EXACT: ESTIMATION–SUPPORT ASYMMETRY IN TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) uses unlabeled target observations to improve model performance under distribution shift, often by estimating target-side statistics from the observed data. Existing methods mainly focus on how these statistics should be estimated or optimized, while paying less attention to whether the available target data are sufficient for reliable estimation. Different target statistics, however, have different data requirements. Global statistics are estimated from all observed target samples, whereas class-specific statistics depend on samples from the corresponding classes. We refer to this difference as estimation–support asymmetry. Our analysis and experiments show that their relative reliability varies with the amount of available target data: global statistics are more reliable when data are limited, whereas class-specific statistics become more effective as target samples accumulate. These findings motivate an inference strategy that ties the use of target statistics to the data available for their estimation. We instantiate this principle in EXACT, a framework that uses global statistics when target data are limited and enables target prototype estimation as observations accumulate. A label-free, coverage-based criterion derived from our class coverage analysis determines when class-specific statistics are used. EXACT requires no source data, target labels, backpropagation, or updates to the pretrained model. Across seven corruption and natural-shift benchmarks, EXACT achieves the highest average accuracy among the compared methods in all four backbone/inference settings. Reliability and ablation analyses provide further evidence for the estimation–support perspective and its coverage-based implementation.

1 INTRODUCTION

Test-time adaptation (TTA) aims to improve model performance under distribution shift using unlabeled observations during inference (Liang et al., 2025; You et al., 2025). Existing approaches use these observations to update model parameters or adjust predictions through target-side statistics and structure (Boudiaf et al., 2022; Li et al., 2024; Dai & Yang, 2025). In online inference, however, the available target data are not fixed (Fuchs et al., 2025). Observations arrive in batches, and the number of target samples for estimating target information grows over time (Zhang et al., 2024; Wang et al., 2025). This raises a practical question: *which target statistics should be used given the amount of available target data for their estimation?*

Fig. 1 illustrates the different sample requirements of target statistics considered in this work. Different target statistics require different amounts of target data for reliable estimation (Kang et al., 2024; Danilowski et al., 2026). Global statistics are estimated from the full observed target set (Niloy et al., 2024; Murphy et al., 2026). Class-specific statistics, such as target prototypes, rely on observations from the corresponding classes (Iwasawa & Matsuo, 2021; Zhang et al., 2025; Dai & Yang, 2025). We refer to this difference as *estimation–support asymmetry*. With limited target data, the observed target set may cover only a subset of the label space. This is a common situation in realistic TTA (Mounsaveng et al., 2024; Zanella et al., 2025; Weerasinghe et al., 2026). Nevertheless, a class prototype computed as a weighted average over the observed target samples can remain numerically defined even when the corresponding class is absent from the observed target set (Dai & Yang, 2025). Such class-specific statistics can therefore be unreliable, but become more reliable as target observations accumulate and class coverage improves.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

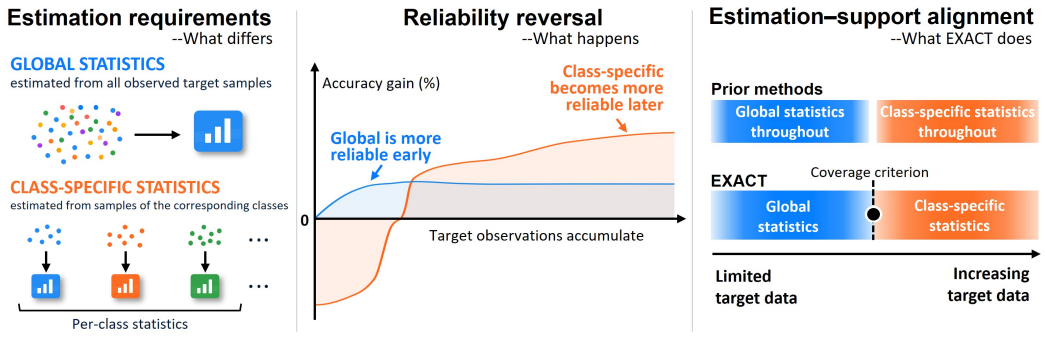

Figure 1: Motivation of EXACT. **Left:** Global statistics and class-specific statistics have different estimation requirements. **Middle:** Reliability reversal on ImageNet-Sketch with ViT-Base. Prediction using global statistics is more reliable with limited target samples, while class-specific statistics become more reliable as target samples accumulate. **Right:** Most representative approaches use the same statistic throughout inference, whereas EXACT aligns the statistic type with the available target samples.

The middle part of Fig. 1 shows that the relative prediction accuracy gain using global and class-specific statistics changes with the available target sample size. In this example, predictions based on global statistics are more accurate when few target samples are available, whereas class-specific predictions achieve higher accuracy at larger sample sizes. The implication is not simply that more data improve estimation: the amount of target data can change which statistics should be used for prediction. As summarized in the right part of Fig. 1, representative TTA methods typically use the same type of target statistics throughout the inference. This fixed use of target statistics does not account for changes in their reliability.

Previous studies have recognized that target estimates can be unreliable when few target samples are available. NEO uses a global feature center to avoid unreliable class-wise estimation (Murphy et al., 2026), while StatA regularizes target statistics toward an initial anchor (Zanella et al., 2025). SCA instead accumulates compact feature statistics over previously observed test samples (Guan et al., 2025). These studies establish the importance of accounting for limited target data. We focus on the difference in sample requirements between global and class-specific statistics, and examine how this difference should affect their use in prediction.

We propose support-dependent inference and instantiate it in EXACT (EXpectation-guided Alignment between Class Coverage and Target Statistics), a frozen-model framework for test-time adaptation. EXACT uses global statistics when target data are limited and enables target prototype estimation as observations accumulate. A label-free, coverage-based criterion derived from our class coverage analysis determines when class-specific statistics are used.

Our contributions are threefold:

- We identify and analyze estimation–support asymmetry in frozen-model TTA, showing that global and class-specific target statistics have different sample requirements and exhibit a reversal in relative reliability.
- We develop EXACT, a framework that uses global and class-specific target statistics according to the available target data, with a label-free criterion based on class coverage analysis determining their use.
- We evaluate EXACT on seven corruption and natural-shift benchmarks and provide systematic analyses of reliability across target support, prototype estimation, coverage-based criterion, class imbalance, and backbone generalization.

2 Related Work

Test-Time Adaptation. Test-time adaptation (TTA) uses unlabeled target observations to improve predictions under distribution shift (Xiao & Snoek, 2024). Early TTA methods optimize normalization parameters (Wang et al., 2021) or prompts (Shu et al., 2022) through entropy minimization. Motivated by the computational overhead of test-time optimization, a growing line of work instead adapts inference without backpropagation or updates to pretrained weights. These works include feature caching (Karmanov et al., 2024), prediction regularization (Boudiaf et al., 2022), or label propagation (Li et al., 2025). EXACT follows this frozen-model setting.

Target Statistics for Adaptation. Different adaptation methods use different forms of target statistics. NEO (Murphy et al., 2026) re-centers features using global target mean, whereas T3A (Iwasawa & Matsuo, 2021) constructs class prototypes from pseudo-labeled observations. LAME (Boudiaf et al., 2022) regularizes predictions using sample affinities, and ECALP (Li et al., 2025) performs context-aware label propagation on an expanding graph. For vision–language models, BCA (Zhou et al., 2025) jointly adapts class embeddings and embedding class priors. D^2O (Luo et al., 2026) instead debiases retrieval features and environment-dependent logits for existing adapters.

Target Estimation with Limited Data. Several studies have recognized that target estimates can be unreliable when target data are limited. StatA (Zanella et al., 2025) anchors class statistics to text-derived priors according to predicted class support, while ADAPT (Zhang et al., 2025) regularizes inference using CLIP priors and a knowledge bank. DOTA (Han et al., 2025) increases the contribution of its distributional classifier as samples accumulate, whereas SCA (Guan et al., 2025) uses prediction uncertainty for adaptive fusion. However, these methods do not explicitly examine how the relative reliability changes with the amount of available target data. Our focus is the relative accuracy of predictions based on global statistics and target prototypes across target sample sizes, and the use of a class-coverage criterion to determine which statistics are used for prediction.

3 Method

3.1 Estimation–Support Asymmetry

Global and class-specific target statistics have different sample requirements. Consider N i.i.d. target samples from a C-class classification problem, and let N_c denote the number of observed samples whose true class is c. The value of N_c is unknown because the target labels are unavailable. A global empirical mean uses all N target observations, whereas a class-specific empirical mean for class c uses only the N_c observations from that class. For $N_c > 0$ and under finite second moments, their root mean squared estimation errors relative to the corresponding population means satisfy

$$\mathrm{RMSE}(\hat{m}) = O(N^{-1/2}), \mathrm{RMSE}(\hat{m}_c) = O(N_c^{-1/2}). \tag{1}$$

Under balanced sampling, N_c concentrates around N/C as N/C grows, giving the class-specific sampling-error scale $O(\sqrt{C/N})$. This difference in sample requirements motivates tying the use of target statistics to the target data available for their estimation.

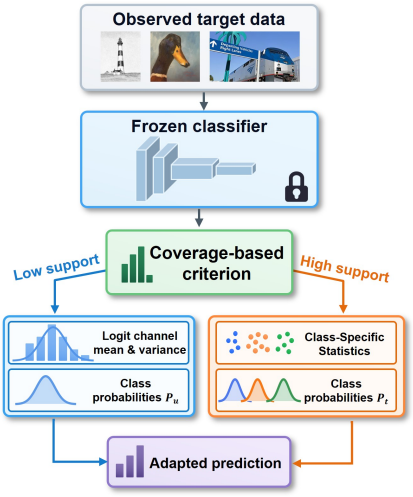

Figure 2: Overview of the proposed method.

3.2 Problem Setup and Overview

Let a pretrained classifier consist of a feature extractor h_θ and a linear classification head (W, b). For an input image $\tilde{x}_i$, the frozen feature and logit vectors are

$$x_i = h_\theta(\tilde{x}_i) \in \mathbb{R}^D, \qquad z_i = Wx_i + b \in \mathbb{R}^C, \tag{2}$$

where D is the feature dimension and C is the number of classes. Both h_θ and (W, b) remain fixed throughout inference. EXACT requires no source data, target labels, gradient computation, or backpropagation, and introduces no additional trainable module.

We consider a target set in which unlabeled observations arrive sequentially as batches $\mathcal{B}_1, \ldots, \mathcal{B}_K$. At step k, the available target observations are

$$\mathcal{D}_{\le k} = \bigcup_{\tau=1}^{k} \mathcal{B}_\tau, \qquad N_k = |\mathcal{D}_{\le k}|. \tag{3}$$

Future batches are inaccessible. The frozen features and logits of the observed samples are collected as

$$X^{(k)} = [x_1, \ldots, x_{N_k}]^\top \in \mathbb{R}^{N_k \times D}, \qquad Z^{(k)} = [z_1, \ldots, z_{N_k}]^\top \in \mathbb{R}^{N_k \times C}. \tag{4}$$

Following the estimation–support asymmetry in Sec. 3.1, EXACT distinguishes two forms of target statistics according to their sample requirements. Global statistics summarize each classifier-response channel over all observed target samples without assigning samples to classes. Class-specific statistics use estimated assignments to construct target class representations, whose estimation depends on observations from the corresponding classes. We use *target support* to denote the amount of observed target data for estimating these statistics.

To account for these different sample requirements, EXACT uses global statistics below a class-coverage-based threshold. Once the threshold is reached, it enables target prototype estimation. Let $P^{(k)} \in [0,1]^{N_k \times C}$ denote the matrix of class probabilities computed by EXACT at step k. Its entry $P^{(k)}_{ic}$ is the probability assigned to class c for observed sample i, and each row sums to one. EXACT computes these probabilities using

$$P^{(k)} = \begin{cases} P^{(k)}_u, & N_k < N_{\mathrm{cov}}(C), \\ P^{(k)}_t, & N_k \ge N_{\mathrm{cov}}(C), \end{cases} \tag{5}$$

where $P^{(k)}_u$ uses global statistics in Sec. 3.4, $P^{(k)}_t$ uses class-specific statistics in Sec. 3.5, and $N_{\mathrm{cov}}(C)$ is the coverage-based threshold defined in Sec. 3.3. We refer to this prediction rule as support-dependent inference.

At each step, the features and logits of the current batch are appended to the cache. Let I_k denote the row indices corresponding to $\mathcal{B}_k$. We output predictions only for the current batch:

$$\hat{y}^{(k)}_i = \arg\max_c P^{(k)}_{ic}, \qquad i \in I_k. \tag{6}$$

Previously emitted predictions are not revised. Algorithm 1 summarizes the online procedure. For readability, we omit the superscript (k) in the following derivations and write $N = N_k$. All statistics and class probabilities are computed using only the observations available at the current step.

3.3 Coverage-based Criterion

Class-specific statistics depend on observations from the corresponding classes, but the true class counts are unavailable during TTA. We therefore use a label-free, coverage-based criterion to determine how to use these statistics. Under uniform class probabilities, let U_N denote the number of classes absent from N i.i.d. target observations. Then

$$\mathbb{E}[U_N] = C\left(1 - \frac{1}{C}\right)^N \le C\exp(-N/C). \tag{7}$$

Thus, reducing the expected number of missing classes to a constant requires a sample-size scale of $C \log C$. Class-specific statistics rely on observations from the corresponding classes. When a class is absent from the observed target set, its estimated statistic can be unreliable. We therefore model this process through class coverage. Since target labels are unavailable during TTA, the resulting reference should depend only on the class-space size C. Let T_{cov} denote the first sample count at which all classes have appeared. By decomposing T_{cov} into the waiting times for successive new classes, we obtain

$$N_{\mathrm{cov}}(C) = \mathbb{E}[T_{\mathrm{cov}}] = CH_C, \qquad H_C = \sum_{j=1}^{C} \frac{1}{j}. \tag{8}$$

The derivation is provided in Appendix A.1.2. We use $N_{\text{cov}}(C) = CH_C$ as the coverage-based threshold and enable class-specific statistics when $N \geq N_{\text{cov}}(C)$. It provides a statistically motivated reference for when class-specific statistics have broader class support. And it does not identify the exact accuracy crossover or full class coverage. Its scope under nonuniform class distributions is discussed in Sec. 4.7 and Appendix A.1.4.

3.4 Prediction Using Global Statistics

When $N < N_{\text{cov}}(C)$, EXACT computes class probabilities using global logit statistics, without requiring target class assignments. We first define $\mathcal{S}(\cdot)$ as row-wise z-score standardization, which independently centers each sample's class-score vector by its mean across classes and scales it by its standard deviation across classes. Let m_c^z and v_c^z denote the mean and variance of logit channel c over the observed target samples. We form the centered and variance-normalized responses

$$\ell_{ic} = z_{ic} - m_c^z, \qquad t_{ic} = \frac{\ell_{ic}}{\sqrt{v_c^z}}. \tag{9}$$

Centering removes the mean response of each logit channel, while variance normalization accounts for differences in response scale across channels. The resulting class probabilities are

$$P_u = \text{softmax}\left(\mathcal{S}(t)\right), \tag{10}$$

where softmax is applied independently to each row.

