Curvature-Aware Zeroth-Order Optimization for Memory-Efficient Test-Time Adaptation

Supplementary Material

8. Convergence Analysis of CAZO

Given a loss function $\mathcal{L}(x;\theta)$, where $x \in \mathbb{R}^n$ is the data and $\theta \in \mathbb{R}^d$ is the parameter, then the definition of the gradient of CAZO is

$$\widetilde{g}(x_t;\theta_t) = \frac{1}{k}\sum_{i=1}^{k} \frac{\mathcal{L}(x_t;\theta_t+\epsilon u_i) - \mathcal{L}(x_t;\theta_t-\epsilon u_i)}{2\epsilon} u_i, \quad \text{with } u_i \sim \mathcal{N}(0, \widetilde{H}_t^{-1}) \tag{7}$$

We further denote $\nabla f(\theta) = \mathbb{E}_x[\nabla f(x;\theta)]$.

$$D_t = (1-\nu)D_{t-1} + \nu\widetilde{g}^2(\theta_{t-1}), \quad \widetilde{H}_t = \operatorname{diag}\left(\frac{D_t}{1-(1-\nu)^t}\right), \tag{8}$$

Assumption (Restatement of L-smoothness)**.** *Assume the loss function $\mathcal{L}(x;\theta)$ is L-smooth respect to parameter θ.*

Assumption (Restatement of Data variance)**.** *Assume the data variance satisfies $\mathbb{E}_x[\|\nabla f(x;\theta) - \nabla f(\theta)\|] \le \sigma^2$.*

Assumption (Restatement of Value range of $\widetilde{H}_t^{-1}$)**.** *Assume for $t \ge 0$, the element in $\widetilde{H}_t^{-1}$ are in range $[\beta_l, \beta_u]$ where $0 < \beta_l \le \beta_u \le \infty$.*

Lemma 1 (Estimation error of the zeroth order gradient estimator)**.** *Given a loss function $\mathcal{L}(\theta_t)$ with parameter θ_t satisfies Assumption 1, and a data x_t sampled from $\mathcal{D}$, a gradient estimator and variance matrix estimator in Eq. (7) and Eq. (6) satisfies Assumption 2 and 3, the following relation holds*

$$\mathbb{E}_{u_i,x_t}\left[\widetilde{g}(x_t;\theta_t)\right] = \widetilde{H}_t^{-1}\nabla\mathcal{L}(\theta_t) + \mathcal{O}(\epsilon)$$
$$\mathbb{E}_{x_t,u_i}\left[\|\widetilde{g}(x_t;\theta_t)\|^2\right] \le 2d(d+2)\beta_u\left(\|\nabla\mathcal{L}(\theta_t)\|^2 + \sigma^2\right) + \mathcal{O}(\epsilon^2).$$

Proof. By the mean value theorem, we have

$$\mathcal{L}(x_t;\theta_t+\epsilon u_i) = \mathcal{L}(x_t;\theta_t) + \epsilon\nabla\mathcal{L}(x_t;\theta_t)^\top u_i + \frac{\epsilon^2}{2}u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^1)u_i,$$
$$\mathcal{L}(x_t;\theta_t-\epsilon u_i) = \mathcal{L}(x_t;\theta_t) - \epsilon\nabla\mathcal{L}(x_t;\theta_t)^\top u_i + \frac{\epsilon^2}{2}u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^2)u_i.$$

So we then have

$$\begin{aligned}&\frac{\mathcal{L}(x_t;\theta_t+\epsilon u_i) - \mathcal{L}(x_t;\theta_t-\epsilon u_i)}{2\epsilon}u_i \\ &= \left(\nabla\mathcal{L}(x_t;\theta_t)^\top u_i\right)u_i + \frac{\epsilon}{4}\left(u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^1)u_i - u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^2)u_i\right)u_i,\end{aligned} \tag{9}$$

where $\xi_t^1 = \lambda_1\theta_t + (1-\lambda_1)(\theta_t+\epsilon u_i)$, $\xi_t^2 = \lambda_2\theta_t + (1-\lambda_2)(\theta_t - \epsilon u_i)$ and $\lambda_1, \lambda_2 \in (0,1)$. We further denote

$$A_i = \left(\nabla\mathcal{L}(x_t;\theta_t)^\top u_i\right)u_i, \quad B_i = \left(u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^1)u_i - u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^2)u_i\right)u_i. \tag{10}$$

