scenarios. In the cross-domain generalization scenario, we evaluate the performance of the model across 10 diverse image classification datasets, each representing a distinct domain with different classes: Aircraft [33], Caltech101 [12], Cars [29], DTD [7], EuroSAT [20], Flower102 [36], Food101 [4], Pets [37], SUN397 [49], and UCF101 [43]. This benchmark provides a comprehensive evaluation of the adaptability of the

Algorithm 1: The pseudocode of Dota.

Input: The embedding of N test samples $\{x_n\}_{n=1}^N$ in an streaming way, zero-shot classification weights $[w_1, \dots, w_K]$;
 Initializing the distribution of different class;
for each test sample x_i **do**
 Obtain the zero-shot probability with Eq. 1;
 Update the distribution of different class with Eq. 6;
 Obtain the test-time classification probability with Eq. 2;
 Obtain the final classification result with Eq. 7.
end for

Table 1: Top-1 accuracy (%) under the cross-domain generalization scenario. “PE” represents the use of prompt enhancement[38].

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
Zero-Shot	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
TPT[41]	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT[13]	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	65.47
TDA[25]	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
BoostAdapter[55]	27.45	94.77	69.30	45.69	61.22	71.66	87.17	89.51	68.09	71.93	68.68
HisTPT[52]	26.90	94.50	69.20	48.90	49.70	71.20	89.30	89.10	67.20	70.10	67.60
ZERO[11]	25.21	93.66	68.04	46.12	34.33	67.68	86.53	87.75	65.03	67.77	64.21
Dota	26.25	94.16	69.56	47.64	62.78	75.23	87.08	92.01	69.80	72.54	69.71
DMN w/PE [56]	30.03	95.38	67.96	55.85	59.43	74.49	85.08	92.04	70.18	72.51	70.30
Dota w/ PE	29.82	94.85	69.06	55.97	58.35	77.06	87.07	92.40	70.97	74.86	71.04

model during test time across various class spaces. For the natural distribution shifts scenario, we utilize multiple datasets including ImageNet [10], ImageNet-A [22], ImageNet-R [21], ImageNet-S [46] and ImageNet-V2 [40], which serve as measures of the robustness of our approach. We also evaluate Dota using medical image datasets and a pathology-pretrained foundation model [23], where the labels were converted into descriptive sentences for evaluation (e.g., transforming “tumor” into “H&E image of a tumor”). The evaluation includes three datasets: the Kather Colon dataset [26], comprising nine different tissue types; the PanNuke dataset [15], focusing on benign versus malignant classifications; and the WSSS4LUAD dataset [18], which distinguishes between tumor and normal samples.

Comparison Method. We compare the proposed method with the following approaches: (1) CLIP’s zero-shot method, which utilizes an ensemble of 80 prompts [39]; (2) Prompt-based training methods, including TPT [41], DiffTPT [13], and Historical Prompt Tuning [52]. These methods focus on optimizing input prompts rather than modifying the model’s parameters. TPT and DiffTPT adapt the prompt at test time, with DiffTPT also introducing more diverse test sample augmentation using a diffusion model. Historical Prompt Tuning uses prior task knowledge to refine prompts for better task adaptation. (3) Efficient test-time adaptation methods, which do not require backpropagation and rely on a cache of representative samples for adaptation, as demonstrated in [11, 25, 56, 55]. The results for the above methods are obtained from the original papers.

Table 2: Performance under the NDS scenario.

Method	ImageNet	ImageNet-A	ImageNet-R	ImageNet-S	Average
Zero-Shot	68.34	49.89	77.65	48.24	61.03
TPT [41]	68.98	54.77	77.06	47.94	62.19
DiffTPT [13]	70.30	55.68	75.00	46.80	61.95
TDA[25]	69.51	60.11	80.24	50.54	65.10
ZERO [11]	69.31	59.61	77.22	48.40	63.64
Dota	70.69	61.50	81.21	51.84	66.31

Table 3: Performance comparison across datasets on PLIP model.

Dataset	Zero-Shot	TDA	CTA
Kather	45.60	49.35	61.92
PanNuke	69.49	69.70	74.68
WSSS4LUAD	70.31	71.96	75.07
average	61.80	63.67	70.56

4.1 Comparison with state-of-the-arts methods

Results under the cross-domain generalization scenario. We first compare Dota with state-of-the-art methods under the cross-domain generalization scenario across 10 diverse image classification datasets, each from a distinct domain with different classes. Tab. 1 presents the experimental results. The proposed method achieved the best performance on most datasets and the top two performance on all datasets. For example, when using the ViT-B/16 backbone network, the average performance was improved by 1.03%.