3.5 Prediction Using Class-Specific Statistics

When $N \geq N_{\text{cov}}(C)$, EXACT uses class-specific target statistics to compute class probabilities. The coverage-based criterion determines when these statistics are used. We therefore estimate target prototypes from the frozen classifier's scores and use them to compute P_t.

Estimating a target prototype requires determining how much each unlabeled sample contributes to the corresponding class. We obtain these contributions from two score constructions of the frozen classifier: its logits and the cosine similarity between target features and classifier weights. The latter measures directional compatibility without the effect of feature or weight magnitudes. Let $w_c \in \mathbb{R}^D$ be the classifier weight for class c, and define $\hat{x}_i = x_i/\|x_i\|_2$ and $\hat{w}_c = w_c/\|w_c\|_2$. The feature–weight compatibility is

$$A_{ic} = \hat{x}_i^\top \hat{w}_c. \tag{11}$$

We use Sinkhorn normalization to obtain soft assignments jointly over the observed target set. Let $\text{Sinkhorn}(M)$ denote this operator for a score matrix $M \in \mathbb{R}^{N\times C}$. The corresponding assignments are

$$P^z = \text{Sinkhorn}(Z), \qquad P^w = \text{Sinkhorn}(A). \tag{12}$$

Because prototype estimation is directly affected by incorrect class contributions, we assign a large construction weight only when the two score constructions agree on the corresponding class:

$$\omega_{ic} = P_{ic}^z P_{ic}^w. \tag{13}$$

The product gives a large construction weight only when both assignments place substantial mass on the same class. The resulting target prototype and its normalized direction are

$$\pi_c = \frac{\sum_{i=1}^N \omega_{ic}\hat{x}_i}{\sum_{i=1}^N \omega_{ic}}, \qquad \hat{\pi}_c = \frac{\pi_c}{\|\pi_c\|_2}. \tag{14}$$

To incorporate these target class directions while retaining the frozen classifier scores, we compute the sample–prototype compatibility

$$G_{ic}^p = \hat{x}_i^\top \hat{\pi}_c \tag{15}$$

and compute the class probabilities

$$P_t = \text{Sinkhorn}\left(S(Z) + S(G^p)\right). \tag{16}$$

Each score matrix is standardized independently before addition to account for their different numerical scales.

Table 1: Comparison of test-time adaptation methods under different distribution shifts. The best result is shown in **bold**.

Method	Corruption Shifts			Natural Shifts				Avg.
	CIFAR10-C	CIFAR100-C	ImageNet-C	Sketch	D109	Drawing	Cartoon	
ONLINE INFERENCE								
ViT-Base								
Frozen	84.37	58.47	55.03	44.89	59.09	58.30	76.36	62.36
T3A	84.52	59.17	56.33	48.50	59.33	61.84	76.87	63.79
LAME	84.36	58.34	54.00	45.28	60.60	57.68	75.86	62.30
NEO	85.28	59.09	58.73	46.83	61.22	62.69	76.68	64.36
EXACT	**86.98**	**60.86**	**61.03**	**54.12**	**64.58**	**65.58**	**77.51**	**67.24**
CLIP-Base								
Frozen	63.30	37.07	26.35	48.38	72.61	38.39	56.34	48.92
ECALP	66.57	39.46	28.96	51.91	73.71	41.41	58.58	51.51
StatA	64.18	34.61	24.17	48.99	74.36	37.48	56.42	48.60
BCA	61.03	37.53	26.92	49.98	72.52	38.97	56.64	49.08
ADAPT	61.55	33.98	26.18	50.87	70.21	38.71	56.21	48.24
SCA	65.30	38.52	27.01	51.63	**75.05**	39.69	57.33	50.65
DOTA	65.40	39.31	27.91	**52.94**	74.70	40.69	**58.59**	51.36
D^2O	61.97	34.09	26.71	51.17	70.39	38.83	55.80	48.42
EXACT	**69.58**	**42.55**	**29.39**	52.81	74.10	**41.68**	57.77	**52.55**
TRANSDUCTIVE INFERENCE								
ViT-Base								
Frozen	84.37	58.47	55.03	44.89	59.09	58.30	76.36	62.36
T3A	84.46	59.15	56.80	49.61	58.63	62.26	76.76	63.95
LAME	84.35	58.41	54.87	46.21	60.98	58.78	76.49	62.87
NEO	85.28	59.10	58.73	46.82	61.22	62.70	76.67	64.36
EXACT	**87.13**	**61.18**	**61.93**	**55.87**	**65.36**	**66.89**	**78.69**	**68.15**
CLIP-Base								
Frozen	63.30	37.07	26.35	48.38	72.61	38.39	56.34	48.92
ECALP	67.36	40.04	29.24	52.62	74.04	41.79	59.23	52.05
StatA	64.80	36.85	25.79	50.27	74.59	39.10	57.81	49.89
BCA	59.82	37.48	26.98	49.61	72.47	38.88	56.59	48.83
ADAPT	66.43	40.40	29.08	52.68	75.42	41.18	59.12	52.04
SCA	65.20	38.43	26.13	51.71	75.17	39.21	57.48	50.48
DOTA	65.49	39.88	28.50	53.39	74.85	40.76	58.85	51.67
D^2O	67.17	40.98	29.54	53.20	**75.85**	41.23	58.86	52.40
EXACT	**69.89**	**43.53**	**30.11**	**54.99**	74.62	**42.96**	**59.43**	**53.65**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOLS

Datasets and Metrics. We evaluate EXACT under two types of distribution shifts. For *corruption shifts*, we use CIFAR-10-C, CIFAR-100-C, and ImageNet-C (Hendrycks & Dietterich, 2019), with all corruption types evaluated at severity level 5. For *natural shifts*, we use ImageNet-Sketch (Wang et al., 2019), ImageNet-D109 (Marsden et al., 2024), ImageNet-Drawing, and ImageNet-Cartoon (Salvador & Oberman, 2022). Top-1 accuracy is used as the evaluation metric. All results are averaged over five random seeds.

Baselines and Evaluation Protocol. We compare EXACT against representative test-time adaptation approaches that introduce no additional trainable parameters: T3A (Iwasawa & Matsuo, 2021), LAME (Boudiaf et al., 2022), and NEO (Murphy et al., 2026) for ViT-based models, and ECALP (Li et al., 2025), StatA (Zanella et al., 2025), BCA (Zhou et al., 2025), ADAPT (Zhang et al., 2025), SCA (Guan et al., 2025), DOTA (Han et al., 2025), and D^2O (Luo et al., 2026) for CLIP-based models. We focus on online inference and additionally report transductive results, with all methods compared under the same data-availability protocol within each setting. We use publicly available pretrained models throughout; further details are provided in Appendix A.3.

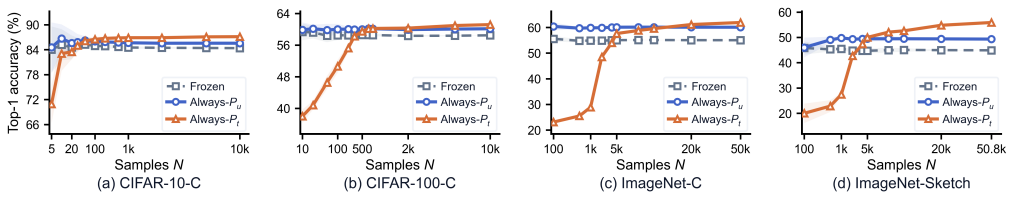

Figure 3: Reliability reversal across target-set sizes. Results are averaged over five random seeds.

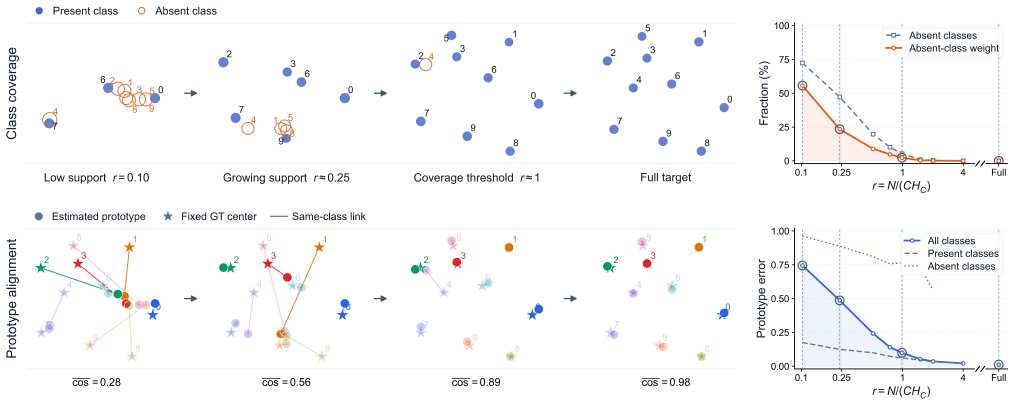

Figure 4: The quantitative curves are averaged over the 15 corruption domains on CIFAR-10-C. The t-SNE panels are constructed using Gaussian noise at severity level 5.

4.2 Main Results

Table 1 summarizes the main comparison across corruption and natural distribution shifts. EXACT achieves the highest average accuracy among the compared methods in all four backbone/inference settings, covering both ViT-Base and CLIP-Base under online and transductive inference. This consistent improvement across backbone families and data-availability protocols supports the effectiveness of EXACT beyond a particular model or inference setting.

4.3 Reliability of Global and Class-Specific Statistics

We examine how the reliability of global and class-specific statistics changes with the available target sample size. We vary the target-set size N and compare frozen model with **Always-**P_u (global statistics) and **Always-**P_t (class-specific statistics), averaging results over five random seeds. Fig. 3 shows a consistent reliability reversal: Always-P_u remains relatively stable, whereas Always-P_t can substantially underperform at low support but eventually surpasses Always-P_u as N increases. The transition shifts from tens of samples on CIFAR-10-C to hundreds on CIFAR-100-C and thousands on the 1000-class ImageNet benchmarks. This pattern is consistent with the support asymmetry analyzed in Sec. 3.1: global statistics use all N observations, whereas class-specific statistics depend on per-class support N_c. Thus, neither statistic type is preferable across all support levels. Results on the remaining benchmarks are provided in Appendix B.4.

4.4 Reliability of Class-Specific Statistics Across Target Support

We examine why class-specific statistics become unreliable under limited target support. To make the amount of available target support comparable across datasets with different numbers of classes, we parameterize N by the normalized target-support level $r = N/(CH_C)$. We measure the prototype-weight mass assigned to absent classes, R_{unsup}, and the prototype error relative to full-target reference centers E_π. Fig. 4 shows substantial weight on absent classes and large prototype errors at low support; both measures decrease as support grows. Prototypes for absent classes are constructed entirely from samples of other classes. Being numerically defined therefore does not imply reliable estimation, consistent with the estimation–support asymmetry. Metric definitions and additional results are provided in Appendix B.5.

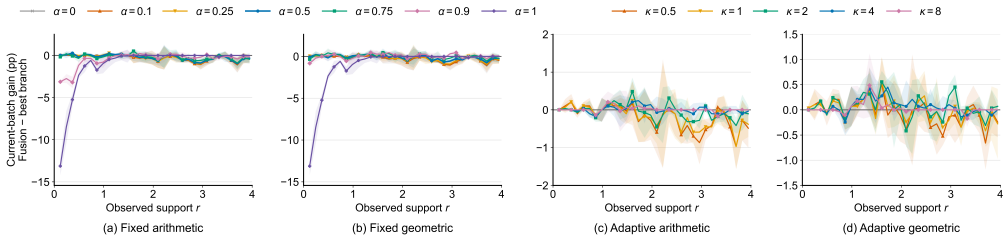

Figure 5: Fixed and adaptive fusion across target support. The preferred fusion weight varies with target support, while positive fusion gains occur mainly near the reliability transition.

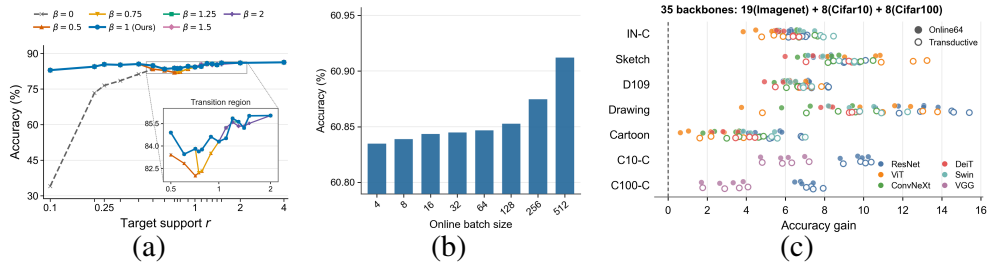

Figure 6: Robustness and generalization of EXACT. (a) Sensitivity to the coverage-based threshold. (b) Sensitivity to the online batch size. (c) Accuracy gain across 35 additional pretrained backbones.

4.5 Analysis of Fusion Strategies

A natural alternative to switching is to combine P_u and P_t throughout inference. We evaluate both fixed and support-adaptive fusion under the online protocol. For fixed fusion, we consider $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1\}$ with both arithmetic and geometric fusion. We further consider support-adaptive fusion, where the weight α increases with the normalized support r as $\alpha_\kappa(r) = \frac{r^\kappa}{1+r^\kappa}$, using different rates κ. For each support level, we report the current-batch accuracy gain of fusion over the better-performing individual branch. Detailed fusion definitions, settings and additional results are provided in Appendix B.6. Fig. 5 shows the results on CIFAR-100-C. Fixed fusion does not consistently improve over the better individual branch, and unsuitable weights can cause substantial degradation when the two branches have different reliability. Support-adaptive weighting reduces these large failures, but the fusion gain still fluctuates around zero for both arithmetic and geometric fusion, with no transition rate consistently producing positive gains across support levels. These results indicate that the tested fusion rules do not consistently outperform the better individual branch at each support level.

4.6 Robustness and Generalization of Support-Dependent Inference

We further examine whether support-dependent inference remains effective under variations in the coverage-based threshold, online batch size, and pretrained backbone.

Threshold sensitivity. We replace the coverage-based threshold CH_C with βCH_C, where $\beta \in \{0, 0.5, 0.75, 1, 1.25, 1.5, 2\}$. As shown in Fig. 6(a), a small β can enable class-specific prediction before sufficient per-class support has accumulated. Errors in these early predictions remain part of the stream-average accuracy because previous predictions are not revised. Conversely, a large β retains P_u after class-specific statistics have become reliable and delays their gains. $\beta = 1$ provides a competitive trade-off across the evaluated settings. More results are in Appendix B.7.

Online batch size. We vary the arrival batch size from 4 to 512 under online inference protocol. Fig. 6(b) shows only minor variation in accuracy across this range. This indicates that EXACT does not depend on a particular online batch size. The result is based on CIFAR-100-C and more results can be found in Appendix B.9.

