Realistic Test-Time Adaptation of Vision-Language Models

Supplementary Material

A. Additional experimental details

Datasets. We follow the setting of previous work [45]. We assess our method on ImageNet [8] and ten other datasets for fine-grained classification of scenes (SUN397 [35]), aircraft types (Aircraft [21]), satellite imagery (EuroSAT [13]), automobiles (StanfordCars [18]), food items (Food101 [2]), pet breeds (OxfordPets [28]), flowers (Flower102 [25]), general objects (Caltech101 [10]), textures (DTD [7]) and human actions (UCF101 [31]). These diverse datasets provide a comprehensive visual classification benchmark. Additional information on the statistics of each dataset is provided in Table 6.

Hyperparameters Generalization to unseen cases is crucial for TTA methods. Therefore, optimizing hyperparameters for each scenario would require access to labels and prior knowledge of the specific scenario encountered during testing, which goes against the core purpose of the TTA approach. For instance, we found that TDA largely relies on dataset-specific hyperparameters without clear guidance on how to tune them for a new dataset. Similarly, DMN conducts an hyperparameter search in order to find the optimal balance between the logits obtained from the text prompts and the logits from their model's memory, using knowledge from ground truth labels. To ensure fairness in comparison, we use hyperparameters optimized for ImageNet for all reported experiments.

Comparative methods. We use the same handcrafted prompts for all methods, which are listed in Table 5. Due to the more versatile scenarios studied in this paper, we find that our centroid initialization based on text embeddings much more robust, especially when the number of effective classes is reduced. Therefore, we implement it for TransCLIP instead of their original initialization based on the top-confident samples for each class. For ZLaP and Dirichlet we follow the hyperparameters of their official implementation. For TDA and DMN, following our discussion, we use the hyperparameters optimized for ImageNet. For TDA, this means the positive logits mixing coefficients is set to 2, while the negative logits mixing coefficient is set to 0.117. For DMN, since we only consider zero-shot scenarios, we only need to set the coefficient relative to the dynamic memory, which is set to 1. For TENT, we use a learning rate of 0.001 and 10 steps of gradient descent per batch.

B. Kullback-Leibler divergence between two multivariate Gaussian distributions

Let $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ two multivariate Gaussian distributions with respective probability density functions p and q. Then, we have

$$\begin{aligned}
\mathrm{KL}(p||q) &= \int_x p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x && (13)\\
&= \mathbb{E}_p[\log(p) - \log(q)] && (14)\\
&= \mathbb{E}_p[\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\mathbf{x} - \boldsymbol{\mu}_q)] && (15)\\
&= \frac{1}{2}\mathbb{E}_p[\log \frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|}] - \frac{1}{2}\mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p)] + \frac{1}{2}\mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\mathbf{x} - \boldsymbol{\mu}_q)] && (16)\\
&= \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{1}{2}\mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p)] + \frac{1}{2}\mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\mathbf{x} - \boldsymbol{\mu}_q)]. && (17)
\end{aligned}$$

We can rewrite the second term as

$$(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p) = \mathrm{Tr}\left((\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p)\right) = \mathrm{Tr}\left((\mathbf{x} - \boldsymbol{\mu}_p)(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}\right) \tag{18}$$

by using

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB). \tag{19}$$

For the third term, since we assume $\mathbf{x}$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, we have (see Matrix cookbook [29] Eq. 380 of Section 8.2)

$$\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\mathbf{x} - \boldsymbol{\mu}_q)] = (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^\top \boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) + \mathrm{Tr}(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p) \tag{20}$$

And therefore

$$\mathrm{KL}(p||q) = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{1}{2}\mathbb{E}_p[\mathrm{Tr}\left((\mathbf{x}-\boldsymbol{\mu}_p)(\mathbf{x}-\boldsymbol{\mu}_p)^\top\boldsymbol{\Sigma}_p^{-1}\right)] + \frac{1}{2}((\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top\boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q) + \mathrm{Tr}\{\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p\}) \tag{21}$$

$$= \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{1}{2}(\mathrm{Tr}\left(\mathbb{E}_p[(\mathbf{x}-\boldsymbol{\mu}_p)(\mathbf{x}-\boldsymbol{\mu}_p)^\top]\boldsymbol{\Sigma}_p^{-1}\right) + \frac{1}{2}((\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top\boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q) + \mathrm{Tr}(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p)) \tag{22}$$

by using the fact that trace and expectation can be interchanged. Moreover,

$$\mathbb{E}_p[(\mathbf{x}-\boldsymbol{\mu}_p)(\mathbf{x}-\boldsymbol{\mu}_p)^\top] = \boldsymbol{\Sigma}_p \tag{23}$$

which simplifies further the second term of the sum and gives

$$\mathrm{KL}(p||q) = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{1}{2}(\mathrm{Tr}(\boldsymbol{I}_k) + \frac{1}{2}((\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top\boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q) + \mathrm{Tr}(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p)) \tag{24}$$

$$= \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - \frac{d}{2} + \frac{1}{2}((\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top\boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q) + \mathrm{Tr}(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p)) \tag{25}$$

$$= \frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_q|}{|\boldsymbol{\Sigma}_p|} - d + (\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top\boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q) + \mathrm{Tr}(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p)\right) \tag{26}$$

with d the number of dimensions.

C. Detailed derivations of the regularized updates of the parameters

We write p_k and q_k the respective probability density functions of the distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\mathcal{N}(\boldsymbol{\mu'}_k, \boldsymbol{\Sigma'}_k)$. With the results of Appendix Section B, we have

$$\mathcal{A}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_k \mathrm{KL}(q_k||p_k) = \frac{1}{2}\sum_k (\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k) + \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma'}_k) + \log\frac{|\boldsymbol{\Sigma}_k|}{|\boldsymbol{\Sigma'}_k|} - d.$$