Results under the natural distribution shifts scenario. We then compare Dota with state-of-the-art methods in the context of natural distribution shifts. Tab. 2 presents the experimental results, revealing the following key observation. Leveraging distribution modeling of the representation of test data, Dota achieves superior performance without requiring gradient backpropagation. For instance, using the CLIP-ViT-B/16 backbone network, Dota outperforms the second-best method by an average of 1.21%, achieving state-of-the-art results across all datasets.

Performance validation on pathological image classification with PLIP. To validate the performance of the proposed method across other pretrained models and application scenarios, we conducted experiments using the PLIP model in the context of pathological image classification. The results are presented in Tab. 3. As shown in the table, the proposed method achieved superior performance, with an average improvement of 6.89% over TDA.

Table 4: Comparisons of our Dota with other methods in terms of efficiency (*Testing Time*) and effectiveness (*Accuracy*).

Method	Testing Time	Accuracy	Gain
Zero-Shot	11.82min	68.34	0
TPT	447min	68.98	+0.64
DiffTPT	1346min	70.30	+1.96
TDA	22min	69.51	+1.17
Dota (Ours)	22min	70.69	+2.35

Table 5: Comparisons of our Dota with other methods on the ImageNetV2 dataset, where each class contains only 10 samples.

Method	ViT-B/16	ResNet-50
Zero-Shot	61.88	52.91
TDA	64.67	55.54
Dota (All test samples)	64.50	55.19
Dota (last 50% test samples)	65.20	55.72

Inference time comparison. To illustrate the efficiency of the proposed method, we conduct evaluation about the inference time using the ViT-B/16 backbone on the ImageNet [10] dataset. The experimental results are shown in Tab. 4. From the table, we can see that the proposed method is faster than the methods that require gradient backpropagation. For example, Dota is 24 times faster than TPT, and 61 times faster than DiffTPT. Therefore, test-time adaptation methods that require gradient backpropagation may not be applicable during deployment due to the performance limitations of the inference device. At the same time, compared with TDA, the speed of the proposed method is comparable, but the performance is higher.

Failure case study. While our approach highlights the advantages of continuously estimating the distribution of test data and adapting to it, it does not consistently outperform TDA across all datasets, particularly those with a limited number of samples. For instance, as shown in Tab. 5, on the ImagenetV2 [40], which contains only 10 samples per class, the performance of Dota is slightly lower than TDA. This is likely because the limited number of samples per class is insufficient to accurately estimate the data distribution online. However, its performance shows a marked improvement on the latter 50% of the test samples. This observation suggests that the proposed model has the potential for further enhancement as the number of available test samples increases.

Table 6: Performance of Dota and TDA, comparing overall accuracy and the last 50% of test samples to show continuous adaptability. The results show that the performance of our method is continual improving. However, TDA is different, and its performance has declined on several datasets.

Method	Aircraft	Caltech101	Cars	DTD	Flower102	Food101	Pets	SUN397	UCF101
TDA (all test samples)	23.91	94.24	67.28	47.40	71.42	86.14	88.63	67.62	70.66
TDA (last 50% test samples)	26.57	93.59	66.95	46.22	71.75	86.02	89.26	67.86	72.20
Dota (All test samples)	26.25	94.16	69.56	47.64	75.23	87.08	92.01	69.80	72.54
Dota (last 50% test samples)	27.65	94.65	69.98	50.47	76.46	87.12	93.30	70.69	73.73

4.2 Ablation studies and further analysis

Analysis of continuous learning ability and test-time forgetting of TDA. We conducted this experiment to validate our research motivation and demonstrate experimentally that Dota possesses the capability for continuous learning. During testing on the ImageNet dataset, we recorded the performance of the most recent 5,000 test samples and compared it with the original zero-shot classifier’s performance. We analyzed the relationship between the improvement in model performance and the number of test samples processed. The results are illustrated in Fig. 3. From the experimental results, it can be shown that the proposed method progressively enhances model performance as the number of test samples increases. In contrast, TDA shows an initial improvement that subsequently declines, indicating its inability to continuously learn from the test data stream. Further analysis on additional datasets, shown in Tab. 6, highlights the performance on the last 50% of test samples as well as all samples. The results clearly demonstrate that the performance of the last 50% of test samples for Dota is significantly higher than the overall performance. This improvement can be attributed to the increasing reliability of the estimated distribution as more test samples are observed. However, TDA exhibits a different pattern, with its performance declining on several datasets.