Backbone generalization. Finally, we apply EXACT to 35 additional pretrained classifiers spanning 6 model families. We evaluate 19 ImageNet classifiers and 8 classifiers for each CIFAR dataset

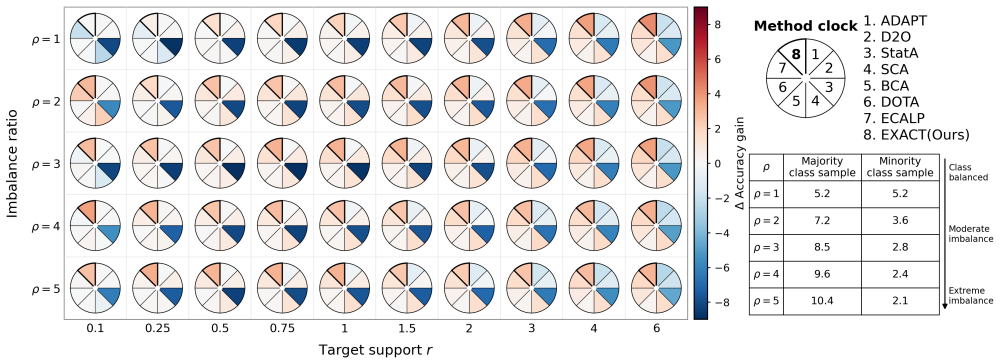

Figure 7: Performance across target support under class imbalance with CLIP-Base. Sector positions identify the methods, and colors indicate Top-1 accuracy gain over the frozen classifier.

and report the accuracy. Fig. 6(c) shows positive gains across all evaluated backbones, target shifts, and inference settings. Thus, the benefit of support-dependent inference is not specific to a particular architecture family or pretrained representation. More results are in Appendix B.2.

4.7 Class Imbalance and the Scope of the Coverage-Based Criterion

The coverage-based threshold $N_{\mathrm{cov}}(C) = CH_C$ is derived under uniform class probabilities, whereas the target class distribution is unknown during TTA. We therefore examine whether EXACT remains effective under imbalanced target class distributions. To construct controlled class imbalance, we randomly permute the C classes and sample target observations with $p_{\pi(j)} \propto \rho^{-j/(C-1)}, j = 0, \ldots, C-1$, giving $p_{\max}/p_{\min} = \rho$. We evaluate $\rho \in \{1, 2, 3, 4, 5\}$ across normalized target-support levels r, with all methods evaluated on the same target streams.

Fig. 7 summarizes the performance trends across target support and class imbalance on CIFAR-100-C. At a fixed target-support level, stronger class imbalance generally makes adaptation more difficult for all compared methods. Nevertheless, EXACT remains consistently competitive across the tested imbalance ratios. Fig. 8 complements this overview with Top-1 accuracy. Additional results on other datasets are provided in Appendix B.8. The results show that, within the nonuniform settings, the same support-dependent inference rule remains effective and retains performance advantage over the compared methods. These results clarify the scope of the uniform-class assumption in our analysis. Under a nonuniform class distribution, class-specific support becomes uneven, so tail classes require more total target observations before their statistics can be estimated reliably. This effect occurs across baselines is not specific to EXACT. We do not interpret CH_C as an exact threshold for arbitrary target class distributions. Rather, it provides a label-free reference that depends only on the number of classes and does not require estimating the unknown target class probabilities.

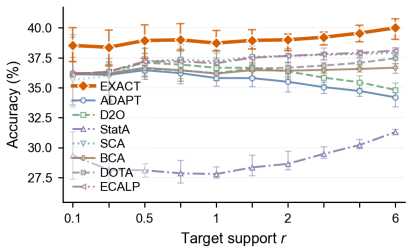

Figure 8: Top-1 accuracy under class imbalance. Markers show mean accuracy, and error bars indicate min–max range.

5 Conclusion

We studied test-time adaptation from the perspective of estimation–support asymmetry, where global and class-specific target statistics have different support requirements. Based on this observation, we introduced EXACT, which uses global statistics under limited target support and enables class-specific statistics once sufficient support accumulates. A label-free criterion derived from class coverage determines this transition. Across diverse distribution shifts, inference settings, and pretrained backbones, the results consistently support this support-dependent inference principle. Overall, our findings highlight the importance of matching target statistics to the support available for their estimation.

AI USE STATEMENT

AI tools were used in this work to assist with writing and language polishing, as well as with the retrieval and discovery of potentially relevant literature. All AI-assisted text was reviewed and revised by the authors. Literature identified through AI-assisted retrieval was independently checked against the original publications before inclusion in the paper.

ETHICS STATEMENT

This work focuses on test-time adaptation using existing image classification benchmarks and publicly available pretrained models. It does not involve human subjects, the collection of personal or sensitive data, or the release of new datasets. The proposed method does not introduce applications or procedures that raise specific privacy, security, fairness, or safety concerns.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce our results throughout the main paper and appendix. Appendix A.2 provides the complete online inference algorithm and its transductive special case. The experimental setup, datasets, evaluation protocols, pretrained models, implementation details, and software and hardware configurations are described in Section 4.1 and Appendices A.3–A.4.

REFERENCES

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8334–8343, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Qiyuan Dai and Sibei Yang. Free on the fly: Enhancing flexibility in test-time adaptation with online em. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9538–9548, 2025.

Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Botta: Benchmarking on-device test time adaptation. In *Proceedings of the 2026 ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems*, SenSys '26, pp. 465–477, New York, NY, USA, 2026. Association for Computing Machinery.

Clément Fuchs, Maxime Zanella, and Christophe De Vleeschouwer. Online gaussian test-time adaptation of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 128–137, 2025.

Zenghao Guan, Yucan Zhou, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 6616–6639, 2025.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 142772–142798. Curran Associates, Inc., 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

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

Juwon Kang, Nayeong Kim, Jungseul Ok, and Suha Kwak. Membn: Robust test-time adaptation via batch norm with statistics memory. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, 2024.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762, 2024.

Yushu Li, Yongyi Su, Adam Goodge, Kui Jia, and Xun Xu. Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. In *International Conference on Learning Representations*, volume 2025, pp. 57322–57342, 2025.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Yihong Luo, Wenwu He, Dong Liang, Yihang Zhou, and Zhuo-Xu Cui. D²O: A dual debiasing operator for training-free test-time adaptation of vision-language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Robert A. Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2543–2553, 2024.

Saypraseuth Mounsaveng, Florent Chiaroni, Malik Boudiaf, Marco Pedersoli, and Ismail Ben Ayed. Bag of tricks for fully test-time adaptation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1925–1934, 2024.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo —no-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations*, volume 2026, pp. 128762–128789, 2026.

Fahim Faisal Niloy, Sk Miraj Ahmed, Dripta S. Raychaudhuri, Samet Oymak, and Amit K. Roy-Chowdhury. Effective restoration of source knowledge in continual test time adaptation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2080–2089, 2024.

Tiago Salvador and Adam M. Oberman. Imagenet-cartoon and imagenet-drawing: two domain shift datasets for imagenet. In *ICML Workshop on Shift happens: Crowdsourcing metrics and test datasets beyond ImageNet.*, 2022.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289. Curran Associates, Inc., 2022.

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.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3): 1106–1139, 2025.

Thilina Weerasinghe, Ruwan Tennakoon, WeiQin Chuah, and Alireza Bab-Hadiashar. Logit-adjusted test-time adaptation under partial class imbalance. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5977–5985, 2026.

Zehao Xiao and Cees GM Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024.

Linjing You, Jiabao Lu, Xiayuan Huang, and Xiangli Nie. Fret: Feature redundancy elimination for test time adaptation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2120–2130, 2025.

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25103–25112, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28718–28728, 2024.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 76280–76310. Curran Associates, Inc., 2025.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29999–30009, 2025.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

This appendix provides a detailed theoretical analysis of our method, along with additional experimental results. The contents are organized as follows:

- **Appendix A: Theoretical Analysis and More Details**
  - A.1 Sample Requirements for Target Statistics
  - A.1.1 Sampling error of global and class-specific statistics
  - A.1.2 Missing classes and the coverage threshold
  - A.1.3 Connection to EXACT's soft-weighted prototypes
  - A.1.4 Scope of the Coverage-Based Threshold
  - A.2 EXACT Inference Algorithm
  - A.3 Implementation Details
  - A.4 Software and Hardware
- **Appendix B: Additional Experimental Results and Analyses**
  - B.1 Fine-Grained Results Across Individual Distribution Shifts
  - B.2 Generalization Across Diverse Pretrained Classifiers
  - B.3 Computational Efficiency
  - B.4 Extended Results on Support-Dependent Reliability
  - B.5 Extended Analysis of Class-Specific Statistics Reliability
  - B.6 Additional Fusion Analysis
  - B.7 Additional Analysis of the Coverage-Based Threshold
  - B.8 Additional Analysis under Class Imbalance
  - B.9 Additional Analysis of the Online Batch Size

A THEORETICAL ANALYSIS AND MORE DETAILS

A.1 SAMPLE REQUIREMENTS FOR TARGET STATISTICS

We analyze i.i.d. observations $(\hat{x}_i, z_i, y_i)$ from a fixed target distribution with $C \geq 2$ classes. Write

$$p_c = \Pr(y_i = c) > 0, \qquad N_c = \sum_{i=1}^{N} \mathbf{1}\{y_i = c\}. \tag{17}$$

Labels are used only in the analysis; EXACT does not observe them. All probability statements below concern a fixed sample size N.

A.1.1 Sampling error of global and class-specific statistics

We first quantify how the statistical reliability of the two statistic types depends on the available target sample size. Global statistics estimate each logit channel using all N observed target samples. For a fixed channel c, define its population mean and variance as

$$m_c = \mathbb{E}[z_{ic}], \qquad \sigma^2_{z,c} = \mathrm{Var}(z_{ic}), \tag{18}$$

and their empirical counterparts as

$$m^z_c = \frac{1}{N}\sum_{i=1}^{N} z_{ic}, \qquad v^z_c = \frac{1}{N}\sum_{i=1}^{N}(z_{ic} - m^z_c)^2. \tag{19}$$

Because only a finite target sample is available, m^z_c and v^z_c fluctuate around their population values. For the empirical mean,

$$\mathbb{E}\big[(m^z_c - m_c)^2\big] = \frac{\sigma^2_{z,c}}{N}, \tag{20}$$

so its root mean squared estimation error decreases as $O(N^{-1/2})$.

Prediction using global statistics also depends on the empirical variance of each logit channel. With divisor N, its deviation from the population variance can be written as

$$v^z_c - \sigma^2_{z,c} = \frac{1}{N}\sum_{i=1}^{N}\big[(z_{ic} - m_c)^2 - \sigma^2_{z,c}\big] - (m^z_c - m_c)^2. \tag{21}$$

The first term is $O(N^{-1/2})$ under the finite fourth-moment assumption, while the second is $O(N^{-1})$. Hence,

$$v^z_c - \sigma^2_{z,c} = O(N^{-1/2}). \tag{22}$$

For class-specific statistics, define the population mean and the true-label sample mean by

$$\nu_c = \mathbb{E}[\hat{x}_i \mid y_i = c], \qquad \bar{m}^{\mathrm{gt}}_c = \frac{1}{N_c}\sum_{i:y_i=c}\hat{x}_i, \quad N_c > 0. \tag{23}$$

Let

$$s^2_c = \mathbb{E}\big[\|\hat{x}_i - \nu_c\|_2^2 \mid y_i = c\big] \le 1. \tag{24}$$

Conditionally on $N_c = n > 0$, independence within the class gives

$$\mathbb{E}\left[\|\bar{m}^{\mathrm{gt}}_c - \nu_c\|_2^2 \mid N_c = n\right] = \frac{s^2_c}{n} \le \frac{1}{n}. \tag{25}$$

Since $N_c \sim \mathrm{Binomial}(N, p_c)$, $N_c/(Np_c) \to 1$ in probability when $Np_c \to \infty$. The class-specific sampling-error scale is therefore $(Np_c)^{-1/2}$, or $\sqrt{C/N}$ under balanced sampling.

A.1.2 Missing classes and the coverage threshold

We next characterize how the available target sample size affects class support under uniform sampling. Suppose $p_c = 1/C$, and let U_N denote the number of classes that are absent from the first N target observations. Let T_{cov} denote the first sample count at which all classes have appeared.

Proposition 1 (Class coverage)**.** *Suppose $p_c = 1/C$. Then, for any $N \ge 1$,*

$$\mathbb{E}[U_N] = C\left(1 - \frac{1}{C}\right)^N \le C\exp\left(-\frac{N}{C}\right). \tag{26}$$

Moreover,

$$\mathbb{E}[T_{\mathrm{cov}}] = CH_C, \qquad H_C = \sum_{j=1}^{C}\frac{1}{j}. \tag{27}$$

Consequently, at the first integer sample count satisfying $N \geq CH_C$, i.e., $N = \lceil CH_C \rceil$,

$$\mathbb{E}[U_N] \leq Ce^{-H_C} < 1. \tag{28}$$

Equivalently, the expected fraction of missing classes satisfies

$$\frac{\mathbb{E}[U_N]}{C} \leq e^{-H_C} = O(C^{-1}). \tag{29}$$

Proof. For any fixed class c, the probability that it is absent from N independent observations is

$$\Pr(N_c = 0) = \left(1 - \frac{1}{C}\right)^N. \tag{30}$$

Therefore, by linearity of expectation,

$$\mathbb{E}[U_N] = C\left(1 - \frac{1}{C}\right)^N \leq C\exp\left(-\frac{N}{C}\right), \tag{31}$$

where we use $1 - x \leq e^{-x}$.

For the coverage time, after j distinct classes have appeared, the probability that the next observation belongs to a previously unseen class is $(C - j)/C$. The corresponding waiting time is geometric with mean $C/(C - j)$. Summing these waiting times gives

$$\mathbb{E}[T_{\text{cov}}] = \sum_{j=0}^{C-1} \frac{C}{C - j} = CH_C. \tag{32}$$

Finally, for $N = \lceil CH_C \rceil$,

$$\mathbb{E}[U_N] \leq C\exp\left(-\frac{N}{C}\right) \leq Ce^{-H_C} < 1, \tag{33}$$

where the last inequality follows from $H_C > \log C$. Dividing by C gives the corresponding bound on the expected missing fraction. □

Since $H_C = \log C + \gamma + o(1)$, the resulting reference has the sample-size scale $CH_C = \Theta(C \log C)$. We therefore use

$$N_{\text{cov}}(C) = CH_C \tag{34}$$

as a label-free reference for enabling class-specific statistics. At this scale, broad class support has accumulated under uniform sampling, with fewer than one class missing in expectation. The criterion is used as a support reference rather than as a fixed-sample guarantee of full class coverage, and it is not intended to identify the exact accuracy crossover between global and class-specific statistics.

A.1.3 Connection to EXACT's soft-weighted prototypes

The preceding analysis uses true-label sample means, whereas EXACT constructs prototypes from soft assignments. We connect the two by separating sampling error from the effect of the prototype weights.

