SUBSPACE-CODED PARALLEL EXPLORATION FOR FORWARD-ONLY TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (ZO) optimization enables forward-only test-time adaptation (TTA), avoiding the computation and memory overhead of backpropagation. However, existing direction-centric approaches face an exploration–exploitation dilemma in capturing gradient energy: exploiting promising directions may overlook complementary gradient components, whereas broader exploration typically requires additional forward evaluations. To address this limitation, we propose Subspace-Coded Optimization via Parallel Exploration (SCOPE), a novel TTA framework that shifts from individual direction selection to higher-dimensional subspace exploration. Specifically, we introduce a perturbation encoding scheme that assigns different samples within a batch complementary combinations of shared basis directions, rather than applying the same perturbation to all samples. A lightweight least-squares estimator jointly decodes their loss differences to estimate batch-average directional responses for a shared prompt update. To support batch-coupled objectives such as feature-statistic alignment, we further decompose their batch-level loss differences into exact sample-wise contributions, preserving the original adaptation objective while making its feedback compatible with the perturbation codes. Together, these designs embed multi-directional measurements into existing batch computation, broadening gradient-energy coverage with only two forward passes. Extensive experiments across mild, continual, and wild TTA settings, together with natural-shift and quantized models, demonstrate the effectiveness and efficiency of SCOPE. Code will be publicly available.

1 INTRODUCTION

In real-world deployment, variations in sensor noise, visual appearance, and environmental conditions can cause test data to deviate from the training distribution (Hendrycks & Dietterich, 2019; Koh et al., 2021). Such shifts undermine the reliability of models that remain fixed after training. Test-Time Adaptation (TTA) addresses this challenge by leveraging unlabeled target data to mitigate performance degradation. Most existing methods (Wang et al., 2021; Niu et al., 2023; Hu et al., 2025b) rely on backpropagation (BP) to optimize unsupervised objectives. However, the associated computational and memory overhead hinders deployment on resource-constrained devices, while the requirement for backward computation limits the applicability in inference-only systems. This motivates BP-free TTA based on forward-only optimization (Niu et al., 2024; Murphy et al., 2026).

Recent BP-free TTA methods increasingly adopt zeroth-order (ZO) optimization, which estimates gradients from loss variations under parameter perturbations. Under a limited forward budget, existing methods mainly improve the selection and utilization of perturbation queries, aiming to identify directions that capture higher gradient energy. For instance, ZOA (Deng et al., 2025) leverages historical domain knowledge as adaptation priors, FOZO (Wang & Wang, 2026) dynamically adjusts the perturbation scale according to the adaptation state, and CAZO (Zhang et al., 2026) exploits curvature information to reshape the perturbation distribution toward more informative directions.

However, this direction-selection paradigm faces an inherent exploration–exploitation dilemma. Exploiting promising directions improves the utility of limited queries but may overlook complementary gradient components, while exploring more diverse directions broadens gradient coverage at the cost of additional forward evaluations. Since each additional direction requires extra perturba-

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.

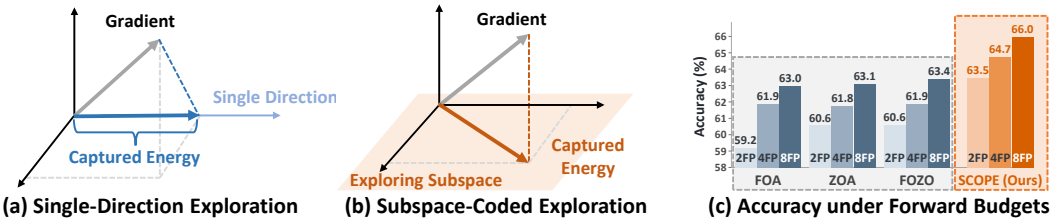

Figure 1: Comparison of exploration strategies. (a) Single-direction exploration captures limited gradient information. (b) SCOPE distributes coded perturbations across samples to explore a higher-dimensional subspace. (c) Accuracy across forward budgets. FP denotes forward passes.

tion queries, a fixed number of forward budget forces adaptation to trade the informativeness of individual queries against the breadth of gradient coverage.

To address this limitation, we propose a novel TTA framework, **S**ubspace-**C**oded **O**ptimization via **P**arallel **E**xploration (**SCOPE**), which shifts the focus from selecting a better direction to capturing gradient information over a higher-dimensional subspace. Our key idea is to let different samples within a batch explore complementary perturbation directions, rather than sharing the same perturbation. Through perturbation encoding and joint decoding, their forward responses are jointly aggregated for a shared prompt update. In this way, multiple directional measurements are embedded into the existing batch computation, enabling higher-dimensional gradient estimation with only two forward passes and capturing more gradient energy under the same forward budget.

Specifically, we introduce a perturbation encoding scheme that assigns each sample a different combination of several basis directions. With only one positive and one negative batch forward, the resulting sample-wise loss differences provide coded measurements of multiple directional responses, which are then jointly decoded through a lightweight least-squares estimator to recover the batch-average responses along the basis directions. To support adaptation objectives containing batch-coupled terms, such as feature-statistic alignment, we further introduce a code-aligned response decomposition that transforms the coupled statistical loss difference into exact sample-wise contributions. Combining these contributions with entropy differences yields complete coded responses, preserving the original adaptation objective while maintaining sample-to-code correspondence. Together, these designs enable multi-direction gradient estimation from only two forward passes, effectively expanding the explored gradient subspace without additional forward evaluations.

We evaluate SCOPE on ImageNet-C, achieving gains of **+2.85** and **+1.90** over the strongest BP-free baselines under mild and continual TTA, respectively, and an average gain of **+1.72** across two wild TTA settings. Beyond corruption shifts, SCOPE achieves the highest accuracy on ImageNet-R/Sketch/V2. We further demonstrate its applicability to W8A8-quantized models, outperforming the strongest baseline by **+2.05**. Moreover, SCOPE slightly surpasses the BP-based ReCAP while using only about **15%** of its peak GPU memory and approximately **20%** less runtime.

Contributions. **1)** We identify the *exploration–exploitation dilemma* underlying direction-centric ZO adaptation and introduce a new perspective for BP-free TTA that shifts from selecting individual perturbation directions to exploring a higher-dimensional gradient subspace under the same forward budget. **2)** We propose a perturbation encoding–decoding framework that distributes complementary direction combinations across samples and recovers multi-directional batch responses from only two forward passes. We further derive an exact sample-wise decomposition for batch-coupled adaptation objectives, making them compatible with the ZO estimator. **3)** Extensive experiments on full-precision and quantized models across diverse distribution shifts demonstrate consistent improvements over BP-free TTA methods and competitive accuracy with BP-based counterparts, achieving a favorable accuracy–efficiency trade-off under stringent forward-query budgets.

2 Related Work

2.1 Test-Time Adaptation (TTA)

TTA adapts pretrained models under distribution shifts using unlabeled target data. Prominent approaches construct unsupervised objectives and update model parameters through backpropagation (BP). Self-supervised test-time training adapts models through auxiliary tasks (Sun et al., 2020; Liu et al., 2021), while normalization-based approaches update target-domain activation

statistics (Schneider et al., 2020). Other methods address temporally correlated or evolving test streams (Gong et al., 2022; Yuan et al., 2023). Entropy minimization promotes confident predictions (Wang et al., 2021; Zhang et al., 2022). To suppress unreliable updates, reliability-aware approaches select samples using confidence- and perturbation-based criteria (Niu et al., 2022; 2023; Hu et al., 2025a;b). Beyond selection, another line enriches adaptation objectives with consistency constraints over local feature regions or semantic–geometric representations (Lee et al., 2024; Hu et al., 2025c). Complementary efforts preserve adaptation knowledge through anti-forgetting regularization, source-parameter restoration, or historical prompt assets for evolving distributions (Niu et al., 2022; Wang et al., 2022; Hu et al., 2026). TTA has been explored in 3D point cloud classification (Wang et al., 2024; Yazdanpanah et al., 2025), vision-language models (Karmanov et al., 2024; Zhou et al., 2025), and video segmentation and restoration (Liu et al., 2024; Yang et al., 2024).

However, BP-based updates incur activation-storage and backward-computation costs, limiting deployment on resource-constrained or inference-only systems (Niu et al., 2024). BP-free alternatives offer different adaptation mechanisms. T3A (Iwasawa & Matsuo, 2021) adjusts the classifier using online class pseudo-prototypes, whereas LAME (Boudiaf et al., 2022) refines output probabilities through a Laplacian-based objective without updating network parameters. NEO (Murphy et al., 2026) re-centers target embeddings using an estimated global centroid, avoiding iterative optimization. FOA (Niu et al., 2024) performs objective-driven prompt search with CMA-ES, complemented by back-to-source activation shifting. These approaches enable adaptation without backpropagation.

2.2 Zeroth-Order Optimization

Zeroth-order (ZO) methods estimate gradients from function evaluations, with classical simultaneous-perturbation and random-direction estimators as foundations (Spall, 1992; Nesterov & Spokoiny, 2017). Query-efficient search uses orthogonal perturbations (Choromanski et al., 2018), surrogate-gradient guidance (Maheswaranathan et al., 2019), or adaptive active subspaces (Choromanski et al., 2019). For language-model fine-tuning, MeZO (Malladi et al., 2023) implements ZO-SGD through in-place perturbations, achieving inference-level memory usage. Within TTA, ZOA (Deng et al., 2025) combines ZO updates with continual domain knowledge learning to accumulate and reuse adaptation knowledge for quantized models. FOZO (Wang & Wang, 2026) employs zeroth-order prompt optimization with dynamic perturbation scaling and deep–shallow feature-statistic alignment. CAZO (Zhang et al., 2026) uses a moving-average diagonal curvature proxy to guide anisotropic perturbation sampling and reduce gradient-estimation variance. Nevertheless, the multi-direction estimators in FOZO and CAZO obtain additional directional feedback through repeated batch evaluations. SCOPE instead distributes coded direction combinations across samples and jointly decodes their responses, expanding the explored gradient subspace within two batch forward passes. Its sample-wise loss-difference decomposition makes batch-coupled adaptation objectives compatible with these parallel measurements.