C.1. With respect to $\boldsymbol{\mu}_k$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\mu}_k} = \frac{\partial}{\partial\boldsymbol{\mu}_k}\left(-\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top\log(\mathbf{p}_i) - \sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_i^\top\mathbf{z}_j + \sum_{i\in\mathcal{Q}}\mathrm{KL}(\mathbf{z}_i||\hat{\mathbf{y}}_i) + \alpha\mathcal{A}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) \tag{27}$$

$$= \frac{\partial}{\partial\boldsymbol{\mu}_k}\left(-\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top\log(\mathbf{p}_i) - \sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_i^\top\mathbf{z}_j + \sum_{i\in\mathcal{Q}}\mathrm{KL}(\mathbf{z}_i||\hat{\mathbf{y}}_i) + \alpha\sum_{l=1}^{K}\mathrm{KL}(\mathbf{q}_l||\mathbf{p}_l)\right) \tag{28}$$

$$= \frac{\partial}{\partial\boldsymbol{\mu}_k}\left(-\sum_{i\in\mathcal{Q}}z_{i,k}\left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{f}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{f}_i-\boldsymbol{\mu}_k)\right) + \frac{\alpha}{2}(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)\right) \tag{29}$$

$$= -\sum_{i\in\mathcal{Q}}z_{i,k}\left(\boldsymbol{\Sigma}_k^{-1}(\mathbf{f}_i-\boldsymbol{\mu}_k)\right) + \alpha\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k) \tag{30}$$

Observe that the term $\alpha\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)$ directly comes from the derivative of our statistical anchor $\mathcal{A}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with regard to $\boldsymbol{\mu}_k$. By setting the derivative to 0

$$-\sum_{i\in\mathcal{Q}}z_{i,k}(\mathbf{f}_i-\boldsymbol{\mu}_k) - \alpha(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k) = 0 \tag{31}$$

$$\sum_{i\in\mathcal{Q}}z_{i,k}\mathbf{f}_i + \alpha\boldsymbol{\mu'}_k = \sum_{i\in\mathcal{Q}}z_{i,k}\boldsymbol{\mu}_k + \alpha\boldsymbol{\mu}_k \tag{32}$$

$$\sum_{i\in\mathcal{Q}}z_{i,k}\mathbf{f}_i + \alpha\boldsymbol{\mu'}_k = \left(\sum_{i\in\mathcal{Q}}z_{i,k} + \alpha\right)\boldsymbol{\mu}_k \tag{33}$$

We then obtain the centroid update

$$\boldsymbol{\mu}_k = \frac{\sum_{i\in\mathcal{Q}} z_{i,k}\mathbf{f}_i + \alpha \boldsymbol{\mu'}_k}{\sum_{i\in\mathcal{Q}} z_{i,k} + \alpha}. \tag{34}$$

If we write

$$\beta_k = \frac{\sum_{i\in\mathcal{Q}} z_{i,k}}{\sum_{i\in\mathcal{Q}} z_{i,k} + \alpha} \tag{35}$$

and

$$\boldsymbol{v}_k = \frac{\sum_{i\in\mathcal{Q}} z_{i,k}\mathbf{f}_i}{\sum_{i\in\mathcal{Q}} z_{i,k}} \tag{36}$$

we get the new centroid update

$$\boldsymbol{\mu}_k = \beta_k \boldsymbol{v}_k + (1-\beta_k)\boldsymbol{\mu'}_k \tag{37}$$

C.2. With respect to $\boldsymbol{\Sigma}_k^{-1}$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k^{-1}} = \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \left(-\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{Q}} w_{ij}\mathbf{z}_i^\top \mathbf{z}_j + \sum_{i\in\mathcal{Q}} \mathrm{KL}(\mathbf{z}_i||\hat{\mathbf{y}}_i) + \alpha\mathcal{A}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \right) \tag{38}$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \left(-\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{Q}} w_{ij}\mathbf{z}_i^\top \mathbf{z}_j + \sum_{i\in\mathcal{Q}} \mathrm{KL}(\mathbf{z}_i||\hat{\mathbf{y}}_i) + \alpha\sum_{l=1}^{K} \mathrm{KL}(\mathbf{q}_l||\mathbf{p}_l) \right) \tag{39}$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \Bigg(-\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \left(-\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{f}_i - \boldsymbol{\mu}_k) \right) + \frac{\alpha}{2}(\log\frac{|\boldsymbol{\Sigma}_k|}{|\boldsymbol{\Sigma'}_k|}$$

$$+ (\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k) + \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma'}_k)\ )\Bigg) \tag{40}$$

Note that the term $\log\frac{|\boldsymbol{\Sigma}_k|}{|\boldsymbol{\Sigma'}_k|} + (\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k) + \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma'}_k)$ directly comes from the derivative of our statistical anchor $\mathcal{A}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with regard to $\boldsymbol{\Sigma}_k^{-1}$. Using the formulas (from Matrix cookbook [29])

$$\frac{\partial}{\partial X}(\log|X|) = (X^{-1})^\top \tag{41}$$

$$\frac{\partial}{\partial X^{-1}}(\log|X|) = -X^\top \tag{42}$$

and

$$\frac{\partial}{\partial X}(\mathrm{Tr}(AXB)) = A^\top B^\top \tag{43}$$

$$\frac{\partial}{\partial X}(\mathrm{Tr}(AX^{-1}B)) = -(X^{-1}BAX^{-1})^\top \tag{44}$$

as well as the fact that covariances are symmetric ($\boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma}$), setting the derivative to 0 yields

$$-\sum_{i\in\mathcal{Q}} z_{i,k}\left(\boldsymbol{\Sigma}_k - (\mathbf{f}_i - \boldsymbol{\mu}_k)(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top\right) + \alpha(-\boldsymbol{\Sigma}_k^\top + (\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)(\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)^\top + \boldsymbol{\Sigma'}_k^\top) = 0 \tag{45}$$

$$-\sum_{i\in\mathcal{Q}} z_{i,k}\left(\boldsymbol{\Sigma}_k - (\mathbf{f}_i - \boldsymbol{\mu}_k)(\mathbf{f}_i - \boldsymbol{\mu}_k)^\top\right) + \alpha(-\boldsymbol{\Sigma}_k + (\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)(\boldsymbol{\mu'}_k - \boldsymbol{\mu}_k)^\top + \boldsymbol{\Sigma'}_k) = 0 \tag{46}$$

$$\tag{47}$$

$$\left(\sum_{i\in\mathcal{Q}} z_{i,k}+\alpha\right)\boldsymbol{\Sigma}_k=\sum_{i\in\mathcal{Q}} z_{i,k}(\mathbf{f}_i-\boldsymbol{\mu}_k)(\mathbf{f}_i-\boldsymbol{\mu}_k)^\top+\alpha(\boldsymbol{\Sigma'}_k+(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)^\top). \tag{48}$$