Then we have

$$\mathbb{E}_{u_i}[A_i] = \mathbb{E}_{u_i}\left[u_iu_i^\top\right]\nabla\mathcal{L}(x_t;\theta_t) = \widetilde{H}_t^{-1}\nabla\mathcal{L}(x_t;\theta_t), \tag{11}$$

$$\begin{aligned}\mathbb{E}_{u_i}[\|A_i\|^2] &= \mathbb{E}_{u_i}\left[\left(\nabla\mathcal{L}(x_t;\theta_t)^\top u_i\right)^2\|u_i\|^2\right] \\ &\le \|\nabla\mathcal{L}(x_t;\theta_t)\|^2\mathbb{E}_{u_i}[\|u_i\|^4]\end{aligned} \tag{12}$$

and because of Assumption 1

$$\begin{aligned}&u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^1)u_i - u_i^\top\nabla^2\mathcal{L}(x_t;\xi_t^2)u_i \le 2L\|u_i\|^2 \\ &\Rightarrow \mathbb{E}_{u_i}[B_i] \le 2L\mathbb{E}_{u_i}[\|u_i\|^3]\end{aligned} \tag{13}$$

$$\mathbb{E}_{u_i}[\|B_i\|^2] \le 4L^2\mathbb{E}_{u_i}[\|u_i\|^6] \tag{14}$$

Now we start to estimate the conclusion

$$\begin{aligned}\mathbb{E}_{u_i}\left[\widetilde{g}(x_t;\theta_t)\right] &= \frac{1}{k}\sum_{i=1}^{k}\left(\mathbb{E}_{u_i}[A_i] + \frac{\epsilon}{4}\mathbb{E}_{u_i}[B_i]\right) \\ &= \mathbb{E}_{u_i}[A_i] + \frac{\epsilon}{4}\mathbb{E}_{u_i}[B_i] \\ &= \widetilde{H}_t^{-1}\nabla\mathcal{L}(x_t;\theta_t) + \frac{\epsilon}{4}\mathbb{E}_{u_i}[B_i],\end{aligned} \tag{15}$$

Then take expectation of x_t on both size of the inequality, we have

$$\mathbb{E}_{x_t,u_i}\left[\widetilde{g}(x_t;\theta_t)\right] = \widetilde{H}_t^{-1}\nabla\mathcal{L}(\theta_t) + \frac{\epsilon}{4}\mathbb{E}_{u_i}[B_i]. \tag{16}$$

By Jensen's inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|\widetilde{g}(x_t;\theta_t)\|^2 &= \left\|\frac{1}{k}\sum_{i=1}^{k}\left(A_i + \frac{\epsilon}{4}B_i\right)\right\|^2 \\ &\le \frac{1}{k}\sum_{i=1}^{k}\left\|A_i + \frac{\epsilon}{4}B_i\right\|^2 \\ &\le \frac{2}{k}\sum_{i=1}^{k}\left(\|A_i\|^2 + \left\|\frac{\epsilon}{4}B_i\right\|^2\right),\end{aligned} \tag{17}$$

and because u_i is I.I.D samples, we have

$$\begin{aligned}\mathbb{E}_{u_i}\left[\|\widetilde{g}(x_t;\theta_t)\|^2\right] &\le \frac{2}{k}\sum_{i=1}^{k}\left(\mathbb{E}_{u_i}[\|A_i\|^2] + \mathbb{E}_{u_i}[\|\tfrac{\epsilon}{4}B_i\|^2]\right) \\ &= 2\mathbb{E}_{u_i}[\|A_i\|^2] + 2\mathbb{E}_{u_i}[\|\tfrac{\epsilon}{4}B_i\|^2] \\ &\le 2\|\nabla\mathcal{L}(x_t;\theta_t)\|^2\mathbb{E}_{u_i}[\|u_i\|^4] \\ &\quad + \frac{\epsilon^2L^2}{2}\mathbb{E}_{u_i}[\|u_i\|^6].\end{aligned} \tag{18}$$

Take the expectation of x_t on both size of the inequality, and using the assumption 2, we have

$$\begin{aligned}\mathbb{E}_{x_t,u_i}\left[\|\widetilde{g}(x_t;\theta_t)\|^2\right] &\leq 2\mathbb{E}_{x_t}[\|\nabla\mathcal{L}(x_t;\theta_t)\|^2]\mathbb{E}_{u_i}[\|u_i\|^4] \\ &\quad + \frac{\epsilon^2 L^2}{2}\mathbb{E}_{u_i}[\|u_i\|^6] \\ &\leq 2\|\nabla\mathcal{L}(\theta_t)\|^2\mathbb{E}_{u_i}[\|u_i\|^4] \\ &\quad + 2\sigma^2\mathbb{E}_{u_i}[\|u_i\|^4] \\ &\quad + \frac{\epsilon^2 L^2}{2}\mathbb{E}_{u_i}[\|u_i\|^6]\end{aligned} \tag{19}$$

Now we start to estimate the upper bound of the moments of u_i. Because $\widetilde{H}_t^{-1}$ is a diagonal matrix and Assumption 3, we have

$$\begin{aligned}\mathbb{E}_{u_i}[\|u_i\|^4] &= (\mathrm{tr}(\widetilde{H}_t^{-1}))^2 + 2\,\mathrm{tr}(\widetilde{H}_t^{-1}) \\ &\leq d^2\beta_u + 2d\beta_u \\ &= d(d+2)\beta_u\end{aligned} \tag{20}$$