For a class with positive total weight, write

$$a_{ic} = \frac{\omega_{ic}}{\sum_{j=1}^N \omega_{jc}}, \qquad \pi_c = \sum_{i=1}^N a_{ic}\hat{x}_i, \qquad \omega_{ic} = P^z_{ic}P^w_{ic}. \tag{35}$$

When $N_c > 0$, define

$$d_c = \frac{1}{2}\sum_{i=1}^N \left| a_{ic} - \frac{\mathbf{1}\{y_i = c\}}{N_c} \right|. \tag{36}$$

This quantity compares EXACT's weights with uniform averaging over the observed class-c samples. It captures both weight assigned to other classes and uneven weighting within class c. The true labels are used only for this analysis.

Soft-weighted prototype error. For any observed set with $N_c > 0$ and positive total class weight, the prototype before direction normalization satisfies

$$\underbrace{\|\pi_c - \nu_c\|_2}_{\text{prototype error}} \le \underbrace{\|\bar{m}_c^{\mathrm{gt}} - \nu_c\|_2}_{\text{sampling error}} + \underbrace{2d_c}_{\text{weight deviation}}, \tag{37}$$

where $\bar{m}_c^{\mathrm{gt}}$ and ν_c are the true-label sample mean and population class mean.

Since $\|\hat{x}_i\|_2 = 1$,

$$\|\pi_c - \bar{m}_c^{\mathrm{gt}}\|_2 = \left\| \sum_{i=1}^{N} \left(a_{ic} - \frac{\mathbf{1}\{y_i = c\}}{N_c} \right) \hat{x}_i \right\|_2 \le 2d_c. \tag{38}$$

The result follows by the triangle inequality.

Proposition 2 (Error of EXACT's soft-weighted prototypes)**.** *Assume $p_c = 1/C$. For $q \in \{z, w\}$, let $M^z = Z$ and $M^w = A$. Suppose the Sinkhorn outputs have the form*

$$P^q = \mathrm{diag}(u^q) \exp(M^q/\tau_q) \,\mathrm{diag}(v^q), \qquad u^q, v^q > 0, \qquad P^q \mathbf{1}_C = \mathbf{1}_N, \tag{39}$$

where $\tau_q > 0$ and the exponential is entrywise. Assume additionally that $\max_c M_{ic}^q - \min_c M_{ic}^q \le R_q < \infty$ almost surely. For fixed $\gamma_q > 0$, define

$$\begin{aligned} g_i^q &= M_{i,y_i}^q - \max_{c \ne y_i} M_{ic}^q, \qquad & \mathcal{L}_q &= \mathbb{E}[(\gamma_q - g_i^q)_+], \\ \varepsilon_q^{\mathrm{SK}} &= \frac{1}{2} \sum_c \left| \frac{1}{N} \sum_i P_{ic}^q - \frac{1}{C} \right|. \end{aligned} \tag{40}$$

Here $(t)_+ = \max\{t, 0\}$, $\mathcal{L}_q$ measures score-margin loss, and $\varepsilon_q^{\mathrm{SK}}$ measures the remaining column-marginal error. Then

$$\begin{aligned} \mathbb{E}\left[\frac{1}{C} \sum_c \|\pi_c - \nu_c\|_2 \right] &\le \min\{2, \mathcal{B}_N\}, \\ \mathcal{B}_N &= \sqrt{\frac{C}{N}} + \left(1 + \sum_q \frac{R_q}{\gamma_q} \right) \sqrt{\frac{C-1}{N}} \\ &\quad + 2 \sum_q \frac{\mathcal{L}_q + \tau_q \log C + R_q \mathbb{E}[\varepsilon_q^{\mathrm{SK}}]}{\gamma_q}. \end{aligned} \tag{41}$$

Proof. Let $Y_{ic} = \mathbf{1}\{y_i = c\}$, $\Delta_N = \frac{1}{2} \sum_c |N_c/N - 1/C|$, and $e_q = N^{-1} \sum_i (1 - P_{i,y_i}^q)$. The positive diagonal-scaling form implies that P^q maximizes $\langle P, M^q \rangle + \tau_q H(P)$ under its achieved marginals, where $H(P) = -\sum_{i,c} P_{ic} \log P_{ic}$ (Cuturi, 2013). Moving mass from surplus to deficit columns of Y, while preserving row sums, constructs a matrix Q with these marginals. The total mass moved is at most $N(\Delta_N + \varepsilon_q^{\mathrm{SK}})$. The score-range bound, optimality of P^q, and $0 \le H(P) \le N \log C$ therefore give

$$\begin{aligned} \gamma_q N e_q - \sum_i (\gamma_q - g_i^q)_+ &\le \langle Y - P^q, M^q \rangle \\ &\le N R_q (\Delta_N + \varepsilon_q^{\mathrm{SK}}) + \tau_q N \log C. \end{aligned} \tag{42}$$

Next, normalizing the construction weights gives $N_c d_c \le \sum_i |\omega_{ic} - Y_{ic}|$ for $N_c > 0$. Since $\omega_{ic} = P_{ic}^z P_{ic}^w$ and both assignments have unit row sums,

$$\begin{aligned} \sum_c |\omega_{ic} - Y_{ic}| &= 1 - P_{i,y_i}^z P_{i,y_i}^w + \sum_{c \ne y_i} P_{ic}^z P_{ic}^w \\ &\le (1 - P_{i,y_i}^z) + (1 - P_{i,y_i}^w). \end{aligned} \tag{43}$$

Hence $\sum_{c:N_c>0} (N_c/N) d_c \le e_z + e_w$, without assuming independence between the two assignments.

Because $\|\pi_c - \nu_c\|_2 \le 2$, including for absent classes, changing the class weights from N_c/N to $1/C$ costs at most $2\Delta_N$. This yields

$$\frac{1}{C}\sum_c \|\pi_c - \nu_c\|_2 \le \frac{1}{N}\sum_{c:N_c>0} N_c\|\overline{m}_c^{\mathrm{gt}} - \nu_c\|_2 + 2e_z + 2e_w + 2\Delta_N. \tag{44}$$

Under uniform sampling, $\mathbb{E}[\Delta_N] \le \frac{1}{2}\sqrt{(C-1)/N}$. Moreover, Eq. (25) gives

$$\mathbb{E}\left[\frac{1}{N}\sum_{c:N_c>0} N_c\|\overline{m}_c^{\mathrm{gt}} - \nu_c\|_2\right] \le \frac{1}{N}\sum_c s_c\,\mathbb{E}[\sqrt{N_c}] \le \sqrt{\frac{C}{N}}. \tag{45}$$

Taking expectations and combining the bounds proves Eq. 41. □

If $\nu_{\min} := \min_c \|\nu_c\|_2 > 0$ and each $\pi_c \neq 0$ almost surely, normalization also gives

$$\mathbb{E}\left[\frac{1}{C}\sum_c \big(1 - \cos(\pi_c, \nu_c)\big)\right] \le \min\left\{2, \frac{2\mathcal{B}_N}{\nu_{\min}}\right\}, \tag{46}$$

since $1 - \cos(\pi_c, \nu_c) \le 2\|\pi_c - \nu_c\|_2/\nu_{\min}$.

For converged balanced assignments, $\varepsilon_q^{\mathrm{SK}} = 0$. The finite-sample terms decrease at scale $O(\sqrt{C/N})$, whereas the score-margin, temperature, and marginal-error terms need not vanish as support increases. Thus, the result controls class-averaged prototype error under explicit score assumptions; it does not make CH_C a guarantee of accurate assignments or prototypes.

A.1.4 Scope of the Coverage-Based Threshold

For nonuniform class probabilities,

$$\mathbb{E}[U_N] = \sum_{c=1}^{C}(1-p_c)^N, \qquad \Pr(U_N > 0) \le \sum_{c=1}^{C} e^{-Np_c} \le Ce^{-Np_{\min}}. \tag{47}$$

Rare classes can therefore require a much larger sample size than the threshold. Without a positive lower bound on class probabilities, no finite threshold based only on C guarantees full coverage uniformly over target distributions. If a class has probability zero, it cannot appear at any sample size.

For a uniformly shuffled finite target set of size M, containing M_c observations of class c, sampling is without replacement. The exact absence probability is instead

$$\Pr(N_c = 0) = \frac{\binom{M-M_c}{N}}{\binom{M}{N}} \le \left(1 - \frac{M_c}{M}\right)^N, \qquad 1 \le N \le M, \tag{48}$$

where the numerator is zero when $N > M - M_c$. Thus, we use CH_C as the same label-free threshold, rather than as an exact waiting-time formula for a shuffled finite dataset.

A.2 EXACT Inference Algorithm

Algorithm 1 summarizes EXACT under the online protocol. At step k, the method appends the current batch's features and logits to the cache, computes class probabilities, and outputs predictions only for the current batch. Transductive evaluation in Algorithm 2 is obtained by applying the same algorithm with $K = 1$ and $\mathcal{B}_1 = \mathcal{D}_t$, where $\mathcal{D}_t$ is the complete unlabeled target set.

Algorithm 1 EXACT: Online Inference

Require: Frozen model (h_θ, W, b), class count C, sequential target batches $\{\mathcal{B}_k\}_{k=1}^{K}$

Initialize empty feature and logit caches $\mathbf{X}$ and $\mathbf{Z}$
Compute $H_C = \sum_{j=1}^{C} \frac{1}{j}$ and $N_{\text{cov}} = CH_C$
for $k = 1, \ldots, K$ **do**
    Extract frozen features $\mathbf{X}_k$ and logits $\mathbf{Z}_k$ for $\mathcal{B}_k$
    Append $\mathbf{X}_k$ and $\mathbf{Z}_k$ to $\mathbf{X}$ and $\mathbf{Z}$
    Set N_k to the accumulated sample count
    Let $\mathcal{I}_k$ index the current batch
    if $N_k < N_{\text{cov}}$ **then**
        Construct $P_u^{(k)}$ using Eq. 10
        $P^{(k)} \leftarrow P_u^{(k)}$
    else
        Compute P^z and P^w using Eqs. 11–12
        Estimate target prototypes Eqs. 13–15
        Construct $P_t^{(k)}$ using Eq. 16
        $P^{(k)} \leftarrow P_t^{(k)}$
    end if
    Emit $\hat{y}_i^{(k)} = \arg\max_c P_{ic}^{(k)},\ i \in \mathcal{I}_k$
end for

Algorithm 2 EXACT: Transductive Inference

Require: Frozen model (h_θ, W, b), class count C, complete unlabeled target set $\mathcal{D}_t$

Extract frozen features $\mathbf{X}$ and logits $\mathbf{Z}$ for $\mathcal{D}_t$
Set $N = |\mathcal{D}_t|$
Compute $H_C = \sum_{j=1}^{C} \frac{1}{j}$ and $N_{\text{cov}} = CH_C$
if $N < N_{\text{cov}}$ **then**
    Construct P_u using Eq. 10
    $P \leftarrow P_u$
else
    Compute P^z and P^w using Eqs. 11–12
    Estimate target prototypes using Eqs. 13–15
    Construct P_t using Eq. 16
    $P \leftarrow P_t$
end if
Emit $\hat{y}_i = \arg\max_c P_{ic}, \quad i = 1, \ldots, N$

A.3 Implementation Details

Pretrained Models. We evaluate EXACT with both supervised ViT and CLIP backbones. For ViT-Base, we use ViT-B/16 classifiers pretrained on ImageNet-21K and subsequently fine-tuned on the corresponding clean classification datasets. Specifically, we use `nateraw/vit-base-patch16-224-cifar10` for CIFAR-10-C, `edumunozsala/vit_base-224-in21k-ft-cifar100` for CIFAR-100-C, and `timm/vit_base_patch16_224.augreg2_in21k_ft_in1k` for ImageNet-C and the ImageNet-derived natural-shift benchmarks. For ImageNet-D109, the corresponding 109 classes are selected from the ImageNet-1K classifier without retraining.

For CLIP-Base, we use the official OpenAI CLIP ViT-B/16 checkpoint for all datasets. Following the common zero-shot classification protocol, the classifier is constructed using seven fixed text templates. Cosine similarities are scaled by a factor of 100 to produce classification logits. For ImageNet-D109, the corresponding 109 text classes are selected from the ImageNet class set. All pretrained model parameters remain frozen during inference, and the same initial checkpoints are used for both online and transductive evaluation.

Baseline Selection. To ensure a controlled and comparable evaluation, we restrict our comparison to test-time adaptation methods that operate on a frozen pretrained model and introduce no additional trainable parameters. This choice matches the setting of EXACT, where adaptation is performed entirely at inference time without retraining or modifying the pretrained model. We therefore exclude methods that require test-time backpropagation, parameter updates or additional trainable parameters, access to source-domain data, architectural modifications incompatible with the evaluated pretrained backbones. These criteria keep the comparison focused on training-free, source-free, and architecture-compatible methods evaluated with the same pretrained models and protocol.

Online inference. Target samples arrive sequentially in batches of 64. At each step, only the current and previously observed batches are available, and predictions are evaluated on the current batch. Previously emitted predictions are not revised using later observations.

Transductive inference. The complete unlabeled target set is available before any predictions are produced.

Baseline Implementations. For all competing methods, we use their publicly released implementations whenever available. Unless otherwise stated, we use the hyperparameters recommended by the original authors. All methods use the same pretrained backbone, input preprocessing, target sample order, and evaluation data.

Table 2: Per-corruption comparison on CIFAR-10-C at severity level 5. All results are Top-1 accuracy (%). Results are grouped by inference protocol and backbone. The best adaptation result within each protocol–backbone setting is shown in **bold**.

Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel.	JPEG	
ONLINE INFERENCE																
ViT-Base																
Frozen	76.23	77.04	74.63	92.84	69.25	90.18	94.00	93.72	92.18	90.21	97.69	92.99	82.88	57.22	84.45	84.37
T3A	76.32	77.21	75.08	92.79	69.84	90.17	93.97	93.68	92.25	90.23	97.67	93.03	82.95	58.15	84.46	84.52
LAME	76.21	76.94	74.36	92.82	69.07	90.28	94.06	93.74	92.36	90.29	97.69	93.06	83.11	56.98	84.44	84.36
NEO	76.99	78.10	77.35	92.91	71.02	90.24	94.06	93.75	92.40	90.48	**97.70**	93.18	83.27	63.21	84.55	85.28
EXACT	**79.76**	**80.44**	**81.81**	**93.21**	**73.20**	**90.74**	**94.40**	**94.09**	**93.37**	**91.01**	97.67	**93.84**	**84.29**	**71.74**	**85.13**	**86.98**
CLIP-Base																
Frozen	40.28	44.82	57.08	72.63	44.01	69.69	74.59	76.42	79.53	72.20	86.25	64.69	55.37	50.79	61.17	63.30
ADAPT	39.60	44.77	55.12	71.05	43.94	67.09	74.40	73.96	76.44	68.74	81.59	64.06	56.74	49.34	56.36	61.55
D^2O	40.17	45.06	55.93	71.08	45.02	67.00	74.87	73.93	76.31	69.17	82.37	64.55	56.56	50.94	56.65	61.97
StatA	38.30	42.57	58.49	75.10	44.90	71.99	78.38	78.52	81.06	71.38	88.54	66.17	57.32	49.87	60.18	64.18
SCA	43.13	47.38	59.52	73.49	45.95	71.94	76.49	78.52	81.78	73.29	87.63	67.62	58.55	51.40	62.78	65.30
BCA	38.48	41.79	54.20	70.89	41.35	67.58	72.64	74.98	77.99	70.66	85.00	63.31	50.77	47.26	58.59	61.03
DOTA	42.48	46.97	60.00	74.38	46.42	71.90	76.64	78.01	80.93	73.78	87.50	67.54	58.16	53.30	62.97	65.40
ECALP	45.49	49.70	62.09	75.99	47.26	72.85	78.14	78.42	81.06	73.37	87.98	68.29	59.99	54.68	63.30	66.57
EXACT	**52.00**	**56.62**	**66.43**	**77.68**	**48.64**	**75.90**	**80.33**	**80.66**	**82.38**	**76.05**	**88.85**	**71.17**	**62.64**	**58.64**	**65.67**	**69.58**
TRANSDUCTIVE INFERENCE																
ViT-Base																
Frozen	76.23	77.04	74.63	92.84	69.25	90.18	94.00	93.72	92.18	90.21	97.69	92.99	82.88	57.22	84.45	84.37
T3A	76.25	77.14	74.90	92.82	69.69	90.17	93.96	93.68	92.25	90.20	97.67	93.02	82.92	57.77	84.49	84.46
LAME	76.20	76.88	74.31	92.73	69.15	90.29	93.90	93.74	92.33	90.25	97.69	93.10	83.10	56.93	84.58	84.35
NEO	76.96	78.09	77.38	92.90	71.04	90.24	94.05	93.75	92.39	90.48	**97.70**	93.18	83.27	63.24	84.55	85.28
EXACT	**79.93**	**80.70**	**82.26**	**93.34**	**73.26**	**90.87**	**94.48**	**94.20**	**93.59**	**91.05**	97.69	**93.87**	**84.49**	**71.92**	**85.26**	**87.13**
CLIP-Base																
Frozen	40.28	44.82	57.08	72.63	44.01	69.69	74.59	76.42	79.53	72.20	86.25	64.69	55.37	50.79	61.17	63.30
ADAPT	46.12	49.06	61.06	75.39	47.01	73.75	78.68	78.77	80.37	73.78	87.54	67.72	59.64	54.78	62.80	66.43
D^2O	47.93	51.95	62.02	75.19	48.27	73.37	78.59	78.71	80.14	73.83	87.92	68.33	60.30	56.98	64.00	67.17
StatA	39.29	43.45	59.50	75.43	46.69	72.45	78.72	78.90	81.44	71.67	88.78	66.35	58.07	50.55	60.73	64.80
SCA	43.08	47.14	59.43	73.21	45.73	72.04	76.49	78.69	81.92	73.16	87.70	67.45	58.75	50.64	62.58	65.20
BCA	36.86	39.91	52.99	69.86	39.58	66.52	71.95	74.27	77.46	70.23	84.73	62.01	48.51	45.04	57.36	59.82
DOTA	42.56	47.01	60.09	74.46	46.52	71.95	76.75	78.13	81.00	73.91	87.58	67.73	58.24	53.38	62.98	65.49
ECALP	46.16	50.45	62.98	76.45	48.44	73.64	78.58	79.26	81.51	74.27	88.52	69.21	60.81	56.07	64.00	67.36
EXACT	**52.23**	**56.72**	**66.74**	**78.04**	**48.85**	**76.17**	**80.69**	**80.86**	**82.70**	**76.42**	**89.14**	**71.51**	**62.96**	**59.18**	**66.19**	**69.89**

Sinkhorn normalization. For all Sinkhorn operations, we use the standard balanced Sinkhorn–Knopp algorithm with a uniform class marginal. Specifically, the total assignment mass is constrained to be equally distributed across the C classes. We run 20 Sinkhorn–Knopp iterations for each normalization and use the same setting throughout all experiments. Since the raw logits or similarity scores may contain negative values, they are first converted into a non-negative kernel through temperature scaling and exponentiation.

A.4 SOFTWARE AND HARDWARE

Our experimental environment uses Python `3.12.7`, PyTorch `2.7.0`, and CUDA `12.8`. Experiments are conducted on a single NVIDIA GeForce RTX 5080 GPU with 16 GB of GPU memory. The workstation is equipped with an Intel Core i7-14700 CPU and 64 GB of system memory.

B ADDITIONAL EXPERIMENTAL RESULTS AND ANALYSES

B.1 FINE-GRAINED RESULTS ACROSS INDIVIDUAL DISTRIBUTION SHIFTS

To complement the aggregate results reported in Table 1, we provide fine-grained performance across individual corruption types and domains. These results examine whether the improvements of EXACT are consistent across heterogeneous distribution shifts rather than being driven by a small subset of conditions. As shown in Table 2, 3 and 4, the per-shift results reveal that the improvement of EXACT is not driven by a small subset of favorable corruptions. Across CIFAR-10-C, CIFAR-100-C, and ImageNet-C, the proposed method remains consistently competitive over heterogeneous

Table 3: Per-corruption comparison on CIFAR-100-C at severity level 5. All results are Top-1 accuracy (%). Results are grouped by inference protocol and backbone. The best adaptation result within each protocol–backbone setting is shown in **bold**.

Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel.	JPEG	
ONLINE INFERENCE																
ViT-Base																
Frozen	40.42	42.15	36.41	75.08	31.18	68.35	75.04	74.32	70.65	62.09	85.92	55.83	50.52	47.40	61.72	58.47
T3A	41.54	43.68	37.48	75.32	32.69	68.56	75.41	74.58	71.06	62.50	85.86	56.50	51.50	48.74	62.06	59.17
LAME	39.75	41.70	35.90	75.00	31.08	68.34	75.07	74.42	70.83	61.98	85.75	55.80	50.47	47.31	61.67	58.34
NEO	41.44	43.76	37.99	75.26	32.58	68.54	75.18	74.53	70.98	62.46	85.93	56.47	51.14	48.21	61.91	59.09
EXACT	**44.43**	**47.02**	**41.47**	**75.56**	**36.56**	**68.97**	**75.73**	**75.33**	**72.14**	**64.05**	**86.05**	**58.98**	**52.96**	**51.06**	**62.53**	**60.86**
CLIP-Base																
Frozen	22.85	24.35	29.51	43.56	20.07	43.63	48.71	48.88	50.36	41.83	58.09	35.24	29.15	26.31	33.51	37.07
ADAPT	19.23	20.85	29.44	40.57	20.41	39.99	45.50	42.61	46.02	38.52	54.18	31.57	27.04	24.35	29.47	33.98
D^2O	19.41	21.68	29.67	40.06	21.22	39.83	44.73	42.68	46.07	38.37	54.36	31.62	27.22	24.85	29.61	34.09
StatA	17.79	20.03	26.48	42.48	17.45	42.40	48.05	46.85	49.24	39.95	57.74	31.86	25.94	22.71	30.12	34.61
SCA	23.99	25.99	31.48	45.26	21.37	44.64	50.07	50.47	52.31	43.37	60.36	36.16	30.11	27.66	34.56	38.52
BCA	23.38	24.74	30.06	44.20	20.68	43.72	49.15	49.02	50.76	42.30	58.27	35.64	29.96	26.85	34.18	37.53
DOTA	24.19	26.05	31.94	46.66	21.78	46.00	51.43	51.24	53.04	44.01	61.10	37.29	31.09	28.51	35.36	39.31
ECALP	24.69	26.36	32.78	46.82	22.14	45.52	51.29	51.05	52.86	44.29	60.92	37.89	31.41	28.91	34.97	39.46
EXACT	**27.63**	**29.96**	**37.43**	**49.94**	**25.97**	**48.54**	**54.50**	**53.80**	**55.56**	**46.96**	**64.38**	**39.92**	**34.27**	**32.55**	**36.90**	**42.55**
TRANSDUCTIVE INFERENCE																
ViT-Base																
Frozen	40.42	42.15	36.41	75.08	31.18	68.35	75.04	74.32	70.65	62.09	85.92	55.83	50.52	47.40	61.72	58.47
T3A	41.60	43.73	38.00	75.26	32.87	68.32	75.36	74.51	70.93	62.33	85.84	56.42	51.45	48.66	61.96	59.15
LAME	39.85	41.82	36.02	75.12	31.22	68.47	74.88	74.42	70.86	62.08	85.87	55.90	50.56	47.43	61.67	58.41
NEO	41.43	43.78	37.99	75.25	32.58	68.55	75.19	74.54	70.99	62.45	85.94	56.48	51.14	48.21	61.93	59.10
EXACT	**44.71**	**47.46**	**41.88**	**75.76**	**37.16**	**69.11**	**76.01**	**75.55**	**72.50**	**64.45**	**86.41**	**59.37**	**53.18**	**51.54**	**62.58**	**61.18**
CLIP-Base																
Frozen	22.85	24.35	29.51	43.56	20.07	43.63	48.71	48.88	50.36	41.83	58.09	35.24	29.15	26.31	33.51	37.07
ADAPT	24.93	26.95	33.86	48.04	24.21	46.70	52.12	51.34	53.63	45.09	61.82	38.21	32.54	30.71	35.88	40.40
D^2O	25.40	27.60	35.02	48.59	24.77	47.19	52.54	51.97	54.02	45.48	62.21	38.82	33.18	31.41	36.53	40.98
StatA	20.11	22.54	29.51	45.11	19.27	44.53	50.36	49.19	51.32	41.97	59.75	34.02	28.11	24.66	32.23	36.85
SCA	23.79	25.83	31.22	45.25	21.08	44.59	50.08	50.44	52.55	43.35	60.69	35.81	29.88	27.47	34.36	38.43
BCA	23.21	24.56	30.08	44.11	20.77	43.71	49.01	49.00	50.77	42.37	58.33	35.56	29.82	26.79	34.18	37.48
DOTA	24.66	26.44	32.64	47.26	22.22	46.72	52.08	52.01	53.75	44.51	61.74	37.89	31.50	28.91	35.85	39.88
ECALP	24.91	26.51	33.37	47.34	22.85	45.92	51.96	51.70	53.67	45.02	61.94	38.40	32.22	29.35	35.38	40.04
EXACT	**28.28**	**30.48**	**38.55**	**51.15**	**26.75**	**49.70**	**55.59**	**54.95**	**56.63**	**47.77**	**65.90**	**40.60**	**35.22**	**33.48**	**37.83**	**43.53**

noise, blur, weather, and digital corruptions, for both supervised ViT and CLIP backbones and under both online and transductive inference. Importantly, the choice between P_u and P_t depends only on the observed sample count N and the number of classes C. Therefore, the consistency across individual shifts provides stronger evidence than the aggregate accuracy alone: the effectiveness of EXACT is associated with the amount of target data available for estimating the target statistics, rather than with a particular corruption mechanism.

B.2 GENERALIZATION ACROSS DIVERSE PRETRAINED CLASSIFIERS

Experimental setup. To examine whether the effectiveness of EXACT depends on the particular backbones used in the main experiments, we further evaluate it on 35 additional pretrained classifiers spanning six architecture families. As summarized in Table 5, the CIFAR experiments include ResNet and VGG models trained on the corresponding clean CIFAR-10 or CIFAR-100 training set, while the ImageNet experiments cover ResNet, ViT, ConvNeXt, DeiT, and Swin architectures with publicly released ImageNet checkpoints. All pretrained parameters and classification heads are kept frozen, and EXACT is directly applied at inference time without retraining the source models.

Following the main evaluation protocol, CIFAR models are evaluated on CIFAR-10-C or CIFAR-100-C at severity level 5, while ImageNet models are evaluated on ImageNet-C, ImageNet-Sketch, ImageNet-D109, ImageNet-Drawing, and ImageNet-Cartoon. We consider both Online64 and transductive inference and report the accuracy improvement over the corresponding frozen classifier,

$$\Delta\mathrm{Acc} = \mathrm{Acc}_{\mathrm{EXACT}} - \mathrm{Acc}_{\mathrm{Frozen}}. \tag{49}$$

Table 4: Per-corruption comparison on ImageNet-C at severity level 5. All results are Top-1 accuracy (%). Results are grouped by inference protocol and backbone. The best adaptation result within each protocol–backbone setting is shown in **bold**.

Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel.	JPEG	
ONLINE INFERENCE																
ViT-Base																
Frozen	55.34	56.23	56.01	46.48	34.78	52.87	44.20	62.39	62.66	65.56	77.70	32.04	45.73	66.72	66.67	55.03
T3A	54.80	55.96	55.78	47.50	36.86	53.85	46.40	63.84	60.79	68.07	78.15	37.97	49.65	67.31	68.05	56.33
LAME	53.89	55.07	54.94	45.62	33.95	52.15	43.68	61.23	61.10	65.63	77.18	27.96	45.37	66.15	66.02	54.00
NEO	56.31	56.89	56.94	49.49	37.41	54.58	46.94	64.85	65.14	70.95	78.14	58.24	49.72	67.41	67.92	58.73
EXACT	**56.56**	**57.63**	**57.53**	**52.41**	**42.49**	**57.46**	**52.17**	**66.48**	**65.87**	**71.80**	**78.73**	**60.01**	**57.34**	**68.99**	**69.99**	**61.03**
CLIP-Base																
Frozen	13.29	14.44	13.76	24.96	16.13	25.98	24.40	34.02	32.34	38.22	56.98	17.97	14.18	33.57	35.08	26.35
ADAPT	13.44	14.43	14.07	24.39	16.24	25.28	23.75	33.57	32.07	38.28	56.98	17.95	15.12	33.01	34.18	26.18
D^2O	14.24	15.22	15.14	24.67	16.69	25.75	24.30	34.04	32.52	38.75	57.01	18.49	16.28	33.31	34.28	26.71
StatA	11.27	12.32	11.91	22.07	13.64	22.32	20.62	31.65	31.29	36.07	57.63	14.76	12.43	30.68	33.88	24.17
SCA	13.66	14.77	14.21	25.39	16.46	26.06	24.57	34.90	33.07	39.39	58.74	18.84	15.06	34.01	36.05	27.01
BCA	13.96	15.04	14.40	25.42	16.64	26.37	24.85	34.80	32.98	38.81	57.24	18.74	14.85	34.18	35.45	26.92
DOTA	14.64	15.75	15.31	26.10	17.23	26.93	25.55	36.07	34.42	40.35	**59.51**	19.35	15.64	35.36	36.41	27.91
ECALP	16.62	17.73	17.20	27.07	18.11	27.59	25.67	36.97	35.03	41.41	59.36	20.44	16.61	**37.68**	**36.93**	28.96
EXACT	**17.08**	**18.03**	**17.89**	**27.28**	**19.05**	**27.98**	**26.99**	**37.08**	**35.04**	**41.65**	59.17	**21.16**	**19.23**	36.46	36.78	**29.39**
TRANSDUCTIVE INFERENCE																
ViT-Base																
Frozen	55.34	56.23	56.01	46.48	34.78	52.87	44.20	62.39	62.66	65.56	77.70	32.04	45.73	66.72	66.67	55.03
T3A	54.21	55.62	55.32	47.71	38.14	53.99	47.56	63.87	59.50	68.53	78.06	42.86	51.20	67.26	68.24	56.80
LAME	54.46	55.64	55.53	46.30	34.63	52.79	44.41	62.06	62.08	67.20	77.95	30.04	46.29	66.94	66.77	54.87
NEO	56.31	56.89	56.94	49.49	37.41	54.58	46.94	64.85	65.16	70.96	78.14	58.24	49.73	67.42	67.91	58.73
EXACT	**57.04**	**58.23**	**58.05**	**53.33**	**43.45**	**58.37**	**53.30**	**67.34**	**66.73**	**72.74**	**79.75**	**61.05**	**58.94**	**69.83**	**70.84**	**61.93**
CLIP-Base																
Frozen	13.29	14.44	13.76	24.96	16.13	25.98	24.40	34.02	32.34	38.22	56.98	17.97	14.18	33.57	35.08	26.35
ADAPT	16.12	17.13	16.64	26.89	18.51	28.10	26.72	37.12	35.35	41.54	60.21	21.03	17.56	36.42	36.89	29.08
D^2O	16.79	17.72	17.40	27.28	19.02	28.38	27.18	37.76	**36.03**	42.02	60.22	21.18	18.52	36.69	36.96	29.54
StatA	12.54	13.79	13.29	24.02	14.99	24.23	22.45	33.68	32.54	38.08	59.00	16.48	13.64	32.80	35.29	25.79
SCA	12.32	13.52	12.74	24.43	15.75	25.09	23.60	34.15	32.38	38.57	58.87	17.72	14.12	32.93	35.83	26.13
BCA	13.95	15.07	14.46	25.48	16.68	26.46	24.95	34.83	33.10	38.84	57.32	18.85	15.00	34.27	35.50	26.98
DOTA	15.28	16.36	16.01	26.49	17.71	27.28	25.96	36.70	35.25	41.13	60.18	19.87	16.37	36.03	36.81	28.50
ECALP	16.73	17.85	17.28	27.41	18.29	27.91	25.90	37.28	35.25	41.80	59.90	20.61	16.94	**38.19**	37.24	29.24
EXACT	**17.44**	**18.55**	**18.22**	**27.87**	**19.51**	**28.50**	**27.70**	**38.13**	35.82	**42.65**	**60.93**	**21.58**	**19.70**	37.52	**37.54**	**30.11**