We get

$$\boldsymbol{\Sigma}_k=\frac{\sum_{i\in\mathcal{Q}} z_{i,k}(\mathbf{f}_i-\boldsymbol{\mu}_k)(\mathbf{f}_i-\boldsymbol{\mu}_k)^\top+\alpha(\boldsymbol{\Sigma'}_k+(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)^\top)}{\sum_{i\in\mathcal{Q}} z_{i,k}+\alpha}. \tag{49}$$

By writing the old $\boldsymbol{\Sigma}_k$-update

$$\boldsymbol{T}_k=\frac{\sum_{i\in\mathcal{Q}} z_{i,k}(\mathbf{f}_i-\boldsymbol{\mu}_k)(\mathbf{f}_i-\boldsymbol{\mu}_k)^\top}{\sum_{i\in\mathcal{Q}} z_{i,k}}, \tag{50}$$

we obtain the new covariance update

$$\boldsymbol{\Sigma}_k=\beta_k\boldsymbol{T}_k+(1-\beta_k)(\boldsymbol{\Sigma'}_k+(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)(\boldsymbol{\mu'}_k-\boldsymbol{\mu}_k)^\top). \tag{51}$$

D. Detailed derivations of the complete formulation

We refer to the derivations and the convergence proof in the TransCLIP paper [40]. The optimization follows a Block Majorize-Minimize (BMM) procedure over three blocks of variables: $\mathbf{z}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. For the Majorize-Minimize (MM) with respect to the $\mathbf{z}$-block (while $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are fixed), both the GMM- and KL-based terms are convex w.r.t $\mathbf{z}_i$. Consequently, we can proceed using similar arguments. For PSD matrix $\boldsymbol{W}$, the Laplacian regularization term in Eq. (11) is concave. To address this, we can replace the quadratic Laplacian term by a linear bound. By introducing simplex constraints $\mathbf{z}_i\in\Delta_K$ (λ_i the corresponding Lagrange multiplier) for $i\in\mathcal{Q}$,

$$\frac{\partial\mathcal{L}}{\partial\mathbf{z}_i}=\frac{\partial}{\partial\mathbf{z}_i}\left(-\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top\log(\mathbf{p}_i)-\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_i^\top\mathbf{z}_j+\sum_{i\in\mathcal{Q}}\mathrm{KL}(\mathbf{z}_i||\hat{\mathbf{y}}_i)+\alpha\mathcal{A}(\boldsymbol{\mu},\boldsymbol{\Sigma})+\sum_{i\in\mathcal{Q}}\lambda_i(z_i^\top\mathbb{1}_K-1)\right) \tag{52}$$

$$=-\log(\mathbf{p}_i)-\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_j-\log(\hat{\mathbf{y}}_i)+\log(\mathbf{z}_i)+(1+\lambda_i)\mathbb{1}_K. \tag{53}$$

Using the constraint

$$\mathbb{1}_K^\top\mathbf{z}_i=1, \tag{54}$$

we solve the Karush-Kuhn-Tucker (KKT) conditions independently for each $\mathbf{z}_i$ and finally obtain

$$\mathbf{z}_i^{(l+1)}=\frac{\hat{\mathbf{y}}_i\odot\exp(\log(\mathbf{p}_i)+\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_j^{(l)})}{(\hat{\mathbf{y}}_i\odot\exp(\log(\mathbf{p}_i)+\sum_{j\in\mathcal{Q}}w_{ij}\mathbf{z}_j^{(l)}))^\top\mathbb{1}_K}. \tag{55}$$

Notice that the obtained $\mathbf{z}$-updates are decoupled, yielding computationally efficient transduction for large-scale datasets (see runtime in Table 3).

E. Text prompts

We use the same text prompts for all our experiments. They are given in Table 5.

F. Implementation details for online test-time adaptation

For generating non i.i.d. data streams, we follow the setup of recent works [36] and adopt a framework based on Dirichlet distributions. Namely, we distribute each class over a fixed number of slots according to proportions drawn following a Dirichlet distribution parametrized by a single scalar parameter γ. Therefore, for large values of γ each class is evenly distributed among slots (i.i.d data stream) while for small values each class is distributed in a single slot (highly correlated data stream). Then, samples are randomly shuffled within each slot. For every dataset and a given batch size, the number of slots is $\min\{K,\left\lfloor\frac{|Q|}{\text{batch size}}\right\rfloor\}$. This is illustrated in Figure 4.

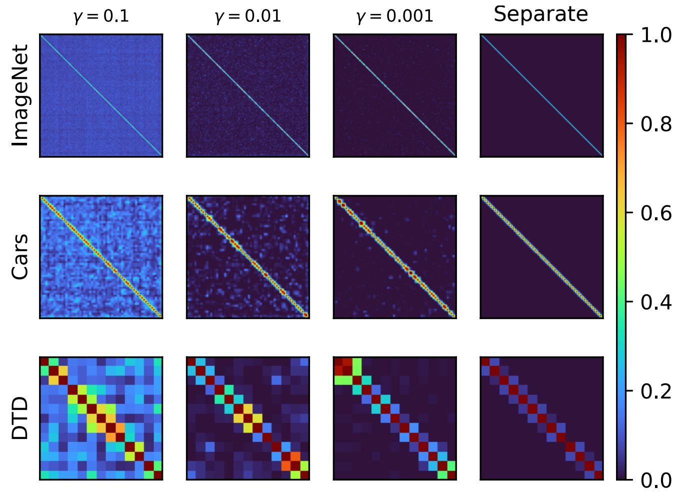

Figure 4. Correlation matrix of per-batch ℓ_2 normalized vectors of class proportions for batch size 128. x and y axis of each plot is the batch index corresponding to the order in which the batches are processed. This illustrates the inter-batch correlation increasing as the Dirichlet parameter γ decreases.

Table 5. Prompt templates used in all experiments.