3 Preliminary

Forward-Only Prompt Adaptation. Given an unlabeled test batch $\mathcal{B}_t = \{x_i\}_{i=1}^{B}$, we prepend shared learnable prompt tokens P_t to a vision transformer (Shu et al., 2022; Jia et al., 2022; Abdul Samadh et al., 2023), forming $[P_t; E(x_i)]$, where $E(x_i)$ contains the classification token and patch embeddings. With the prompt vectorized as $P_t \in \mathbb{R}^d$, adaptation minimizes the unsupervised batch objective $\mathcal{L}(P; \mathcal{B}_t)$ through forward evaluations, while all pretrained model parameters remain frozen (Niu et al., 2024; Wang & Wang, 2026). Following prior ZO analyses (Wang & Wang, 2026; Zhang et al., 2026), we assume local smoothness of the batch objective, under which perturbation-induced loss differences provide local gradient information without backpropagation.

Assumption 1 (**Local Smoothness**). *The batch objective has an L-Lipschitz gradient in a neighborhood of P_t containing all perturbation segments:*

$$\|\nabla_P \mathcal{L}(P; \mathcal{B}_t) - \nabla_P \mathcal{L}(P'; \mathcal{B}_t)\|_2 \le L\|P - P'\|_2. \tag{1}$$

This condition controls local gradient variation for the finite-difference approximation below.

Zeroth-Order Gradient Estimation. We use a Gaussian symmetric two-point ZO estimator (Nesterov & Spokoiny, 2017; Wang & Wang, 2026), sampling $Z \sim \mathcal{N}(0, I_d)$ with perturbation scale $\epsilon_t > 0$. The positive–negative loss difference, normalized by $2\epsilon_t$, defines the finite-difference response $\widehat{a}(Z)$, which under Assumption 1 approximates the true directional response $\bar{a}(Z)$:

$$\widehat{a}(Z) = \frac{\mathcal{L}(P_t + \epsilon_t Z; \mathcal{B}_t) - \mathcal{L}(P_t - \epsilon_t Z; \mathcal{B}_t)}{2\epsilon_t} \approx \bar{a}(Z) := Z^\top \nabla_P \mathcal{L}(P_t; \mathcal{B}_t). \tag{2}$$

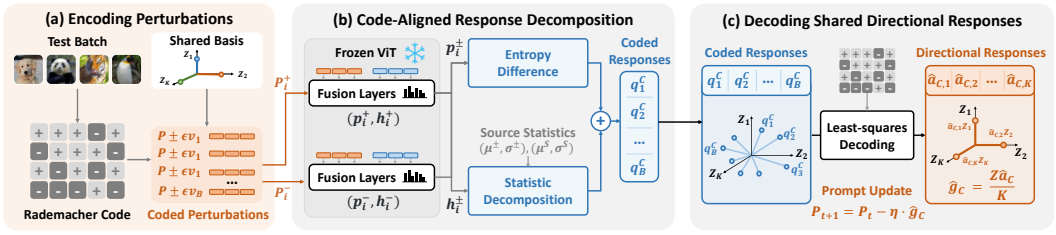

Figure 2: Overview of our SCOPE framework. (a) Shared basis directions are Rademacher-coded into complementary sample-wise prompt perturbations across the batch. (b) Two batch forward passes yield code-aligned responses by combining entropy differences with exact sample-wise statistic decomposition for batch-coupled objectives. (c) Least-squares decoding recovers batch-average directional responses, which are aggregated to estimate the gradient for updating the shared prompt.

The response $\widehat{a}(Z)$ provides the coefficient of a single-direction gradient estimate $\widehat{a}(Z)Z$. Aggregating K independently sampled directions gives

$$\widehat{\nabla}\mathcal{L}(P_t;\mathcal{B}_t) = \frac{1}{K}\sum_{k=1}^{K}\widehat{a}(Z_k)Z_k, \qquad P_{t+1} = P_t - \eta\widehat{\nabla}\mathcal{L}(P_t;\mathcal{B}_t), \tag{3}$$

where $\eta > 0$ is the learning rate. Since each direction is shared across the batch and requires one positive and one negative evaluation, estimating K directional responses requires $2K$ batch forward passes, coupling broader gradient coverage with linearly increasing evaluation cost.

4 Methodology

In this section, we present SCOPE, which turns batch computation into parallel subspace exploration. Its perturbation encoding–decoding scheme obtains multi-directional feedback, while sample-wise loss-difference decomposition makes the complete adaptation objective compatible with these measurements. We describe both components and then summarize the overall procedure.

4.1 Subspace-Coded Parallel Exploration

To move beyond the single-direction updates of existing zeroth-order TTA methods (Deng et al., 2025; Zhang et al., 2026; Wang & Wang, 2026), we use the batch dimension as a parallel exploration channel, assigning each sample a coded combination of K random basis directions within the same two-forward budget. The resulting finite-difference responses are then jointly decoded into batch-average directional coefficients for updating the shared prompt.

Encoding Complementary Perturbations. To broaden gradient coverage without additional full-batch evaluations, SCOPE distributes complementary direction combinations across samples. At test step t, we sample K independent Gaussian directions, $Z_k \sim \mathcal{N}(0, I_d)$, which define the exploration subspace $\mathcal{S} = \mathrm{span}\{Z_1, \ldots, Z_K\}$. To distribute this subspace across the batch, the encoding must satisfy two requirements. First, different samples should receive signed combinations of the basis directions while maintaining a consistent perturbation scale. Second, these combinations must collectively preserve all K directions so that no dimension of $\mathcal{S}$ is lost during encoding. We therefore sample a scaled Rademacher code matrix from the set of full-column-rank matrices:

$$v_i = \sum_{k=1}^{K} c_{ik}Z_k, \qquad P_i^{\pm} = P_t \pm \epsilon_t v_i, \tag{4}$$

where the code row $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})$ determines how sample i combines the shared basis directions. Positive perturbations $\{P_i^+\}_{i=1}^B$ are evaluated in one batch forward pass, and negative perturbations $\{P_i^-\}_{i=1}^B$ in another. The following proposition establishes that the full-rank encoding preserves the exploration subspace and quantifies the gradient energy within K-dimensional span.

Proposition 1 (Gradient-Energy Coverage). *Fix the current batch and prompt center, and let* $g = \nabla_P\mathcal{L}(P_t;\mathcal{B}_t)$. *Since* C *has full column rank, the encoded perturbations satisfy* $\mathrm{span}\{v_1, \ldots, v_B\} = \mathcal{S}$. *For the* K *independent Gaussian directions,*

$$\mathbb{E}_{Z_1,\ldots,Z_K}\|\Pi_{\mathcal{S}}g\|_2^2 = \frac{K}{d}\|g\|_2^2, \tag{5}$$

where $\Pi_{\mathcal{S}} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ *denotes the orthogonal projection onto* $\mathcal{S}$, *with* $\mathbf{Z} = [Z_1, \ldots, Z_K]$.

This result clarifies the complementary roles of the code matrix and the Gaussian basis. The encoded perturbations collectively span a space containing K times the expected projected gradient energy of a single isotropic direction. SCOPE therefore converts the batch dimension into a K-fold increase in expected gradient-energy coverage while retaining the same two forward passes, providing a richer set of directional responses for the subsequent decoder.

Decoding Shared Directional Responses. The coded perturbations broaden the explored subspace, but their sample-wise loss differences cannot be used directly to update the shared prompt. Each observed difference corresponds to a mixture of basis directions, whereas the prompt update requires a separate batch-level response along each Z_k. We therefore treat the sample-wise finite differences as coded measurements and jointly decode their shared directional components.

Under classical Assumption 1, the positive–negative loss difference for sample i approximates its directional response along v_i, with a bounded finite-difference error. Because each perturbation combines multiple basis directions, the resulting observation jointly carries their directional responses:

$$q_i^C = \frac{\mathcal{L}(x_i, P_i^+) - \mathcal{L}(x_i, P_i^-)}{2\epsilon_t} = \sum_{k=1}^{K} c_{ik} a_i(Z_k) + r_i, \qquad |r_i| \le \frac{L\epsilon_t \|v_i\|_2^2}{2}, \tag{6}$$

where $a_i(Z_k) = Z_k^\top \nabla_P \mathcal{L}(x_i, P_t)$ denotes the response of sample i along Z_k, and r_i is the finite-difference residual. While these measurements are obtained sample-wise, the shared prompt should be updated according to the gradient of the batch objective. The desired coefficient for each basis direction is therefore the batch-average directional response:

$$\bar{a}(Z_k) = \frac{1}{B} \sum_{i=1}^{B} a_i(Z_k) = Z_k^\top g, \qquad k = 1, \dots, K, \tag{7}$$

where $\bar{\mathbf{a}} = (\bar{a}(Z_1), \dots, \bar{a}(Z_K))^\top$ forms the decoding target. The observed responses, however, do not directly measure $\bar{\mathbf{a}}$. Each q_i^C is generated by applying the code row $\mathbf{c}_i$ to the local directional responses of sample i, which may deviate from their batch averages. Decomposing each local response as its batch-average component and sample-specific deviation gives

$$q^C = C\bar{\mathbf{a}} + \xi + r, \qquad \xi_i = \sum_{k=1}^{K} c_{ik} \big(a_i(Z_k) - \bar{a}(Z_k)\big), \tag{8}$$

where $C\bar{\mathbf{a}}$ is the coded observation predicted by the shared batch-level responses, ξ captures inter-sample variation around these responses, and r collects the finite-difference residuals. Because $K \le B$ and C has full column rank, the shared K-dimensional response vector can be uniquely estimated in the least-squares sense from the B coded observations:

$$\widehat{\mathbf{a}}_C = \arg\min_{\mathbf{a} \in \mathbb{R}^K} \|q^C - C\mathbf{a}\|_2^2 = (C^\top C)^{-1} C^\top q^C. \tag{9}$$

Together, the encoding–decoding design converts sample-level perturbation diversity into shared multi-directional feedback, turning the batch dimension into a parallel exploration channel under the same two-forward budget. We next extend this mechanism to the batch-coupled adaptation objective through sample-wise loss-difference decomposition.

