Neutralizing Token Aggregation via Information Augmentation for Efficient Test-Time Adaptation

Yizhe Xiong[1†], Zihan Zhou[1†], Yiwen Liang[1†], Hui Chen[2⋆], Zijia Lin[1], Xinhao Xu[1], Tianxiang Hao[1], Fan Zhang[1], Jungong Han[2,3], and Guiguang Ding[1,2⋆]

[1] School of Software, Tsinghua University, China
[2] BNRist, Tsinghua University, China
[3] Department of Automation, Tsinghua University, Beijing, China
xiongyizhe2001@163.com

Abstract. Test-Time Adaptation (TTA) has emerged as an effective solution for adapting Vision Transformers (ViT) to distribution shifts without additional training data. However, existing TTA methods often incur substantial computational overhead, limiting their applicability in resource-constrained real-world scenarios. To reduce inference cost, plug-and-play token aggregation methods merge redundant tokens in ViTs to lower computation demands. Albeit efficient, it suffers from significant performance degradation when directly integrated with existing TTA methods. We formalize this problem as Efficient Test-Time Adaptation (ETTA), seeking to preserve the adaptation capability of TTA while reducing inference latency. In this paper, we start by providing an analysis showing that token aggregation inherently leads to information loss, which cannot be fully mitigated by conventional norm-tuning-based TTA methods. Guided by this insight, we propose to **N**eutralize Token **A**ggregation **v**ia **I**nformation **A**ugmentation (**NAVIA**). Specifically, we propose an information augmentation mechanism that employs an input-level [CLS] embedding augmentation to compensate for domain-level information loss, and a feature-level [CLS] bias augmentation for fine-grained layerwise information compensation. Theoretical analysis provides a principled motivation that augmenting a global information carrier before aggregation can improve the learnable-information upper bound, with entropy minimization serving as a practical surrogate objective. Extensive experiments across various out-of-distribution benchmarks and model backbones demonstrate that NAVIA outperforms all competing methods. Notably, NAVIA improves performance by up to 3.1% while achieving 14% to 26% wall-clock speedup over the TTA state-of-the-art, effectively addressing the ETTA challenge. We provide our official implementation at https://github.com/Bostoncake/NAVIA.

Keywords: Test-Time Adaptation · Token Aggregation

† Equal contribution.
⋆ Corresponding author.

1 Introduction

Vision transformers (ViT) [10, 47] have significantly advanced the field of computer vision, bringing substantial advancements across various tasks [5, 22]. Typically, ViTs are evaluated on data drawn from the same distribution as the training dataset. However, real-world applications [63, 65, 66] often present data exhibiting a *distribution shift* [46], which can differ markedly from the original training distribution. To address those practical challenges, Unsupervised Domain Adaptation (UDA) [32,34,44,58,61] has been proposed to learn domain-invariant image representations from unlabeled data drawn from the target distribution. Nonetheless, UDA methods typically require extra training, leading to inefficiency and limited deployment flexibility. Recently, Test-Time Adaptation (TTA) [37–39, 43, 52] has emerged as an alternative, adapting models on out-of-distribution (OOD) test data during inference via parameter-efficient fine-tuning [25, 39, 43, 52] or gradient-free update strategies [3, 11, 20, 26, 37, 64, 70]. Compared to traditional UDA, TTA eliminates the entire training stage, rendering it particularly suitable for real-world scenarios.

Although TTA achieves promising OOD performance, its practicality is often constrained by the inference cost of the underlying backbone, where the *forward pass* dominates the latency [3]. In modern TTA pipelines, strong backbones such as ViTs are widely adopted for their robustness under distribution shifts, which already incur substantial per-batch compute. Worse still, many methods further even amplify this cost by performing extra forward cycles [37, 39] or evaluating multiple augmentations on the same test batch [11,23,43]. For example, SAR [39] computes second-order gradients across multiple inference cycles and DeYO [23] forwards the same sample using varying patch sequences. Consequently, TTA effectiveness is increasingly bounded by inference efficiency, motivating approaches that *directly reduce backbone forward cost* while preserving adaptation quality.

To address those limitations, we focus on the challenge of **Efficient Test-Time Adaptation** (**ETTA**), aiming to improve the *inference efficiency* of TTA methods during inference. We focus on Vision Transformers (ViT) [10] as ViTs demonstrate better generalization under distribution shifts, and are widely adopted in TTA [11, 37]. Prior studies have extensively explored model compression methods, including structural compression [14, 35, 50, 59, 67] and token aggregation [2, 21, 27], to enhance inference efficiency in ViTs. While model compression methods typically require extensive retraining with substantial data and computational resources to recover the performance lost during compression, token aggregation methods can be adopted for TTA in a plug-and-play manner as they do not alter ViT parameters and do not interfere with the optimization process within TTA. Naturally, a straightforward solution for ETTA would be a combination of representative TTA approaches, namely Tent [52], FOA [37], SAR [39], DeYO [23], and CoTTA [54] with a widely-adopted parameter-free token aggregation method, ToMe [2][4]. As shown in Fig. 1, even at a modest

[4] In Section 4, we demonstrate that ToMe serves as a neat and strong choice for token aggregation methods.

compression rate (12.5% sparsity, reducing r=2 tokens in each layer), existing TTA methods suffer significant performance loss of over 3% on average. Further increasing the compression rate (i.e., r=4) exacerbates this issue, highlighting that a direct combination of existing methods cannot effectively address the efficiency-performance trade-offs central to ETTA.