So we proved the result. □

Theorem 1 (Convergence rate of CAZO)**.** *Given a loss function $\mathcal{L}$ satisfies Assumption 1, a gradient estimator and variance matrix estimator in Eq. (7) and Eq. (6) satisfies Assumption 2 and 3, with training time T, learning rate $\eta = \frac{\beta_l}{2Ld(d+2)\beta_u\sqrt{T}}$, the convergence rate of the CAZO is*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta_t)\|^2\right] &\leq \frac{4Ld(d+2)\beta_u\left(\mathcal{L}(\theta_t)-\mathcal{L}(\theta^*)\right)}{\beta_l^2\left(\sqrt{T}-1\right)} \\ &\quad + \frac{\sigma^2}{\sqrt{T}-1} + \mathcal{O}(\epsilon^2).\end{aligned} \tag{21}$$

Proof. Because of Assumption 1, we have $f(\theta) = \mathbb{E}_x[f(x;\theta)]$ is also a L-smooth function. So $f(\theta)$ satisfies the following inequality

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) \leq& \mathcal{L}(\theta_t) - \eta\nabla\mathcal{L}(\theta_t)^\top\widetilde{g}(x_t;\theta_t) \\ &+ \frac{L\eta^2}{2}\|\widetilde{g}(x_t;\theta_t)\|^2\end{aligned} \tag{22}$$

Using the result from Lemma 1, take expectation on u_i and x_t and choose a suitable ϵ, we can have

$$\begin{aligned}\mathbb{E}_{x_t}[\mathcal{L}(\theta_{t+1})|\theta_t] &\leq \mathcal{L}(\theta_t) - \eta\nabla\mathcal{L}(\theta_t)^\top\widetilde{H}_t^{-1}\nabla\mathcal{L}(\theta_t) \\ &\quad + \eta\mathcal{O}(\epsilon\|\nabla\mathcal{L}(\theta_t)\|) \\ &\quad + L\eta^2 d(d+2)\beta_u\left(\|\nabla\mathcal{L}(\theta_t)\|^2 + \sigma^2\right) \\ &\quad + \mathcal{O}(\epsilon^2) \\ &\leq \mathcal{L}(\theta_t) - \frac{\eta}{2}\|\nabla\mathcal{L}(\theta_t)\|^2_{\widetilde{H}_t^{-1}} \\ &\quad + L\eta^2 d(d+2)\beta_u\|\nabla\mathcal{L}(\theta_t)\|^2 \\ &\quad + L\eta^2 d(d+2)\beta_u\sigma^2 + \mathcal{O}(\epsilon^2) \\ &\leq \mathcal{L}(\theta_t) - \frac{\eta\beta_l}{2}\|\nabla\mathcal{L}(\theta_t)\|^2 \\ &\quad + L\eta^2 d(d+2)\beta_u\|\nabla\mathcal{L}(\theta_t)\|^2 \\ &\quad + L\eta^2 d(d+2)\beta_u\sigma^2 + \mathcal{O}(\epsilon^2) \\ &= \mathcal{L}(\theta_t) + L\eta^2 d(d+2)\beta_u\sigma^2 \\ &\quad - \left(\frac{\eta\beta_l}{2} - L\eta^2 d(d+2)\beta_u\right)\|\nabla\mathcal{L}(\theta_t)\|^2 \\ &\quad + \mathcal{O}(\epsilon^2).\end{aligned} \tag{23}$$

The last inequality holds because Assumption 3 and $\widetilde{H}_t^{-1}$ is a diagonal matrix. Then we have

$$\begin{aligned}&\left(\frac{\eta\beta_l}{2} - L\eta^2 d(d+2)\beta_u\right)\|\nabla\mathcal{L}(\theta_t)\|^2 \\ &\quad\leq \mathcal{L}(\theta_t) - \mathbb{E}_{x_t}[\mathcal{L}(\theta_{t+1})|\theta_t] \\ &\quad + L\eta^2 d(d+2)\beta_u\sigma^2 + \mathcal{O}(\epsilon^2).\end{aligned} \tag{24}$$

Sum the term from $t=1$ to $t=T$, take expectation, and using the telescoping sum, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\eta\beta_l}{2} - L\eta^2 d(d+2)\beta_u\right)\|\nabla\mathcal{L}(\theta_t)\|^2\right] \\ &\quad\leq \mathcal{L}(\theta_0) - \mathbb{E}[\mathcal{L}(\theta_T)] \\ &\quad + TL\eta^2 d(d+2)\beta_u\sigma^2 + \mathcal{O}(T\epsilon^2).\end{aligned} \tag{25}$$

Because $\eta = \frac{\beta_l}{2Ld(d+2)\beta_u\sqrt{T}}$, we have

$$\frac{\eta\beta_l}{2} - L\eta^2 d(d+2)\beta_u = \frac{\beta_l^2\left(\sqrt{T}-1\right)}{4Ld(d+2)\beta_u T}, \tag{26}$$

and

$$TL\eta^2 d(d+2)\beta_u\sigma^2 = \frac{\beta_l^2\sigma^2}{4Ld(d+2)\beta_u} \tag{27}$$