4.2 Code-aligned Response Decomposition

The shared directional decoder requires one scalar response per sample, paired with the corresponding perturbation code. Sample-separable losses provide this interface, whereas shared batch statistics couple all samples within each coded branch, causing the loss difference to entangle the effects of different perturbations. To recover the sample-to-code correspondence, we transform the coupled statistic-alignment difference into an additive form that exposes sample-level contributions.

Statistic-Decoupling Transformation. Following prior TTA methods (Niu et al., 2024; Wang & Wang, 2026), we adopt an adaptation objective combining entropy with statistic alignment:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} H(p_i) + \frac{\lambda}{B} \left(\|\mu - \mu^S\|_2^2 + \|\sigma - \sigma^S\|_2^2 \right), \tag{10}$$

where p_i and h_i denote the prediction and feature of sample i, H is prediction entropy, (μ, σ) and (μ^S, σ^S) denote the target-batch and source statistics, and λ balances two terms. The entropy is directly separable across samples. In contrast, the statistic-alignment terms are batch-coupled as μ and σ aggregate the features within each coded branch. Their loss differences therefore mix the effects of different perturbation codes and cannot be assigned to individual samples directly.

Let $+$ and $-$ denote quantities obtained from the positive and negative coded branches, respectively. To decouple the batch-level alignment difference, we apply the bilinear factorization to both quadratic alignment terms. The mean change can be expanded over sample features, whereas the standard-deviation change is rationalized into a variance difference and then expanded sample-wise. This transformation expresses the coupled alignment difference as an average of sample-indexed inner products, separating the shared statistical coefficients from the sample-specific feature changes:

$$m_i = \left\langle (\mu^+ + \mu^- - 2\mu^S) + \frac{\sigma^+ + \sigma^- - 2\sigma^S}{\sigma^+ + \sigma^-} \odot (h_i^+ + h_i^- - \mu^+ - \mu^-),\, h_i^+ - h_i^- \right\rangle. \tag{11}$$

The first term inside the leading factor is the shared coefficient induced by the mean-alignment difference. The second combines the standard-deviation alignment coefficient with the centered feature sum, thereby attributing the corresponding variance change to sample i. Crucially, this decoupling is an exact algebraic identity rather than a heuristic attribution: averaging m_i over the batch recovers the complete statistic-alignment difference between the two coded branches.

Sample-Wise Contribution Construction. Since prediction entropy provides a sample-wise difference, we combine it with the statistical contribution m_i to construct the complete coded response:

$$q_i^C = \frac{H(p_i^+) - H(p_i^-) + \frac{\lambda}{B} m_i}{2\epsilon_t}, \qquad \frac{1}{B}\sum_{i=1}^{B} q_i^C = \frac{\mathcal{L}^+ - \mathcal{L}^-}{2\epsilon_t}, \tag{12}$$

where $\mathcal{L}^{\pm}$ denotes the original objective evaluated on the complete positive or negative coded batch. The second equality establishes an exact conservation property: the batch average of the constructed responses recovers the finite difference of the original adaptation objective. The decomposition therefore preserves both the original objective and the correspondence between each response and its perturbation code. The resulting responses can be directly supplied to the decoder in Eq. 9, extending multi-directional estimation to the complete batch-coupled objective.

4.3 Overall Procedure of SCOPE

Combining parallel exploration with sample-wise loss-difference decomposition yields the complete SCOPE procedure. At each test step, SCOPE samples K basis directions and a code matrix to form complementary sample-wise perturbations, then evaluates the positive and negative coded batches. Averaging the two branches' logits yields classification predictions (Wang & Wang, 2026), while branch-specific predictions and features produce sample-wise responses, jointly decoded into batch-average directional coefficients for prompt adaptation:

$$\widehat{\nabla}_C \mathcal{L}(P_t; \mathcal{B}_t) = \frac{1}{K}\sum_{k=1}^{K} \widehat{a}_{C,k} Z_k, \qquad P_{t+1} = P_t - \eta \widehat{\nabla}_C \mathcal{L}(P_t; \mathcal{B}_t). \tag{13}$$

Pretrained model parameters remain frozen throughout adaptation. SCOPE explores and aggregates multiple directional responses within only two batch forward passes, achieving broader gradient coverage under the same forward budget without backpropagation or batch expansion.

5 Experiments

5.1 Experimental Setup

Datasets and Evaluation Protocols. Following prior protocols (Hu et al., 2025b; Zhang et al., 2026; Wang & Wang, 2026), we evaluate SCOPE on ImageNet-C (Hendrycks & Dietterich, 2019) at severity level 5, covering 15 corruption types across noise, blur, weather, and digital categories. Each corruption is evaluated using 50,000 test images. We consider the following settings: 1) Mild TTA, which adapts independently to each corruption and resets the adaptation state between domains;

Table 1: Top-1 accuracy (%) on ImageNet-C (severity 5) under **Mild TTA**.

Method	FP	Noise			Blur				Weather				Digital				Avg↑
		Gaus.	Shot	Impu.	Defo.	Gla.	Moti.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pixel	JPEG	
NoAdapt	1	56.76	56.79	57.50	46.88	35.58	53.13	44.82	62.23	62.58	65.72	77.66	32.64	46.03	66.95	67.60	55.52
T3A (NeurIPS'21)	1	56.34	56.87	57.31	47.90	37.73	54.20	47.00	63.62	60.78	68.31	78.11	38.42	49.86	67.52	68.94	56.86
LAME (CVPR'22)	1	56.47	56.50	57.24	46.36	34.83	52.69	44.17	58.55	61.53	63.07	77.41	24.63	44.59	66.58	67.29	54.13
FOA (ICML'24)	2	58.00	58.21	58.62	51.10	39.07	55.48	48.08	64.69	62.86	69.70	77.96	56.59	50.41	68.02	69.19	59.20
ZOA (ACMMM'25)	2	58.90	59.37	60.33	53.29	44.10	57.58	52.15	65.72	62.11	67.02	78.44	56.41	54.24	69.42	69.43	60.57
NEO (ICLR'26)	1	57.49	57.55	58.27	49.73	38.12	54.83	47.52	64.64	65.04	**71.23**	78.11	58.37	49.94	67.64	68.80	59.15
CAZO (CVPR'26)	3	55.79	56.42	57.33	51.32	45.11	57.43	53.19	65.03	63.59	60.72	78.56	56.38	54.64	69.93	67.96	59.56
FOZO (CVPR'26)	2	58.93	59.13	59.56	53.34	42.31	57.48	50.51	65.99	62.74	67.80	78.37	60.86	53.40	68.99	69.91	60.62
SCOPE (Ours)	2	**60.57**	**61.60**	**62.26**	**56.78**	**48.34**	**59.56**	**53.66**	**67.66**	**65.73**	69.18	**80.03**	**65.97**	**57.83**	**71.25**	**71.61**	**63.47**

Table 2: Top-1 accuracy (%) on ImageNet-C (severity 5) under **Continual TTA**.

Method	FP	Noise			Blur				Weather				Digital				Avg↑
		Gaus.	Shot	Impu.	Defo.	Gla.	Moti.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pixel	JPEG	
NoAdapt	1	56.76	56.79	57.50	46.88	35.58	53.13	44.82	62.23	62.58	65.72	77.66	32.64	46.03	66.95	67.60	55.52
T3A (NeurIPS'21)	1	56.34	56.72	56.74	45.32	34.42	51.82	43.54	60.42	62.76	60.48	77.02	45.86	44.49	66.71	68.41	55.40
LAME (CVPR'22)	1	56.47	56.54	57.24	46.37	34.79	52.75	44.20	58.45	61.56	63.03	77.42	24.67	44.61	66.62	67.29	54.13
FOA (ICML'24)	2	57.89	60.94	62.41	49.82	40.80	56.79	52.35	65.73	66.06	65.19	79.77	61.32	52.68	69.89	71.11	60.85
ZOA (ACMMM'25)	2	58.82	60.97	62.27	53.33	46.27	**58.93**	53.89	65.40	62.49	66.17	78.59	60.03	55.36	69.89	69.37	61.45
NEO (ICLR'26)	1	57.49	57.55	58.29	49.72	38.14	54.85	47.54	64.66	65.07	**71.24**	78.12	58.40	49.95	67.66	68.80	59.17
CAZO (CVPR'26)	3	54.87	55.14	57.57	50.39	43.00	58.17	53.26	65.42	66.02	70.16	79.08	62.07	50.55	70.65	**71.79**	60.54
FOZO (CVPR'26)	2	59.06	62.14	**63.34**	51.11	43.02	57.21	52.47	66.13	66.39	66.89	79.43	61.70	54.67	68.29	69.85	61.45
SCOPE (Ours)	2	**60.53**	**62.74**	63.28	**53.83**	**48.93**	58.40	**55.78**	**66.40**	**68.24**	68.16	**80.42**	**64.52**	**56.77**	**70.85**	71.37	**63.35**

2) Continual TTA, which processes corruption domains sequentially without resets; 3) Quantized-model Adaptation (W8A8), which follows the mild protocol with 8-bit weight and activation quantization. Additional results under wild TTA settings (mixed testing domains and imbalanced label shifts) and natural distribution shifts (ImageNet-R/Sketch/V2) (Hendrycks et al., 2021; Wang et al., 2019; Recht et al., 2019) are provided in Appendix.

Compared Methods. We compare SCOPE with seven BP-free and three BP-based TTA methods, using NoAdapt as the unadapted reference. The BP-free baselines include T3A (Iwasawa & Matsuo, 2021), LAME (Boudiaf et al., 2022), and NEO (Murphy et al., 2026), alongside forward-only optimization methods: FOA (Niu et al., 2024) performs evolutionary prompt search, whereas ZOA (Deng et al., 2025), FOZO (Wang & Wang, 2026), and CAZO (Zhang et al., 2026) employ zeroth-order updates. The BP-free comparisons use tight forward budgets: SCOPE, FOA, ZOA, and FOZO use two forward passes per batch, CAZO uses three, and T3A, LAME, and NEO use one. We additionally evaluate BP-based TENT (Wang et al., 2021), SAR (Niu et al., 2023), and ReCAP (Hu et al., 2025b), comparing accuracy, runtime and memory to assess the accuracy–efficiency trade-off.