Table 5: Pretrained classifiers used for the backbone-generalization experiments in Fig. 9. All checkpoints are publicly released and remain frozen during inference.

Training / Pretraining	Family	Variants	Checkpoint source
CIFAR-10/100	ResNet	20, 32, 44, 56	chenyaofo/pytorch-cifar-models
CIFAR-10/100	VGG-BN	11, 13, 16, 19	chenyaofo/pytorch-cifar-models
ImageNet-1K	ResNet	18, 34, 50, 101, 152	torchvision via timm
ImageNet-21K $\rightarrow$ 1K	ViT	T/16, S/16, B/32, L/16	Google ViT AugReg via timm
ImageNet-1K	ConvNeXt	T, S, B, L	official ConvNeXt via timm
ImageNet-1K	DeiT	T, S, B	official DeiT via timm
ImageNet-1K	Swin	T, S, B	official Swin via timm

Figure 9 shows a remarkably consistent pattern: EXACT improves the corresponding frozen classifier for every evaluated backbone, target shift, and inference setting. The gains persist across conventional CNNs (ResNet and VGG), modern convolutional architectures (ConvNeXt), and multiple Transformer families (ViT, DeiT, and Swin), demonstrating that the improvement is not tied to a particular backbone architecture or representation family. Overall, these results support the effectiveness of support-dependent inference across the evaluated backbone families.

B.3 COMPUTATIONAL EFFICIENCY

We further compare the computational efficiency of EXACT with the corresponding TTA baselines on ImageNet-C. To provide a controlled comparison, all methods are evaluated on Gaussian noise

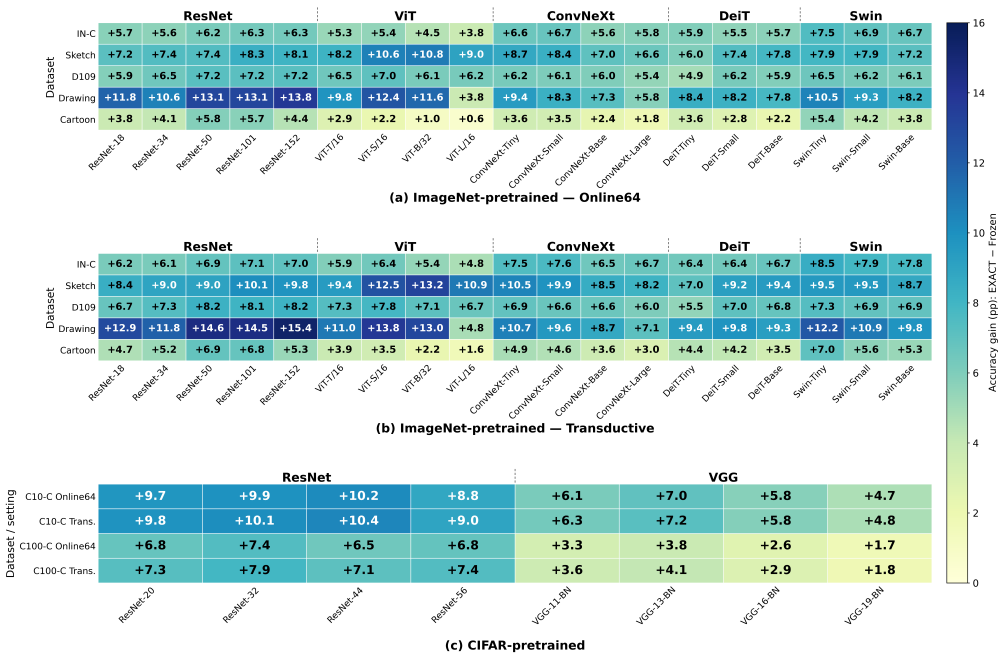

Figure 9: Backbone generalization of EXACT. Each cell reports the Top-1 accuracy gain of EXACT over the corresponding frozen classifier. (a–b) Results on 19 additional ImageNet-pretrained classifiers spanning ResNet, ViT, ConvNeXt, DeiT, and Swin under Online64 and transductive inference. (c) Results on CIFAR-pretrained ResNet and VGG classifiers on CIFAR-10-C and CIFAR-100-C. EXACT yields positive gains for every evaluated backbone, dataset, and inference setting.

Table 6: Computational efficiency on ImageNet-C. All methods are evaluated on Gaussian noise at severity 5 using the complete 50,000-image corruption domain, with an online batch size of 64 and a single NVIDIA GeForce RTX 5080 GPU. We report wall-clock time, per-image latency, peak GPU memory, adaptation FLOPs, and total FLOPs per image. Lower is better for all efficiency metrics.

Method	**Adaptation Runtime**		**Memory & Compute**		
	Time (s/domain)	**Adaptation Latency** (ms/image)	**Peak Mem.** (MiB)	**Adapt.** (GFLOPs/image)	**Total** (GFLOPs/image)
ViT-Base					
T3A	14.63	0.2927	960.8	0.0016	35.2214
LAME	270.86	5.4173	2085.5	2.1277	37.3475
NEO	0.18	0.0036	895.1	0.0015	35.2213
EXACT	18.91	0.3782	2426.8	1.8831	37.1029
CLIP-Base					
ADAPT	17.59	0.3517	1098.7	0.0862	35.2700
D^2O	44.35	0.8870	2181.0	0.9975	36.1813
StatA	615.39	12.3079	2617.7	23.8151	58.9989
SCA	1.73	0.0345	1061.1	0.0123	35.1960
BCA	12.24	0.2448	1059.1	0.0061	35.1898
DOTA	586.68	11.7335	4485.0	1.8486	37.0324
ECALP	6523.66	130.4731	13367.6	0.6512	35.8350
EXACT	17.74	0.3547	2331.2	1.2937	36.4774

at severity level 5 using the full 50,000-image corruption domain. We use the same pretrained backbone and input preprocessing for each backbone family, with an online batch size of 64. We report wall-clock time for one corruption domain, average latency per image, peak GPU memory, adaptation FLOPs, and total FLOPs per image.

Table 6 shows that EXACT does not rely on minimal adaptation computation to obtain its performance gains. On ViT-Base, EXACT requires 0.3782 ms per image and 2426.8 MiB of peak GPU

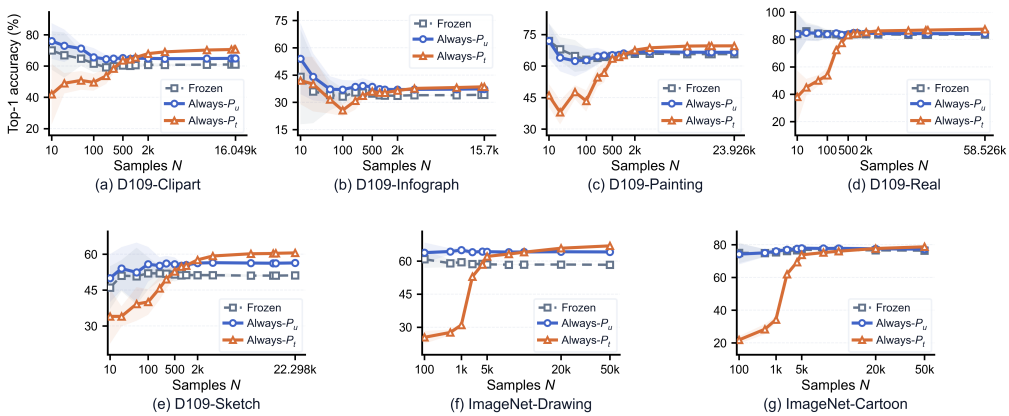

Figure 10: Support-dependent reliability of global and class-specific statistics on the remaining natural-shift benchmarks. The same support-dependent reliability reversal persists across ImageNet-D109, ImageNet-Drawing, and ImageNet-Cartoon.

memory, with 1.8831 adaptation GFLOPs per image. Its runtime is lower than LAME (5.4173 ms/image), but higher than the lightweight T3A and NEO baselines. On CLIP-Base, EXACT requires 0.3547 ms per image and 2331.2 MiB of peak memory. It is substantially faster than StatA, DOTA, and ECALP, whose per-image latencies are 12.3079, 11.7335, and 130.4731 ms, respectively, while methods such as ADAPT, SCA, and BCA incur lower computational cost.

The additional cost of EXACT mainly comes from prototype estimation and the computation of P_t, which are enabled when the coverage-based criterion is met. Importantly, wall-clock latency and FLOPs do not induce the same ordering across methods. For example, some baselines use fewer arithmetic operations but require considerably longer measured runtime. Reporting both quantities therefore separates algorithmic computation from realized execution cost. Overall, these results show that EXACT trades additional inference-time computation for the accuracy improvements reported in the main experiments, while avoiding the substantially larger runtime or memory cost observed for several competing methods.

B.4 Extended Results on Support-Dependent Reliability

B.4.1 Additional Natural-Shift Benchmarks

Figure 10 extends the support-dependent reliability analysis to the remaining natural-shift benchmarks. The same reliability reversal is consistently observed across the five ImageNet-D109 domains, ImageNet-Drawing, and ImageNet-Cartoon: global statistics remain comparatively stable under scarce target support, whereas class-specific statistics are initially unreliable but improve rapidly as more target observations become available. Importantly, this behavior persists across substantially different visual shifts, indicating that the phenomenon is not specific to a particular corruption process or target domain.

This reversal admits a simple statistical interpretation. Global statistics are estimated from all N target observations. In contrast, class-specific statistics depend on observations supporting the individual classes. Under the uniform-occupancy approximation, its typical effective sample size is only $N_c \approx N/C$, giving an uncertainty on the order of $O(N_c^{-1/2}) = O(\sqrt{C/N})$. More critically, at low support some classes have $N_c = 0$. The expected number of unobserved classes is

$$\mathbb{E}[U_N] = C\left(1 - \frac{1}{C}\right)^N \approx Ce^{-N/C}. \tag{50}$$

Thus, soft assignments can numerically define class-specific statistics even for classes with no observed samples. The threshold CH_C is derived from the expected time required to observe all classes under uniform class probabilities. These results therefore expose a distinction central to EXACT: *A numerically defined class-specific statistic need not be reliably estimated from the observed target data.* Rather than always using class-specific statistics, EXACT matches the statistic type to the

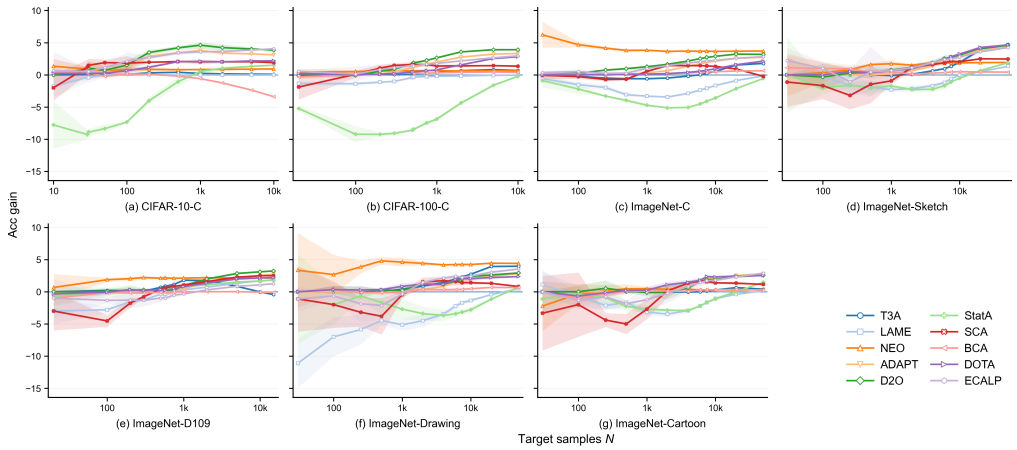

Figure 11: Performance of competing TTA methods across target-set sizes. We vary the available target sample size N and report the Top-1 accuracy gain over the corresponding frozen classifier across the seven evaluation benchmarks. Results are averaged over five random seeds.

available support. The consistent reversal provides evidence that this support-dependent reliability pattern extends across the evaluated natural shifts.

B.4.2 Performance of competing methods across target support

To examine whether performance dependence on target support is specific to EXACT, we evaluate the competing methods across different target-set sizes. Fig. 11 shows that many methods vary substantially with N: several yield negative gains under limited target support and improve as target observations accumulate, while others remain comparatively stable or show non-monotonic changes. This behavior is consistently observed across the evaluated corruption and natural-shift benchmarks. These results indicate that the amount of available target data is a general factor affecting TTA performance, although the underlying mechanism can differ across methods.