Dataset	Prompt template
ImageNet	`"a photo of a [].”`
SUN397	`"a photo of a [].”`
Aircraft	`"a photo of a [], a type of aircraft.”,`
EuroSAT	`"a centered satellite photo of [].”,`
Cars	`"a photo of a [].”,`
Food101	`"a photo of [], a type of food.”,`
Pets	`"a photo of [], a type of pet.”,`
Flower102	`"a photo of a [], a type of flower.”,`
Caltech101	`"a photo of a [].”,`
DTD	`"[] texture.”,`
UCF101	`"a photo of a person doing [].”,`

Table 6. Additional information on the datasets.

Dataset name	Other given name	# classes	# test samples	task description
SUN397	Sun397	397	19,850	scenes classification
Aircraft	FGVCAircraft	100	3,333	aircraft classification
EuroSAT	EuroSAT	10	8,100	satellite images classification
Cars	StanfordCars	196	8,041	cars classification
Food101	Food101	101	30,300	food classification
Pets	OxfordPets	37	3,669	pets classification
Flowers102	OxfordFlowers	102	2,463	flowers classification
Caltech101	Caltech101	101	2,465	objects classification
DTD	DTD	47	1,692	textures classification
UCF101	UCF101	101	3,783	actions classification
ImageNet	ImageNet	1000	50,000	objects classification

G. Additional results with other backbones

We present results on four additional CLIP encoders: two convolutional neural networks (ResNet-50 and ResNet-101) and two transformer-based architectures (ViT-B/32 and ViT-L/14), aiming to demonstrate that the findings in the main paper generalize well to other model choices. For batch test-time adaptation (see Tables 7, 8, and 9), we observe consistent improvements across various architectures and model sizes. Similarly, for online test-time adaptation (see Table 10), the results show that the observed improvements remain consistent regardless of the architecture or model size.

H. Additional results with other batch sizes

We present results on intermediate batch sizes (128, 256 and 500) with the ViT-B/16 backbone in Table 11 to demonstrate the robustness of Stat$\mathcal{A}$.

I. Additional results for another realistic scenario

We present an additional scenario (Table 12) in which the number of effective classes K_{eff} is randomly selected within a range from one to the minimum between the batch size and total number of classes in the dataset ($K_{\text{eff}} \in (1, \min(\text{batch\_size}, \text{total\_classes}))$). It can be seen as an average of the different scenarios `Low`, `Medium` and `High` introduced in the main paper. This new scenario is evaluated under various batch sizes (64, 128, 256, 500, 1000, 2000). We can observe that even in this challenging scenario, Stat$\mathcal{A}$ provides stable improvements in the vast majority of cases.

Table 7. Comparison of various CLIP encoders for the batch test-time adaptation setting with a batch size of 64. Each reported accuracy is averaged over 1,000 tasks. Subscripts indicate improvement or degradation compared to zero-shot.

(a) ResNet-50.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	58.7	58.2	58.9	17.0	36.2	55.8	77.4	85.7	66.1	85.7	42.8	61.8
`Very Low`	Stat$\mathcal{A}$	$\mathbf{65.8}_{+7.1}$	$68.2_{+10.0}$	$63.7_{+4.8}$	$21.1_{+4.1}$	$43.3_{+7.1}$	$71.3_{+15.5}$	$87.1_{+9.7}$	$93.1_{+7.4}$	$74.1_{+8.0}$	$90.1_{+4.4}$	$45.3_{+2.5}$	$66.2_{+4.4}$
`Low`	Stat$\mathcal{A}$	$\mathbf{62.3}_{+3.7}$	$65.0_{+6.8}$	$62.8_{+3.9}$	$17.8_{+0.8}$	$31.7_{-4.5}$	$67.1_{+11.3}$	$83.6_{+6.2}$	$88.2_{+2.5}$	$71.7_{+5.6}$	$89.0_{+3.3}$	$44.9_{+2.1}$	$64.0_{+2.2}$
`Medium`	Stat$\mathcal{A}$	$\mathbf{58.6}_{-0.1}$	$61.1_{+2.9}$	$59.5_{+0.6}$	$16.4_{-0.6}$	$27.3_{-8.9}$	$62.1_{+6.3}$	$78.7_{+1.3}$	$81.4_{-4.3}$	$64.1_{-2.0}$	$87.9_{+2.2}$	$43.8_{+1.0}$	$62.0_{+0.2}$

(b) ResNet-101.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	59.5	61.3	59.0	17.9	32.9	63.2	80.7	86.9	64.3	89.9	37.3	61.1
`Very Low`	Stat$\mathcal{A}$	$\mathbf{66.2}_{+6.7}$	$73.0_{+11.7}$	$66.5_{+7.5}$	$22.5_{+4.6}$	$30.4_{-2.5}$	$76.2_{+13.0}$	$89.5_{+8.8}$	$95.2_{+8.3}$	$74.6_{+10.3}$	$91.9_{+2.0}$	$42.9_{+5.6}$	$65.1_{+4.0}$
`Low`	Stat$\mathcal{A}$	$\mathbf{65.1}_{+5.6}$	$71.2_{+9.9}$	$65.9_{+6.9}$	$20.0_{+2.1}$	$29.6_{-3.3}$	$73.1_{+9.9}$	$88.1_{+7.4}$	$92.9_{+6.0}$	$74.9_{+10.6}$	$92.8_{+2.9}$	$42.9_{+5.6}$	$64.4_{+3.3}$
`Medium`	Stat$\mathcal{A}$	$\mathbf{62.5}_{+3.0}$	$67.0_{+5.7}$	$62.7_{+3.7}$	$18.6_{+0.7}$	$28.7_{-4.2}$	$69.6_{+6.4}$	$84.9_{+4.2}$	$88.8_{+1.9}$	$70.1_{+5.8}$	$92.1_{+2.2}$	$40.9_{+3.6}$	$63.6_{+2.5}$