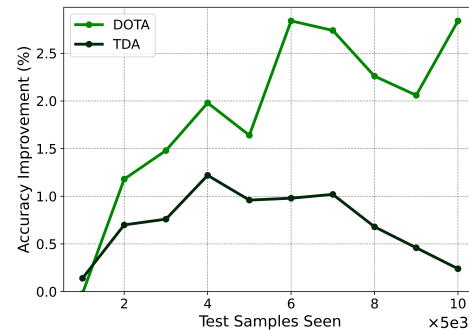


Figure 3: Improvement of different methods in model performance as the number of encountered test samples increases.

Table 7: Ablation study to compare the performance of Dota with two variants: (1) rely solely on high-confidence samples and their predictive labels for estimating the distribution; (2) using only the mean, without updating the covariance matrix in the estimation of the Gaussian distribution.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
Dota	26.25	94.16	69.56	47.64	62.78	75.23	87.08	92.01	69.80	72.54	69.71
high-confidence samples only	24.63 -1.62	94.12 -0.04	67.11 -2.45	46.34 -1.30	53.28 -9.50	72.07 -3.16	86.47 -0.61	90.68 -1.33	68.24 -1.56	69.28 -3.26	67.22 -2.49
w/o covariance	25.29 -0.96	94.16 0.00	67.47 -2.09	45.62 -2.02	55.06 -7.72	71.34 -3.89	86.44 -0.64	90.57 -1.44	67.88 -1.92	69.34 -3.20	67.32 -2.39

The Importance of distribution estimation using predictive probabilities of all test samples in an EM framework. We compared the performance of the Dota with a simplified version that only uses high-confidence samples and their corresponding predictive labels for estimating the distribution. The experimental results are shown in Tab. 7. The experimental results demonstrate that updating the data distribution using the proposed zero-shot predictive probabilities achieves better performance compared to using pseudo-labels from only high-confidence samples. For instance, overall performance sees a decline of approximately 2.49% when relying solely on high-confidence pseudo-label.

The necessity of distribution estimation. We compared the performance of the Dota with a simplified version that uses only the mean, excluding the estimation of the Gaussian distribution by removing the updates to the covariance matrices. This experiment aimed to understand the necessity of continual distribution estimation in enhancing model accuracy. The experimental results are shown in Tab. 7. The fifth row in the table presents the accuracy reductions across different datasets when the covariance matrix is not updated. The results indicate a consistent decrease in accuracy across all datasets, with a particularly notable drop of 7.72% on the EuroSAT dataset. These findings highlight the importance of continual distribution estimation.

Table 8: Hyperparameters analysis on the σ^2 and (ρ, η) combinations.

σ^2	0.0001	0.001	0.002	0.004	0.008	0.02
Acc	70.72	70.72	70.69	70.70	70.60	70.42
$\eta \backslash \rho$	0.005	0.01	0.02	0.03		
0.2	70.69	70.66	70.59	70.54		
0.3	70.64	70.55	70.36	70.28		
0.4	70.64	70.51	70.24	70.13		
0.5	70.64	70.48	70.15	70.00		

Hyperparameters analysis. To validate the sensitivity of our model to hyperparameters, we conduct systematic experiments and analyses. First, we evaluate the hyperparameter σ^2 while keeping other fixed. The results showed minimal impact on accuracy, with performance ranging from 70.42 to 70.72.

Next, we test different ρ and η combinations, observing stable performance across combinations. For instance, accuracy ranged from 70.69 to 70.00 as ρ and η varied. Another hyperparameter ω we consistently set to 0.001. Notably, all hyperparameter combinations show that the proposed method outperforms the original zero-shot classifier, indicating that TTA can usually significantly enhance performance even without a validation set for hyperparameter tuning.

The necessity of adaptive fusion of zero-shot and test-time classifier. We conduct ablation study to show that adaptive fusion of zero-shot and test-time classifier is necessary. The specific experimental results are shown on the Tab. 8. It can be observed that as ρ increases (indicating the diminishing effect of dynamic fusion), the performance of Dota consistently decreases.

5 Conclusion and Limitations

In this paper, we proposed DistributiOnal Test-time Adaptation (Dota), a method that overcomes the limitations of cache-based test-time adaptation by continuously estimating the underlying distribution of test data. Unlike traditional methods, which are constrained by fixed cache sizes and suffer from catastrophic forgetting, Dota dynamically updates class distributions, leading to more efficient and adaptive test-time inference. Our experiments show that Dota significantly reduces forgetting, improves performance, and offers over 20 times faster inference compared to test-time prompt training methods. This approach provides an effective solution for deploying vision-language models in dynamic environments, offering a scalable and computationally efficient alternative to traditional adaptation techniques.

