ON THE INTERACTION BETWEEN MODEL COMPRESSION AND TEST-TIME ADAPTATION

Francesco Corti
Graz University of Technology
Graz, Austria
francesco.corti@tugraz.at

Dong Wang
Graz University of Technology
Graz, Austria
dong.wang@tugraz.at

Young D. Kwon
Samsung AI Center-Cambridge
University of Cambridge
Cambridge, United Kingdom
yd.kwon@samsung.com

Cecilia Mascolo
University of Cambridge
Cambridge, United Kingdom
cm542@cam.ac.uk

Olga Saukh
Graz University of Technology
Complexity Science Hub
Graz, Vienna, Austria
saukh@tugraz.at

ABSTRACT

Deep neural networks deployed in the wild must be both efficient and adaptable, requiring model compression and test-time adaptation (TTA). While both are well studied in isolation, their interaction remains poorly understood. We systematically analyze how structured compression affects a model's ability to adapt under distribution shift. Using ResNet-18 and ViT-Base on CIFAR-10-C and ImageNet-C, we evaluate multiple compression methods combined with standard TTA techniques. We introduce a diagnostic framework that examines representational expressivity and adaptation subspace compatibility. Our results reveal a consistent gap: although compressed models retain high accuracy under supervised adaptation, their TTA performance degrades significantly with increasing compression. We show that this stems from reduced representational diversity and structural constraints that limit recoverability. These effects strongly depend on the compression method, highlighting the need to design compression strategies that preserve adaptability.

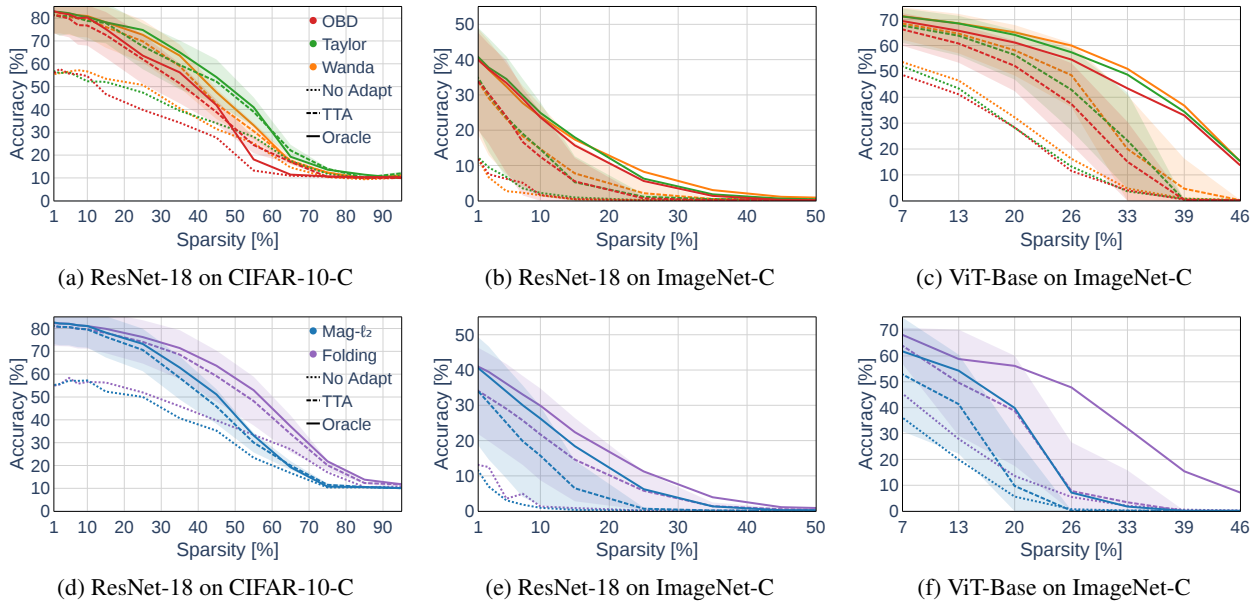
1 INTRODUCTION

Deployed deep neural networks (DNNs) often operate under resource constraints and distribution shifts, necessitating both compression to meet memory and latency budgets and adaptation to cope with changing environments (Han et al., 2015; Sun et al., 2020; Wang et al., 2021). While compression and model adaptation have been extensively studied in isolation, their interaction remains largely unexplored (Liang et al., 2025; Choudhary et al., 2020). In particular, it is unclear whether compression strategies can preserve the capacity of a model to adapt to distribution shifts or task variations (Liebenwein et al., 2021).

This interaction is especially critical in edge settings, where adaptation incurs substantial memory overhead due to gradient computation and the need to store intermediate activations (Dong et al., 2026; Gomez et al., 2017; Xiao et al., 2026a). For example, on resource-constrained devices such as the Raspberry Pi Zero, adaptation via entropy minimization leads to out-of-memory errors even at small batch sizes (Jia et al., 2024). Consequently, compression is not only desirable for efficient inference but also a prerequisite for enabling on-device adaptation (Lin et al., 2022).

Structured compression methods reduce model size by removing or merging entire computational units (Li et al., 2017; Wang et al., 2025). These methods can be divided into data-dependent methods, which leverage calibration data to estimate unit importance (Sun et al., 2024; Molchanov et al., 2019; LeCun et al., 1989), and data-free methods, which rely on weight statistics (Li et al., 2017) or structural redundancies (Wang et al., 2025).

Test-time adaptation (TTA) has emerged as a promising paradigm for adapting deep models using only unlabeled incoming data (Liang et al., 2025; 2020; Wang et al., 2021). Existing approaches can be broadly categorized into backpropagation-based (BP) and backpropagation-free (BP-free) methods. BP-based methods update model parameters via gradient-based optimization, typically fine-tuning a subset of parameters (Wang et al., 2021; Niu et al., 2023; Gong et al., 2023). While achieving strong adaptation performance, they incur substantial memory overhead due to gradient computation and activation storage, limiting their applicability on resource-constrained devices (Jia et al., 2024).



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 1: **The gap between test-time adaptation (TTA) and supervised fine-tuning (Oracle) widens with increasing compression, particularly on ImageNet-C.** We compare post-compression TTA against supervised fine-tuning across architectures and datasets, averaged across all corruptions. At low sparsity levels, TTA remains competitive, suggesting that adaptation signals are largely preserved. However, as compression increases, the gap between TTA and supervised fine-tuning grows, suggesting that compressed architectures increasingly hinder effective adaptation with TTA. This effect is especially pronounced on ImageNet-C. Shaded regions indicate $\pm\sigma$ across corruption types. **Top row** (a–c): data-dependent pruning (Wanda, Taylor, OBD). **Bottom row** (d–f): data-free compression (Folding, Mag-ℓ_2). TTA is performed using SAR (Niu et al., 2023) for ResNet-18 and SPA (Niu et al., 2025) for ViT-Base.

In contrast, BP-free methods avoid gradient computation by adapting parameters during the forward pass (Mirza et al., 2022), modifying lightweight input prompts (Niu et al., 2024) or correcting embedding distortions via layer-wise alignment (Xiao et al., 2026b), making them more suitable for edge deployment. Despite this progress, existing studies predominantly assume dense models and do not account for compression (Mirza et al., 2022; Wang et al., 2021; Niu et al., 2023; Wang et al., 2022; Niu et al., 2022; Sun et al., 2020). This assumption is misaligned with resource-constrained deployment scenarios, where models are routinely compressed to satisfy hardware and efficiency constraints (Fang et al., 2023; Ashkboos et al., 2024; Ma et al., 2023).

Recent studies have shown that compression degrades learned representation quality (Corti et al., 2022), and erodes robustness under distribution shift (Liebenwein et al., 2021). Yet existing work that explicitly considers the compression and adaptation interaction remains limited to quantization (Xiao et al., 2024; Deng et al., 2025). Whether structurally pruned or folded models preserve the representational and geometrical properties required for test-time adaptation remains an open question. In this work, we address the following question:

> *Does structured compression degrade a model's ability to adapt at test time, or can suitable compression strategies preserve adaptability under distribution shift and task variation?*

This question is closely related to continual learning: models deployed at the edge must retain *plasticity*, *i.e.*, the ability to adapt to incoming data over time (Wang et al., 2022). However, compression applied to meet deployment constraints may inadvertently erode this capacity. If compression reduces feature diversity, misaligns the unsupervised adaptation signal from the supervised gradient direction, or restricts the useful update subspace, the standard compress-and-deploy pipeline can yield models that maintain high accuracy on source data but lose the ability to adapt. We refer to this phenomenon as *silent plasticity loss*.

Figure 1 provides empirical evidence of this effect: while compressed models retain strong performance under supervised adaptation, their test-time adaptation performance degrades substantially as compression increases, leading to a widening gap between the two. This gap indicates that, despite preserved predictive performance, compression can impair the model's ability to effectively utilize unlabeled test-time signals. A per-corruption decomposition of this gap (Fig. 6) shows that the widening trend generalizes across most corruption types rather than being driven by a subset. To

isolate the compression-induced component of this widening from the supervised-unsupervised signal-quality difference already present for the dense model, Fig. 10 reports the Oracle-TTA gap after subtracting its zero-compression value.

We introduce a framework to diagnose how compression interacts with TTA along two complementary dimensions (Sec. 3). First, *diversity/expressivity*: compression may reduce the variability and richness of internal representations, limiting the range of functions the model can realize during TTA. We quantify this using representation similarity between dense and compressed models (Kornblith et al., 2019), as well as activation entropy across feature maps and tokens (Skean et al., 2025). Second, *subspace compatibility*: TTA updates are restricted to a small parameter subspace (*e.g.*, normalization layers), and compression may misalign this subspace with useful update directions. We quantify this via gradient cosine similarity between unsupervised and supervised updates, and provide a theoretical analysis of objective-induced gradient degeneracy in both entropy-based and consistency-based TTA objectives.

Our empirical analysis across convolutional and transformer architectures reveals the following insights:

- The gap between TTA and supervised adaptation (Oracle) widens with increasing compression, suggesting that TTA depends on structural properties of dense models that are not preserved under compression.
- TTA degrades representations already at low compression levels and fails to recover them at higher sparsity, whereas supervised adaptation consistently restores them, highlighting TTA 's sensitivity to compression-induced representational distortions.
- Compression induces two distinct failure modes of gradient alignment between TTA and supervised updates: *gradient degeneracy*, where near-uniform predictions attenuate the TTA signal, and *active divergence*, where the TTA gradient opposes the supervised gradient direction even at minimal sparsity. We show theoretically that both entropy- and consistency-based TTA objectives exhibit gradient collapse under compression, while the supervised cross-entropy gradient can remain more informative at the same sparsity.
- The severity of these effects depends on the compression method: Wanda and Taylor better preserve recoverable structure, while OBD and Mag-ℓ_2 lead to pronounced degradation; folding achieves more stable recovery.
- These effects are substantially more pronounced on ImageNet-C than on CIFAR-10-C, highlighting the increased difficulty of adaptation under complex tasks.

2 Related Work

Structured Model Compression. Structured pruning removes entire computational units (*e.g.*, convolutional filters (Li et al., 2017) or attention heads (Michel et al., 2019)), enabling efficient inference on commodity hardware (Li et al., 2017). Selection criteria range from magnitude-based heuristics (Li et al., 2017) to activation-aware scores (Sun et al., 2024), first-order Taylor approximations (Molchanov et al., 2019), and second-order Hessian-based methods (LeCun et al., 1989). Beyond pruning, model folding clusters similar channels and replaces them with shared representations (Wang et al., 2025; Saukh et al., 2026; Son et al., 2018). Existing benchmarks primarily evaluate compressed models by their accuracy on the source distribution (Liebenwein et al., 2021; Hooker et al., 2020). This protocol does not capture whether compressed models retain the capacity to adapt under distribution shift. As reported in Fig. 1, these properties can diverge, indicating that standard compression benchmarks provide an incomplete view of model quality.

Test-Time Adaptation. Test-time adaptation (TTA) adapts deployed models to distribution shifts using only unlabeled data. TENT (Wang et al., 2021) updates batch-normalization parameters via entropy minimization; SAR (Niu et al., 2023) extends this with sharpness-aware optimization and reliable sample filtering; EATA (Niu et al., 2022) incorporates an anti-forgetting regularizer to preserve important weights. For vision transformers, SPA (Niu et al., 2025) enforces consistency between clean and augmented inputs. In contrast, BP-free methods eliminate gradient computation at test time: FOA (Niu et al., 2024) uses a derivative-free covariance matrix adaptation of new prompt embeddings; PEA (Xiao et al., 2026b) corrects the embedding distortions at each intermediate layer through a covariance alignment procedure.