(c) ViT-B/32.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	61.9	62.0	62.1	19.1	45.4	60.2	80.4	87.3	66.6	91.4	42.7	63.5
`Very Low`	Stat$\mathcal{A}$	$\mathbf{67.3}_{+5.4}$	$68.1_{+6.1}$	$65.6_{+3.5}$	$23.0_{+3.9}$	$53.2_{+7.8}$	$71.9_{+11.7}$	$85.8_{+5.4}$	$94.3_{+7.0}$	$74.9_{+8.3}$	$93.4_{+2.0}$	$45.2_{+2.5}$	$64.5_{+1.0}$
`Low`	Stat$\mathcal{A}$	$\mathbf{66.4}_{+4.5}$	$67.2_{+5.2}$	$65.7_{+3.6}$	$21.9_{+2.8}$	$50.1_{+4.7}$	$69.3_{+9.1}$	$84.5_{+4.1}$	$92.5_{+5.2}$	$75.3_{+8.7}$	$93.2_{+1.8}$	$46.1_{+3.4}$	$64.6_{+1.1}$
`Medium`	Stat$\mathcal{A}$	$\mathbf{64.3}_{+2.4}$	$65.5_{+3.5}$	$64.8_{+2.7}$	$20.0_{+0.9}$	$45.4_{0.0}$	$65.0_{+4.8}$	$82.6_{+2.2}$	$89.4_{+2.1}$	$70.6_{+4.0}$	$92.9_{+1.5}$	$46.7_{+4.0}$	$64.4_{+0.9}$

(d) ViT-L/14.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	72.6	73.5	67.7	32.5	60.3	76.9	90.9	93.5	79.5	95.2	53.5	74.9
`Very Low`	Stat$\mathcal{A}$	$\mathbf{77.3}_{+4.7}$	$78.9_{+5.4}$	$71.3_{+3.6}$	$40.4_{+7.9}$	$71.4_{+11.1}$	$84.4_{+7.5}$	$94.2_{+3.3}$	$97.1_{+3.6}$	$82.9_{+3.4}$	$97.0_{+1.8}$	$55.3_{+1.8}$	$77.1_{+2.2}$
`Low`	Stat$\mathcal{A}$	$\mathbf{76.1}_{+3.5}$	$78.2_{+4.7}$	$71.6_{+3.9}$	$38.4_{+5.9}$	$65.6_{+5.3}$	$82.4_{+5.5}$	$93.1_{+2.2}$	$96.3_{+2.8}$	$82.8_{+3.3}$	$96.1_{+0.9}$	$55.4_{+1.9}$	$76.8_{+1.9}$
`Medium`	Stat$\mathcal{A}$	$\mathbf{74.5}_{+2.0}$	$76.6_{+3.1}$	$70.0_{+2.3}$	$36.4_{+3.9}$	$62.6_{+2.3}$	$80.6_{+3.7}$	$92.1_{+1.2}$	$93.9_{+0.4}$	$80.8_{+1.3}$	$95.6_{+0.4}$	$54.6_{+1.1}$	$77.1_{+2.2}$

Table 8. Comparison of various CLIP encoders for the batch test-time adaptation setting with a batch size of 1,000. Each reported accuracy is averaged over 1,000 tasks. Subscripts indicate improvement or degradation compared to zero-shot.

(a) ResNet-50.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	58.7	58.2	58.9	17.0	36.2	55.8	77.4	85.7	66.1	85.7	42.8	61.8
`Medium`	Stat$\mathcal{A}$	**64.1$_{+5.4}$**	65.2$_{+7.0}$	61.5$_{+2.6}$	18.6$_{+1.6}$	51.2$_{+15.0}$	67.2$_{+11.4}$	80.9$_{+3.5}$	89.1$_{+3.4}$	70.7$_{+4.6}$	88.5$_{+2.8}$	46.8$_{+4.0}$	65.3$_{+3.5}$
`High`	Stat$\mathcal{A}$	**63.5$_{+4.8}$**	65.4$_{+7.2}$	63.1$_{+4.2}$	16.5$_{-0.5}$	51.7$_{+15.5}$	65.4$_{+9.6}$	81.0$_{+3.6}$	84.4$_{-1.3}$	70.0$_{+3.9}$	88.3$_{+2.6}$	47.2$_{+4.4}$	66.0$_{+4.2}$
`Very High`	Stat$\mathcal{A}$	**61.8$_{+3.1}$**	63.5$_{+5.3}$	62.4$_{+3.5}$	14.8$_{-2.2}$	51.7$_{+15.5}$	60.8$_{+5.0}$	77.8$_{+0.4}$	83.5$_{-2.2}$	66.2$_{+0.1}$	87.9$_{+2.2}$	46.6$_{+3.8}$	64.5$_{+2.7}$

(b) ResNet-101.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	59.5	61.3	59.0	17.9	32.9	63.2	80.7	86.9	64.3	89.9	37.3	61.1
`Medium`	Stat$\mathcal{A}$	**65.0$_{+5.5}$**	70.5$_{+9.2}$	65.3$_{+6.3}$	20.5$_{+2.6}$	33.6$_{+0.7}$	73.9$_{+10.7}$	85.4$_{+4.7}$	91.1$_{+4.2}$	73.1$_{+8.8}$	92.2$_{+2.3}$	43.2$_{+5.9}$	66.5$_{+5.4}$
`High`	Stat$\mathcal{A}$	**64.3$_{+4.8}$**	71.4$_{+10.1}$	66.2$_{+7.2}$	18.6$_{+0.7}$	32.8$_{-0.1}$	72.2$_{+9.0}$	85.1$_{+4.4}$	87.9$_{+1.0}$	71.9$_{+7.6}$	92.2$_{+2.3}$	42.5$_{+5.2}$	66.5$_{+5.4}$
`Very High`	Stat$\mathcal{A}$	**62.6$_{+3.1}$**	70.1$_{+8.8}$	65.4$_{+6.4}$	16.9$_{-1.0}$	32.9$_{0.0}$	68.2$_{+5.0}$	82.4$_{+1.7}$	87.2$_{+0.3}$	68.7$_{+4.4}$	91.3$_{+1.4}$	41.9$_{+4.6}$	63.8$_{+2.7}$

(c) ViT-B/32.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	61.9	62.0	62.1	19.1	45.4	60.2	80.4	87.3	66.6	91.4	42.7	63.5
`Medium`	Stat$\mathcal{A}$	**65.9$_{+4.0}$**	65.9$_{+3.9}$	63.3$_{+1.2}$	21.9$_{+2.8}$	51.3$_{+5.9}$	69.3$_{+9.1}$	82.2$_{+1.8}$	90.3$_{+3.0}$	74.1$_{+7.5}$	92.6$_{+1.2}$	47.4$_{+4.7}$	66.1$_{+2.6}$
`High`	Stat$\mathcal{A}$	**66.0$_{+4.1}$**	67.0$_{+5.0}$	65.0$_{+2.9}$	20.2$_{+1.1}$	51.1$_{+5.7}$	68.5$_{+8.3}$	82.7$_{+2.3}$	88.5$_{+1.2}$	73.7$_{+7.1}$	92.5$_{+1.1}$	49.5$_{+6.8}$	66.9$_{+3.4}$
`Very High`	Stat$\mathcal{A}$	**65.1$_{+3.2}$**	66.6$_{+4.6}$	66.0$_{+3.9}$	18.8$_{-0.3}$	51.0$_{+5.6}$	65.1$_{+4.9}$	81.5$_{+1.1}$	88.0$_{+0.7}$	70.6$_{+4.0}$	91.9$_{+0.5}$	49.5$_{+6.8}$	66.5$_{+3.0}$