Implementation Details. We use a pretrained ViT-B/16 (Dosovitskiy et al., 2021) model from the timm repository (Wightman, 2019) and obtain its W8A8-quantized counterpart using PTQ4ViT (Jacob et al., 2018; Yuan et al., 2022). Following FOZO (Wang & Wang, 2026), we use 3 prompt tokens, batch size $B = 64$, learning rate $\eta = 0.08$, and alignment weight $\lambda = 0.5$. Source feature statistics are precomputed from ImageNet-1K (Deng et al., 2009; Russakovsky et al., 2015) and kept fixed during adaptation. For subspace-coded exploration, we sample $K = 16$ basis directions per batch. The perturbation scale is fixed at $\epsilon_t = 0.5$ for full-precision models and $\epsilon_t = 2.0$ for W8A8 models. We update the shared prompt once per batch while keeping all pretrained model parameters frozen. To mitigate the effects of random fluctuations, we report the mean performance over 5 independent runs. Additional implementation details are provided in Appendix.

5.2 Evaluation Results

Mild Test-Time Adaptation. To evaluate the effectiveness of SCOPE under individual corruption shifts, we compare BP-free methods in the mild TTA setting. As shown in Tab. 1, SCOPE substantially improves adaptation performance, achieving 63.47% average accuracy and surpassing FOZO by +2.85 with the same two batch forward passes. Beyond ranking first on 14 of 15 corruptions, SCOPE achieves the highest category-wise average across all four corruption groups, whereas the

Table 3: Top-1 accuracy (%) on ImageNet-C (severity 5) under **W8A8-quantized Mild TTA**.

Method	FP	Noise			Blur				Weather				Digital				Avg↑
		Gaus.	Shot	Impu.	Defo.	Gla.	Moti.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pixel	JPEG	
NoAdapt	1	56.02	56.24	56.98	46.96	34.79	52.66	42.94	61.11	61.50	66.33	77.13	21.11	44.62	65.70	66.45	54.04
T3A (NeurIPS'21)	1	55.79	56.28	56.84	47.85	36.72	53.58	45.19	62.63	59.59	68.74	77.60	28.77	48.30	66.43	67.87	55.48
LAME (CVPR'22)	1	55.78	56.01	56.73	46.61	34.06	52.24	42.24	57.40	60.34	65.11	76.90	10.36	43.16	65.43	66.10	52.56
FOA (ICML'24)	2	56.58	56.92	57.57	49.85	37.32	54.36	45.82	63.50	62.67	71.45	77.41	58.62	48.30	66.38	67.90	58.31
ZOA (ACMMM'25)	2	57.30	58.33	58.85	52.29	43.28	57.03	**50.72**	64.72	61.91	66.96	77.83	51.70	52.99	68.01	68.39	59.36
NEO (ICLR'26)	1	56.85	56.91	57.62	49.41	37.16	54.29	45.70	63.58	64.15	71.03	77.59	55.72	48.37	66.43	67.66	58.17
CAZO (CVPR'26)	3	55.08	55.31	56.40	49.21	**44.02**	55.96	49.57	64.29	62.04	59.22	78.14	51.67	51.75	67.77	67.46	57.86
FOZO (CVPR'26)	2	56.72	57.03	57.63	49.78	38.32	55.04	47.36	64.21	**64.82**	69.40	77.66	49.51	50.93	66.80	68.38	58.24
SCOPE (Ours)	2	**60.02**	**60.32**	**60.95**	**55.61**	42.28	**57.57**	50.59	**66.39**	63.19	**71.85**	**78.20**	**61.33**	**53.34**	**69.35**	**70.19**	**61.41**

Table 4: Comparison with BP-based TTA methods on ImageNet-C (severity 5) under Continual TTA. We report Top-1 accuracy (%), runtime per batch (ms), and peak GPU memory (MiB).

Method	BP	Noise			Blur				Weather				Digital				Time (ms)	Mem. (MiB)	Avg↑
		Gaus.	Shot	Impu.	Defo.	Gla.	Moti.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pixel	JPEG			
NoAdapt	×	56.76	56.79	57.50	46.88	35.58	53.13	44.82	62.23	62.58	65.72	77.66	32.64	46.03	66.95	67.60	35.3	819	55.52
TENT(ICLR'21)	✓	60.34	**63.83**	**64.73**	**54.82**	56.30	**60.92**	55.72	61.73	62.83	**69.90**	78.60	**65.07**	28.66	24.06	23.77	83.9	5495	55.42
SAR (ICLR'23)	✓	59.25	61.36	61.65	54.29	55.44	58.90	56.01	60.78	61.93	64.89	76.80	60.47	**57.78**	68.78	69.07	166.3	5495	61.83
ReCAP (ICML'25)	✓	59.22	62.07	61.68	53.33	**56.72**	59.84	**58.48**	63.18	63.86	68.97	77.39	59.50	57.72	**72.23**	70.86	87.8	5699	63.00
SCOPE (Ours)	×	**60.53**	62.74	63.28	53.83	48.93	58.40	55.78	**66.40**	**68.24**	68.16	**80.42**	64.52	56.77	70.85	**71.37**	69.4	834	**63.35**

strongest baseline varies by group. These results demonstrate broad adaptation gains across distinct corruption families under a tight forward budget.

Continual Test-Time Adaptation. To evaluate adaptation under evolving distributions, we compare methods across sequential corruption domains without resetting adaptation states. In Tab. 2, SCOPE achieves 63.35% average accuracy, outperforming ZOA and FOZO by +1.90. SCOPE achieves the highest group-wise average across four corruption categories. These results demonstrate SCOPE's consistent advantage across heterogeneous shifts in the continual adaptation stream.

Quantized-Model Adaptation. To assess applicability to resource-constrained deployment, we evaluate SCOPE on W8A8-quantized models, where limited support for backpropagation motivates forward-only adaptation. In Tab. 3, SCOPE achieves 61.41% average accuracy, surpassing ZOA, designed for quantized-model adaptation, by +2.05. These results highlight SCOPE's ability to retain a strong adaptation advantage under reduced precision using only two batch forward passes.

Comparison with BP-based Methods. Beyond quantized-model compatibility, we compare SCOPE with BP-based methods under continual TTA to assess its accuracy–efficiency trade-off. As shown in Tab. 4, SCOPE achieves the highest average accuracy of 63.35%, slightly surpassing the strongest BP-based method, ReCAP. Meanwhile, SCOPE requires only 834 MiB of memory and 69.4 ms per batch, corresponding to about **15%** of ReCAP's memory usage and **20%** lower runtime. These results demonstrate that SCOPE achieves competitive accuracy with BP-based adaptation while maintaining near-inference memory requirements and lower latency.

Additional evaluations under wild TTA (mixed testing domains and imbalanced label shifts) and on ImageNet-R/Sketch/V2 are provided in Appendix.

6 Ablation Study and Visualization

In this section, we conduct ablation studies and visualizations on ImageNet-C (severity 5) with ViT-B/16, focusing on the effectiveness of subspace-coded exploration and the robustness of its key hyperparameters. Additional experiments and analyses are provided in Appendix.

6.1 Effectiveness of Subspace-Coded Exploration

To investigate whether broader exploration provides more useful gradient information, we compare SCOPE with FOZO, CAZO, and ZOA throughout adaptation. Gradient coverage measures the fraction of reference-gradient energy within the queried span, while gradient alignment measures the cosine similarity between the estimated and reference gradients. As shown in Fig. 3(a), SCOPE

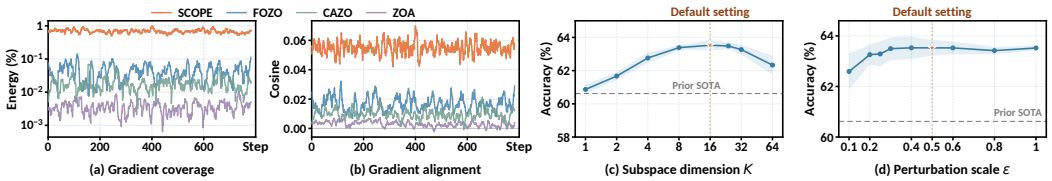

Figure 3: Mechanism analysis and hyperparameter sensitivity of SCOPE. (a) Gradient-energy coverage (%, log scale) and (b) gradient alignment during adaptation. Accuracy with varying (c) subspace dimension K and (d) fixed perturbation scale ϵ. Stars mark the default settings.

consistently achieves higher gradient-energy coverage than the compared methods, maintaining this advantage across the recorded adaptation steps. To examine whether this broader coverage translates into useful feedback, we further inspect the gradient alignment in Fig. 3(b). SCOPE produces estimates with greater cosine similarity to the reference gradients computed by backpropagation, indicating that its broader gradient coverage translates into more accurate update directions. Together, these observations support the use of complementary sample-wise perturbations to obtain multi-directional feedback for a shared update, with SCOPE requiring only two batch forward passes.

6.2 Sensitivity Analysis of Hyperparameters

We analyze two key hyperparameters of SCOPE: the subspace dimension K, which controls the breadth of parallel exploration, and the perturbation scale ϵ, which determines the local probing range. As shown in Fig. 3(c), increasing K consistently improves accuracy up to $K = 16$, confirming the benefit of exploring a broader gradient subspace. Further expansion leads to performance degradation, suggesting that excessively large subspaces make reliable recovery of directional feedback more difficult under a fixed batch budget. We therefore set $K = 16$ by default. Fig. 3(d) further shows that SCOPE remains consistently strong across a broad range of fixed perturbation scales. Accuracy peaks around $\epsilon = 0.5$ and changes only mildly over $[0.2, 1.0]$. We use $\epsilon = 0.5$, which lies in this stable high-performing region. These results indicate that SCOPE benefits from sufficiently broad subspace exploration while remaining robust to the perturbation magnitude.

6.3 Effectiveness of Components

We investigate subspace-coded exploration and sample-wise loss decomposition using variants in Tab. 5. From entropy-only shared-direction ZO (A), coded exploration (C) consistently improves accuracy across all four settings. Adding feature-statistic alignment to A (B) yields substantial gains. Replacing uniform allocation with the proposed sample-wise decomposition (E) improves all four settings, highlighting the importance of preserving sample-to-code correspondence for batch-coupled feedback. The complete framework (E) consistently outperforms B under the same adaptation objective and forward budget. These comparisons demonstrate the benefit of coded exploration and how sample-wise decomposition extends it to batch-coupled objectives.