Loss Landscape Geometry and Mode Connectivity. Solutions found by SGD in overparameterized networks are often connected by low-loss curves in weight space (Garipov et al., 2018; Draxler et al., 2018), and become linearly connected when accounting for permutation symmetries (Entezari et al., 2022; Ainsworth et al., 2023). Frankle et al. (2020) formalize *linear mode connectivity* (LMC) and show that pruned subnetworks recover full accuracy only when they are stable to SGD noise. Recent work by Saukh et al. (2026) extends this perspective to model folding, showing how clustering-based compression reshapes the loss landscape geometry. While these works characterize how compression alters the loss landscape, they do not address whether adaptation can compensate for these changes.

Compression-Adaptation Interaction. Plasticity loss in deep networks has been linked to loss landscape properties even without compression (Lyle et al., 2023; Dohare et al., 2024). Compression can further amplify representa-

tional biases, *e.g.*, toward low-frequency subgroups (Hooker et al., 2020). Liebenwein et al. (2021) suggest that overparameterization may be necessary to preserve robustness under distribution shift.

Despite these insights, the mechanistic question remains open: *which structural properties must a compression method preserve to enable effective adaptation?*

3 Diagnostic Framework

Given a classifier $f(x;\theta)$ with pretrained parameters θ_0, trained on a source distribution $\mathcal{P}$, we consider deployment under a shifted target distribution $\mathcal{Q} \neq \mathcal{P}$. To improve performance on $\mathcal{Q}$, test-time adaptation (TTA) updates the model parameters as:

$$\theta^* = \theta_0 + \Delta, \quad \Delta \in \mathcal{U}, \tag{1}$$

where $\mathcal{U}$ denotes the set of admissible update directions (*e.g.*, affine normalization parameters). In contrast, we refer to *Oracle* as supervised adaptation over the same subspace $\mathcal{U}$, using labeled target data.

In practice, models are often compressed prior to deployment. We formalize structured compression as a projection $\Pi : \mathbb{R}^D \to \mathbb{R}^d$ with $d \ll D$, mapping the original parameters to a compressed representation $\phi_0 = \Pi(\theta_0)$. This formulation captures two common compression strategies:

- **Structured Pruning:** Π acts as a coordinate selection operator, applying binary masks to channels or attention heads and retaining only a subset of parameters indexed by $S \subset \{1, \ldots, D\}$.
- **Folding:** Π performs cluster-based aggregation (Wang et al., 2025; Son et al., 2018), grouping similar channels and replacing each cluster with its centroid. This reduces dimensionality via linear combinations rather than explicit removal.

To enable a controlled comparison, we use equal per-layer compression ratios, resulting in matched parameter budgets across methods. The central question is how structured compression affects the model's ability to adapt to $\mathcal{Q}$. While compression reduces model capacity, its interaction with adaptation is governed by more subtle mechanisms, which we analyze next.

Models and Datasets. We evaluate the diagnostic framework on both convolutional and self-attention architectures. Specifically, we use ResNet-18 (He et al., 2016) checkpoints trained on CIFAR-10 (11.1M parameters) and adapted to CIFAR-10-C (Hendrycks & Dietterich, 2019), as well as ResNet-18 (11.7M parameters) and ViT-Base (Dosovitskiy et al., 2021) (86M parameters) checkpoints trained on ImageNet and adapted to ImageNet-C (Hendrycks & Dietterich, 2019). All experiments are conducted at severity level 5 across the 15 corruption types defined in CIFAR-10-C and ImageNet-C. Multi-seed and severity-3 experiments are reported in Appendix A.

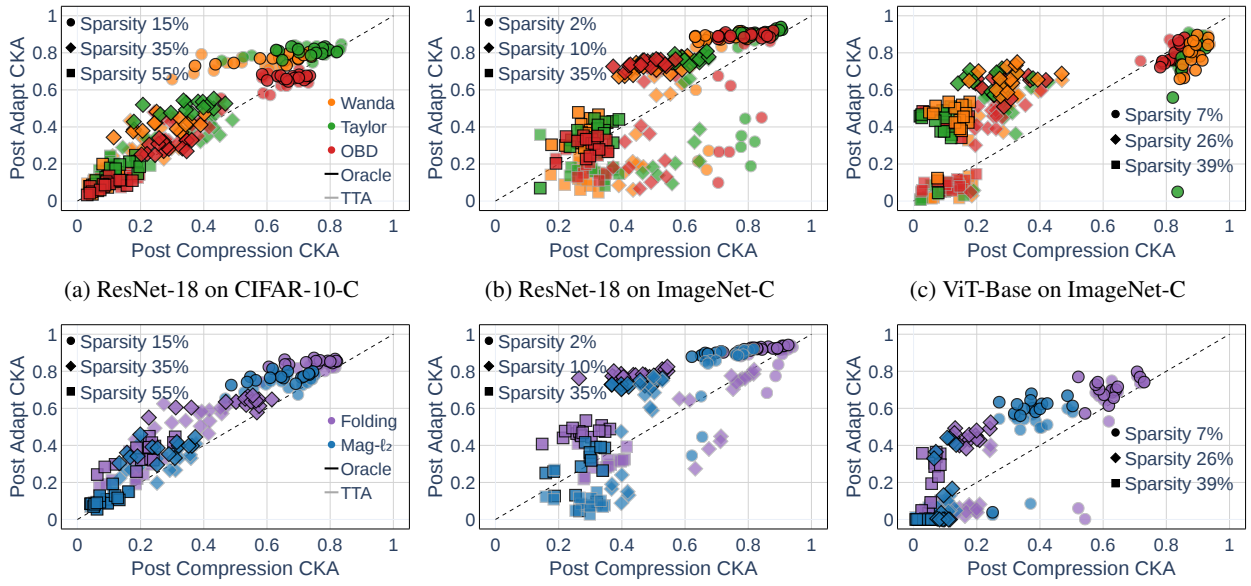
Compression Methods. We analyze five structured compression strategies: magnitude-based ℓ_2 pruning (Li et al., 2017), Wanda (Sun et al., 2024) (combining weight magnitudes with activation norms), Taylor-based pruning (Molchanov et al., 2019), Optimal Brain Damage (OBD) (LeCun et al., 1989) using Hessian-based importance scores, and model folding via k-means channel clustering (Wang et al., 2025).

The *compression ratio* r denotes the uniform fraction of channels removed per layer. For ResNet-18, r corresponds to the fraction of convolutional filters (output channels) removed. For ViT-Base, r specifies the fraction of hidden units removed from the feed-forward network layers in each transformer block. We evaluate $r \in \{0.0, 0.01, \ldots, 0.95\}$ for ResNet-18 (sparsity $s \in [0\%, 94.9\%]$) and $r \in \{0.0, 0.1, \ldots, 0.7\}$ for ViT-Base (sparsity $s \in [0\%, 45.8\%]$). An analysis of the memory use, latency, FLOPs and adaptation time of the compressed models is reported in Fig. 12.

For compressed ResNet-18 models, we apply REPAIR (Jordan et al., 2023) to re-estimate batch normalization statistics using four batches of training data, following Wang et al. (2025). In contrast, ViT-Base uses layer normalization (Ba et al., 2016), which computes statistics per sample and does not require post-compression re-estimation.

Adaptation Methods. We focus on backpropagation-based TTA methods, as our diagnostic framework relies on gradient-based analysis (*e.g.*, gradient cosine similarity and objective-induced degeneracy), which requires backpropagation gradients to exist. For ResNet-18, we use Sharpness-Aware and Reliable (SAR) entropy minimization (Niu et al., 2023) with the original hyperparameters. For ViT-Base, we employ Self-Bootstrapping Adaptation (SPA) (Niu et al., 2025), which optimizes a consistency objective between clean and augmented inputs. Together, these two methods cover TTA entropy-based and consistency-based objectives; whether backpropagation-free methods (Niu et al., 2024; Xiao et al., 2026b) exhibit different sensitivity to compression is an important direction for future work.

To isolate the effect of structured compression from the adaptation objective, we introduce supervised Oracle variants that share the *same* optimization settings (optimizer, trainable parameters, learning rate, and number of adaptation



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

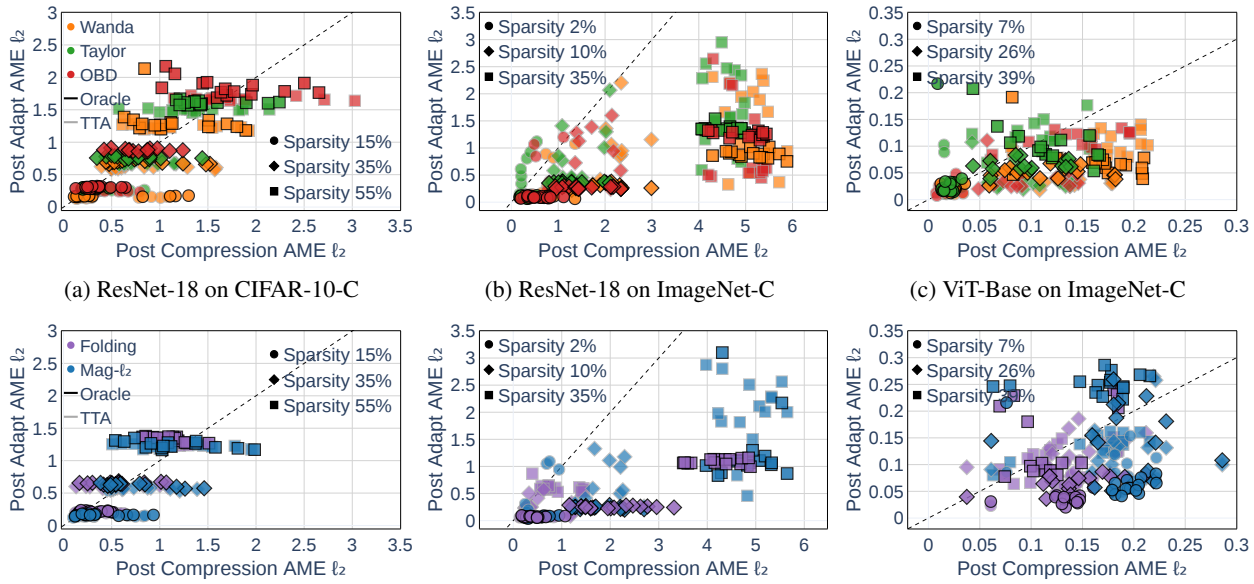
Figure 2: **Test-time adaptation (TTA) degrades representations more than supervised adaptation (Oracle), with the gap depending on the compression method.** We measure worst-layer CKA between dense and compressed models before (x-axis) and after adaptation (y-axis). Points above the identity line ($y{=}x$) indicate that adaptation brings representations closer to the dense model. Across all settings, both TTA and Oracle generally improve representations relative to the compressed model (most points lie above the identity line). However, TTA exhibits substantially weaker recovery than Oracle, with many points lying closer to or below the identity line, especially as compression increases. Notably, this degradation is already visible at low sparsity levels, indicating early loss of adaptable representations. The extent of this effect depends on the compression method: Wanda and Taylor preserve representations that remain partially recoverable, whereas OBD and Mag-ℓ_2 exhibit markedly weaker recovery under TTA. Folding retains higher recoverability at larger sparsity. This gap is particularly strong on ImageNet-C, where TTA fails to recover representations even at moderate sparsity. *Top row* (a–c): data-dependent pruning (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression (Folding, Mag-ℓ_2).

steps) as their analyzed TTA counterparts. For ResNet-18, Oracle-SAR replaces the entropy-based objectives in both inner and outer loops with supervised cross-entropy. For ViT-Base, Oracle-SPA replaces all self-generated signals with supervised cross-entropy losses across all views. We emphasize that the Oracle is not intended as a fair comparison with TTA, but as a diagnostic upper bound on the recoverability of the adaptation subspace $\mathcal{U}$.