(d) ViT-L/14.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	72.6	73.5	67.7	32.5	60.3	76.9	90.9	93.5	79.5	95.2	53.5	74.9
`Medium`	Stat$\mathcal{A}$	**76.0$_{+3.4}$**	76.2$_{+2.7}$	69.4$_{+1.7}$	39.1$_{+6.6}$	71.0$_{+10.7}$	81.9$_{+5.0}$	91.7$_{+0.8}$	94.8$_{+1.3}$	81.9$_{+2.4}$	95.6$_{+0.4}$	56.9$_{+3.4}$	77.6$_{+2.7}$
`High`	Stat$\mathcal{A}$	**76.3$_{+3.7}$**	77.2$_{+3.7}$	70.9$_{+3.2}$	36.8$_{+4.3}$	71.2$_{+10.9}$	82.0$_{+5.1}$	92.3$_{+1.4}$	94.3$_{+0.8}$	81.9$_{+2.4}$	95.3$_{+0.1}$	58.7$_{+5.2}$	78.8$_{+3.9}$
`Very High`	Stat$\mathcal{A}$	**75.7$_{+3.1}$**	77.3$_{+3.8}$	71.6$_{+3.9}$	33.7$_{+1.2}$	71.2$_{+10.9}$	79.5$_{+2.6}$	91.7$_{+0.8}$	94.1$_{+0.6}$	80.7$_{+1.2}$	94.9$_{-0.3}$	59.0$_{+5.5}$	78.7$_{+3.8}$

Table 9. Comparison of various CLIP encoders for the batch test-time adaptation setting on whole datasets. Each reported accuracy is averaged over 1,000 tasks. Subscripts indicate improvement or degradation compared to zero-shot.

(a) ResNet-50.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	58.7	58.2	58.9	17.0	36.2	55.8	77.4	85.7	66.1	85.7	42.8	61.8
`All`	Stat$\mathcal{A}$	**62.4**$_{+3.7}$	60.4$_{+2.2}$	64.3$_{+5.4}$	16.0$_{-1.0}$	50.5$_{+14.3}$	58.2$_{+2.4}$	77.9$_{+0.5}$	87.7$_{+2.0}$	67.7$_{+1.6}$	87.3$_{+1.6}$	48.5$_{+5.7}$	67.5$_{+5.7}$

(b) ResNet-101.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	59.5	61.3	59.0	17.9	32.9	63.2	80.7	86.9	64.3	89.9	37.3	61.1
`All`	Stat$\mathcal{A}$	**63.6**$_{+4.1}$	64.4$_{+3.1}$	64.9$_{+5.9}$	18.3$_{+0.4}$	43.3$_{+10.4}$	66.2$_{+3.0}$	81.9$_{+1.2}$	88.8$_{+1.9}$	69.1$_{+4.8}$	91.5$_{+1.6}$	42.9$_{+5.6}$	67.8$_{+6.7}$

(c) ViT-B/32.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	61.9	62.0	62.1	19.1	45.4	60.2	80.4	87.3	66.6	91.4	42.7	63.5
`All`	Stat$\mathcal{A}$	**65.5**$_{+3.7}$	64.6$_{+2.6}$	67.1$_{+5.0}$	19.7$_{+0.6}$	54.3$_{+8.9}$	62.5$_{+2.3}$	81.4$_{+1.0}$	89.2$_{+1.9}$	71.5$_{+4.9}$	91.5$_{+0.1}$	50.8$_{+8.1}$	68.4$_{+4.9}$

(d) ViT-L/14.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	72.6	73.5	67.7	32.5	60.3	76.9	90.9	93.5	79.5	95.2	53.5	74.9
`All`	Stat$\mathcal{A}$	**76.7**$_{+4.1}$	76.9$_{+3.4}$	72.8$_{+5.1}$	35.4$_{+2.9}$	76.7$_{+16.4}$	78.0$_{+1.1}$	91.8$_{+0.9}$	94.6$_{+1.1}$	81.8$_{+2.3}$	95.0$_{-0.2}$	59.7$_{+6.2}$	81.3$_{+6.4}$

Table 10. Comparison of various CLIP encoders for the online test-time adaptation setting with a batch size of 128. Each reported accuracy is averaged over 100 tasks. Subscripts indicate improvement or degradation compared to zero-shot.

(a) ResNet-50.

Scenario	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	58.7	58.2	58.9	17.0	36.2	55.8	77.4	85.7	66.1	85.7	42.8	61.8
`Low`	Stat$\mathcal{A}$	58.4$_{-0.3}$	54.6$_{-3.6}$	56.6$_{-2.3}$	15.1$_{-1.9}$	39.7$_{+3.5}$	57.6$_{+1.8}$	79.4$_{+2.0}$	85.1$_{-0.6}$	60.7$_{-5.4}$	87.8$_{+2.1}$	44.4$_{+1.6}$	61.7$_{-0.1}$
`Medium`	Stat$\mathcal{A}$	**62.8**$_{+4.1}$	59.6$_{+1.4}$	60.8$_{+1.9}$	17.7$_{+0.7}$	43.5$_{+7.3}$	65.9$_{+10.1}$	84.5$_{+7.1}$	90.6$_{+4.9}$	68.1$_{+2.0}$	89.3$_{+3.6}$	45.5$_{+2.7}$	64.5$_{+2.7}$
`High`	Stat$\mathcal{A}$	**64.3**$_{+5.6}$	64.7$_{+6.5}$	62.6$_{+3.7}$	18.5$_{+1.5}$	43.6$_{+7.4}$	68.5$_{+12.7}$	85.8$_{+8.4}$	92.2$_{+6.5}$	70.1$_{+4.0}$	89.8$_{+4.1}$	45.9$_{+3.1}$	65.2$_{+3.4}$
`Separate`	Stat$\mathcal{A}$	**65.1**$_{+6.4}$	66.6$_{+8.4}$	62.6$_{+3.7}$	19.8$_{+2.8}$	44.3$_{+8.1}$	69.5$_{+13.7}$	85.6$_{+8.2}$	93.8$_{+8.1}$	71.9$_{+5.8}$	90.2$_{+4.5}$	46.0$_{+3.2}$	65.3$_{+3.5}$

(b) ResNet-101.