Table 5: Impact of component variations on ImageNet-C. **Coding** denotes subspace-coded parallel exploration, **Stat.** denotes feature-statistic alignment, and **Decomp.** denotes the proposed sample-wise loss-difference decomposition. All variants use prediction entropy.

ID	Coding	Stat.	Decomp.	Mild	CTTA	Mixed	W8A8
A				58.17	36.19	36.61	59.15
B		✓		60.93	61.67	60.14	60.17
C	✓			59.27	38.24	37.99	59.36
D	✓	✓		60.16	40.63	39.63	59.54
E	✓	✓	✓	**63.47**	**63.35**	**61.13**	**61.41**

7 Conclusion

In this paper, we propose SCOPE, a forward-only TTA framework that turns batch computation into parallel subspace exploration for broader gradient-energy coverage. Our perturbation encoding–decoding scheme distributes complementary direction combinations across samples and jointly estimates batch-average directional responses for a shared prompt update, using only two batch forward passes. We further derive a code-aligned response decomposition that expresses batch-coupled statistic-alignment loss differences as exact sample-wise contributions, preserving the original adaptation objective and sample-to-code correspondence. Extensive experiments across diverse distribution shifts and quantized-model settings demonstrate SCOPE's effectiveness and favorable accuracy–efficiency trade-off. We hope this work inspires future research to move beyond individual direction selection toward sample-wise parallel exploration of higher-dimensional gradient subspaces under tight forward budgets.

AI USE STATEMENT

We used LLMs as an auxiliary tool to refine and improve the clarity of our writing. They were not used to generate experimental data or any results reported in this work. All AI-assisted revisions were carefully reviewed, further edited, and verified by the authors, who assume full responsibility for the final content of the manuscript. We advocate for the transparent and responsible use of LLMs in academic and research writing.

ETHICS STATEMENT

This study relies on publicly available image-classification benchmarks, including ImageNet-C, ImageNet-R, ImageNet-Sketch, and ImageNet-V2, and uses them in compliance with their respective licenses and terms of use. The work does not involve human participants or the collection of new data, and the models evaluated in our experiments neither rely on demographic information nor make decisions concerning individual persons. Accordingly, we identify no direct human-subject or privacy risks arising from the experimental design. The authors report no financial, professional, or personal conflicts of interest associated with this work. We have reviewed and complied with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To support reproducibility, the main text and appendix describe the methodology, theoretical derivations, experimental procedures, and implementation configurations. In particular, Appendix A supplies full proofs of the theoretical results; Appendix C.3 presents the pseudocode for our SCOPE; Sec. 5.1 outlines the experimental setup and hyperparameter settings; and Appendix C provides further details on the datasets, TTA scenarios, and implementation details.

REFERENCES

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36:80396–80413, 2023.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8344–8353, 2022.

Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pp. 970–978. PMLR, 2018.

Krzysztof M Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, and Vikas Sindhwani. From complexity to simplicity: Adaptive es-active subspaces for blackbox optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Zeshuai Deng, Guohao Chen, Shuaicheng Niu, Hui Luo, Shuhai Zhang, Yifan Yang, Renjie Chen, Wei Luo, and Mingkui Tan. Test-time model adaptation for quantized neural networks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7258–7267, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Zixuan Hu, Yichun Hu, and Ling-Yu Duan. Seva: Leveraging single-step ensemble of vicinal augmentations for test-time adaptation. *arXiv preprint arXiv:2505.04087*, 2025a.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. In *International Conference on Machine Learning*, 2025b.

Zixuan Hu, Dongxiao Li, Xinzhu Ma, Shixiang Tang, Xiaotong Li, Wenhan Yang, and Ling-Yu Duan. Adaptive dual uncertainty optimization: Boosting monocular 3d object detection under test-time shifts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7273–7283. IEEE, 2025c.

Zixuan Hu, Xuantuo Huang, Yancheng Li, Yichun Hu, Shengyong Xu, and Ling-Yu Duan. Turning adaptation into assets: Cross-domain bridging for online vision-language navigation. *International Conference on Machine Learning*, 2026.

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

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 2704–2713. IEEE, 2018.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

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. IEEE, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024.

Weihuang Liu, Xi Shen, Haolun Li, Xiuli Bi, Bo Liu, Chi-Man Pun, and Xiaodong Cun. Depth-aware test-time training for zero-shot video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19218–19227. IEEE, 2024.

Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *International Conference on Machine Learning*, pp. 4264–4273. PMLR, 2019.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

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.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

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

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38298–38315. PMLR, 21–27 Jul 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

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

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

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 P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Xingyu Wang and Tao Wang. Fozo: Forward-only zeroth-order prompt optimization for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7936–7945, June 2026.

Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shafin Rahman, David Ahmedt-Aristizabal, Xuesong Li, Lars Petersson, and Mehrtash Harandi. Backpropagation-free network for 3d test-time adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23231–23241. IEEE, 2024.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Yijun Yang, Hongtao Wu, Angelica I Aviles-Rivero, Yulun Zhang, Jing Qin, and Lei Zhu. Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25606–25616. IEEE, 2024.

Moslem Yazdanpanah, Ali Bahri, Mehrdad Noori, Sahar Dastani, Gustavo Adolfo Vargas Hakim, David Osowiechi, Ismail Ben Ayed, and Christian Desrosiers. Purge-gate: Backpropagation-free test-time adaptation for point clouds classification via token purging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27640–27649, 2025.

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

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pp. 191–207. Springer, 2022.

Junming Zhang, Shuyu Yin, Peilin Liu, Rendong Ying, and Fei Wen. Curvature-aware zeroth-order optimization for memory-efficient test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 836–846, June 2026.

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

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. IEEE, 2025.

Subspace-Coded Parallel Exploration for Forward-Only Test-Time Adaptation

————Appendix————

A Theoretical Analysis and Proofs

Below, we will provide detailed proofs of the theoretical results in the main manuscript.

Notation. We first recall the notation used in the main manuscript and this appendix. At step t, let $\mathcal{B}_t = \{x_i\}_{i=1}^B$ be the test batch, $P_t \in \mathbb{R}^d$ the shared prompt, and $\mathbf{Z}_t = [Z_{t,1}, \dots, Z_{t,K}]$ the matrix of independent standard Gaussian directions, with $1 \le K \le B$ and $K \le d$. The code matrix C_t is sampled uniformly from

$$\mathcal{C} = \left\{ C \in \{\pm 1/\sqrt{K}\}^{B\times K} : \operatorname{rank}(C) = K \right\},$$

independently of the batch and basis directions. We write $C_t^\dagger = (C_t^\top C_t)^{-1} C_t^\top$. Each code row $\mathbf{c}_{i,t}$ has unit norm and defines $v_{i,t} = \mathbf{Z}_t \mathbf{c}_{i,t}^\top$.

Let $g_t = \nabla_P \mathcal{L}(P_t; \mathcal{B}_t)$ and let $g_{i,t}$ denote the first-order gradient contribution of sample i, satisfying $B^{-1}\sum_i g_{i,t} = g_t$. Thus, $\mathbf{a}_{i,t} = \mathbf{Z}_t^\top g_{i,t}$ and $\bar{\mathbf{a}}_t = \mathbf{Z}_t^\top g_t$. For sample-separable objectives, $g_{i,t} = \nabla_P \mathcal{L}(x_i, P_t)$; the contributions for the complete batch-coupled objective are derived in Subsection A.3.

We separate the coded responses into their first-order part and finite-difference residual: $q_t^C = q_t^0 + r_t$, where $(q_t^0)_i = v_{i,t}^\top g_{i,t}$. The SCOPE estimator consequently admits the decomposition

$$\widehat{g}_t = \widetilde{g}_t + e_t, \qquad \widetilde{g}_t = \frac{1}{K}\mathbf{Z}_t C_t^\dagger q_t^0, \qquad e_t = \frac{1}{K}\mathbf{Z}_t C_t^\dagger r_t. \tag{14}$$

Here, $\widehat{g}_t$ is the gradient estimate used in Eq. 13, $\widetilde{g}_t$ is its first-order counterpart, and e_t is the reconstructed finite-difference error. We use $\mathbb{E}_t$ for expectation conditional on the history before step t. For fixed-step arguments, we omit t when unambiguous; $\mathbb{E}_C$ then averages only over the code matrix.

A.1 Convergence Analysis

Following the analysis structure of prior ZO methods (Wang & Wang, 2026; Zhang et al., 2026), we establish convergence through estimator moments and expected descent. We consider a stationary test-domain segment with batches sampled independently from a fixed distribution $\mathcal{D}_B$. The objective is

$$F(P) = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}_B}[\mathcal{L}(P;\mathcal{B})], \qquad \nabla F(P) = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}_B}[\nabla_P \mathcal{L}(P;\mathcal{B})], \tag{15}$$

where differentiation and expectation are assumed interchangeable. This definition retains the batch-coupled terms of the original adaptation objective.

Assumptions. Assume that F is L-smooth and bounded below by $F_\star$. In addition to the local regularity used for finite differences, suppose that, for all iterations and the deterministic perturbation scales considered, there exist finite constants G, R such that

$$\mathbb{E}_t\left[\frac{1}{B}\sum_{i=1}^B \|g_{i,t}\|_2^2\right] \le G^2, \qquad \mathbb{E}_t\|e_t\|_2^2 \le R^2\epsilon_t^2. \tag{16}$$

These conditions control the gradient contributions and the finite-difference error uniformly over the optimization trajectory. The basis directions and code matrices are independently resampled at every step.

Proposition 2 (Convergence of SCOPE). *Under the preceding conditions, let*

$$V = \frac{B(dK+2)G^2}{K^2}\mathbb{E}_C\|C^\dagger\|_2^2.$$

For the update $P_{t+1} = P_t - \eta\widehat{g}_t$ with $0 < \eta \le 1/(2L)$, SCOPE satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(P_t)\|_2^2 \le \frac{2(F(P_0)-F_\star)}{\eta T} + 2L\eta V + \frac{2R^2}{T}\sum_{t=0}^{T-1}\epsilon_t^2. \tag{17}$$

Proof. We first control the first-order estimator, then account for finite-difference errors in the expected loss decrease.