3.1 Expressivity and Diversity

Structured compression can reduce the diversity of internal representations and collapse their effective rank, thereby restricting the range of functions the model can realize during adaptation. We ask: *To what extent can adaptation recover the representational expressivity and diversity lost due to structured compression?* To answer this question, we quantify representational recovery using Centered Kernel Alignment (CKA) to measure alignment with the dense model and Activation Map Entropy (AME) to capture the diversity of layer-wise activations. We additionally test whether the representational collapse is an artifact of the similarity index, and whether model size explains the loss of adaptability.

Centered Kernel Alignment. We quantify representational similarity between dense and compressed models using Centered Kernel Alignment (CKA) (Kornblith et al., 2019), computed on post-residual hidden representations and summarized by the worst-layer CKA across all layers. This worst-layer statistic identifies the most degraded layer under compression. As reported in Tishby & Zaslavsky (2015), information lost at any intermediate layer cannot be recovered by subsequent layers; the most degraded layer acts as a representational bottleneck that may constrain the model's capacity for adaptation. The information-bottleneck principle (Tishby & Zaslavsky, 2015) governs the mutual information between the input and the layer-l activation, not the kernel alignment that CKA measures; we use the information-bottleneck argument as an intuition for the monotone non-increase of recoverable information along the



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 3: **Test-time adaptation (TTA) only partially restores output expressivity, with recovery strongly dependent on the compression method.** We measure expressivity via Activation Map Entropy (AME) and report the ℓ_2 distance between entropy vectors of compressed and dense models before (x-axis) and after adaptation (y-axis). Lower values indicate outputs closer to the dense model. Points below the identity line ($y{=}x$) correspond to successful recovery of output expressivity, the lower the better. Across all settings, both TTA and supervised adaptation (Oracle) reduce the AME distance, indicating partial restoration of output diversity. However, Oracle consistently achieves stronger and more stable recovery, while TTA exhibits weaker and less consistent improvements, especially at higher compression levels and on ImageNet-C. The recovery behavior varies across compression methods. Wanda and Taylor enable moderate recovery under TTA, whereas OBD and Mag-ℓ_2 show unstable recovery, often failing to restore output expressivity. In contrast, Folding yields more consistent recovery patterns, but still remains below Oracle. *Top row* (a–c): data-dependent pruning (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression (Folding, Mag-ℓ_2).

forward pass and motivate the worst-layer statistic empirically by reporting its correlation with adaptation accuracy (Fig. 7).

On ResNet-18 (Fig. 2b, 2e), we observe that TTA exhibits representational collapse already at very low sparsity (as early as 2% on ImageNet-C), while Oracle consistently recovers representations closer to the dense model. This gap persists and widens with increasing compression, indicating that TTA fails to restore the most degraded layers. This behavior is consistent with a self-reinforcing degradation under entropy minimization (Niu et al., 2023): since gradients are derived from the model's predictions damaged by compression, prediction errors may influence the TTA adaptation signal, reinforcing representational distortions rather than correcting them. In contrast, Oracle relies on ground-truth labels, which may provide a more robust gradient signal and explain its consistently stronger representation recovery. We formalize this hypothesis in the gradient degeneracy analysis (Sec. 3.2).

On ViT-Base (Fig. 2f), the impact of the compression method becomes more pronounced. At 39% sparsity, Mag-ℓ_2 leads to near-complete representational collapse (CKA ≈ 0), whereas Folding recovers substantially higher post-adaptation alignment with the dense model. This difference may arise from how the two methods transform the activation Gram matrices analyzed by CKA (Kornblith et al., 2019): Mag-ℓ_2 removes entire units and their contributions, while Folding aggregates correlated units via clustering (Wang et al., 2025), better preserving the variance structure of the representations. This observation is consistent with Hu et al. (2024), who show that pruning in the original feature dimension disrupts principal components that span multiple dimensions, motivating projection-based approaches.

Activation Map Entropy. While CKA captures representational *expressivity* via alignment with the dense model (Kornblith et al., 2019), it does not characterize the spectral diversity of layer-wise activations. To complement this, we define Activation Map Entropy (AME) building on the matrix-based entropy framework of Skean et al. (2025) as a measure of representational *diversity*. For each layer l, AME is defined as the Shannon entropy of the normalized eigenvalues of the activation Gram matrix $G_l \in \mathbb{R}^{C_l \times C_l}$, quantifying the effective rank of the representation (Roy & Vetterli, 2007).

Structured compression reduces G_l from $C_l \times C_l$ to $C'_l \times C'_l$, imposing an upper bound on entropy at $\log C'_l$. Since TTA methods (Niu et al., 2023; Wang et al., 2021) update only normalization parameters, they cannot increase the rank of G_l beyond C'_l. Compression therefore imposes an upper bound on the recoverable diversity of representations. Consistent with prior observations that pruning reduces activation entropy (Liao et al., 2024), we expect this reduction to propagate to the eigenspectrum of G_l.

Empirically, on CIFAR-10-C (Fig. 3a, 3d), both Oracle and TTA recover AME at low sparsity (15%), but diverge at higher sparsity (35%, 55%), where points form horizontal saturation bands. These bands reflect the entropy ceiling imposed by reduced dimensionality, limiting further recovery regardless of the adaptation method. On ImageNet-C with ResNet-18 (Fig. 3b, 3e), this limitation becomes more pronounced: at 35% sparsity, most points lie below the identity line, indicating partial recovery, with Oracle consistently achieving tighter and lower-distance clusters than TTA. This suggests that while both methods operate under the same dimensionality constraint, Oracle more effectively redistributes the available eigenvalue mass toward the dense reference.

On ViT-Base (Fig. 3c, 3f), AME distances are substantially smaller and less sensitive to compression. This can be attributed to FFN-only compression, which preserves the post-residual dimensionality fixed at 768 (Dosovitskiy et al., 2021), keeping the size of G_l invariant. As a result, the entropy ceiling remains largely unchanged, and residual connections further mitigate the impact of compression on the activation spectrum. Nevertheless, compression methods still differ: Folding achieves lower post-adaptation AME distances than Mag-ℓ_2 and consistently outperforms TTA, aligning with the trends observed in the CKA analysis (Fig. 2f).

Metric Similarity Analysis. To further support our analysis, we adopt the spectral-ratio pseudo-distance d_{SR} (Cayco-Gajic & Pellegrino, 2026), which compares the intrinsic geometry of neural representations under the manifold hypothesis, whereas CKA compares their extrinsic geometry. Following Cayco-Gajic & Pellegrino (2026), we apply d_{SR} to the Riemannian pullback metrics of the dense and the compressed model, which measure how strongly each model's output responds to small perturbations of the input, rather than to the activation Gram matrices G_l analyzed by CKA and AME. d_{SR} ranks post-compression representation consistently with worst-layer CKA (Spearman $\rho = -0.925$ on ResNet-18/ImageNet-C and -0.665 on ViT-Base/ImageNet-C; the negative sign is expected when a distance metric is compared with a similarity; see Fig. 17), indicating that the representational collapse observed on ImageNet-C is not an artifact of the similarity index (Davari et al., 2023).

Disentangling compression from capacity. We compare compressed against *smaller-dense* models of equal architecture trained from scratch: at increasing sparsity, TTA and its Oracle adaptation show an earlier collapse compared to the smaller-dense models (Fig. 19), indicating that the loss of adaptability is induced by the structured compression rather than by the reduced model size.

Structured compression imposes intrinsic limits on both representational alignment and diversity that adaptation cannot fully overcome. While Oracle approaches these limits, TTA remains constrained by degraded signals and reduced capacity, resulting in systematically weaker recovery.

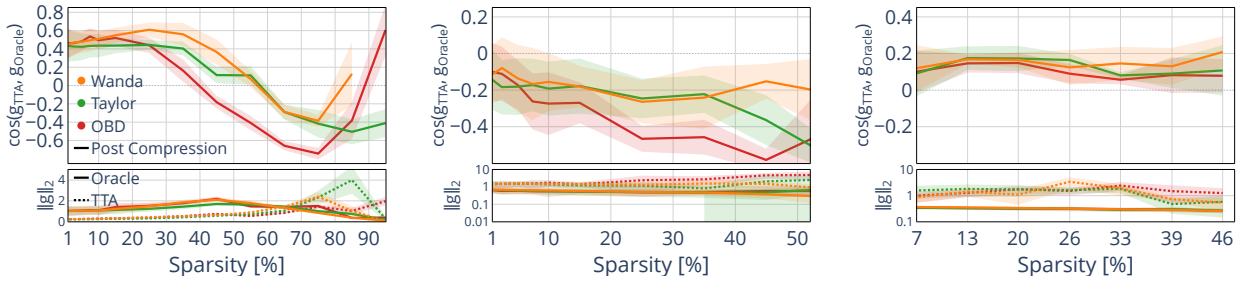
3.2 Subspace Compatibility

Structured compression reduces the set of trainable parameters available during adaptation and may also alter how well the remaining update directions align with the target-domain loss. We ask: *Does compression reduce the fraction of useful adaptation signal accessible within the admissible update subspace, and can this help explain the growing gap between unsupervised test-time adaptation and supervised target-domain adaptation?*

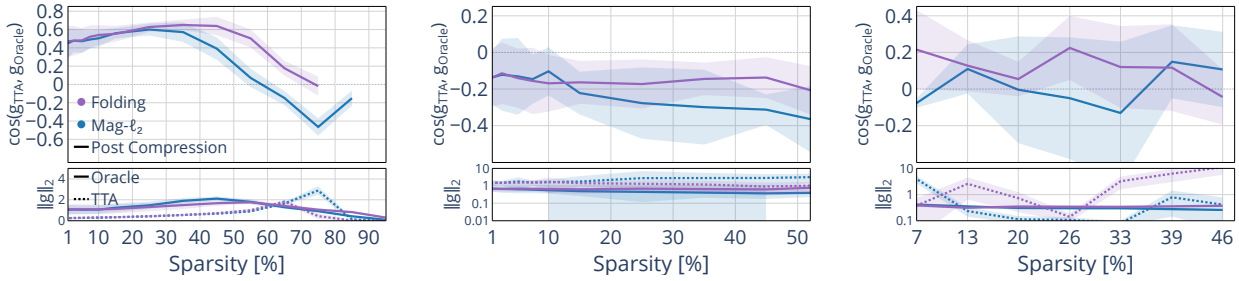
Gradient cosine similarity. Structured compression reduces the dimensionality of the admissible update $\Delta \in \mathcal{U}$ (Eq. 1) by eliminating adaptable parameters corresponding to pruned (Li et al., 2017) or folded (Wang et al., 2025) units. Since pruned deep networks suffer disproportionately under distribution shift (Liebenwein et al., 2021), adaptability may depend not only on the preserved $\dim(\mathcal{U})$ but also on whether the adaptation gradient directions preserve their alignment with the gradient of the supervised loss on the target domain. We quantify this alignment via the cosine similarity between the accumulated TTA and Oracle gradients, restricted to $\mathcal{U}$:

$$\cos(\mathbf{g}_{\mathrm{TTA}}, \mathbf{g}_{\mathrm{Oracle}}) = \frac{\mathbf{g}_{\mathrm{TTA}} \cdot \mathbf{g}_{\mathrm{Oracle}}}{\|\mathbf{g}_{\mathrm{TTA}}\| \cdot \|\mathbf{g}_{\mathrm{Oracle}}\|}, \tag{2}$$

where $\mathbf{g}_{\mathrm{TTA}} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\phi \mathcal{L}_{\mathrm{TTA}}(x_i; \phi_0)$ and $\mathbf{g}_{\mathrm{Oracle}} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\phi \mathcal{L}_{\mathrm{CE}}(x_i, y_i; \phi_0)$ are the adaptation set gradients averaged over all N corruption samples. We compute a single cosine similarity metric on these adaptation set gradients, rather than averaging per-sample (Chen et al., 2025) or per-task (Yu et al., 2020) similarities, to assess whether the TTA loss provides a meaningful proxy of the supervised optimization direction at the pre-adaptation parameters ϕ_0. The



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C



(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 4: **Structured compression increasingly misaligns the TTA gradient from the supervised gradient direction (Oracle).** We report cosine similarity $\cos(\mathbf{g}_{\text{TTA}}, \mathbf{g}_{\text{Oracle}})$ (Eq. 2, top panels) and gradient ℓ_2 norms (bottom panels), at the post-compression, pre-adaptation parameters ϕ_0 averaged across all corruptions. At low sparsity, the TTA and Oracle gradients are positively aligned, suggesting that the TTA objective initially provides a reliable surrogate for the supervised adaptation direction. As sparsity increases, this alignment deteriorates and can reverse sign, indicating that the TTA updates conflict with the supervised ones, a more severe failure mode than the loss of adaptation signal ($\cos \approx 0$). TTA gradient norms vanish (degeneracy) or remain large while misaligned (active divergence), both leading to adaptation collapse reflecting the TTA accuracy collapse observed in Fig. 1; lines terminate when the metric becomes undefined at high sparsities. The SPA objective on ViT-Base preserves weakly positive alignment across all tested sparsities, consistent with the smaller Oracle-TTA gap in Fig. 1. *Top row* (a–c): data-dependent pruning (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression (Folding, Mag-ℓ_2).