This method has three main limitations that suggest directions for future work. First, its single Gaussian distribution assumption restricts its ability to model complex data, which could be addressed by exploring more flexible models. Second, the assumption of uniform class priors is inconsistent with real-world class imbalances and could be improved with an adaptive prior estimation mechanism. Finally, the heuristic fusion strategy could be replaced by a more principled approach that dynamically adjusts weights based on real-time sample uncertainty.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (No. 624B2100, No. 62376193, No. 61925602, No. T2522008, and No. 62272055), New Cornerstone Science Foundation through the XPLOER PRIZE. We thank Prof. Qinghua Hu for his help with the manuscript. Mike Shou does not receive any funding for this work. The authors appreciate the valuable feedback from anonymous reviewers.

References

- [1] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.
- [2] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17480–17489, 2024.
- [3] Abhijit Bendale and Terrance Boulton. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014.
- [5] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022.

- [6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.
- [7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.
- [8] Sanjoy Dasgupta and Daniel Hsu. On-line estimation with the multivariate gaussian distribution. In *International Conference on Computational Learning Theory*, pages 278–292. Springer, 2007.
- [9] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8250–8259, 2021.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.
- [11] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.
- [12] Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pages 178–178. IEEE, 2004.
- [13] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2704–2714, 2023.
- [14] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.
- [15] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019.
- [16] 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*, 132(2):581–595, 2024.
- [17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- [18] Chu Han, Xipeng Pan, Lixu Yan, Huan Lin, Bingbing Li, Su Yao, Shanshan Lv, Zhenwei Shi, Jinhai Mai, Jiatai Lin, et al. Wsss4lud: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. *arXiv preprint arXiv:2204.06455*, 2022.
- [19] Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):155–176, 1996.
- [20] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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, 2019.
- [21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8349, 2021.

- [22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.
- [23] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.
- [24] 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.
- [25] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.
- [26] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.
- [27] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.
- [28] Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. *arXiv preprint arXiv:2112.02355*, 2021.
- [29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 554–561, 2013.
- [30] Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mido Assran, Andrew Gordon Wilson, Aaron C. Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [31] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024.
- [32] Hyesu Lim, Byeongeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [34] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024.
- [35] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D’Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.
- [36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.
- [37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012.

- [38] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15691–15701, 2023.
- [39] 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*, pages 8748–8763. PMLR, 2021.
- [40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.
- [41] 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*, 35:14274–14289, 2022.
- [42] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- [43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012.
- [44] 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*, 2021.
- [45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations*, 2021.
- [46] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.
- [47] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned clip. In *International Conference on Machine Learning*, pages 51734–51754. PMLR, 2024.
- [48] Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.
- [49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.
- [50] 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*, pages 10899–10909, 2023.
- [51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11986, 2023.
- [52] Jingyi Zhang, Jiaying Huang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Historical test-time prompt tuning for vision foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pages 12872–12896, 2024.
- [53] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022.

- [54] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.
- [55] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. In *Advances in Neural Information Processing Systems*, volume 37, pages 67795–67825, 2024.
- [56] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28718–28728, 2024.
- [57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022.

A Appendix

A.1 Proof of Proposition 3.1

The EM algorithm is an iterative method used for finding maximum likelihood estimates of parameters in statistical models with latent (unobserved) variables. It consists of two main steps:

- (1) Expectation Step (E-step): Compute the expected value of the log-likelihood function, with respect to the current estimate of the distribution of the latent variables.
- (2) Maximization Step (M-step): Maximize this expected log-likelihood to update the parameter estimates.

In the context of Gaussian Discriminant Analysis (GDA) for classification, the latent variables correspond to the true class labels of the data points, which are unobserved during testing.

Assumption A.1 (Class-Conditional Distributions). For each class $k \in \{1, 2, \dots, K\}$, the data embedding $\mathbf{x}$ follows a Gaussian distribution:

$$P(\mathbf{x} \mid y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of class k , respectively.

Assumption A.2 (Prior Probabilities). The prior probability for each class is uniform: $P(y = k) = \frac{1}{K}$.

The objective is how to estimate the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ using the EM algorithm, leveraging the zero-shot predictive probabilities $P_k^{\text{zs}}(y = k \mid \mathbf{x}_n)$ as part of the expectation step.