Scenario	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	59.5	61.3	59.0	17.9	32.9	63.2	80.7	86.9	64.3	89.9	37.3	61.1
`Low`	Stat$\mathcal{A}$	**61.3**$_{+1.8}$	60.5$_{-0.8}$	59.3$_{+0.3}$	16.9$_{-1.0}$	32.7$_{-0.2}$	65.5$_{+2.3}$	84.9$_{+4.2}$	91.0$_{+4.1}$	67.8$_{+3.5}$	92.2$_{+2.3}$	41.1$_{+3.8}$	62.8$_{+1.7}$
`Medium`	Stat$\mathcal{A}$	**64.6**$_{+5.1}$	66.1$_{+4.8}$	64.2$_{+5.2}$	19.7$_{+1.8}$	33.3$_{+0.4}$	72.2$_{+9.0}$	88.1$_{+7.4}$	94.1$_{+7.2}$	72.1$_{+7.8}$	93.2$_{+3.3}$	42.9$_{+5.6}$	65.2$_{+4.1}$
`High`	Stat$\mathcal{A}$	**65.7**$_{+6.2}$	70.5$_{+9.2}$	65.9$_{+6.9}$	20.6$_{+2.7}$	33.5$_{+0.6}$	74.1$_{+10.9}$	88.7$_{+8.0}$	94.4$_{+7.5}$	73.1$_{+8.8}$	93.4$_{+3.5}$	43.0$_{+5.7}$	65.7$_{+4.6}$
`Separate`	Stat$\mathcal{A}$	**65.8**$_{+6.3}$	71.4$_{+10.1}$	65.7$_{+6.7}$	22.1$_{+4.2}$	32.2$_{-0.7}$	74.9$_{+11.7}$	88.5$_{+7.8}$	94.2$_{+7.3}$	73.9$_{+9.6}$	93.4$_{+3.5}$	41.9$_{+4.6}$	65.7$_{+4.6}$

(c) ViT-B/32.

Scenario	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	61.9	62.0	62.1	19.1	45.4	60.2	80.4	87.3	66.6	91.4	42.7	63.5
`Low`	Stat$\mathcal{A}$	**63.9**$_{+2.0}$	61.4$_{-0.6}$	62.7$_{+0.6}$	19.2$_{+0.1}$	51.0$_{+5.6}$	61.8$_{+1.6}$	82.6$_{+2.2}$	91.0$_{+3.7}$	69.0$_{+2.4}$	92.9$_{+1.5}$	46.4$_{+3.7}$	64.4$_{+0.9}$
`Medium`	Stat$\mathcal{A}$	**65.8**$_{+3.9}$	64.6$_{+2.6}$	64.8$_{+2.7}$	21.4$_{+2.3}$	49.9$_{+4.5}$	68.1$_{+7.9}$	84.4$_{+4.0}$	92.8$_{+5.5}$	72.5$_{+5.9}$	93.5$_{+2.1}$	46.4$_{+3.7}$	65.5$_{+2.0}$
`High`	Stat$\mathcal{A}$	**66.4**$_{+4.6}$	66.9$_{+4.9}$	64.9$_{+2.8}$	22.0$_{+2.9}$	50.1$_{+4.7}$	69.9$_{+9.7}$	84.6$_{+4.2}$	93.2$_{+5.9}$	73.5$_{+6.9}$	93.7$_{+2.3}$	46.3$_{+3.6}$	65.6$_{+2.1}$
`Separate`	Stat$\mathcal{A}$	**65.9**$_{+4.0}$	67.0$_{+5.0}$	63.8$_{+1.7}$	22.9$_{+3.8}$	44.9$_{-0.5}$	70.4$_{+10.2}$	84.1$_{+3.7}$	92.8$_{+5.5}$	74.6$_{+8.0}$	94.0$_{+2.6}$	45.1$_{+2.4}$	65.0$_{+1.5}$

(d) ViT-L/14.

Scenario	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	72.6	73.5	67.7	32.5	60.3	76.9	90.9	93.5	79.5	95.2	53.5	74.9
`Low`	Stat$\mathcal{A}$	**74.3**$_{+1.7}$	73.3$_{-0.2}$	68.2$_{+0.5}$	34.1$_{+1.6}$	68.8$_{+8.5}$	77.7$_{+0.8}$	92.0$_{+1.1}$	95.0$_{+1.5}$	80.2$_{+0.7}$	95.6$_{+0.4}$	55.4$_{+1.9}$	76.9$_{+2.0}$
`Medium`	Stat$\mathcal{A}$	**76.0**$_{+3.4}$	75.8$_{+2.3}$	70.6$_{+2.9}$	38.3$_{+5.8}$	68.9$_{+8.6}$	81.9$_{+5.0}$	93.2$_{+2.3}$	96.3$_{+2.8}$	81.5$_{+2.0}$	95.8$_{+0.6}$	55.6$_{+2.1}$	77.6$_{+2.7}$
`High`	Stat$\mathcal{A}$	**76.4**$_{+3.8}$	77.6$_{+4.1}$	71.1$_{+3.4}$	39.6$_{+7.1}$	68.9$_{+8.6}$	82.9$_{+6.0}$	93.5$_{+2.6}$	96.5$_{+3.0}$	81.9$_{+2.4}$	95.7$_{+0.5}$	55.5$_{+2.0}$	77.5$_{+2.6}$
`Separate`	Stat$\mathcal{A}$	**76.1**$_{+3.6}$	77.6$_{+4.1}$	70.5$_{+2.8}$	41.3$_{+8.8}$	66.3$_{+6.0}$	83.2$_{+6.3}$	93.5$_{+2.6}$	96.3$_{+2.8}$	82.0$_{+2.5}$	95.8$_{+0.6}$	54.5$_{+1.0}$	76.8$_{+1.9}$

Table 11. Comparison of different batch sizes with the ViT-B/16 backbone. Each reported accuracy is averaged over 1,000 tasks.