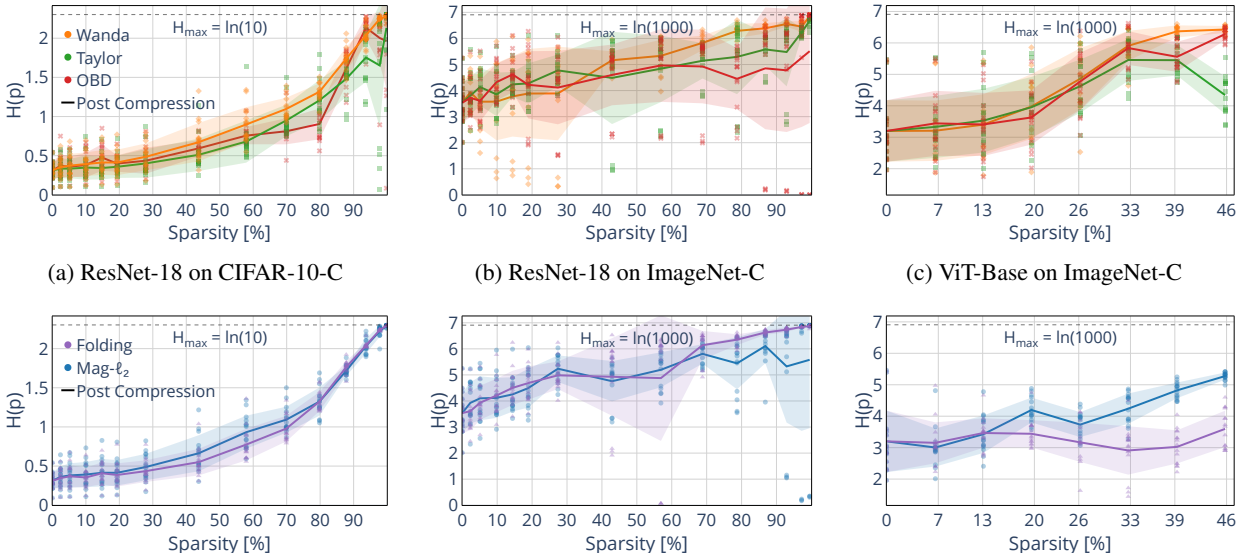
metric is undefined when either gradient norm vanishes, which implies an adaptation collapse regime, a condition we analyze theoretically below.

Empirically, Fig. 4 reveals three distinct regimes of gradient alignment under structured compression. On CIFAR-10-C (Fig. 4a, 4d), the TTA and Oracle gradients are positively aligned at low sparsity, indicating that the TTA updates initially point in a direction consistent with the supervised optimization objective. As sparsity increases, however, the cosine similarity crosses zero and becomes strongly negative, reaching its minimum value between 70% and 80% sparsity across all compression criteria. This sign reversal suggests that the TTA update actively drives the model *away* from the supervised optimum, constituting a more severe failure mode than the simple loss of an adaptation signal, which would correspond to $\cos \approx 0$. This phenomenon is compounded by the progressively widening discrepancy in gradient norms. The emergence of this misalignment coincides with the sparsity level at which TTA accuracy begins to deteriorate sharply (Fig. 1), offering a gradient lens explanation for the observed adaptation degradation. At extreme sparsities ($\geq 75\%$), the metric becomes undefined when the TTA gradient norm vanishes for all corruptions, *e.g.*, Mag-ℓ_2 beyond 85%. These results support our entropy gradient analysis (Eq. 4, 5): at equal sparsity, compressed models driven to near-uniform predictions produce weak adaptation signals, in contrast to the supervised signals which do not vanish.

On ImageNet-C with ResNet-18 (Fig. 4b, 4e), gradient alignment is negative already at 1%, suggesting that SAR's loss objective may be a poor surrogate of the supervised loss on this harder distribution shift, even before significant model capacity is removed by compression. The gradient norms in the bottom panel reveal a failure mode distinct from CIFAR-10-C: $\|\mathbf{g}_{\text{TTA}}\|$ systematically exceeds $\|\mathbf{g}_{\text{Oracle}}\|$ across all sparsities. This result, combined with the negative cosine similarity, indicates an *active divergence* regime: the SAR-based objective produces large updates in a direction that conflicts with the supervised one. This adaptation regime, associated with structured compression, may actively

damage the model consistent with the severe Oracle-TTA gap previously observed (Fig. 1). A causal ablation of this mechanism (Fig. 18) indicates that the confident-but-wrong samples admitted by SAR's reliability filter drive the misalignment: excluding them with a correctness filter computed from the ground-truth labels restores the alignment to $\approx +0.9$ across all compression criteria and architectures.

On ViT-Base with SPA (Fig. 4c, 4f), for data-dependent criteria, cosine similarity remains weakly positive across all sparsities, suggesting that the consistency-based SPA objective maintains partial alignment with the supervised direction even at 46% sparsity. No sign reversal occurs, consistent with the smaller Oracle-TTA accuracy gap observed in Fig. 1. Data-free compression exhibits more unstable adaptation behavior: Mag-ℓ_2 produces negative mean alignment at several sparsity levels, and both Folding and Mag-ℓ_2 exhibit high variance across corruptions. Gradient alignment remains fragile and corruption-dependent under data-free compression, where no calibration data guides unit selection. Decomposing these curves by corruption family (Figs. 13 –16) shows similar gradient alignment regimes within each family.



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 5: **Prediction entropy $H(\mathbf{p})$ increases with sparsity in most settings, empirically supporting the gradient degeneracy mechanism of Eq. 3 and Eq. 5.** Each dot represents the mean prediction entropy $H(\mathbf{p}) = -\sum_c p_c \ln p_c$ of a single corruption type at severity 5; solid lines and shaded bands show the mean and $\pm\sigma$ across all corruptions. The dashed line marks $H_{max} = \ln C$ (uniform distribution). **Top row** (a–c): data-dependent pruning (Wanda, Taylor, OBD). **Bottom row** (d–f): data-free compression (Mag-ℓ_2, Folding).

Objective-Induced Gradient Degeneracy. Even if the adaptation subspace is aligned with the target-domain loss, the TTA objective may fail to produce informative gradients under structured compression. We show that this degeneracy arises in both entropy- and consistency-based TTA objectives.

Entropy minimization. TTA methods such as TENT (Wang et al., 2021) and SAR (Niu et al., 2023) optimize:

$$\mathcal{L}_{\text{ent}}(p) = -\sum_{c=1}^{C} p_c \log p_c, \tag{3}$$

where $p = (p_1, \ldots, p_C)$ is the predictive distribution. Using $\sum_c p_c = 1$ and thus $\sum_c \nabla_\phi p_c = 0$, differentiating gives:

$$\nabla_\phi \mathcal{L}_{\text{ent}} = -\sum_{c=1}^{C} (\log p_c + 1) \nabla_\phi p_c = -\sum_{c=1}^{C} \log(C p_c)\, \nabla_\phi p_c. \tag{4}$$

When $p = (1/C, \ldots, 1/C)$, all coefficients $\log(C p_c)$ vanish and $\nabla_\phi \mathcal{L}_{\text{ent}} = \mathbf{0}$. More generally:

$$\|\nabla_\phi \mathcal{L}_{\text{ent}}\|_2 \leq \left(\sum_{c=1}^{C} |\log(C p_c)| \right) \max_c \|\nabla_\phi p_c\|_2. \tag{5}$$

If compression pushes predictions toward a near-uniform regime, the entropy signal weakens (*gradient degeneracy*), as observed on CIFAR-10-C (Fig. 4a, 4d). Conversely, when compression preserves confident but incorrect predictions, $|\log(Cp_c)|$ remains large, producing misaligned high-norm gradients (*active divergence*), as observed on ImageNet-C (Fig. 4b, 4e). Prediction entropy measurements indicate that $H(\mathbf{p})$ increasingly approaches H_{max} with higher sparsity in most settings (Fig. 5), further supporting the gradient degeneracy mechanism under compression.

KL-consistency. SPA (Niu et al., 2025) minimizes the KL divergence between a detached clean prediction $q = \mathrm{softmax}(f(x;\phi))$ and an augmented-view prediction $p_{\mathrm{aug}} = \mathrm{softmax}(g(x_{\mathrm{aug}};\phi))$:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{c=1}^{C} q_c \log \frac{q_c}{p_{\mathrm{aug},c}}. \tag{6}$$

Since q is detached, $\nabla_\phi \mathcal{L}_{\mathrm{KL}} = -\sum_c (q_c/p_{\mathrm{aug},c})\, \nabla_\phi p_{\mathrm{aug},c}$. When $q \approx p_{\mathrm{aug}}$, each ratio $q_c/p_{\mathrm{aug},c} \approx 1$, and together with $\sum_c \nabla_\phi p_{\mathrm{aug},c} = \mathbf{0}$ this implies that the gradient vanishes. Compression triggers this by driving both views toward uniform predictions.

Supervised cross-entropy (Oracle). By contrast, the supervised gradient for a labeled example (x, y):

$$\nabla_\phi \mathcal{L}_{\mathrm{CE}} = -\frac{1}{p_y} \nabla_\phi p_y, \tag{7}$$

concentrates on a single direction scaled by $1/p_y$, which grows as confidence drops. Even when compression degrades predictions, supervised updates thus remain informative where entropy or consistency signals vanish. Together, Eqs. 4–7 reveal two complementary mechanisms behind the post-compression Oracle-TTA gap: (i) compression removes or misaligns parameter directions useful for target-domain adaptation, and (ii) unsupervised TTA objectives simultaneously lose their gradient signal as predictions approach an uninformative regime.

4 Discussion, Outlook and Limitations

This paper analyzes the interaction between structured compression and test-time adaptation, a combination that is common in edge deployment but has received limited attention. Our study reveals that the standard compress-and-deploy pipeline can induce *silent plasticity loss*: compressed models that preserve source-domain accuracy yet progressively lose the ability to adapt at test time via unsupervised objectives. This finding extends prior observations that compression harms robustness under distribution shift (Liebenwein et al., 2021) and disproportionately affects underrepresented subgroups (Hooker et al., 2020), by showing that degradation extends to adaptability. Our theoretical analysis establishes that the gradients of both entropy-based and consistency-based TTA objectives can degenerate under compression, while the supervised cross-entropy gradient remains more informative at the same sparsity (Eq. 4, Eq. 7).

Practical guidelines. Our results suggest two actionable considerations. First, the compression criterion should be selected for adaptability, not source accuracy alone; data-dependent criteria (Wanda, Taylor) preserve higher representational recoverability (Fig. 2) and gradient alignment (Fig. 4) than Mag-ℓ_2 and OBD, while Folding retains more adaptability among data-free methods. Second, the compression budget and TTA objective should be co-designed: on CIFAR-10-C, gradient alignment reverses sign between 45% and 65% sparsity for the pruning criteria (near 75% for Folding); on ImageNet-C, entropy-based objectives report active divergence already at minimal sparsity, whereas consistency-based objectives maintain higher alignment across all analyzed sparsities.

Outlook. There are several directions to extend this work. First, although our empirical evaluation covers both backpropagation-based and backpropagation-free TTA approaches (FOA (Niu et al., 2024), PEA (Xiao et al., 2026b)), our diagnostics are mainly gradient-based and do not yet cover backpropagation-free adaptation methods; we leave this as an open direction. The corresponding accuracy curves are reported in Fig. 9: the benefit of backpropagation-free adaptation likewise collapses toward the no-adaptation baseline as compression increases. Second, designing compression-aware TTA objectives that preserve adaptability under compression is left for future work. Third, extending this analysis to unstructured pruning, quantization, and their combinations would provide a more complete picture of the compression-adaptation interaction.