E-Step: Compute Expected Log-Likelihood. In the E-step, we compute the expectation of the log-likelihood with respect to the posterior distribution of the latent variables given the current parameter estimates. Given the zero-shot predictive probabilities $P_k^{\text{zs}}(y = k \mid \mathbf{x}_n)$, we treat them as the responsibilities (i.e., the posterior probabilities) in the E-step:

$$\pi_{nk} = P_k^{\text{zs}}(y = k \mid \mathbf{x}_n)$$

The expected log-likelihood Q is then:

$$Q(\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \mid \text{current estimates}) = \sum_{n=1}^N \sum_{k=1}^K \pi_{nk} \left[\log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \frac{1}{K} \right]$$

M-Step: Maximize the Expected Log-Likelihood. In the M-step, we maximize Q with respect to $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. **Maximizing with Respect to $\boldsymbol{\mu}_k$.** To find the optimal $\boldsymbol{\mu}_k$, take the derivative of Q with respect to $\boldsymbol{\mu}_k$ and set it to zero. First, expand the Gaussian log-probability:

$$\log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{d}{2} \log(2\pi)$$

where d is the dimensionality of $\mathbf{x}_n$. Focusing on terms involving $\boldsymbol{\mu}_k$:

$$Q_{\boldsymbol{\mu}} = -\frac{1}{2} \sum_{n=1}^N \pi_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

Take the derivative with respect to $\boldsymbol{\mu}_k$ and set to zero:

$$\frac{\partial Q_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^N \pi_{nk} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

Solving for $\boldsymbol{\mu}_k$:

$$\begin{aligned} \sum_{n=1}^N \pi_{nk} \mathbf{x}_n &= \left(\sum_{n=1}^N \pi_{nk} \right) \boldsymbol{\mu}_k \\ \boldsymbol{\mu}_k &= \frac{\sum_{n=1}^N \pi_{nk} \mathbf{x}_n}{\sum_{n=1}^N \pi_{nk}} \end{aligned}$$

This corresponds directly to the provided update formula for $\hat{\mu}_k$:

$$\hat{\mu}_k = \frac{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n)}$$

Maximizing with Respect to Σ_k . Similarly, to find the optimal Σ_k , consider the terms in Q involving Σ_k :

$$Q_{\Sigma} = -\frac{1}{2} \sum_{n=1}^N \pi_{nk} [(\mathbf{x}_n - \mu_k)^{\top} \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) + \log |\Sigma_k|]$$

Take the derivative with respect to Σ_k^{-1} (using matrix derivative identities) and set to zero:

$$\frac{\partial Q_{\Sigma}}{\partial \Sigma_k^{-1}} = \frac{1}{2} \sum_{n=1}^N \pi_{nk} [(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\top} - \Sigma_k] = 0$$

Solving for Σ_k :

$$\Sigma_k = \frac{\sum_{n=1}^N \pi_{nk} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\top}}{\sum_{n=1}^N \pi_{nk}}$$

This aligns with the provided update formula for $\hat{\Sigma}_k$:

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n) (\mathbf{x}_n - \hat{\mu}_k)(\mathbf{x}_n - \hat{\mu}_k)^{\top}}{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n)}$$

Finally, the provided update equation is:

$$\hat{\mu}_k = \frac{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n)}, \quad \hat{\Sigma}_k = \frac{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n) (\mathbf{x}_n - \hat{\mu}_k)(\mathbf{x}_n - \hat{\mu}_k)^{\top}}{\sum_{n=1}^N P_k^{zs}(y = k | \mathbf{x}_n)}$$

are precisely the M-step updates obtained by maximizing the expected log-likelihood in the EM algorithm, where the E-step uses the zero-shot predictive probabilities as responsibilities. This demonstrates that the parameter estimation process described is equivalent to performing a single EM iteration. This single EM iteration leverages the zero-shot predictions to adjust the Gaussian parameters, effectively “reweighting” the contributions of different samples based on their inferred class probabilities.

A.2 Justification for the Gaussian Assumption

Our assumption that the embedding distribution of each class k follows a Gaussian distribution, i.e., $P(\mathbf{x} | y = k) = \mathcal{N}(\mu_k, \Sigma_k)$, is grounded in both theoretical principles and empirical evidence.

- **Theoretical support:** The Central Limit Theorem (CLT) provides a general justification. A high-dimensional embedding can be viewed as an aggregation of numerous lower- and mid-level features. The CLT posits that the sum of many weakly correlated random variables will tend toward a normal distribution, making it reasonable to expect that the composite feature for a single class will approach a multivariate Gaussian distribution.
- **Empirical validation:** Our own ablation study validates the importance of this assumption; omitting covariance updates, which relies on the Gaussian model, significantly degrades performance (e.g., a 7.72% drop on EuroSAT), confirming the practical utility of Gaussian modeling.