Moments of the First-Order Estimator. The row-exchangeability argument in Eq. 26 gives $\mathbb{E}_C[C_t^\dagger q_t^0 \mid \mathcal{B}_t, P_t, \mathbf{Z}_t] = \bar{\mathbf{a}}_t$. Since $\mathbb{E}_{\mathbf{Z}_t}[\mathbf{Z}_t\mathbf{Z}_t^\top] = KI_d$,

$$\mathbb{E}_t[\widetilde{g}_t] = \mathbb{E}_t\left[\frac{1}{K}\mathbf{Z}_t\mathbf{Z}_t^\top g_t\right] = \nabla F(P_t). \tag{18}$$

For any fixed C, batch, and sample contribution g_i, the Gaussian fourth-moment identity yields

$$\mathbb{E}_{\mathbf{Z}}\left[\|\mathbf{Z}\|_F^2(\mathbf{c}_i\mathbf{Z}^\top g_i)^2\right] = (dK+2)\|g_i\|_2^2, \tag{19}$$

because $\|\mathbf{c}_i\|_2 = 1$. Using $\|\mathbf{Z}\|_2 \le \|\mathbf{Z}\|_F$ and the independence of C_t from the batch, we obtain

$$\begin{aligned}\mathbb{E}_t\|\widetilde{g}_t\|_2^2 &\le \frac{1}{K^2}\mathbb{E}_t\left[\|C_t^\dagger\|_2^2\|\mathbf{Z}_t\|_F^2\sum_{i=1}^{B}(\mathbf{c}_{i,t}\mathbf{Z}_t^\top g_{i,t})^2\right]\\ &\le \frac{B(dK+2)G^2}{K^2}\mathbb{E}_C\|C^\dagger\|_2^2 = V.\end{aligned} \tag{20}$$

The constant V is finite because $\mathcal{C}$ is a finite set of full-column-rank matrices.

Expected Descent. By smoothness, Eq. 14, and $\|u+v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2$,

$$\begin{aligned}\mathbb{E}_t F(P_{t+1}) &\le F(P_t) - \eta\|\nabla F(P_t)\|_2^2 - \eta\langle\nabla F(P_t), \mathbb{E}_t e_t\rangle + L\eta^2(V + R^2\epsilon_t^2)\\ &\le F(P_t) - \frac{\eta}{2}\|\nabla F(P_t)\|_2^2 + L\eta^2 V + \eta R^2\epsilon_t^2.\end{aligned} \tag{21}$$

The second inequality follows from $\|\mathbb{E}_t e_t\|_2^2 \le R^2\epsilon_t^2$, Young's inequality, and $\eta \le 1/(2L)$.

Accumulation over Iterations. Taking total expectations and summing Eq. 21 gives

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(P_t)\|_2^2 \le F(P_0) - F_\star + L\eta^2 VT + \eta R^2\sum_{t=0}^{T-1}\epsilon_t^2. \tag{22}$$

Dividing by $\eta T/2$ proves Eq. 17. □

Choosing $\eta = \eta_0/\sqrt{T}$ with $0 < \eta_0 \le 1/(2L)$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(P_t)\|_2^2 = O(T^{-1/2}) + O\left(\frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t^2\right). \tag{23}$$

Thus, a fixed perturbation scale ϵ gives $O(T^{-1/2}) + O(\epsilon^2)$, whereas $\epsilon_t = O((t+1)^{-1/4})$ gives an $O(T^{-1/2})$ averaged stationarity bound. With both η and ϵ fixed, the bound instead retains $O(\eta) + O(\epsilon^2)$ residual terms. Each iteration requires only two batch forward passes.

A.2 Analysis of the Decoding Bias

We now fix $\mathcal{B}_t$, P_t, and $\mathbf{Z}$, and analyze the expectation over C. Let $\mathbf{a}_i = (a_i(Z_1), \dots, a_i(Z_K))^\top$ and $\bar{\mathbf{a}} = B^{-1}\sum_i \mathbf{a}_i$. The complete objective admits this first-order representation under the regularity conditions established in Subsection A.3.

Proposition 3 (Asymptotic Unbiasedness of Decoding). *Under the specified code sampling, the first-order decoded responses satisfy $\mathbb{E}_C[C^\dagger q^0] = \bar{\mathbf{a}}$. If the residual additionally satisfies $\sup_{C\in\mathcal{C}}\|r(C,\epsilon_t)\|_2 \le R_0\epsilon_t$ for sufficiently small ϵ_t, where R_0 is independent of C and ϵ_t, then*

$$\left\|\mathbb{E}_C[\widehat{\mathbf{a}}_C] - \bar{\mathbf{a}}\right\|_2 = O(\epsilon_t), \qquad \epsilon_t \to 0. \tag{24}$$

Proof. We separate sample heterogeneity from finite-difference error and show that only the latter contributes to the expected bias.

Cancellation of Sample Heterogeneity. Define $\delta_i = \mathbf{a}_i - \bar{\mathbf{a}}$, so that $\sum_i \delta_i = 0$. Eqs. 8 and 9 give

$$\widehat{\mathbf{a}}_C - \bar{\mathbf{a}} = \sum_{i=1}^{B} M_i(C)\delta_i + C^\dagger r, \qquad M_i(C) = (C^\top C)^{-1}\mathbf{c}_i^\top \mathbf{c}_i. \tag{25}$$

The uniform distribution on $\mathcal{C}$ is invariant under row permutations, which preserve both the sign constraints and full column rank. Hence all $M_i(C)$ have the same expectation. Since $\sum_i M_i(C) = I_K$ for every admissible C,

$$\mathbb{E}_C[M_i(C)] = \frac{1}{B} I_K, \qquad \mathbb{E}_C[C^\dagger \xi] = \sum_{i=1}^{B} \mathbb{E}_C[M_i(C)]\delta_i = \frac{1}{B}\sum_{i=1}^{B}\delta_i = 0. \tag{26}$$

This establishes $\mathbb{E}_C[C^\dagger q^0] = \bar{\mathbf{a}}$ without assuming independence between C and ξ.

Control of the Finite-Difference Bias. The remaining bias is bounded by

$$\left\|\mathbb{E}_C[\widehat{\mathbf{a}}_C] - \bar{\mathbf{a}}\right\|_2 = \left\|\mathbb{E}_C[C^\dagger r]\right\|_2 \le R_0 \epsilon_t \, \mathbb{E}_C\left[\frac{1}{\sigma_{\min}(C)}\right] = O(\epsilon_t). \tag{27}$$

The expectation is finite because every matrix in the finite set $\mathcal{C}$ has positive smallest singular value. □

For sample-separable objectives with locally Lipschitz sample-loss gradients, Eq. 6 provides the required residual bound: $\|v_i\|_2 \le \|\mathbf{Z}\|_2$ implies that $\|r\|_2 = O(\epsilon_t)$ uniformly over $\mathcal{C}$. The analogous property for the complete objective is established below.

A.3 Proof of the Statistic-Decoupling Transformation

Let $h_i^\pm$ be the features from the two coded branches. We use population statistics,

$$\mu^\pm = \frac{1}{B}\sum_{i=1}^{B} h_i^\pm, \qquad (\sigma^\pm)^{\odot 2} = \frac{1}{B}\sum_{i=1}^{B}(h_i^\pm - \mu^\pm)^{\odot 2}, \tag{28}$$

and assume that every entry of $\sigma^+ + \sigma^-$ is positive. The following identities also hold when the same fixed stabilizer is added inside both square roots. Vector squares, products, and quotients below are element-wise, except for inner products.

For brevity, write $d_i = h_i^+ - h_i^-$, $s_i = h_i^+ + h_i^-$, and $\Delta\mathcal{L}_{\text{stat}} = \mathcal{L}_{\text{stat}}^+ - \mathcal{L}_{\text{stat}}^-$, where $\mathcal{L}_{\text{stat}}^\pm = \|\mu^\pm - \mu^S\|_2^2 + \|\sigma^\pm - \sigma^S\|_2^2$.

Proposition 4 (**Exact Statistic Decoupling**). *The contributions m_i in Eq. 11 and the responses q_i^C in Eq. 12 satisfy*

$$\frac{1}{B}\sum_{i=1}^{B} m_i = \Delta\mathcal{L}_{\text{stat}}, \qquad \frac{1}{B}\sum_{i=1}^{B} q_i^C = \frac{\mathcal{L}^+ - \mathcal{L}^-}{2\epsilon_t}. \tag{29}$$

These identities hold at finite perturbation scales without a first-order approximation.

Proof. We decompose the mean and standard-deviation alignment differences separately, then combine them with the entropy differences.

Mean-Alignment Difference. The quadratic difference factorizes as

$$\|\mu^+ - \mu^S\|_2^2 - \|\mu^- - \mu^S\|_2^2 = \langle \mu^+ + \mu^- - 2\mu^S, \mu^+ - \mu^- \rangle = \frac{1}{B}\sum_{i=1}^{B}\langle \mu^+ + \mu^- - 2\mu^S, d_i\rangle. \tag{30}$$

Thus, the shared mean-alignment coefficient weights each sample's feature change.

Standard-Deviation Alignment Difference. Expanding the population variances gives

$$(\sigma^+)^{\odot 2} - (\sigma^-)^{\odot 2} = \frac{1}{B}\sum_{i=1}^{B}\left[(h_i^+)^{\odot 2} - (h_i^-)^{\odot 2}\right] - \left[(\mu^+)^{\odot 2} - (\mu^-)^{\odot 2}\right]$$
$$= \frac{1}{B}\sum_{i=1}^{B}(s_i - \mu^+ - \mu^-) \odot d_i. \tag{31}$$

Any common variance stabilizer cancels in this difference. Since $(\sigma^+)^{\odot 2} - (\sigma^-)^{\odot 2} = (\sigma^+ + \sigma^-) \odot (\sigma^+ - \sigma^-)$, we obtain

$$\sigma^+ - \sigma^- = \frac{1}{B}\sum_{i=1}^{B}\frac{s_i - \mu^+ - \mu^-}{\sigma^+ + \sigma^-} \odot d_i. \tag{32}$$

Applying the quadratic factorization again yields

$$\|\sigma^+ - \sigma^S\|_2^2 - \|\sigma^- - \sigma^S\|_2^2 = \frac{1}{B}\sum_{i=1}^{B}\left\langle \frac{\sigma^+ + \sigma^- - 2\sigma^S}{\sigma^+ + \sigma^-} \odot (s_i - \mu^+ - \mu^-), d_i \right\rangle. \tag{33}$$

Adding Eqs. 30 and 33 gives precisely $B^{-1}\sum_i m_i = \Delta\mathcal{L}_{\text{stat}}$.