Limitations. Our study considers two architectures (ResNet-18 and ViT-Base) and two (one backpropagation-based, one backpropagation-free) TTA methods per architecture, covering both entropy- and consistency-based objectives; the generalization to larger models (*e.g.*, large language models) remains for future work. While the experiments are primarily conducted at maximum corruption severity (*i.e.*, 5), we additionally report results at intermediate severity 3 (Fig. 11) and across multiple seeds (Fig. 8). We hope this work sparks interest in designing structured compression strategies that explicitly preserve adaptability, narrowing the gap between efficient deployment and test-time adaptation.

ACKNOWLEDGMENTS

This research was funded in part by the Austrian Science Fund (FWF) within the DENISE doctoral school (grant DOI: 10.55776/DFH5) and the FFG COMET K1 Center "Pro^{2}Future II" (Cognitive and Sustainable Products and Production Systems of the Future), Contract No. 911655. The results presented in this paper were computed using the computational resources of the HLR Zentralen Informatikdienstes of Graz University of Technology and the Austrian Scientific Computing (ASC) infrastructure.

REFERENCES

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

N. Alex Cayco-Gajic and Arthur Pellegrino. Geometry-aware similarity metrics for neural representations on riemannian and statistical manifolds. *arXiv preprint arXiv:2603.28764*, 2026.

Ziyang Chen, Yiwen Ye, Yongsheng Pan, and Yong Xia. Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2429–2437, 2025.

Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration: T. choudhary et al. *Artificial Intelligence Review*, 53(7):5113–5155, 2020.

Francesco Corti, Rahim Entezari, Sara Hooker, Davide Bacciu, and Olga Saukh. Studying the impact of magnitude pruning on contrastive learning methods. *arXiv preprint arXiv:2207.00200*, 2022.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

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.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

Jiaheng Dong, Hong Jia, Soumyajit Chatterjee, Abhirup Ghosh, James Bailey, and Ting Dang. E-bats: Efficient backpropagation-free test-time adaptation for speech foundation models. *Advances in neural information processing systems*, 38:141099–141126, 2026.

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.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16091–16101. IEEE, 2023.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International conference on machine learning*, pp. 3259–3269. PMLR, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. *Advances in neural information processing systems*, 30, 2017.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36:14070–14093, 2023.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Yuxuan Hu, Jing Zhang, Zhe Zhao, Chen Zhao, Xiaodong Chen, Cuiping Li, and Hong Chen. Sp3: Enhancing structured pruning via pca projection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3150–3170, 2024.

Hong Jia, Young D Kwon, Alessio Orsino, Ting Dang, Domenico Talia, and Cecilia Mascolo. Tinytta: Efficient test-time adaptation via early-exit ensembles on edge devices. *Advances in Neural Information Processing Systems*, 37:43274–43299, 2024.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gU5sJ6ZggcX.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJqFGTslg.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

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

Zhu Liao, Victor Quétu, Van-Tam Nguyen, and Enzo Tartaglione. Nepenthe: Entropy-based pruning as a neural network depth's reducer. *arXiv e-prints*, pp. arXiv–2404, 2024.

Lucas Liebenwein, Cenk Baykal, Brandon Carter, David Gifford, and Daniela Rus. Lost in pruning: The effects of pruning neural networks beyond test accuracy. *Proceedings of Machine Learning and Systems*, 3:93–138, 2021.

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256kb memory. *Advances in Neural Information Processing Systems*, 35:22941–22954, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International conference on machine learning*, pp. 23190–23211. PMLR, 2023.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14745–14755. IEEE, 2022.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11256–11264. IEEE, 2019.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=qz1Vx1v9iK.

Shuaicheng Niu, Guohao Chen, Peilin Zhao, Tianyi Wang, Pengcheng Wu, and Zhiqi Shen. Self-bootstrapping for versatile test-time adaptation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Li4rieeClO.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Olga Saukh, Dong Wang, Haris Šikić, Yun Cheng, and Lothar Thiele. Cut less, fold more: Model compression through the lens of projection geometry. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JV9CEtKLQF.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Georg Slamanig, Francesco Corti, and Olga Saukh. From llms to edge: Parameter-efficient fine-tuning on edge devices, 2025. URL https://arxiv.org/abs/2507.23536.

Sanghyun Son, Seungjun Nah, and Kyoung Mu Lee. Clustering convolutional kernels to compress deep neural networks. In *European Conference on Computer Vision*, pp. 225–240. Springer, 2018.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

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.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pp. 1–5. Ieee, 2015.

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. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Dong Wang, Haris Šikić, Lothar Thiele, and Olga Saukh. Forget the data and fine-tuning! just fold the network to compress. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=W2Wkp9MQsF.

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.

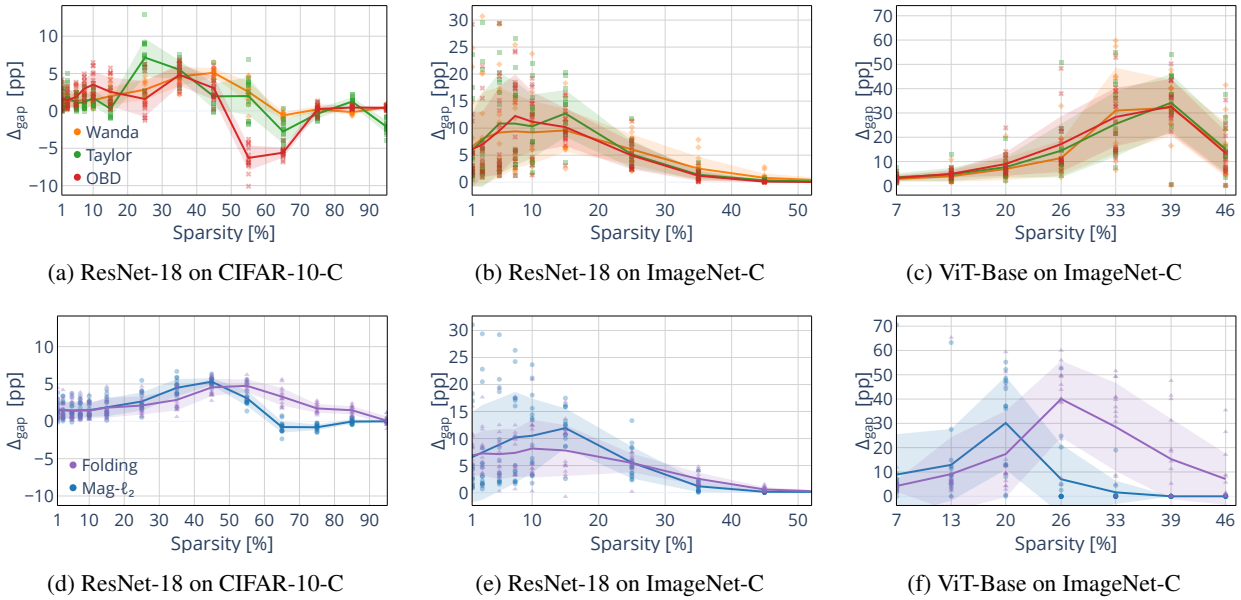
Junrui Xiao, Zhikai Li, Lianwei Yang, Yiduo Mei, and Qingyi Gu. Ttaq: Towards stable post-training quantization in continuous domain adaptation. *arXiv preprint arXiv:2412.09899*, 2024.

MA Xiao, Young D. Kwon, and Dong Ma. Efficient Test-Time Adaptation via Decoupled BN Update For Edge Devices. In *Third Workshop on Test-Time Updates (Main Track)*, 2026a. URL https://openreview.net/forum?id=35oSXGUjDH.

MA Xiao, Young D. Kwon, Pan Zhou, and Dong Ma. Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=7kLNGaAHaw.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

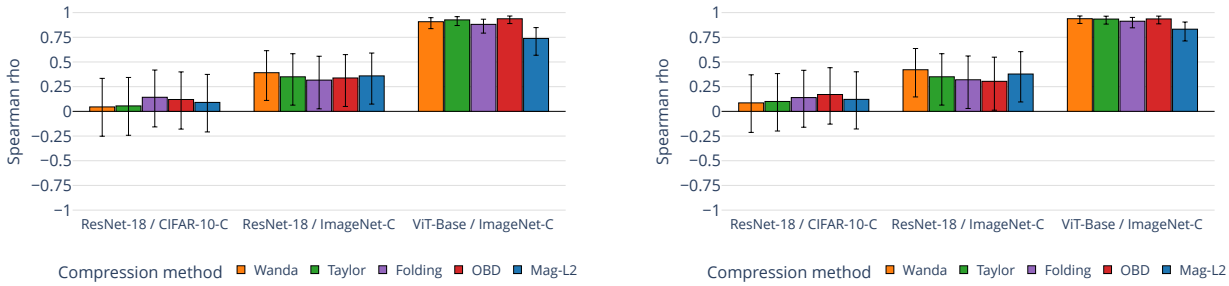
A APPENDIX



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 6: **The per-corruption adaptation gap $\Delta_{\text{gap}} = (\textbf{Oracle} - \textbf{TTA})_{\textbf{compressed}}$ generally widens with increasing compression, providing empirical support for *silent plasticity loss*.** Each dot represents the adaptation gap for a single corruption type at a given sparsity level; solid lines and shaded bands show the mean and $\pm\sigma$ across all corruptions. *Top row* (a–c): data-dependent pruning criteria (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression criteria (Folding, Mag-ℓ_2).

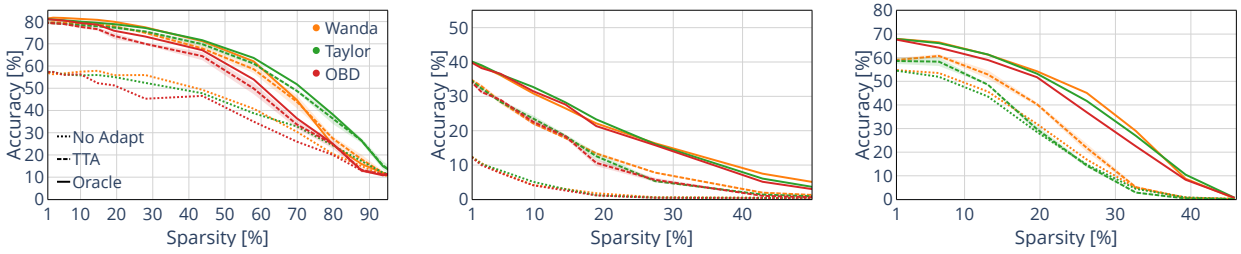


(a) Worst-layer CKA (severity 3). (b) Worst-layer CKA (severity 5).

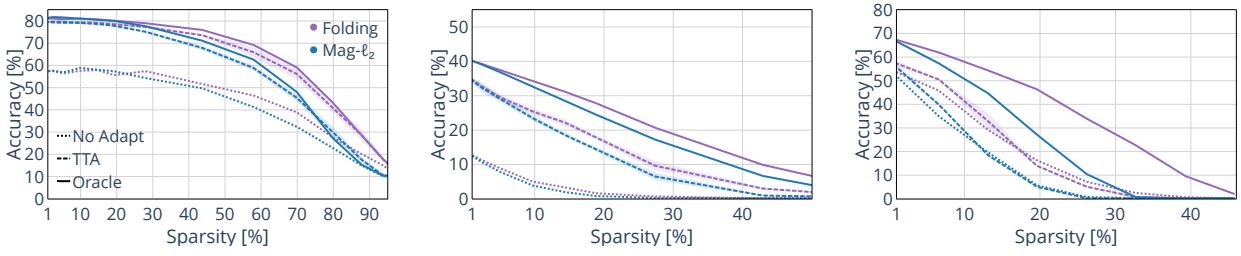
Figure 7: Before adaptation, the worst-layer CKA of the compressed model is positively correlated with its post-adaptation accuracy, strongly on ViT-Base/ImageNet-C and only weakly on ResNet-18. Each bar reports the Spearman correlation ρ between the pre-TTA worst-layer CKA at ϕ_0 and the post-adaptation accuracy of SAR ResNet-18 or SPA ViT-Base over all corruptions. Black error bars are 95% confidence intervals. *Left to right*: severity 3, severity 5 corruptions level.