Because $\mathbb{E}[\mathcal{L}(\theta_T)] \leq \mathcal{L}(\theta^*)$ where θ^* is the global minimum and we divide both side by $\left(\frac{\eta\beta_l}{2} - L\eta^2 d(d+2)\beta_u\right)T$, we can obtain the final resultas follow

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta_t)\|^2\right] &\leq \frac{4Ld(d+2)\beta_u\left(\mathcal{L}(\theta_0)-\mathcal{L}(\theta^*)\right)}{\beta_l^2\left(\sqrt{T}-1\right)} \\ &\quad + \frac{\sigma^2}{\sqrt{T}-1} + \mathcal{O}(\epsilon^2).\end{aligned} \tag{28}$$

The proof is finished.

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta_t)\|^2\right] \leq \mathcal{O}\left(\frac{d^2\beta_u}{\beta_l^2\left(\sqrt{T}-1\right)}+\frac{\sigma^2}{\sqrt{T}-1}\right)+\mathcal{O}(\mu^2). \tag{29}$$

□

9. Experimental Settings

9.1. Dataset and Model

We evaluate our method on ImageNet-C, a benchmark dataset for test-time adaptation [19]. ImageNet-C contains 15 common corruption types, each with five severity levels, yielding 75 corrupted versions of the ImageNet validation set. The corruptions include noise, blur, weather-related distortions, and other real-world degradations. We use ViT-Base/16 [9] as the source model. The ViT-Base model consists of 12 layers, with a hidden dimension of 768. The model is pretrained on the original ImageNet datasetv[6], and we evaluate its performance in the test-time adaptation setting on ImageNet-C. Furthermore, we evaluate our approach on three domain-shifted benchmark datasets: ImageNet-R [20], ImageNet-V2 [42], and ImageNet-Sketch [53].

ImageNet-V2 is a newly curated test dataset derived from the same underlying distribution as the original ImageNet. This dataset consists of three distinct test sets, each containing 10,000 images, and collectively spans 1,000 classes found in ImageNet. In line with prior methods for test-time augmentation (TTA) [34], we specifically employ the Matched-Frequency subset of ImageNet-V2 for our evaluation. This subset is designed such that the images are sampled to align with the class frequency distributions of the original ImageNet validation dataset.

ImageNet-R (Renditions) provides a specialized benchmark for evaluating robustness against *concept shifts*. It contains 30,000 images spanning 200 ImageNet classes, curated from diverse non-photorealistic renditions including artwork, cartoons, graffiti, sculptures, and video game renders. Crucially, all images are selected from samples misclassified by ResNet-50 models, emphasizing challenging semantic variations. The label space aligns with ImageNet-2012, enabling direct compatibility with standard evaluation pipelines. This dataset is particularly effective for testing model adaptability to abstract representations where texture and shape biases significantly impact performance [21].

ImageNet-Sketch focuses on *structural generalization* by exclusively using hand-drawn sketches of ImageNet objects[53]. Each sketch replicates the original 1,000-class structure, providing a domain-shifted testbed devoid of photographic textures. Models relying heavily on texture cues exhibit significant performance drops on this dataset, making it a critical benchmark for evaluating shape-based reasoning capabilities. Its construction mirrors the ImageNet validation set in class distribution and scale, ensuring comparable statistical rigor.

9.2. Baseline Methods

We compare CAZO with two categories of methods: non-backpropagation-based methods and backpropagation-based methods. The non-backpropagation methods include: LAME [1], a post-training adaptation technique that refines the model's output probabilities; T3A [23], which adjusts the model's linear classifier during inference; and FOA [40], which employs a covariance matrix adaptation evolution strategy to update prompts. In addition, we include ZOA baseline [7] that estimates gradients via two-point random perturbations without backpropagation; each adaptation step uses multiple forward-only evaluations to form a ZO gradient estimate. Backpropagation-based methods include TENT [51], which minimizes entropy to fine-tune normalization layer parameters; CoTTA [54], which combines knowledge distillation with data augmentation; and SAR [39], which selects reliable samples to stabilize predictions. EATA [38] further improves entropy-minimization based TTA by filtering unreliable samples and adding a lightweight regularization to reduce forgetting, typically updating only BN-related parameters with BP under an entropy loss; this yields efficient, stable adaptation under streaming data. DeYO [27] argues that pure entropy minimization can be deceptive on domain-shifted targets; it disentangles/regularizes latent factors to curb degeneration, combining entropy terms with factor-aware constraints under BP-based updates to improve robustness in continual settings. RoTTA [56] focuses on temporally varying test streams, using teacher–student consistency, distribution-aware augmentation and memory mechanisms to remain stable across dynamic shifts; adaptation proceeds with BP while controlling drift. LCoTTA [11] identifies *entropy-deceptive (ED)* samples as the cause of degeneration in continual TTA, reveals that entropy-minimization gradients possess a low-dimensional principal subspace dominated by *entropy-truthful (ET)* samples, and constrains weight updates by *tracking* this principal subspace online and *projecting* gradients into it. This subspace-projected BP update suppresses ED gradients, stabilizing long-horizon adaptation across many cycles.