Conservation of the Complete Loss Difference. Combining the statistical contributions with the sample-wise entropy differences gives

$$\frac{1}{B}\sum_{i=1}^{B} q_i^C = \frac{1}{2\epsilon_t}\left[\frac{1}{B}\sum_{i=1}^{B}\big(H(p_i^+) - H(p_i^-)\big) + \frac{\lambda}{B^2}\sum_{i=1}^{B} m_i\right]$$
$$= \frac{1}{2\epsilon_t}\left[\frac{1}{B}\sum_{i=1}^{B}\big(H(p_i^+) - H(p_i^-)\big) + \frac{\lambda}{B}\Delta\mathcal{L}_{\text{stat}}\right] = \frac{\mathcal{L}^+ - \mathcal{L}^-}{2\epsilon_t}. \tag{34}$$

This proves both identities in Eq. 29. □

First-Order Compatibility with Coded Decoding. To connect exact conservation with the preceding estimator analysis, assume additionally that the model processes samples independently, the entropy gradients and feature Jacobians are locally Lipschitz, and the standard deviations are positive near P_t. All quantities below are evaluated at the shared prompt center. Let J_i be the feature Jacobian and define

$$b_i = \mu - \mu^S + \frac{\sigma - \sigma^S}{\sigma} \odot (h_i - \mu).$$

Differentiating the complete objective gives the sample contributions

$$g_i = \nabla_P H(p_i) + \frac{2\lambda}{B} J_i^\top b_i, \qquad \frac{1}{B}\sum_{i=1}^{B} g_i = \nabla_P \mathcal{L}(P_t; \mathcal{B}_t). \tag{35}$$

For fixed batch and basis directions, $h_i^\pm = h_i \pm \epsilon_t J_i v_i + O(\epsilon_t^2)$. Substituting these expansions into Eq. 11, and combining with the entropy expansion, yields

$$\frac{m_i}{2\epsilon_t} = 2\langle b_i, J_i v_i\rangle + O(\epsilon_t), \qquad q_i^C = v_i^\top g_i + O(\epsilon_t) = \sum_{k=1}^{K} c_{ik} a_i(Z_k) + O(\epsilon_t), \tag{36}$$

where $a_i(Z) = Z^\top g_i$. Their batch averages satisfy $B^{-1}\sum_i a_i(Z) = Z^\top \nabla_P \mathcal{L}(P_t; \mathcal{B}_t)$. Because the code set is finite, the remainder is uniform over $C \in \mathcal{C}$ for fixed batch and basis directions. The complete responses therefore satisfy the coded observation model required by Proposition 3.

B Further Experiments

In this section, we extend the main experiments to wild TTA scenarios on ImageNet-C and natural distribution shifts on ImageNet-R/Sketch/V2. We compare with the same BP-free baselines using ViT-B/16, with SCOPE retaining two batch forward passes throughout.

Table 6: Top-1 accuracy (%) on ImageNet-C (severity 5) under **Imbalanced Label Shift**.

Method	FP	Noise			Blur				Weather				Digital				Avg↑
		Gaus.	Shot	Impu.	Defo.	Gla.	Moti.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pixel	JPEG	
NoAdapt	1	56.77	56.80	57.51	46.80	35.58	53.02	44.70	62.30	62.59	65.86	77.69	32.51	46.03	66.98	67.59	55.51
T3A (NeurIPS'21)	1	55.69	56.19	56.49	46.76	37.09	53.26	46.19	63.05	58.41	66.59	77.66	37.39	49.75	67.00	68.33	55.99
LAME (CVPR'22)	1	56.34	56.31	57.03	46.07	34.60	52.36	43.89	59.42	61.41	63.23	77.20	26.20	44.62	66.35	67.02	54.14
FOA (ICML'24)	2	58.91	59.77	60.25	52.46	40.68	55.93	48.65	65.40	63.15	69.43	78.04	59.57	51.98	68.85	69.49	60.17
ZOA (ACMMM'25)	2	59.61	60.56	61.35	54.58	46.54	59.06	54.18	66.38	62.64	68.03	78.81	59.77	56.40	70.20	69.84	61.86
NEO (ICLR'26)	1	57.52	57.58	58.28	49.69	38.12	54.74	47.40	64.72	65.11	**71.31**	78.14	58.42	49.97	67.67	68.81	59.17
CAZO (CVPR'26)	3	55.92	56.36	57.64	51.70	44.30	57.36	52.70	65.61	63.11	66.65	78.33	55.61	53.30	69.23	68.47	59.75
FOZO (CVPR'26)	2	59.38	60.16	60.92	54.59	44.84	58.24	51.09	66.52	63.78	69.22	78.65	61.70	54.17	70.13	70.14	61.57
SCOPE (Ours)	2	**60.72**	**62.20**	**62.33**	**57.21**	**50.33**	**60.29**	**54.99**	**68.64**	**66.75**	70.87	**80.61**	**66.66**	**59.57**	**72.91**	**72.26**	**64.42**

B.1 Evaluation on Wild TTA Scenarios

Imbalanced Label Shift. To evaluate adaptation under uneven class distributions, we compare BP-free methods on ImageNet-C at severity 5. As shown in Tab. 6, SCOPE achieves 64.42% average accuracy, outperforming the strongest baseline, ZOA, by +2.56. Beyond the average gain, SCOPE ranks first on 14 of 15 corruptions and achieves the highest group-wise average across noise, blur, weather, and digital corruptions. These results demonstrate that its advantage under label imbalance extends across corruption families rather than relying on a few favorable domains.

Mixed Shifts. To assess adaptation to heterogeneous test streams, we evaluate mixed corruption domains at severity levels 3, 4, and 5. The left block of Tab. 7 shows that SCOPE achieves the highest accuracy at all three severities, surpassing FOZO by +0.47, +0.44, and +0.87, respectively. Its advantage persists as corruption severity increases, with the largest margin occurring at severity 5. These results extend SCOPE's adaptation benefits from individual corruption domains to mixed-domain streams.

Across the two wild scenarios at severity 5, SCOPE improves over the respective strongest baselines by an average of +1.72. These findings complement the mild and continual results by demonstrating its effectiveness under both imbalanced class composition and mixed corruption domains.

B.2 Evaluation on Natural Distribution Shifts

To examine whether SCOPE's gains extend beyond synthetic corruptions, we further evaluate its performance on ImageNet-R, ImageNet-Sketch, and ImageNet-V2. As shown in the right block of Tab. 7, SCOPE achieves the highest accuracy among the compared BP-free methods on all three benchmarks, reaching 61.19%, 49.27%, and 75.73%, respectively.

The results also reveal that adaptation gains are not uniform across datasets. T3A improves upon NoAdapt on ImageNet-Sketch but degrades performance on ImageNet-R, while FOZO improves on both yet falls slightly below NoAdapt on ImageNet-V2. In contrast, SCOPE improves over the unadapted model across all three datasets. Notably, it retains a modest positive gain on ImageNet-V2, where most compared methods fail to improve upon NoAdapt. Together, these results support SCOPE's applicability beyond ImageNet-C to additional distribution-shift benchmarks.

B.3 Additional Hyperparameter Analysis

We extend the main-text sensitivity analysis to batch size B and feature-statistic alignment weight λ, together with the fixed perturbation scale ϵ in W8A8 adaptation.

As shown in Fig. 4(a), accuracy improves as B increases from 16 to 64, but further expansion provides no additional benefit. Since SCOPE distributes coded measurements across samples, increasing the batch size supplies more observations for joint decoding. The initial improvement is consistent with the benefit of richer within-batch feedback, while the subsequent decline indicates that larger batches do not necessarily improve adaptation. We retain $B = 64$ by default, with all tested batch sizes outperforming the prior-SOTA reference. Fig. 4(b) shows that removing feature-

Table 7: Top-1 accuracy (%) under mixed shifts on ImageNet-C and natural distribution shifts on ImageNet-R/Sketch/V2 with ViT-B/16. Level denotes corruption severity.

Method	FP	Mixed Shifts			Natural Distribution Shifts		
		Level 3	Level 4	Level 5	R	Sketch	V2
NoAdapt	1	71.79	65.89	55.52	59.47	44.91	75.44
T3A (NeurIPS'21)	1	71.82	65.50	53.87	57.80	48.48	75.52
LAME (CVPR'22)	1	71.49	65.58	55.09	58.89	44.39	75.24
FOA (ICML'24)	2	73.34	68.05	59.56	60.42	46.73	75.31
ZOA (ACMMM'25)	2	73.57	68.11	59.82	60.67	47.03	75.45
NEO (ICLR'26)	1	72.43	66.70	56.72	60.45	46.80	75.39
CAZO (CVPR'26)	3	71.06	68.24	59.31	60.14	45.01	74.88
FOZO (CVPR'26)	2	73.60	68.50	60.26	61.00	48.16	75.35
SCOPE (Ours)	2	**74.07**	**68.94**	**61.13**	**61.19**	**49.27**	**75.73**

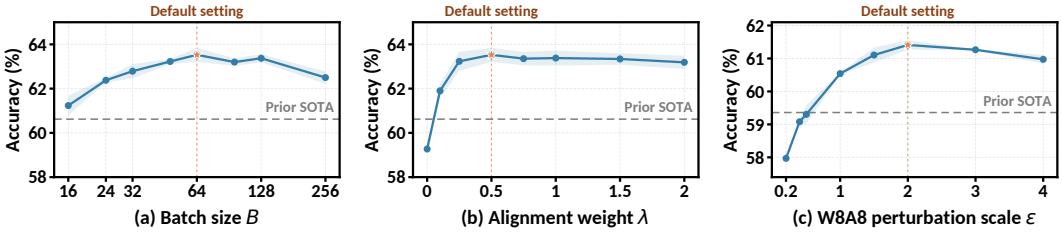

Figure 4: Additional hyperparameter analysis of SCOPE. Accuracy with varying (a) batch size B, (b) feature-statistic alignment weight λ, and (c) fixed perturbation scale ϵ for W8A8-quantized models. Stars mark the default settings; horizontal dashed lines indicate the corresponding prior-SOTA accuracy.

statistic alignment ($\lambda = 0$) substantially degrades performance. Introducing alignment sharply improves accuracy, which then stabilizes across the larger tested weights. This contrast highlights the importance of incorporating batch-coupled statistical feedback, while the broad performance plateau indicates that its effectiveness does not depend on a narrowly tuned alignment weight. Finally, Fig. 4(c) examines perturbation scales for W8A8-quantized models, with each scale held fixed throughout adaptation. Accuracy improves as ϵ increases toward 2.0, followed by a modest decline at larger values. The weaker results at small scales suggest that a sufficiently large perturbation magnitude is important for effective adaptation under reduced precision. We therefore use $\epsilon = 2.0$ by default. SCOPE also exceeds the quantized prior-SOTA reference at all tested scales from 1.0 to 4.0, demonstrating a broad effective range around this setting.