Fig. 6 decomposes the averaged Oracle–TTA adaptation gap observed in Fig. 1 into individual corruptions. In Fig. 20 and Fig. 21, marker shape encodes sparsity, color encodes post-adaptation accuracy; thick black edge denotes Oracle, gray edge denotes TTA. Points above the identity line indicate representational recovery after adaptation in Fig. 20, in Fig. 21 points *below* the identity line indicate entropy-distance adaptation recovery. Across all figures, *Top row* (a–c): data-dependent pruning criteria (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression criteria (Folding, Mag-ℓ_2).

Multi-seed evaluation. We train three checkpoints per (architecture, dataset, pre-training paradigm) with different seeds. Fig. 11 reports the multi-seed analysis of CKA for severity 3 corruptions, supporting our previous severity 5

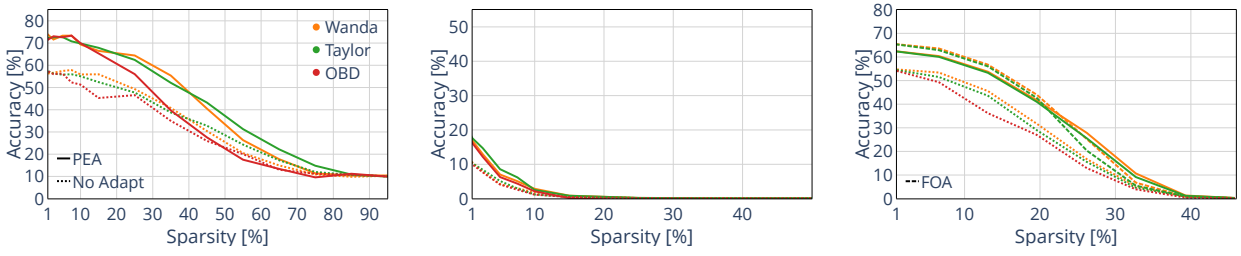


(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

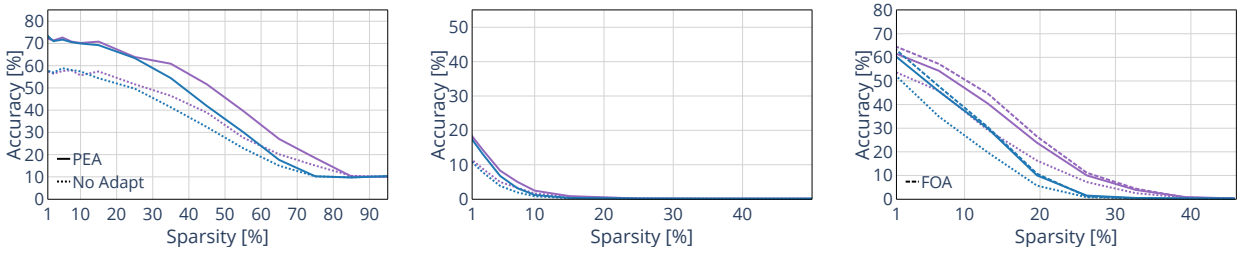


(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 8: **Multi-seed gap between test-time adaptation (SAR) and the matched supervised baseline (Oracle-SAR), with SAR applied to both architectures.** SAR and Oracle-SAR share the same optimization settings and differ only in the loss definitions. Shaded regions indicate $\pm\sigma$ across three random seeds of the per-seed corruption-averaged accuracy. Same layout as Fig. 1.



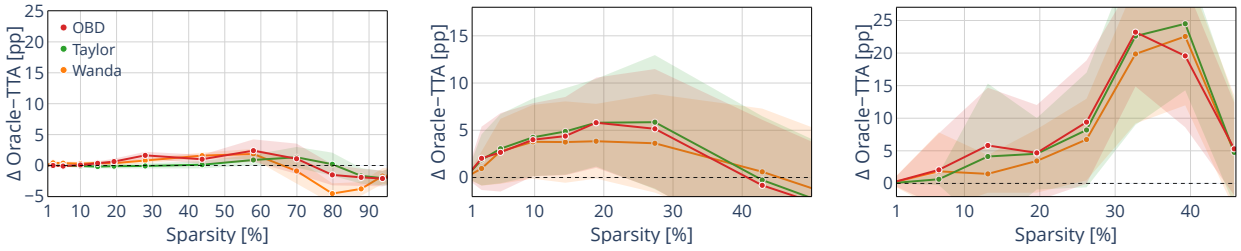
(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C



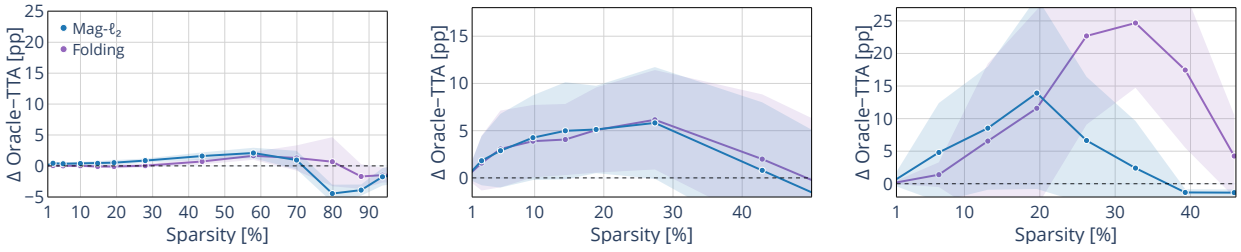
(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 9: **Adaptation accuracy curves under backpropagation-free TTA (PEA and FOA) methods.** PEA (Xiao et al., 2026b) and FOA (Niu et al., 2024) post-adaptation accuracy averaged across all corruptions with severity 5. PEA is applied to both ResNet-18 and ViT-Base; FOA is applied to ViT-Base only since it relies on learnable prompt embeddings, not available for the ResNet-18 models analyzed in this work. Same layout as Fig. 1: Top row (a–c): data-dependent pruning (Wanda, Taylor, OBD). *Bottom row* (d–f): data-free compression (Folding, Mag-ℓ_2).

findings. Shaded bands of Fig. 8 are the across-seed standard deviation of the per-seed corruption-averaged accuracy. These checkpoints follow a different pre-training recipe than the single checkpoint of Fig. 1, shifting the *absolute* post-adaptation accuracy of Fig. 8 relative to Fig. 1. Nevertheless, the compression-induced Oracle-TTA gap and

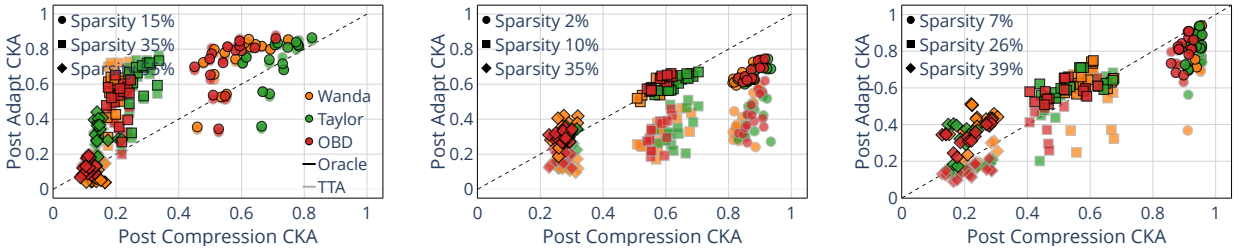


(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

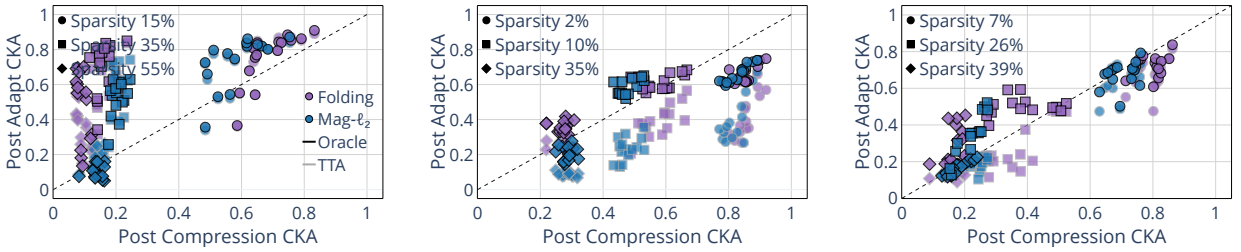


(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 10: **Compression-induced Oracle–TTA gap** $\Delta G(r) = [\text{Oracle} - \text{TTA}](r) - [\text{Oracle} - \text{TTA}](r = 0.0)$. Solid lines are corruption-averaged means with $\pm\sigma$ bands across all corruptions; a positive value indicates that compression amplifies the Oracle–TTA gap beyond the dense baseline. Same layout as Fig. 1.



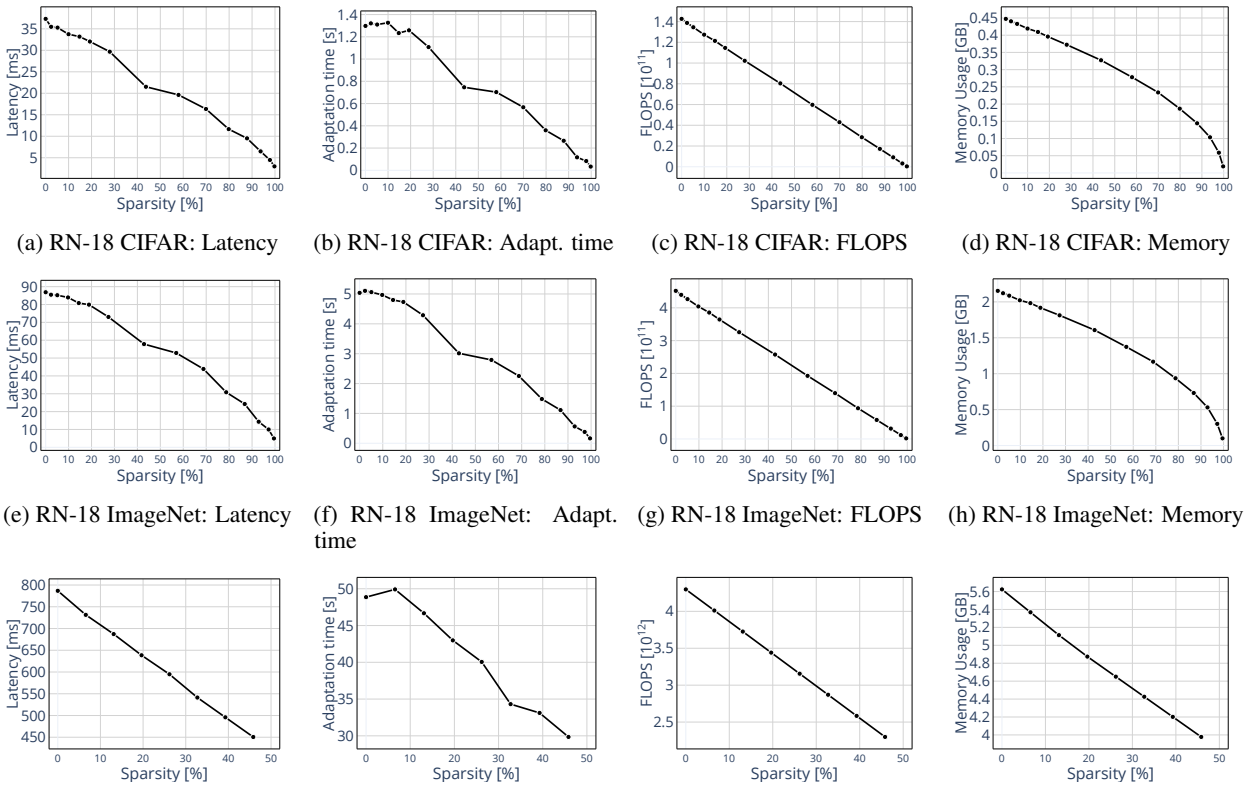
(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C



(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 11: **Multi-seed worst-layer CKA, *dense-prune-TTA*, severity 3.** Same layout as Fig. 2.

its widening trend persist under both recipes, indicating that the effect reflects structured compression rather than a particular pre-training run.