B.4.3 Low-Support Prediction with P_u in Representative TTA Methods

To examine whether the low-support benefit of P_u is specific to EXACT, we apply P_u to five representative TTA methods: T3A, ADAPT, D2O, SCA, and ECALP. Under online inference, for $N < N_{\text{cov}}(C) = CH_C$, we replace the original predictions of each method with P_u, while leaving the remaining inference procedure unchanged. We report the Top-1 accuracy gain over the corresponding original method, evaluated only on predictions emitted before the coverage-based threshold.

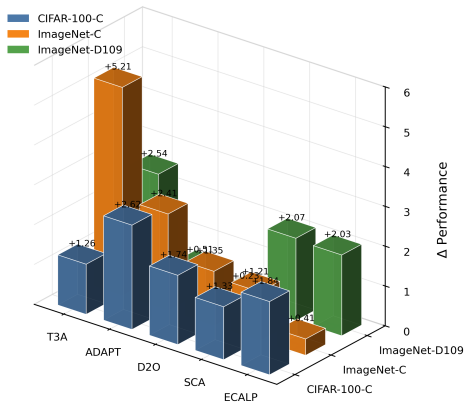

Figure 12: Low-support gains from using P_u in representative TTA methods.

As shown in Fig. 12, using P_u yields positive gains across all evaluated methods and datasets. The improvements are observed for methods with different inference rules and across both ViT-based and CLIP-based settings. These results indicate that the benefit of P_u under limited target support is not specific to the other components of EXACT. Rather, using global statistics during the early stage of online inference can also reduce low-support errors in existing TTA methods.

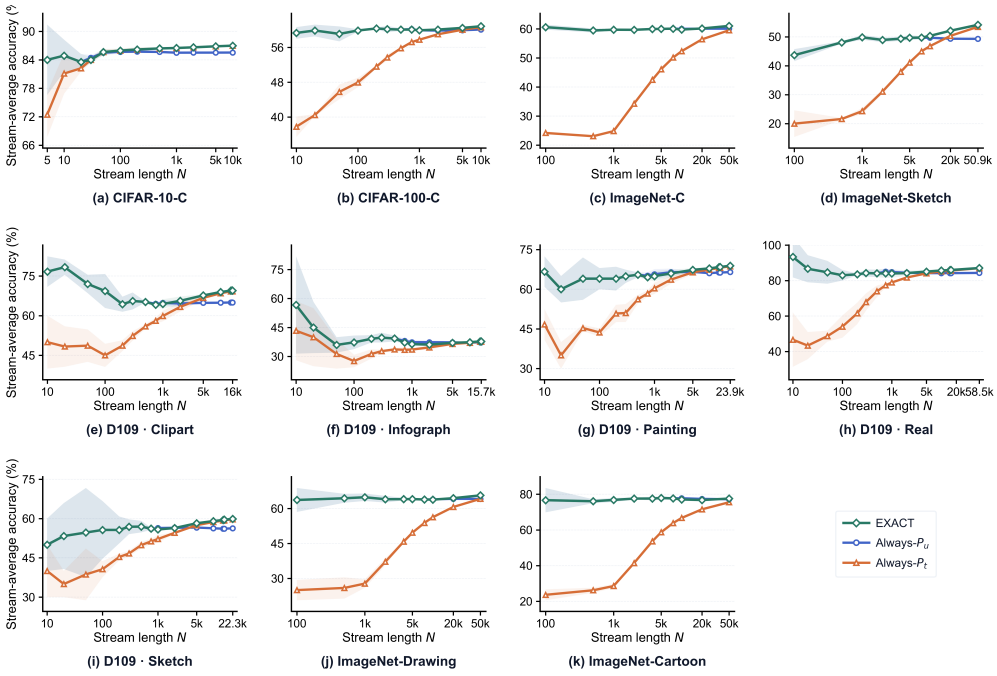

Figure 13: Online stream-average accuracy across increasing stream lengths. Always-P_u remains stable under limited target support, while Always-P_t improves as target observations accumulate but retains the effect of early low-support errors in the cumulative metric.

B.4.4 Online Stream-Level Evaluation

We further evaluate support-dependent inference under the online protocol. Target samples arrive sequentially in batches of 64, and predictions at each inference step are produced using only the target observations available up to that step. For each stream length N, we evaluate the corresponding target-stream samples and report the stream-average Top-1 accuracy, where each prediction is evaluated when it is emitted and is not recomputed using future observations. We compare EXACT with Always-P_u and Always-P_t, which compute class probabilities using global and class-specific statistics, respectively, throughout the stream. EXACT uses P_u when $N_k < N_{\mathrm{cov}}(C)$ and P_t otherwise. We evaluate all seven benchmarks using the same stream-length grid as the support-dependent reliability analysis, with the complete target stream additionally included as the final evaluation point.

As shown in Fig. 13, the support-dependent reliability difference between the two branches translates directly into cumulative online performance. At short stream lengths, Always-P_t suffers substantial degradation because class-specific statistics are used before sufficient per-class support has accumulated, whereas Always-P_u remains comparatively stable. As more target observations become available, the performance of Always-P_t improves steadily, with the recovery occurring at larger stream lengths for datasets with larger class spaces. Unlike the fixed-support results, this recovery is delayed in stream-average accuracy because prediction errors made earlier in the stream remain part of the cumulative metric. EXACT avoids accumulating these low-support errors by using P_u before the coverage-based switching point and then exploiting P_t after additional target support becomes available. Consequently, it retains the stable early-stream behavior of the global-statistics branch while benefiting from class-specific statistics later in the stream, and finishes above the two always variants across the evaluated benchmarks.

B.5 Extended Analysis of Class-Specific Statistics Reliability

B.5.1 Metrics

Absent-class prototype-weight ratio. Let $\mathcal{M}_N$ denote the set of classes with zero GT samples among the current N observations. We measure the fraction of total prototype-construction weight

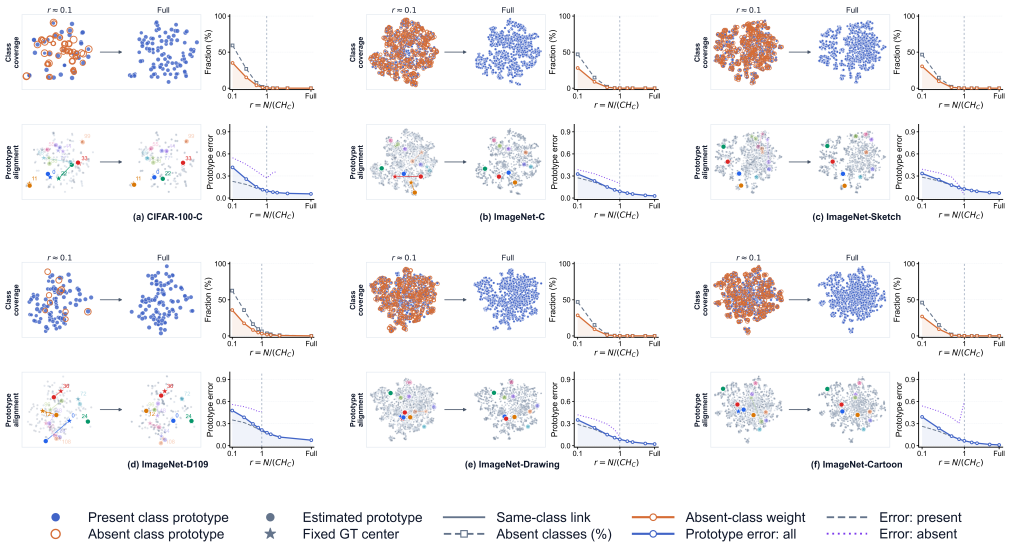

Figure 14: Extended reliability analysis of class-specific statistics. Under limited target support, positive prototype weight is assigned to absent classes, and prototype errors are large. As target samples accumulate, the absent-class prototype-weight ratio and prototype error generally decrease.

assigned to these classes as

$$R_{\text{unsup}} = \frac{\sum_{c \in \mathcal{M}_N} \sum_i \omega_{ic}}{\sum_c \sum_i \omega_{ic}}, \tag{51}$$

where ω_{ic} is the prototype-construction weight used in Section 3.5.

Average prototype error. Using the complete target domain only for post-hoc analysis, we define a fixed center $\pi_c^\star$ for each class and compute the average prototype error

$$E_\pi = 1 - \frac{1}{C} \sum_{c=1}^{C} \cos(\pi_c(N), \pi_c^\star)\,. \tag{52}$$

R_{unsup} measures the fraction of prototype-construction weight assigned to absent classes, while E_π measures average prototype error relative to the full-target reference centers.

B.5.2 Additional Results

Fig. 14 extends the reliability analysis of class-specific statistics to six additional corruption and natural-shift benchmarks. A consistent pattern emerges across all datasets. Under limited support ($r \ll 1$), a substantial fraction of classes is still absent from the observed target set, yet soft prototype construction assigns non-zero prototype weight to these absent classes. Consequently, the absent-class prototype-weight ratio closely follows the fraction of absent classes, while the estimated prototypes exhibit large geometric deviations from their full-target oracle centers.

As target samples accumulate, class coverage improves, and the absent-class prototype-weight ratio and prototype error generally decrease. These observations support using the available target data to determine when to enable class-specific statistics.

B.6 Additional Fusion Analysis

We provide the detailed formulation and experimental setup for the fusion analysis in Sec. 4.5. The experiments follow the same online inference protocol as EXACT.

B.6.1 FIXED FUSION

We first consider a fixed fusion weight $\alpha \in [0, 1]$ throughout the target stream. For arithmetic fusion, the fused class-probability matrix is

$$P_{\mathrm{A}}^{(k)}(\alpha) = (1-\alpha)P_u^{(k)} + \alpha P_t^{(k)}. \tag{53}$$

For geometric fusion, the probability assigned to class c for sample i is

$$P_{\mathrm{G},ic}^{(k)}(\alpha) = \frac{\left(P_{u,ic}^{(k)}\right)^{1-\alpha}\left(P_{t,ic}^{(k)}\right)^{\alpha}}{\sum_{c'}\left(P_{u,ic'}^{(k)}\right)^{1-\alpha}\left(P_{t,ic'}^{(k)}\right)^{\alpha}}. \tag{54}$$

We evaluate

$$\alpha \in \{0,\ 0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9,\ 1\}, \tag{55}$$

where $\alpha = 0$ and $\alpha = 1$ recover P_u and P_t, respectively.

B.6.2 SUPPORT-ADAPTIVE FUSION

A fixed fusion weight does not account for the changing reliability of the two branches as target observations accumulate. We therefore also consider a support-adaptive weight

$$\alpha_\kappa(r) = \frac{r^\kappa}{1+r^\kappa}, \tag{56}$$

where $\kappa > 0$ controls how rapidly the weight shifts from the global branch toward the class-specific branch around $r = 1$. Larger κ produces a sharper transition, while smaller κ gives a smoother change.

The adaptive arithmetic fusion is

$$P_{\mathrm{A}}^{(k)}(\kappa) = \left(1-\alpha_\kappa(r_k)\right)P_u^{(k)} + \alpha_\kappa(r_k)P_t^{(k)}, \tag{57}$$

and the adaptive geometric fusion is

$$P_{\mathrm{G},ic}^{(k)}(\kappa) = \frac{\left(P_{u,ic}^{(k)}\right)^{1-\alpha_\kappa(r_k)}\left(P_{t,ic}^{(k)}\right)^{\alpha_\kappa(r_k)}}{\sum_{c'}\left(P_{u,ic'}^{(k)}\right)^{1-\alpha_\kappa(r_k)}\left(P_{t,ic'}^{(k)}\right)^{\alpha_\kappa(r_k)}}. \tag{58}$$

We evaluate

$$\kappa \in \{0.5,\ 1,\ 2,\ 4,\ 8\}. \tag{59}$$

B.6.3 CURRENT-BATCH FUSION GAIN

The purpose of this analysis is to determine whether fusion provides a consistent benefit at the support available at the current online step. Therefore, although both class-probability matrices are computed from all N_k observations available so far, accuracy is computed only over the current batch $\mathcal{B}_k$. For a fused class-probability matrix $P_f^{(k)}$, we define

$$\Delta_{\mathrm{batch}}^{(k)} = \mathrm{Acc}\left(P_f^{(k)}; \mathcal{B}_k\right) - \max\left\{\mathrm{Acc}\left(P_u^{(k)}; \mathcal{B}_k\right), \mathrm{Acc}\left(P_t^{(k)}; \mathcal{B}_k\right)\right\}. \tag{60}$$

A positive value indicates that fusion improves over both individual branches on the current batch, whereas a negative value indicates that fusion is worse than the better branch.

B.6.4 EXTENDED FUSION RESULTS ACROSS BENCHMARKS

Figs. 15-18 extend the fusion analysis to additional benchmarks. The results exhibit the same qualitative behavior as in the main paper. For fixed fusion, no single value of α provides a consistent advantage across target-support levels. Under limited support, assigning a large weight to P_t can lead to substantial negative fusion gains, reflecting the unreliability of class-specific statistics before sufficient per-class evidence has accumulated. This penalty gradually decreases as r increases,

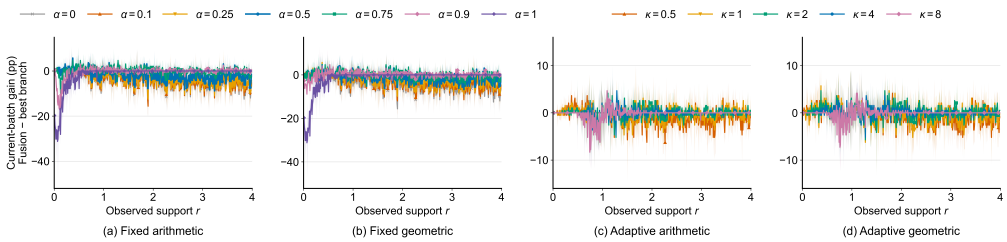

Figure 15: Fusion gain across target support on ImageNet-Sketch.

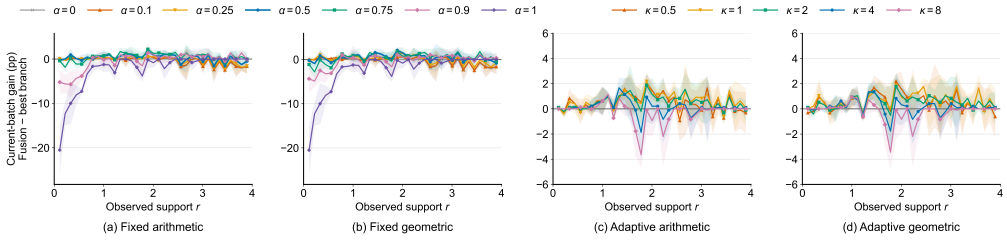

Figure 16: Fusion gain across target support on ImageNet-D109.