C Implementation Details

C.1 More Details on Datasets

ImageNet-C Dataset. ImageNet-C (Hendrycks & Dietterich, 2019) is constructed by applying common visual corruptions to the ImageNet validation set (Deng et al., 2009). Its standard evaluation suite contains 15 corruption types across four categories: noise, blur, weather, and digital transformations, each with five severity levels. These include Gaussian, shot, and impulse noise; defocus, glass, motion, and zoom blur; snow, frost, fog, and brightness; and contrast, elastic transformation, pixelation, and JPEG compression. The benchmark enables controlled evaluation across different corruption types and intensities. We evaluate 50,000 images per corruption at severity 5 in the main experiments and additionally consider severities 3 and 4 in the mixed-shift evaluation.

ImageNet-R Dataset. ImageNet-R (Hendrycks et al., 2021) contains 30,000 renditions of 200 ImageNet classes, including paintings, cartoons, sculptures, embroidery, origami, and toys. Rather than applying synthetic corruptions to existing photographs, it introduces substantial variations in visual style and object appearance through independently collected renditions. We use this benchmark to evaluate adaptation beyond common corruptions, under shifts from photographic imagery to diverse visual representations of the same semantic categories.

ImageNet-Sketch Dataset. ImageNet-Sketch (Wang et al., 2019) contains approximately 50,000 sketch images covering all 1,000 ImageNet classes. The images were collected through class-specific image searches and manually filtered, with collection restricted to black-and-white imagery. Compared with source-domain photographs, sketches provide fewer color and texture cues while emphasizing object contours and shape. This benchmark evaluates adaptation under a pronounced photographic-to-sketch domain gap.

ImageNet-V2 Dataset. ImageNet-V2 (Recht et al., 2019) provides newly collected test images for the original 1,000 ImageNet categories, following a collection process designed to resemble that of ImageNet. It comprises three test-set variants, each containing 10,000 images, which differ in their image-selection criteria. Unlike the explicit corruption and style changes in the preceding benchmarks, ImageNet-V2 examines generalization to independently collected photographs. We use it to assess adaptation under distribution differences arising from data recollection rather than synthetic image transformations.

C.2 More Details on TTA Scenarios

Across the following evaluations, target labels are unavailable to the adaptation algorithm. SCOPE updates only the shared prompt, while pretrained model parameters and precomputed source feature statistics remain fixed.

Mild TTA Setting. The mild setting evaluates online adaptation within an individual target domain (Wang et al., 2021). On ImageNet-C, each corruption type at a fixed severity is treated as a separate domain. Adaptation state is retained across batches within that domain and reset before evaluating the next corruption. For SCOPE, this means carrying the updated prompt forward within each domain and restoring its initialization between domains. This protocol isolates adaptation to individual corruption shifts.

Continual TTA Setting. The continual setting processes corruption domains sequentially without resetting the adaptation state at domain transitions (Wang et al., 2022). SCOPE therefore carries its learned prompt across successive corruptions rather than restarting from its initialization. This protocol evaluates adaptation when previously accumulated updates interact with subsequent distribution changes.

Wild TTA Setting. Following the wild TTA scenarios studied by SAR and ReCAP (Niu et al., 2023; Hu et al., 2025b), we evaluate imbalanced label distributions and mixed testing domains. These scenarios extend individual-domain evaluation by changing the class composition or domain composition of the incoming test stream.

Imbalanced Label Shift. The incoming batches exhibit uneven class proportions that vary over time, so adaptation feedback may be dominated by frequently occurring classes. We evaluate this setting on ImageNet-C at severity 5. It tests whether SCOPE remains effective when online updates are driven by imbalanced class composition rather than balanced target batches.

Mixed Shifts. Samples from different corruption domains are mixed within the test stream rather than processed as separate, homogeneous domains. We evaluate corruption mixtures separately at severity levels 3, 4, and 5. This setting tests adaptation to heterogeneous domain composition, complementing continual TTA, where corruption domains occur sequentially.

Quantized Models. Resource-constrained deployment motivates low-precision models to reduce storage and computational demands. Following FOZO (Deng et al., 2025; Zhang et al., 2026; Wang & Wang, 2026), we evaluate adaptation on W8A8-quantized ViT-B/16 models obtained using

PTQ4ViT (Yuan et al., 2022), where weights and activations are quantized to 8 bits. Limited support for standard backpropagation in such deployments makes forward-only adaptation particularly relevant. Our quantized experiments follow the mild protocol on ImageNet-C at severity 5, resetting the prompt between corruption domains. SCOPE updates the prompt while keeping the quantized backbone frozen, using two batch forward passes and a fixed perturbation scale of $\epsilon = 2.0$. This evaluates whether its adaptation benefits persist under reduced numerical precision without requiring backward computation.

C.3 Pseudocode of SCOPE

Algorithm 1 outlines the online execution of SCOPE. For each incoming batch, SCOPE encodes shared basis directions into complementary sample-wise perturbations and evaluates the positive and negative branches using two batch forward passes. Each branch returns logits $z_i^{\pm}$ and features $h_i^{\pm}$. The corresponding probabilities $p_i^{\pm} = \mathrm{softmax}(z_i^{\pm})$ and features are used to construct code-aligned responses for the complete adaptation objective. These responses are jointly decoded into batch-average directional coefficients, which form the gradient estimate for updating the shared prompt. For prediction, SCOPE averages the two branches' logits and applies softmax to obtain the fused probabilities, reusing the same two forward evaluations without an additional pass. The updated prompt is carried forward to subsequent batches, with resets applied only when required by the evaluation protocol. All pretrained parameters and source statistics remain fixed throughout adaptation.

Algorithm 1: **S**ubspace-**C**oded **O**ptimization via **P**arallel **E**xploration (SCOPE).

Input: Unlabeled test batches $\{\mathcal{B}_t\}_{t=0}^{T-1}$ of size B, pretrained model f, initial prompt $P_0 \in \mathbb{R}^d$, source feature statistics (μ^S, σ^S), subspace dimension $K \le B$ with $K \le d$, fixed perturbation scale $\epsilon > 0$, learning rate η, alignment weight λ.

Output: Fused prediction probabilities $\{p_i\}_{i=1}^B$ for each batch and final shared prompt P_T.

Freeze all pretrained parameters of f and disable gradient tracking;

for $t = 0, \dots, T-1$ **do**

  if *the evaluation protocol specifies a reset before* $\mathcal{B}_t$ **then**

    $P_t \leftarrow P_0$;

  end

  `// 1. Subspace-coded parallel exploration`

  Sample K independent directions $Z_k \sim \mathcal{N}(0, I_d)$;

  Independently sample $C = (c_{ik}) \in \{\pm 1/\sqrt{K}\}^{B \times K}$ uniformly, resampling until $\mathrm{rank}(C) = K$;

  Form sample-wise perturbations: $v_i \leftarrow \sum_{k=1}^K c_{ik} Z_k$, $P_i^{\pm} \leftarrow P_t \pm \epsilon v_i$; `// Eq. 4`

  $\{(z_i^+, h_i^+)\}_{i=1}^B \leftarrow f(\mathcal{B}_t, \{P_i^+\}_{i=1}^B)$; `// One positive batch forward`

  $\{(z_i^-, h_i^-)\}_{i=1}^B \leftarrow f(\mathcal{B}_t, \{P_i^-\}_{i=1}^B)$; `// One negative batch forward`

  Compute branch probabilities: $p_i^{\pm} \leftarrow \mathrm{softmax}(z_i^{\pm}), i = 1, \dots, B$;

  `// 2. Code-aligned response decomposition`

  Compute the means and population standard deviations $(\mu^{\pm}, \sigma^{\pm})$ from the corresponding batch features $\{h_i^{\pm}\}_{i=1}^B$;

  for $i = 1, \dots, B$ **do**

    Construct the statistical loss-difference contribution m_i from $h_i^{\pm}$, $(\mu^{\pm}, \sigma^{\pm})$, and (μ^S, σ^S) ; `// Eq. 11`

    Combine with the entropy difference: $q_i^C \leftarrow \dfrac{H(p_i^+) - H(p_i^-) + \frac{\lambda}{B} m_i}{2\epsilon}$; `// Eq. 12`

  end

  `// 3. Shared directional decoding and prompt update`

  Set $q^C \leftarrow (q_1^C, \dots, q_B^C)^\top$;

  Solve $(C^\top C)\widehat{\mathbf{a}}_C = C^\top q^C$ for the batch-average directional coefficients ; `// Eq. 9`

  Reconstruct the gradient estimate: $\widehat{g}_t \leftarrow \dfrac{1}{K} \sum_{k=1}^K \widehat{a}_{C,k} Z_k$;

  Update only the shared prompt: $P_{t+1} \leftarrow P_t - \eta \widehat{g}_t$; `// Eq. 13`

  `// 4. Prediction fusion`

  Fuse the two branches by averaging logits: $p_i \leftarrow \mathrm{softmax}\left(\dfrac{z_i^+ + z_i^-}{2}\right), i = 1, \dots, B$;

  Emit the fused predictions $\{p_i\}_{i=1}^B$ for $\mathcal{B}_t$;

end