Additionally, to demonstrate the effectiveness of the curvature-aware sampling approach in CAZO, we consider a ZO baseline algorithm using the standard RGE perturbation form [12, 36] within the same framework of CAZO. This ZO baseline employs standard Gaussian distribution perturbations and estimates gradients through multiple double-point perturbations, which maintains the BP-free nature of our approach.

9.3. Implementation Details

In our experiments, we set the batch size to 64 and the **learning rate** to 0.01 for all algorithms. We use the ViT-B/16 model with 12 transformer layers, integrating an adapter into an early layer for adaptation. The adapter employs a scaling factor of 384 for upsampling and downsampling, with additional tests on multiples of 768 to better align with the ViT model dimensions. The adapter is initialized using Kaiming initialization for downsampling and zero initialization for upsampling to ensure stable adaptation. The number of perturbations is set to 20 for all ZO-based method. The adapter scaling factor is 0.1, based on the configuration in AdaptFormer [3]. For the ImageNet-C dataset, we evaluate on the most severe corruption level (level 5). And we set five **random-seed**: $seed \in \{42, 2020, 2025, 1234, 888\}$.

Table 6. Experiment hyperparameters

Hyperparameter	Value
Batch size	64
Learning rate	0.01
Vision Transformer model	ViT-B/16 (12 transformer layers)
Adapter integration location	Early layer(layer = 3)
Adapter downsampling scaling factor	Main: 384 Test variants: Multiples of 768
Adapter initialization	Downsampling: Kaiming Upsampling: Zero
Adapter residual scaling factor	0.1
Number of perturbations (ZO-based methods)	20
ImageNet-C corruption level	5 (most severe)
Random seeds	{42, 2020, 2025, 1234, 888}

9.4. Evaluation Metrics

In our experiments, we evaluate the performance of our model using two primary metrics: classification accuracy (ACC) and Expected Calibration Error (ECE). ACC measures the proportion of correctly classified samples on the out-of-distribution (OOD) data, providing an indication of the model's robustness in handling data that differs from the training distribution. ECE quantifies the calibration of predicted probabilities, reflecting the difference between the predicted confidence and the actual accuracy of predictions [35]. A lower ECE indicates better calibration, meaning the model's predicted probabilities align more closely with the true likelihood of correct predictions.

10. More Experiment Details

10.1. Analysis of Hessian Matrix Properties in TTA

Motivation Experiment Description This pilot study investigates the structural properties of Hessian matrices during test-time adaptation (TTA). During experimentation, we employed a batch size of 32 and executed 100 optimization steps. At each step t, we recorded the Hessian matrix alongside critical quantities including weight updates, gradient statistics, and the effective rank derived from spectral analysis. The effective rank computation followed Shannon entropy principles: *effective rank* $= \exp\left(-\sum_i p_i \log p_i\right)$ where $p_i = |\lambda_i| / \sum_{j=1}^{n} |\lambda_j|$ represents the normalized magnitude of the i-th eigenvalue λ_i of the Hessian matrix. A lightweight single-layer adapter module was incorporated, with all parameters continuously updated through standard backpropagation throughout the test-time adaptation (TTA) process. Through comprehensive eigenvalue spectrum analysis shown in Figure 7, we examine some critical aspects:

- **Eigenvalue Decay Pattern** (Fig. 7a): Demonstrates exponential magnitude decay (10^{-1} to 10^{-9}) across optimization steps. Notably, the first 10 eigenvalues contain $> 99\%$ of total spectral energy, with decay slopes remaining consistent throughout adaptation.
- **Cumulative Explained Variance** (Fig. 7b): Confirms that fewer than 15 principal components capture $> 90\%$ variance at all optimization stages (consistently above both 90% and 95% thresholds), indicating dimensionality saturation.
- **Effective Rank Evolution** (Fig. 7c): Quantifies stable low-dimensional structure, where effective rank plateaus at $\sim$12 after initial optimization phase.

Hessian Matrix Projection Ratio Analysis in TTA This section presents a detailed analysis of the projection ratio $\rho_t^{(r)}$, which quantifies how much the principal curvature directions of the Hessian are preserved across test-time adaptation steps. As shown in Figure 7, we investigate this ratio under different batch sizes, using a batch size of 512 for the primary experiments (Figure 7a-c), and a batch size of 200 for comparison (Figure 7d). The projection ratio for the larger batch size remains relatively stable throughout the adaptation process, reflecting a smoother and more consistent estimation of the Hessian's principal subspace. In contrast, with batch size 200, the projection ratio shows greater fluctuations, indicating higher instability in the Hessian estimation. This observation suggests that using larger batch sizes helps reduce the noise in curvature estimates, leading to more stable adaptation dynamics.