(a) Batch Size: 128.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
`Medium`	Stat$\mathcal{A}$	**68.5**$_{\mathbf{+3.3}}$	72.0$_{+5.4}$	66.5$_{+4.0}$	26.6$_{+1.9}$	48.2$_{-0.1}$	72.7$_{+7.1}$	88.7$_{+2.8}$	92.1$_{+3.0}$	75.4$_{+4.7}$	93.7$_{+0.5}$	47.7$_{+4.2}$	70.0$_{+2.5}$
`High`	Stat$\mathcal{A}$	**66.3**$_{\mathbf{+1.1}}$	69.4$_{+2.8}$	64.9$_{+2.4}$	23.6$_{-1.1}$	47.2$_{-1.1}$	68.0$_{+2.4}$	87.0$_{+1.1}$	88.2$_{-0.9}$	72.0$_{+1.3}$	94.0$_{+0.8}$	46.9$_{+3.4}$	68.2$_{+0.7}$

(b) Batch Size: 256.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
`Medium`	Stat$\mathcal{A}$	**69.5**$_{\mathbf{+4.3}}$	72.0$_{+5.4}$	66.7$_{+4.2}$	27.1$_{+2.4}$	56.0$_{+7.7}$	74.1$_{+8.5}$	88.9$_{+3.0}$	92.9$_{+3.8}$	76.0$_{+5.3}$	93.6$_{+0.4}$	47.0$_{+3.5}$	70.5$_{+3.0}$
`High`	Stat$\mathcal{A}$	**68.1**$_{\mathbf{+2.9}}$	71.1$_{+4.5}$	66.3$_{+3.8}$	24.2$_{-0.5}$	55.5$_{+7.2}$	70.6$_{+5.0}$	87.6$_{+1.7}$	88.9$_{-0.2}$	73.7$_{+3.0}$	94.1$_{+0.9}$	47.0$_{+3.5}$	69.9$_{+2.4}$

(c) Batch Size: 500.

K_{eff}	Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
	CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
`Medium`	Stat$\mathcal{A}$	**69.8**$_{\mathbf{+4.5}}$	71.5$_{+4.9}$	65.5$_{+3.0}$	27.8$_{+3.1}$	59.3$_{+11.0}$	74.9$_{+9.3}$	88.3$_{+2.4}$	93.1$_{+4.0}$	76.8$_{+6.1}$	93.1$_{-0.1}$	47.1$_{+3.6}$	69.9$_{+2.4}$
`High`	Stat$\mathcal{A}$	**69.3**$_{\mathbf{+4.1}}$	72.1$_{+5.5}$	67.3$_{+4.8}$	25.1$_{+0.4}$	60.0$_{+11.7}$	72.3$_{+6.7}$	88.2$_{+2.3}$	90.3$_{+1.2}$	75.5$_{+4.8}$	93.8$_{+0.6}$	47.2$_{+3.7}$	70.7$_{+3.2}$

Table 12. Comparison of different batch sizes. Scenario with $K_{\text{eff}} \in (1, \min(\text{batch\_size}, \#\text{total\_classes}))$. The best average accuracy for each configuration is highlighted in **bold**, while the second-best is indicated with underline. Subscripts indicate improvement or degradation compared to zero-shot. Each reported performance is averaged over 1,000 tasks.

(a) Batch Size: 64.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**66.9**$_{+1.7}$	68.7$_{+2.1}$	64.8$_{+2.3}$	24.1$_{-0.6}$	50.6$_{+2.3}$	68.9$_{+3.3}$	87.1$_{+1.2}$	90.8$_{+1.7}$	72.1$_{+1.4}$	93.8$_{+0.6}$	46.6$_{+3.1}$	68.7$_{+1.2}$

(b) Batch Size: 128.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**66.8**$_{+1.6}$	68.6$_{+2.0}$	64.3$_{+1.8}$	23.6$_{-1.1}$	53.3$_{+5.0}$	67.5$_{+1.9}$	86.9$_{+1.0}$	91.1$_{+2.0}$	71.5$_{+0.8}$	93.8$_{+0.6}$	46.9$_{+3.4}$	67.9$_{+0.4}$

(c) Batch Size: 256.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**67.3**$_{+2.1}$	68.2$_{+1.6}$	64.1$_{+1.6}$	24.1$_{-0.6}$	55.3$_{+7.0}$	66.3$_{+0.7}$	87.4$_{+1.5}$	91.9$_{+2.8}$	73.2$_{+2.5}$	93.8$_{+0.6}$	47.4$_{+3.9}$	68.7$_{+1.2}$

(d) Batch Size: 500.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**67.6**$_{+2.3}$	68.1$_{+1.5}$	64.2$_{+1.7}$	24.9$_{+0.2}$	54.5$_{+6.2}$	67.2$_{+1.6}$	87.5$_{+1.6}$	92.5$_{+3.4}$	74.2$_{+3.5}$	93.5$_{+0.3}$	47.1$_{+3.6}$	69.4$_{+1.9}$

(e) Batch Size: 1000.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**68.0**$_{+2.8}$	67.5$_{+0.9}$	65.2$_{+2.7}$	25.4$_{+0.7}$	55.0$_{+6.7}$	68.0$_{+2.4}$	87.2$_{+1.3}$	93.0$_{+3.9}$	75.3$_{+4.6}$	93.3$_{+0.1}$	47.8$_{+4.3}$	70.7$_{+3.2}$

(f) Batch Size: 2000.

Method	AVERAGE	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
CLIP	65.2	66.6	62.5	24.7	48.3	65.6	85.9	89.1	70.7	93.2	43.5	67.5
Stat$\mathcal{A}$	**68.7**$_{+3.5}$	68.1$_{+1.5}$	66.1$_{+3.6}$	26.3$_{+1.6}$	56.7$_{+8.4}$	69.6$_{+4.0}$	86.7$_{+0.8}$	93.0$_{+3.9}$	77.0$_{+6.3}$	93.3$_{+0.1}$	47.4$_{+3.9}$	71.5$_{+4.0}$