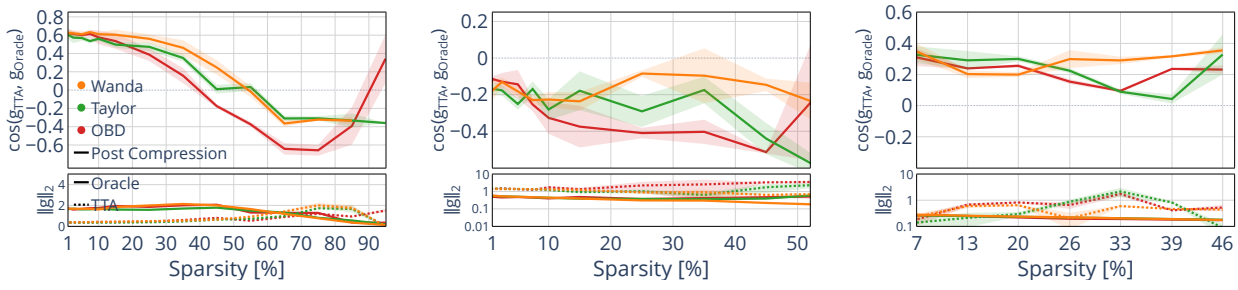


(a) RN-18 CIFAR: Latency (b) RN-18 CIFAR: Adapt. time (c) RN-18 CIFAR: FLOPS (d) RN-18 CIFAR: Memory

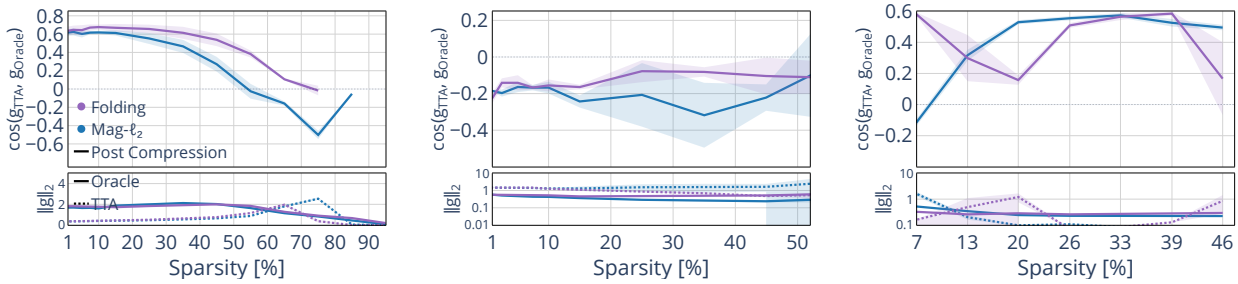
(e) RN-18 ImageNet: Latency (f) RN-18 ImageNet: Adapt. time (g) RN-18 ImageNet: FLOPS (h) RN-18 ImageNet: Memory

(i) ViT-Base: Latency (j) ViT-Base: Adapt. time (k) ViT-Base: FLOPS (l) ViT-Base: Memory

Figure 12: **Compute, memory and latency profile across compression ratios.** For each architecture we report, as a function of sparsity, the CPU inference latency (ms), the full adaptation-step time (s). The analytical FLOPs and memory usage (GB) per adaptation step are also reported. All measurements use batch size B=64 on three channel inputs at the pretraining data resolution. The measurements are obtained on a single AMD EPYC 9654 96-core processor single-threaded. We use the evaluation protocol described in Slamanig et al. (2025). Under the uniform per-layer compression, for each compression ratio all the five structured compression methods studied in the paper yield architecturally equal smaller-dense ResNet-18 and ViT-Base networks, thus a single curve per architecture suffices; we report Magnitude-ℓ_2 as a representative structured compression method.

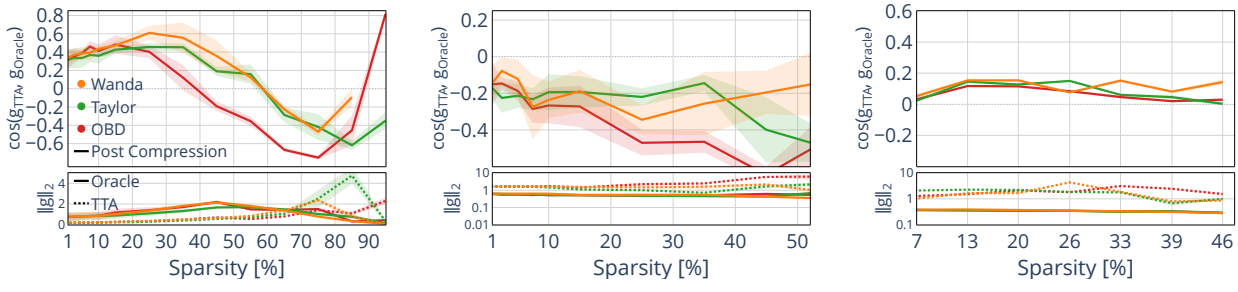


(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C. (c) ViT-Base / ImageNet-C.

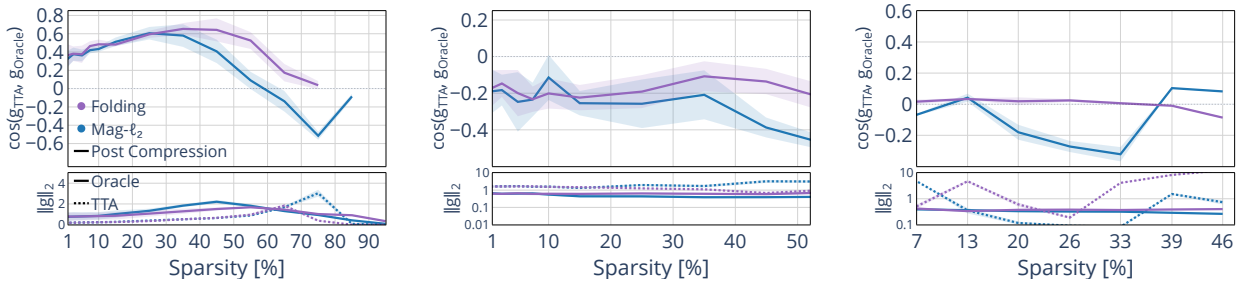


(d) ResNet-18 / CIFAR-10-C. (e) ResNet-18 / ImageNet-C. (f) ViT-Base / ImageNet-C.

Figure 13: **Gradient alignment restricted to the *Noise* corruption family** (Gaussian, Shot, Impulse). Same visualization scheme as Fig. 4; $\pm\sigma$ band across the three Noise corruptions.

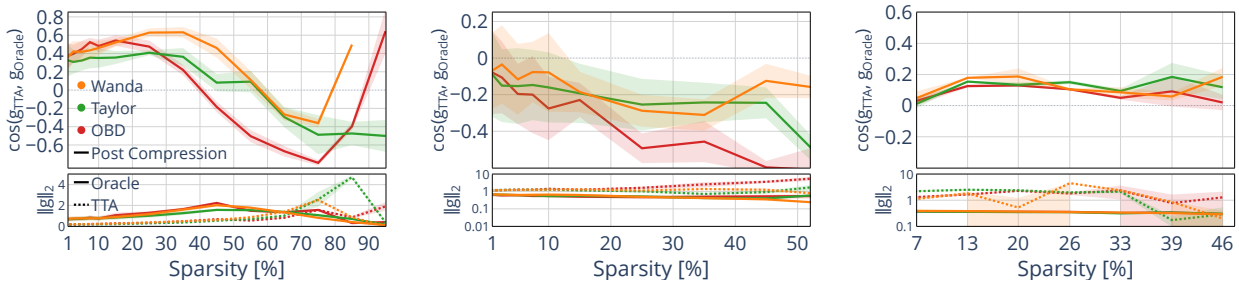


(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C. (c) ViT-Base / ImageNet-C.

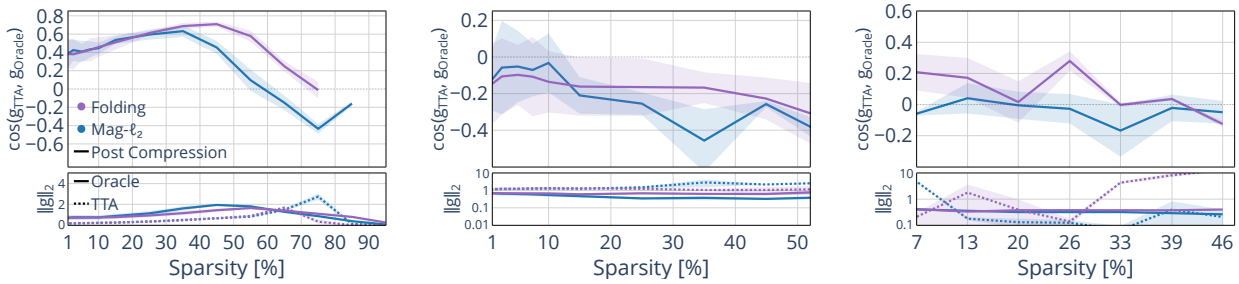


(d) ResNet-18 / CIFAR-10-C. (e) ResNet-18 / ImageNet-C. (f) ViT-Base / ImageNet-C.

Figure 14: **Gradient alignment restricted to the *Blur* corruption family** (Defocus, Glass, Motion, Zoom). Same visualization scheme as Fig. 4; $\pm\sigma$ band across the four Blur corruptions.

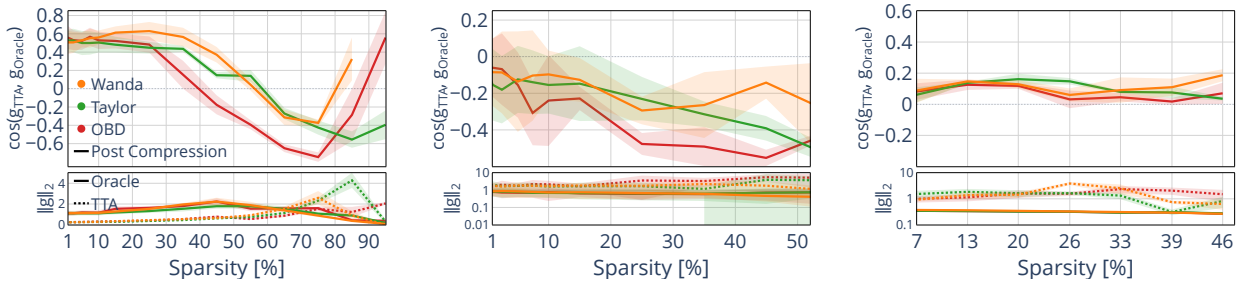


(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C. (c) ViT-Base / ImageNet-C.

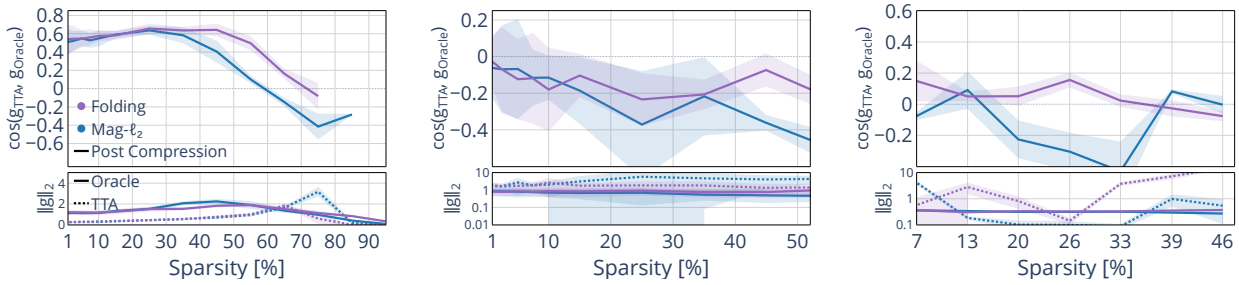


(d) ResNet-18 / CIFAR-10-C. (e) ResNet-18 / ImageNet-C. (f) ViT-Base / ImageNet-C.

Figure 15: **Gradient alignment restricted to the *Weather* corruption family** (Snow, Frost, Fog, Brightness). Same visualization scheme as Fig. 4; $\pm\sigma$ band across the four Weather corruptions.