- **Projection Ratio with Batch Size 200** (Fig. 7d): When using batch size 200, the projection ratio exhibits more significant fluctuations, suggesting that a larger batch size selection can better fit the Loss landscape, thereby providing a more powerful explanation of the slow changing nature of curvature. Therefore, batch size 512 is used as the main experiment in the main text.

Key Findings Three fundamental observations from this analysis:

- **Consistent Low-Rank Structure**: The persistent eigenvalue decay patterns across optimization steps (across steps 0, 25, 99) reveals intrinsic low-rank properties independent of adaptation progress.
- **Dimensional Inertia**: The stabilized effective rank ($r_{\text{eff}} \approx 12$) and variance concentration demonstrate fixed intrinsic dimensionality, despite continuous parameter updates.
- **Slow-Varying Curvature**: The Hessian matrix exhibits a slow-varying property over time, as evidenced by the stable projection

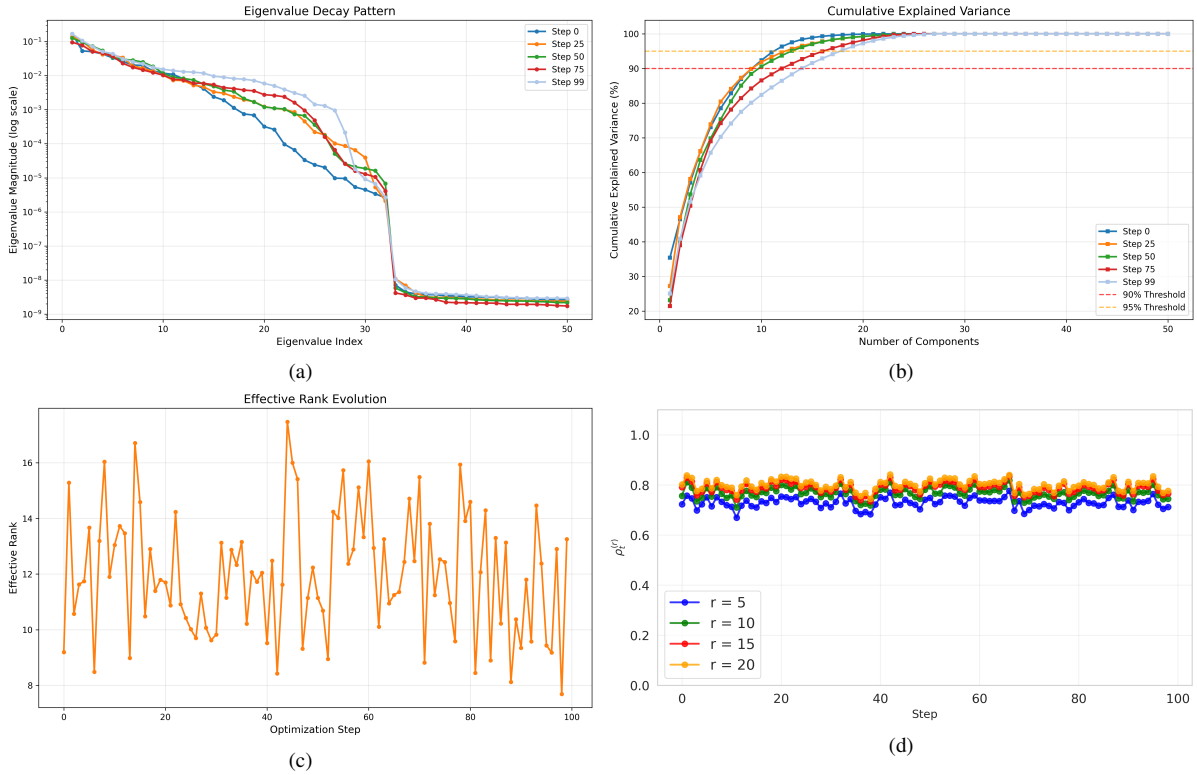


Figure 7. Hessian eigenvalue distribution analysis during test-time adaptation process. Four complementary perspectives are presented: (a) Eigenvalue decay pattern across optimization steps; (b) Cumulative explained variance of principal components; (c) Evolution of effective rank; (d) Projection rate with batch size 200 and 100 steps

Table 7. Entropy-only loss results under the same protocol (ImageNet-C, severity-5).

Method	FOA	ZO (RGE)	CAZO
Acc. (%, ↑)	44.90	47.32	56.52

ratio and effective rank. The principal curvature directions in the Hessian remain largely consistent across adaptation steps, suggesting that the underlying loss landscape does not undergo abrupt changes during the adaptation process.