A.3 Analysis of Catastrophic Forgetting

A key motivation for our work is to mitigate the catastrophic forgetting inherent in cache-based methods like TDA. In the context of TTA, forgetting manifests at a “sample-level”: due to a fixed-size cache, adapting to new samples necessitates discarding information from older ones. To rigorously validate this issue, we conducted two targeted experiments. The core idea is to have the model, after

a period of continuous adaptation, revisit and re-evaluate its performance on historical test samples. Our experiments revealed two levels of forgetting.

Forgetting without Significant Distribution Change. To assess this phenomenon, we designed a two-stage experiment. This process first involves adapting a model on ImageNet and recording its performance (Stage 1), after which we freeze the model’s parameters and re-evaluate it on the early 25,000 test samples in the dataset (Stage 2). We hypothesize that if no forgetting occurs, this initial adaptation stage should directly lead to an increase in the model’s performance. Conversely, a performance decrease between stage 1 and stage 2 would suggest that the model has experienced forgetting. Experimental results show that TDA suffers from obvious performance degradation, while our method does not.

Table 9: Performance comparison in a two-stage experiment.

Samples Seen	10000	20000	25000
TDA Stage 1	0.42	0.71	0.76
TDA Stage 2 (Frozen)	0.17	0.57	0.29
DOTA Stage 1	0.58	1.16	1.25
DOTA Stage 2 (Frozen)	2.34	2.30	2.25

Domain-Level Forgetting. We challenged the model by adapting it first to an original domain (ImageNet), then to a new, shifted domain (ImageNet-C), and finally re-evaluating its performance on the original domain. TDA’s performance failed to improve upon returning to ImageNet, indicating it had forgotten the original domain’s features. DOTA, however, showed improvement, demonstrating it retained this knowledge.

Table 10: Domain-level forgetting experiment.

Method	ImageNet	ImageNet-C brightness	ImageNet (Frozen model)
Zero-Shot	68.34	56.98	68.34
TDA	69.55	58.22	69.35
DOTA	70.76	60.64	70.93

We reinterpreted Fig. 3 and added additional experiments. We found that catastrophic forgetting degrades a model’s generalization capabilities by eroding previously learned information. More strikingly, we found that under certain conditions, continued adaptation is actively counterproductive, yielding worse results than if adaption had stopped. This is because as we continue to adapt, we forget more than we learn. The Tab. 11 illustrates this clearly. The standard TDA model’s performance peaked after 20,000 samples and then degraded. A model whose training was frozen at that point (TDA frozen) consistently outperformed the continuously trained model thereafter. Our proposed method, DOTA, effectively mitigates this degradation and delivers robustly superior performance.

Table 11: Performance degradation during continuous adaptation.

Samples Seen	5k	10k	15k	20k	25k	30k	35k	40k	45k	50k	Ave
TDA	0.14	0.70	0.76	1.22	0.96	0.98	1.02	0.68	0.46	0.24	0.72
TDA (frozen after 20k)	0.14	0.70	0.76	1.22	1.00	1.54	1.04	0.82	0.66	0.68	0.86
DOTA	-0.02	1.18	1.48	1.98	1.64	2.84	2.74	2.26	2.06	2.84	1.90

A.4 Results across multiple data ordering

We conducted experiments on five test data in different ordering on multiple datasets. The experimental results are shown in the Tab. 12. From the experimental results, we can see that the order of test data has little effect on the prediction performance.

Table 12: Experimental results across multiple datasets and test data orderings.

Dataset	Method\Data ordering	1	2	3	4	5	Average
ImageNet	Dota	70.54	70.66	70.72	70.67	70.63	70.64
eurosat	Dota	62.74	62.56	63.01	62.46	63.28	62.81
OxfordPets	Dota	92.04	91.93	92.29	92.04	92.12	92.08

A.5 Performance on non-i.i.d. data streams

We conducted additional experiments to evaluate the model’s performance under non-i.i.d. data distribution during testing, using the ImageNet dataset as a benchmark. By employing a Dirichlet distribution, we simulated varying degrees of non-i.i.d. data streams, adjusting the concentration parameter and dividing the dataset into 5 and 10 time slices for analysis. The details of the experiments are shown as follows:

Time Slices: We divided the ImageNet dataset into 5 and 10 time slices, where each slice contains varying numbers of samples and class distributions.