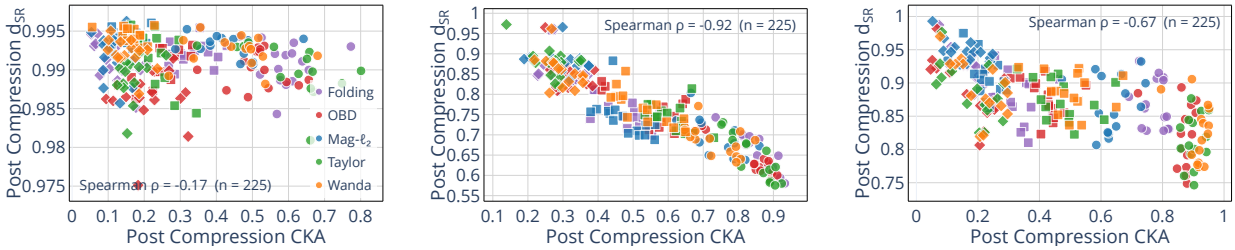


(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C. (c) ViT-Base / ImageNet-C.



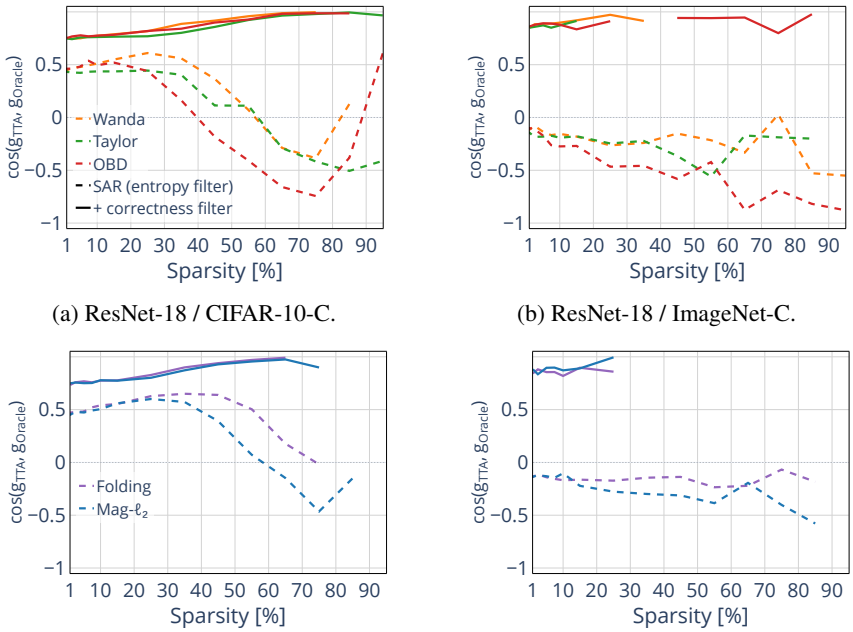
(d) ResNet-18 / CIFAR-10-C. (e) ResNet-18 / ImageNet-C. (f) ViT-Base / ImageNet-C.

Figure 16: **Gradient alignment restricted to the *Digital* corruption family** (Contrast, Elastic, Pixelate, JPEG). Same visualization scheme as Fig. 4; $\pm\sigma$ band across the four Digital corruptions.



(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C. (c) ViT-Base / ImageNet-C.

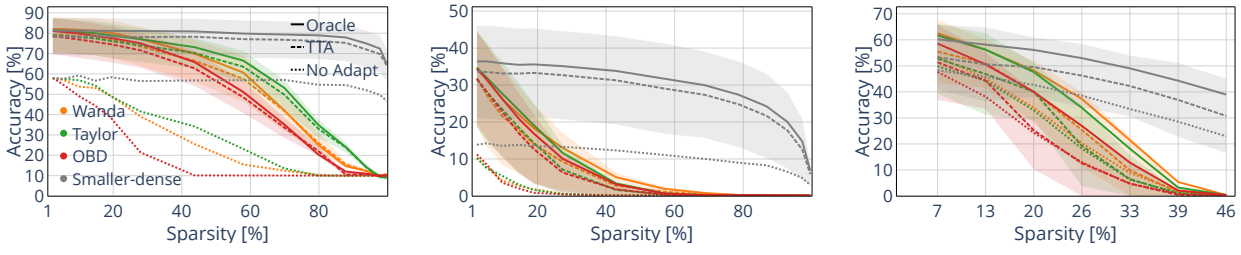
Figure 17: **Geometry-aware similarity agrees with worst-layer CKA representation metric on ImageNet-C.** Each point is one (criterion, sparsity, corruption) cell at severity 5 post-compression pre-adaptation (n equals to the number of total samples considered per panel; colors denote the compression criterion). The x-axis is the worst-layer CKA between the dense and the compressed model, the y-axis is the spectral-ratio pseudo-distance d_{SR} (Cayco-Gajic & Pellegrino, 2026) computed on the same batches; since d_{SR} is a distance, agreement corresponds to negative rank correlation.



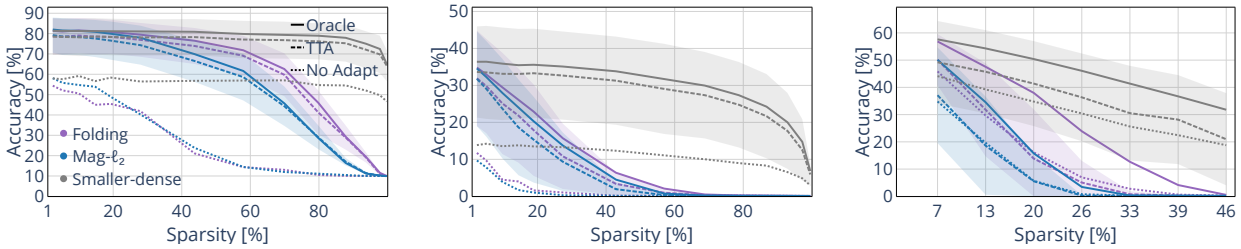
(a) ResNet-18 / CIFAR-10-C. (b) ResNet-18 / ImageNet-C.

(c) ResNet-18 / CIFAR-10-C. (d) ResNet-18 / ImageNet-C.

Figure 18: **Removing the confident-but-wrong samples that pass SAR's reliability filters recovers gradient alignment.** We report $\cos(\mathbf{g}_{\mathrm{TTA}}, \mathbf{g}_{\mathrm{Oracle}})$ at ϕ_0 (Eq. 2), averaged across all corruptions, with colors as in Fig. 4. Dashed lines reproduce the cosine-similarity curves of Fig. 4 unchanged; solid lines restrict the reliable set to correctly classified samples and evaluate the Oracle-SAR gradient over the same set. Excluding the confident-wrong samples restores the alignment to $\approx +0.9$ in every considered scenario.

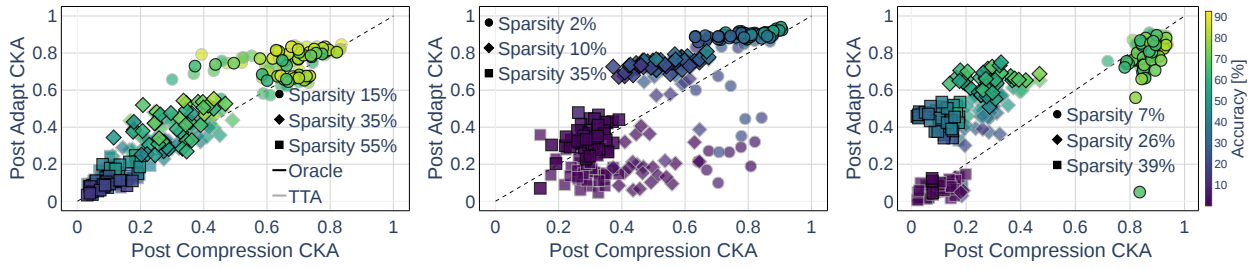


(a) ResNet-18 / CIFAR-10-C (b) ResNet-18 / ImageNet-C (c) ViT-Base / ImageNet-C

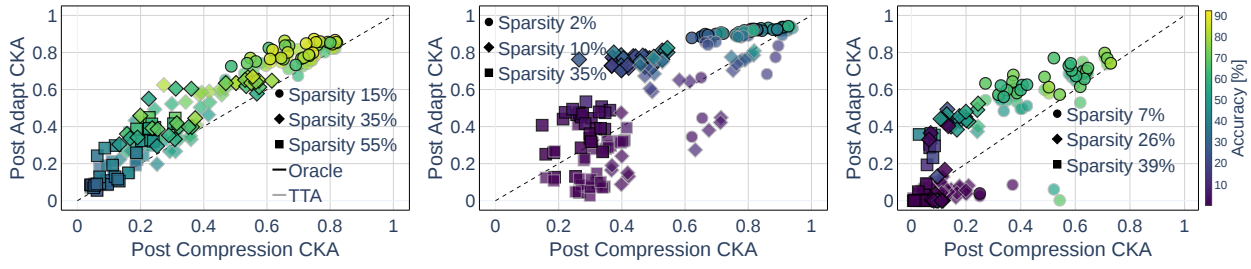


(d) ResNet-18 / CIFAR-10-C (e) ResNet-18 / ImageNet-C (f) ViT-Base / ImageNet-C

Figure 19: **Test-time adaptation collapses with increasing compression for the pruned models, unlike the smaller-dense models of equal parameter budget trained from scratch.** Same layout as Fig. 1 using SAR and Oracle-SAR on all the displayed architectures. The gray curves show the method-agnostic *smaller-dense* TTA reference. At low sparsity, TTA on the compressed models remains competitive with the references; as compression increases, TTA and its Oracle both collapse on every criterion, while the smaller-dense models retain adaptation gain. Together, these observations suggest that the loss of adaptability is induced by the structured compression rather than by the reduced model size.

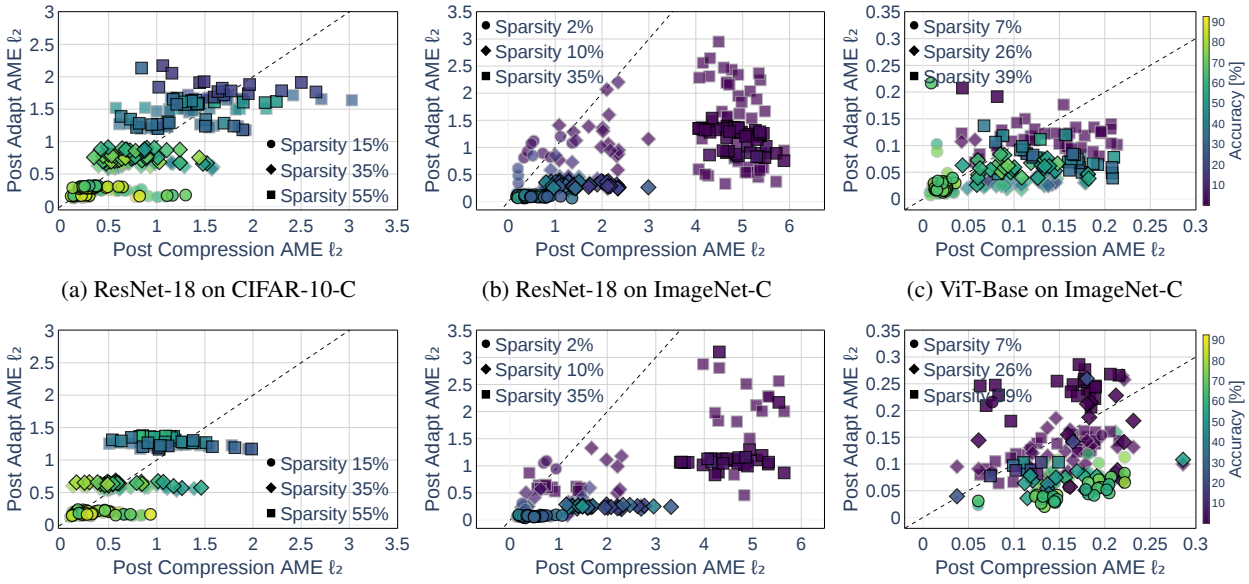


(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C



(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 20: **Test-time adaptation (TTA) degrades representations more than supervised fine-tuning (Oracle), with the gap depending on the compression method.**



(a) ResNet-18 on CIFAR-10-C (b) ResNet-18 on ImageNet-C (c) ViT-Base on ImageNet-C

(d) ResNet-18 on CIFAR-10-C (e) ResNet-18 on ImageNet-C (f) ViT-Base on ImageNet-C

Figure 21: **Test-time adaptation (TTA) only partially restores output expressivity, with recovery strongly dependent on the compression method.**