10.2. Entropy-Only Loss Supplement

We further evaluate CAZO under an *entropy-only* objective, i.e., L_{ent} without the feature-alignment term used in our default composite loss $L_{\text{com}} = L_{\text{ent}} + L_{\text{align}}$. This setting is known to be challenging for BP-free TTA, and we observe a substantial performance degradation for forward-only baselines. Nevertheless, CAZO remains significantly more robust than FOA and vanilla ZO (RGE) under the same protocol.

In addition, our Hessian observations are not loss-specific: we also ran the Hessian analysis using the composite loss L_{com} and still observe a persistent low-rank spectrum as well as a slowly varying dominant subspace (Fig. 8).

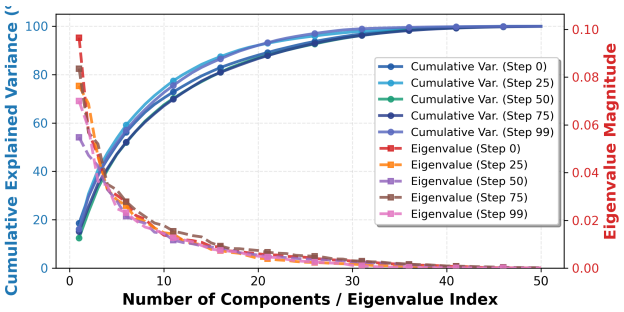
10.3. Adapter Layer Selection Across Transformer Backbones

In the main paper, we select an early layer (layer-3) to insert the adapter for adaptation efficiency. To verify the feasibility and generality of this choice beyond ViT-B/16, we conduct additional experiments on DeiT[48] and Swin-Tiny[32]. As shown in Fig. 9, we observe a consistent trend across architectures: inserting the adapter into relatively early layers (typically the 3rd or 4th block) yields the best performance. Therefore, selecting layer-3 (or layer-4 when preferred) provides a practical and universal default for different Transformer backbones.

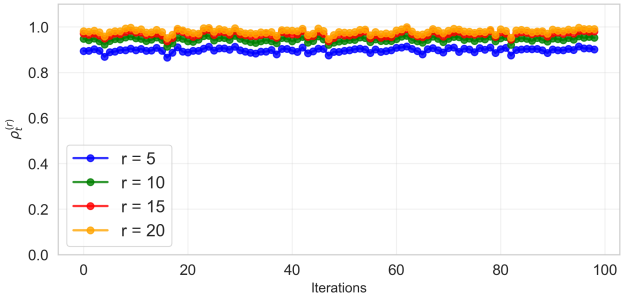
10.4. Controlled Experiments on Optimization Efficiency Under TTA Constraints

Test-time adaptation is a constrained unsupervised setting where each test sample is seen only once, and the adaptation must be completed within a limited number of update steps. Therefore, optimization efficiency under a limited update budget is crucial. We conduct controlled experiments comparing (i) standard backpropagation-based adaptation (BP), (ii) vanilla zeroth-order optimization (ZO; RGE), and (iii) CAZO, under: *(a) unsupervised TTA* and *(b) supervised adaptation (fine-tuning)* with the same number of parameter updates.

As expected, BP adapts fastest. Vanilla ZO suffers from slow convergence due to the well-known variance scaling with dimen-



(a) Low-rank Hessian spectrum



(b) Slow-varying subspace $\rho_t^{(r)}$

Figure 8. Hessian analysis for the composite loss $\mathcal{L}_{\text{com}}$.

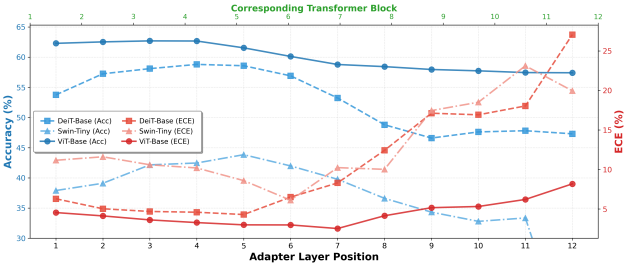


Figure 9. Adapter layer position ablation across Transformer backbones (ViT, DeiT, Swin-Tiny). Early layers (e.g., 3rd/4th) tend to perform best.

sion $\mathcal{O}(d)$, which becomes more severe under the short-horizon TTA constraint. By leveraging curvature information through anisotropic perturbation sampling, CAZO reduces the variance of ZO updates and achieves much faster learning, leading to significant improvement over vanilla ZO in the limited-step TTA regime. Figure 10 summarizes the learning dynamics under these controlled settings.

10.5. Extended Ablation Study

Downsampling Ratio We evaluate the effect of the downsampling ratio of the adapter. Table 8 presents the results of different downsampling ratios of the adapter when it is applied to the first transformer layer of ViT.