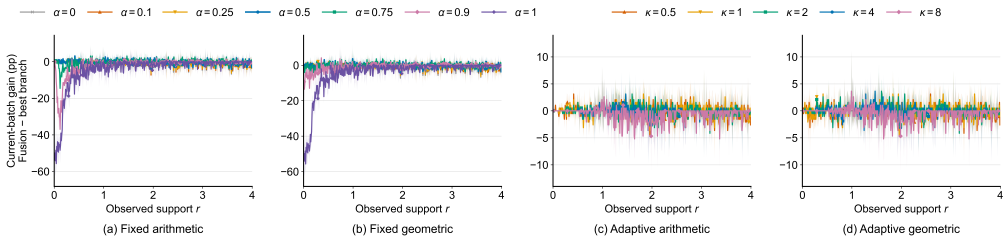

Figure 17: Fusion gain across target support on ImageNet-Cartoon.

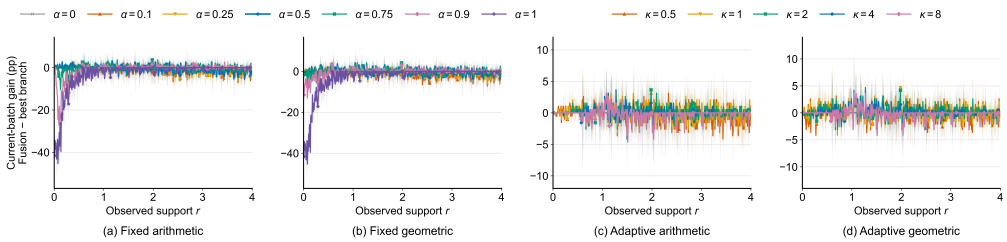

Figure 18: Fusion gain across target support on ImageNet-Drawing.

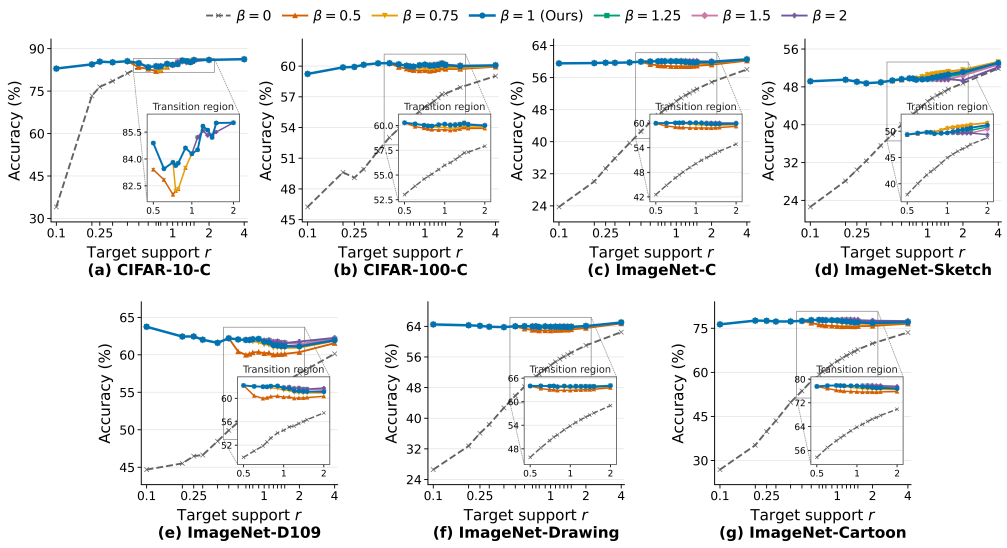

Figure 19: Online threshold ablation across all seven benchmarks. Stream-average Top-1 accuracy is reported.

but the preferred fusion weight also changes with support. Consequently, a fixed mixture cannot consistently track the evolving relative reliability of the two branches.

Support-adaptive fusion alleviates the severe low-support degradation by assigning more weight to P_u early in the stream and progressively increasing the contribution of P_t. However, across the evaluated transition rates κ, the current-batch fusion gains remain concentrated around zero and fluctuate between positive and negative values.

B.7 Additional Analysis of the Coverage-Based Threshold

Fig. 19 provides the full online threshold analysis across all seven benchmarks used in our evaluation: CIFAR-10-C, CIFAR-100-C, ImageNet-C, ImageNet-Sketch, ImageNet-D109, ImageNet-Drawing, and ImageNet-Cartoon. We follow the same online evaluation protocol and report stream-average Top-1 accuracy.

A consistent support-dependent pattern is observed across the seven benchmarks. When $\beta = 0$, class-specific statistics are used from the beginning of the stream and can cause substantial accuracy degradation under limited target support. Using a positive threshold multiplier keeps P_u active for a longer period and reduces these early low-support errors. However, a threshold that is too large delays the switch to P_t, postponing the gains from class-specific statistics after sufficient target support has accumulated.

The exact transition point varies across datasets, as the relative accuracy of P_u and P_t depends not only on the class-space size but also on the target distribution and feature geometry. Nevertheless, $\beta = 1$ remains consistently competitive around the transition region across all seven benchmarks. These results support CH_C as a common label-free reference for switching between global and class-specific statistics, without requiring target labels or dataset-specific threshold tuning.

B.8 Additional Analysis under Class Imbalance

We further examine support-dependent inference when the target class distribution is nonuniform. We construct controlled long-tailed target streams with imbalance ratio $\rho \in \{1, 2, 3, 4, 5\}$. For a dataset with C classes, we first sample a class permutation π and define the class probabilities as

$$p_{\pi(j)} = \frac{q^j}{\sum_{k=0}^{C-1} q^k}, \qquad q = \rho^{-1/(C-1)}, \qquad j = 0, \ldots, C-1. \tag{61}$$

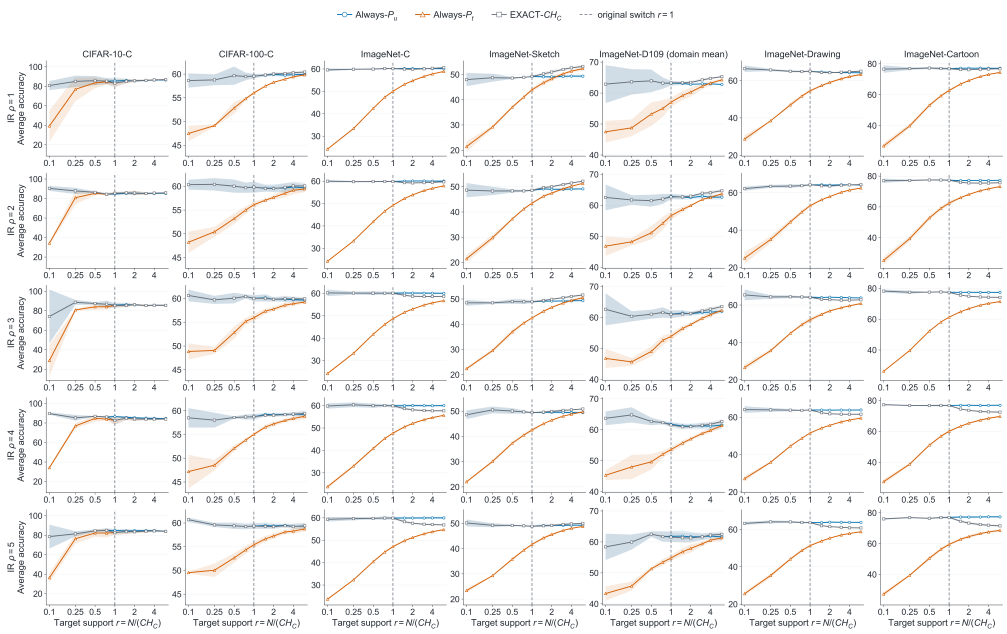

Figure 20: Support-dependent online performance under controlled class imbalance. Rows correspond to imbalance ratios $\rho \in \{1, 2, 3, 4, 5\}$, and columns correspond to the seven evaluation benchmarks. Curves show the stream-average Top-1 accuracy of Always-P_u, Always-P_t, and EXACT using $N_{\text{cov}} = CH_C$. The dashed vertical line marks the default switching point $r = N/(CH_C) = 1$.

This construction gives $p_{\max}/p_{\min} = \rho$, with $\rho = 1$ corresponding to the uniform class distribution. We evaluate the same seven benchmarks used in the main experiments with ViT-Base.

Fig. 20 shows the support-dependent behavior under increasing class imbalance. Across the seven benchmarks, P_u remains comparatively stable as target support increases, whereas P_t improves gradually as more target observations become available. Increasing ρ generally delays this recovery. Under a long-tailed class distribution, tail classes receive fewer observations than under uniform sampling, so sufficient class-specific support is reached later even when the total number of observed samples is unchanged. This effect is more visible on benchmarks with large class spaces, where P_t can remain below P_u beyond $r = 1$ under stronger imbalance.

These results clarify the role of the coverage-based threshold under class imbalance. Its main purpose is to avoid using class-specific statistics when target support is clearly insufficient. A nonuniform class distribution changes how the available observations are distributed across classes and can therefore shift the reliability transition toward larger target support. Nevertheless, CH_C remains a practical label-free reference because it depends only on the class-space size and does not require estimating the unknown target class distribution or dataset-specific threshold tuning.

To complement the branch-level analysis above, Figs. 21 and 22 extend the method-level comparison to ViT-Base. EXACT maintains positive gains over the frozen classifier across the tested imbalance ratios and target-support levels. Increasing ρ tends to reduce the gain at a fixed support level, with the effect becoming more visible around and beyond $r = 1$, where class-specific statistics are enabled. This behavior is consistent with the support analysis: under a nonuniform class distribution, minority classes receive fewer observations and therefore require more total target samples for reliable class-specific estimation. These results further support CH_C as a label-free support reference rather than an exact reliability transition under class imbalance.

B.9 Additional Analysis of the Online Batch Size

We further examine the sensitivity of EXACT to the arrival batch size under online inference. Using ViT-Base, we vary the online batch size over $\{4, 8, 16, 32, 64, 128, 256, 512\}$ and evaluate all seven benchmarks used in the main experiments. For each batch size, target samples are processed

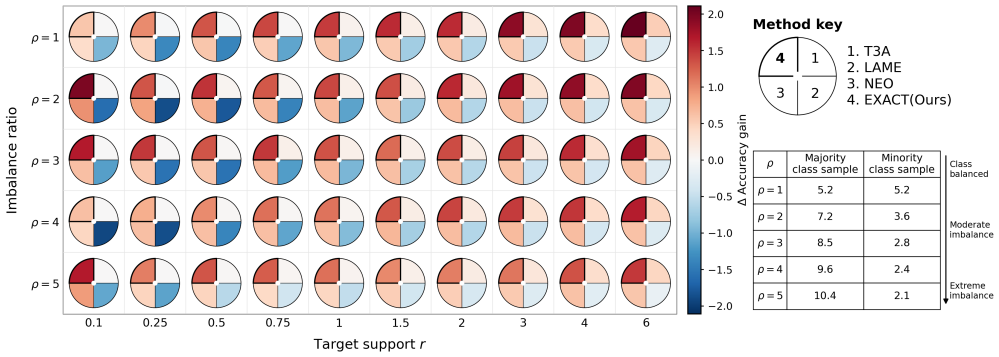

Figure 21: Top-1 accuracy gains across target support under controlled class imbalance. Results are shown for ViT-Base on CIFAR-100-C.

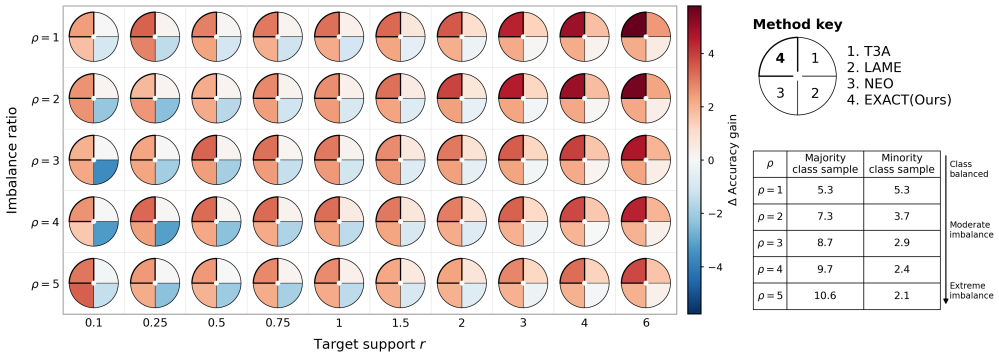

Figure 22: Top-1 accuracy gains across target support under controlled class imbalance. Results are shown for ViT-Base on ImageNet-D109.

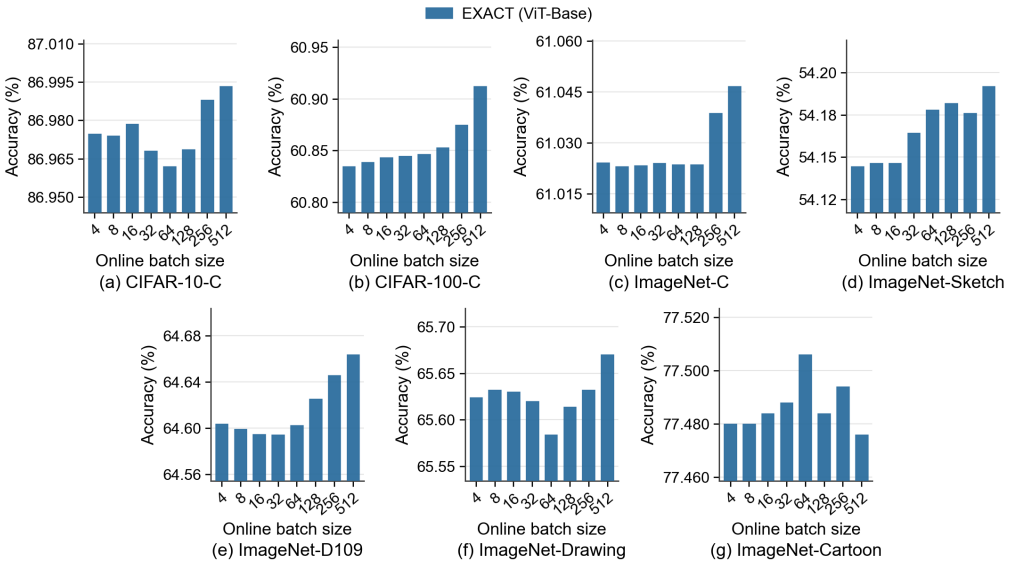

Figure 23: Sensitivity of EXACT to the online batch size. We vary the batch size from 4 to 512 under online inference with ViT-Base and report Top-1 accuracy on all seven benchmarks. Results are averaged over five random seeds.

sequentially following the same online protocol, while the pretrained model and all other inference settings are kept unchanged. We report Top-1 accuracy averaged over five random seeds.

As shown in Fig. 23, EXACT is insensitive to the online batch size. Across all seven benchmarks, the variation in accuracy from a batch size of 4 to 512 remains below 0.1 percentage points. Larger batches yield modest improvements on some datasets, such as CIFAR-100-C and ImageNet-D109; comparable performance is also obtained with substantially smaller batches. These results indicate that the performance of EXACT is not tied to a particular choice of online batch size.