Concentration Parameter ($[\alpha]_K$): The concentration parameter of the Dirichlet distribution controls the uniformity of class distributions across slices. Smaller α values (e.g., 0.1) create highly uneven distributions, while larger values (e.g., 0.5 and 1) result in more uniform distributions.

Evaluation Setting: Since the sizes of sub-datasets for each time slice are unequal, the final average accuracy is a weighted average based on the number of samples in each slice. The experimental results are summarized below.

Table 13: Performance on non-i.i.d. data streams (5 slices).

α	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Average
0.1	68.18	70.06	71.60	71.09	70.91	70.41
0.5	69.55	70.78	69.43	71.95	70.92	70.58
1	69.41	71.64	71.05	70.01	71.81	70.83

Table 14: Performance on non-i.i.d. data streams (10 slices).

α	Slice 1	Slice 2	Slice 3	Slice 4	Slice 5	Slice 6	Slice 7	Slice 8	Slice 9	Slice 10	Average
0.1	69.99	69.31	67.34	70.23	71.49	68.8	73.13	70.35	69.89	72.39	70.39
0.5	68.74	69.32	72.17	71.27	69.97	69.8	70.3	70.78	72.48	70.78	70.61
1	67.33	70.29	70.92	68.43	69.85	71.14	72.03	72.52	71.29	71.66	70.68

From the experimental results, we can see that the model shows strong robustness to non-i.i.d. data streams, with only minimal accuracy decline under small α (e.g., $\alpha = 0.1$).

A.6 More details and explanation about the $f_k(\mathbf{x})$ in Eq. 2.

The function $f_k(\mathbf{x})$, often referred to as the *discriminant function*, measures how well a data point $\mathbf{x}$ fits the distribution of class k . It is derived from Gaussian Discriminant Analysis and consists of two main components. The first component is the Mahalanobis distance, $-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x} - \mu_k)$, which calculates the squared distance between $\mathbf{x}$ and the class mean μ_k , scaled by the inverse of the covariance matrix Σ_k . This term captures the similarity of $\mathbf{x}$ to the center of the class, considering feature correlations. The second component is the normalization term, $-\frac{1}{2} \log |\Sigma_k|$, which accounts for the determinant of the covariance matrix Σ_k and reflects the spread (or volume) of the Gaussian distribution for class k . This ensures that classes with larger variances are normalized appropriately. Intuitively, a larger value of $f_k(\mathbf{x})$ indicates a higher likelihood that $\mathbf{x}$ belongs to class k . In classification, $f_k(\mathbf{x})$ is used within the softmax function to compute the posterior probability $P(y = k | \mathbf{x})$, which determines the most likely class for $\mathbf{x}$:

$$P(y = k | \mathbf{x}) = \frac{\exp(f_k(\mathbf{x}))}{\sum_{k=1}^K \exp(f_k(\mathbf{x}))}$$

A.7 CLIP’s Superior Calibration Performance

When estimating the distribution, we use zero-shot probabilities, which we argue here to be relatively well-calibrated. Specifically, we provide comparative data on the Expected Calibration Error (ECE) metric to substantiate CLIP’s superior calibration performance. We draw on data from [47] and [17] to present the following comparative table of ECE metrics on the ImageNet dataset for different models.

Table 15: Datasets details.

Dataset	Classes	Validation Size	Test Size	Task
ImageNet	1,000	N/A	50,000	Classification
ImageNet-V2	1,000	N/A	10,000	Generalization
ImageNet-S	1,000	N/A	50,000	Generalization
ImageNet-A	200	N/A	7,500	Generalization
ImageNet-R	200	N/A	30,000	Generalization
Aircraft	100	3,333	3,333	Aircraft recognition
Caltech101	100	1,649	2,465	Object recognition
Cars	196	1,635	8,041	Car recognition
DTD	47	1,128	1,692	Texture classification
EuroSAT	10	5,400	8,100	Remote sensing classification
Flowers102	102	1,633	2,463	Flower recognition
Food101	101	20,200	30,300	Food classification
Pets	37	736	3,669	Pet classification
SUN397	397	3,970	19,850	Scene recognition
UCF101	101	1,898	3,783	Action recognition
Kather	2	10,718	32,154	Colon classification
PanNuke	9	623	1,888	Tissue classification
WSSS4LUAD	2	1,009	3,028	Tissue classification

Table 16: Comparison of Expected Calibration Error (ECE) on the ImageNet dataset. Lower values indicate better calibration. Note that CLIP achieves the lowest ECE without post-hoc calibration methods like Temperature Scaling (TS) or Vector Scaling (VS).