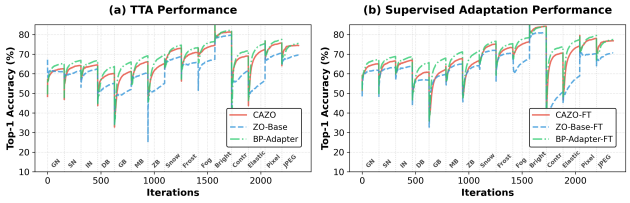


Figure 10. Controlled experiments quantifying the impact of TTA constraints on optimization efficiency: (a) unsupervised TTA; (b) supervised adaptation with the same update budget.

Table 8. Adapter downsampling with different ratios in CAZO. The values in parentheses represent the downsampled dimensions.

Downsampling Ratio	384(2)	256(3)	192(4)	128(6)	96(8)	48(16)
Accuracy	62.7	62.6	62.8	62.7	62.5	62.3
ECE	3.28	3.20	3.14	3.29	3.27	3.21

EMA Update Interval We study the sensitivity to the EMA update interval τ for the curvature proxy. Specifically, τ denotes the interval (in steps) at which we update the EMA-based diagonal curvature estimate; larger τ reduces the update frequency but may lag behind curvature drift. Table 9 shows that $\tau \in \{1, 2, 3\}$ yields nearly identical performance, while overly infrequent updates (e.g., $\tau = 10$) lead to a noticeable degradation. We use $\tau = 1$ by default.

Table 9. Ablation on EMA update interval.

Update interval τ					
τ	1	2	3	5	10
Acc.	**69.01**	69.00	68.98	68.80	68.34

Perturbation Scale We also evaluate the perturbation magnitude ϵ used in symmetric finite-difference gradient estimation. As shown in Table 10, a moderate ϵ (around 0.1) achieves the best accuracy. Too small ϵ can be dominated by numerical noise and yield weak signal, while too large ϵ introduces bias by probing a non-local region of the loss landscape. We use $\epsilon = 0.1$ in all experiments.

Table 10. Ablation on perturbation scale.

Perturbation scale ϵ							
ϵ	0.01	0.05	0.08	0.10	0.20	0.50	1.00
Acc.	61.01	67.75	68.82	**69.01**	61.78	49.25	44.08

Exploration of Visual Prompt Tuning (VPT) We also consider adapting VPT and evaluate the differences among different parameter efficient fine-tuning methods as in Table 11.

Our comprehensive comparison between visual prompt tuning (VPT) and adapter-based efficient tuning methods reveals no discernible performance advantage for the VPT paradigm. Under equivalent parameter budgets (VPT implemented with 3 prompts),

Table 11. Performance comparison of VPT versus Adapter methods on ImageNet-C (severity level 5) with ViT: Accuracy (%, ↑). Batch size is fixed at 64.

	Noise			Blur Defoc.				Weather				Digital				Average
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Acc.
NoAdapt	56.8	56.8	57.5	46.9	35.6	53.1	44.8	62.2	62.5	65.7	77.7	32.6	46.0	67.0	67.6	55.5
ZO (Adapter)	61.7	62.5	63.2	54.2	51.9	59.9	52.7	67.9	67.7	68.8	79.8	65.5	57.4	70.7	71.1	63.6
FOA (VPT)	61.6	62.5	63.2	58.5	54.0	61.2	57.0	69.6	68.6	74.1	80.9	67.3	63.3	73.6	72.7	65.9
CAZO (VPT)	61.8	62.7	63.1	58.0	51.6	62.6	57.4	70.5	66.2	73.2	81.2	67.8	68.4	74.7	72.3	66.1
CAZO (Adapter)	62.7	64.3	64.1	60.0	61.5	66.0	64.8	72.7	71.0	74.3	81.3	69.5	72.7	75.7	74.5	69.0

CAZO-VPT (66.1%) exhibits lower average accuracy than CAZO-Adapter (69.0%) on ImageNet-C severe corruption benchmarks. This performance gap persists across all corruption categories including noise, blur, weather, and digital distortions. The consistent accuracy deficit relative to adapting adapter informs our methodological selection. Consequently, we adopt adapter as our primary tuning framework, prioritizing its superior corruption robustness as evidenced by the comprehensive benchmark results.

Different ν Influence in CAZO Table 12 shows the performance of CAZO under different ν values (with $\epsilon = 0.1$). The results are reported on ImageNet-C (severity level 5) with ViT-B/16, and both accuracy (Acc.) and expected calibration error (ECE) are averaged over all corruptions.

Table 12. Performance of CAZO for different ν on ImageNet-C (severity level 5) with ViT-B/16.

ν	Acc. (%, ↑)	ECE (%, ↓)
0.01	68.3	4.6
0.05	68.5	4.5
0.1	68.5	4.5
0.2	68.5	4.4
0.3	68.6	4.4
0.4	68.6	4.4
0.5	68.7	4.4
0.6	68.9	4.3
0.7	68.9	4.3
0.8	69.0	4.3
0.9	68.8	4.2
0.95	64.4	5.0
1	0.1	*

The results in Table 12 show that choosing a value of ν in the range from 0.6 to 0.9 can yield satisfactory performance. We use $\nu = 0.8$ in all experiments.