Method	DenseNet 161	DenseNet 161+TS	DenseNet 161+VS	ResNet 152	ResNet 152+TS	ResNet 152+VS	CLIP
ECE	6.28%	1.99%	2.24%	5.48%	1.86%	2.23%	1.51%

In the table, “TS” refers to Temperature Scaling, and “VS” refers to Vector Scaling. These are calibration strategies applied to DenseNet and ResNet to reduce Expected Calibration Error (ECE), whereas CLIP adopts no such strategies. Experimental data were obtained on the ImageNet dataset.

The table demonstrates that CLIP’s ECE is significantly lower than that of DenseNet and ResNet. Even after applying scaling strategies to reduce ECE for the baseline models, CLIP’s ECE remains lower, confirming its superior calibration performance.

A.8 Statistical significance for results

To evaluate the stability and robustness of our proposed method, Dota, we conducted experiments across all datasets using five different random seeds. The aggregated results are presented in Tab. 17, Tab. 18, and Tab. 19, which detail the mean accuracy, standard deviation (stddev), and 95% confidence intervals. As shown, Dota exhibits highly consistent performance. This analysis confirms the robustness and reliability of our findings.

Table 17: Statistical analysis for cross-domain generalization results.

Dataset	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flowers	Food101	Pets	Sun397	UCF101	Average
Dota (seed=1)	26.25	94.16	69.56	47.64	62.78	75.23	87.08	92.01	69.87	72.54	69.71
Dota (mean)	26.18	94.59	69.57	47.88	62.71	75.32	87.06	91.91	69.82	72.69	69.77
stddev	0.26	0.29	0.17	0.36	0.28	0.21	0.06	0.28	0.07	0.21	0.22
confidence interval	[25.86, 26.51]	[94.22, 94.96]	[69.35, 69.78]	[47.44, 48.33]	[62.37, 63.06]	[75.06, 75.58]	[86.99, 87.13]	[91.56, 92.26]	[69.73, 69.91]	[72.43, 72.94]	[69.50, 70.05]

A.9 Implementation details

All the models in our experiments are built upon the pre-trained CLIP model [39] that consists of an image encoder and a text encoder. Test-time adaptation is set for single-image scenarios, using a batch size of 1. For natural distribution shifts scenario, we tune our hyperparameters using the validation set. For the cross-domain generalization scenario, we perform hyperparameter search using the corresponding validation sets. We adjust σ^2 within [0.001, 0.002, 0.004], then search for the best

Table 18: Statistical analysis for Natural Distribution Shift (NDS) results.

Dataset	ImageNet	ImageNet-A	ImageNet-R	ImageNet-S	Average
Dota (seed=1)	70.69	61.50	81.21	51.84	66.31
Dota (mean)	70.66	61.32	81.13	51.85	66.24
stddev	0.07	0.20	0.11	0.11	0.12
confidence interval	[70.57, 70.75]	[61.06, 61.57]	[81.00, 81.26]	[51.71, 51.99]	[66.08, 66.39]

Table 19: Statistical analysis for medical scenario results.

Dataset	Kather	PanNuke	WSSS4LUAD	Average
Dota (seed=1)	61.92	74.68	75.07	70.56
Dota (mean)	61.94	74.65	75.13	70.57
stddev	0.14	0.14	0.15	0.14
confidence interval	[61.76, 62.11]	[74.49, 74.82]	[74.94, 75.32]	[70.40, 70.75]

η across [0.2, 0.3, 0.4, 0.5] and ρ across [0.005, 0.01, 0.02, 0.03], with the shrinkage parameter ϵ set to 0.0001. We use top-1 accuracy (%) as our evaluation metric. All experiments are conducted using a single NVIDIA RTX 4090 GPU and a 12-core Intel Xeon Platinum 8352V CPU.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”.**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: the paper discusses the limitations of the work performed by the authors.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: the code has been released.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The code has been released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The full details have been provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: we provide appropriate information about the statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: we have provided time of execution.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research carried out in the paper complies in all respects with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Foundational models are being widely deployed, but they do not always adapt perfectly to the distribution of test data. Collecting new data and fine-tuning models for specific applications can be costly and slow in response. Therefore, allowing models to adapt to unseen data during test time can enhance their generalization and adaptability.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: The creators or original owners of assets (e.g. code, data, models) used in the paper are properly credited, and the licence and terms of use are explicitly stated and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: the paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.