Building on those observations, we take a closer look at why token aggregation breaks down in TTA. Through an analysis in Sec. 3.2, we analyze information propagation within the model and show that performance drops can be traced to mutual information loss introduced by token aggregation, which cannot be adequately compensated by norm-tuning strategies commonly used in TTA. To effectively mitigate these issues, we propose to **N**eutralize Token **A**ggregation **V**ia **I**nformation **A**ugmentation (**NAVIA**). Specifically, at the input-level, adaptation jointly refines [CLS] representations together with norm parameters to inject domain-level information and increase the upper bound for learnable information under token aggregation. Additionally, at the feature-level, a set of learnable bias vectors is fine-tuned to enable a more granular compensation of feature distortions with negligible overhead in forward computation. Extensive experiments (see Sec. 4 and Fig. 1) across 4 OOD benchmarks with both ViT-B/16 and ViT-L/16 show that NAVIA consistently outperforms all competing baselines. Notably, NAVIA improves performance by up to 3.1% while achieving a 14% to 26% wall-clock speedup over the TTA state-of-the-art, further highlighting its effectiveness and efficiency.

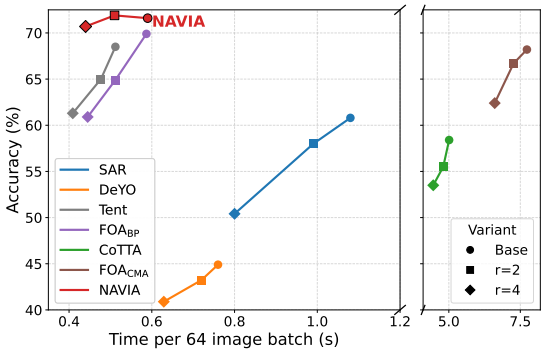


Fig. 1: ViT-L/16 performance of our proposed NAVIA and other ETTA methods in terms of latency-accuracy trade-offs. We apply ToMe with varying token reduction counts (r) to existing TTA methods to obtain their ETTA variants.

Overall, our contributions are summarized as follows:

- We introduce a novel challenge termed Efficient Test-Time Adaptation (ETTA), emphasizing the simultaneous optimization of inference efficiency and prediction performance for test-time adaptation.
- Through analyzing existing methods from a novel mutual information perspective, we show that simply applying token aggregation methods to TTA for efficiency introduces information loss which cannot be compensated by norm-tuning adopted in existing TTA methods.
- We propose NAVIA, featuring an input-level [CLS] embedding augmentation and a feature-level [CLS] bias augmentation to compensate for information

loss. Theoretical analysis motivates augmenting a global information carrier before token aggregation.
– Extensive experiments show that NAVIA significantly outperforms existing methods, improving both inference efficiency and prediction performance. Notably, NAVIA yields a better accuracy-efficiency balance.

2 Related Works

2.1 Test-Time Adaptation

Test-Time Adaptation (TTA) enables models to adapt to evolving environments exhibiting distribution shifts during deployment [57]. TTA methods generally fall into two categories [24]: instance-level and batch-level. Instance-level adaptation leverages zero-shot priors from vision-language models via prompt tuning or memory retrieval [20, 43], but requires repeated updates or substantial storage, limiting their practicality in resource-constrained environments. Batch-level methods primarily rely on normalization-based approaches, updating normalization statistics with the current test data [23, 37, 39, 52]. For instance, Tent [52] minimizes entropy over batches, EATA [38] selects diverse and reliable samples, and SAR [39] employs sharpness-aware minimization to reduce noisy gradients. However, batch-level approaches often demand significant computational resources, restricting their use in real-time or edge-device scenarios. To address these limitations, we pioneer the application of token aggregation in batch-level TTA, significantly enhancing computational efficiency while preserving model adaptability.

2.2 Model Compression

Model compression has been widely explored to enhance inference efficiency, particularly for deployment in resource-limited environments. Two prominent approaches are structural compression and token aggregation. Structural compression reduces model complexity by directly compressing redundant structures of the model backbone, including model compression [7, 33, 35, 51, 55, 62, 67], knowledge distillation [1, 13, 17, 49], and model quantization [4, 8, 28, 29, 31, 59]. However, these methods typically require extensive retraining, which is infeasible in TTA due to the limited availability of test-time data. Token aggregation [2, 6], in contrast, enhances inference efficiency by merging redundant tokens during inference without retraining. Notable methods include ToMe [2], which progressively merges similar tokens using lightweight bipartite matching, EViT [27], which utilizes attention scores to select critical tokens, and ToFu [21], which combines pruning and aggregation across layers. Despite their effectiveness, token aggregation methods still suffer from significant information loss when directly integrated into TTA. We propose to augment mutual information, leading to a new state-of-the-art ETTA approach.

3 Methodology

3.1 Preliminaries

ViT Model. In this paper, we mainly focus on conducting ETTA with Vision Transformers (ViT). ViTs are encoder-based transformers, where each encoder consists of a Multi-Head Self-Attention (MHSA) module and a Feed-Forward Network (FFN). In the forward pass, an input image $\mathbf{x}$ is reshaped and projected to N tokens $\mathbf{T}_{\text{img}}^{\text{Em}} = [t_1^{\text{Em}}, t_2^{\text{Em}}, \cdots, t_N^{\text{Em}}]$[5], each with a hidden dimension d. A classification ([CLS]) token $t_{\text{CLS}}^{\text{Em}}$ is prepended to the token sequence, resulting in the ViT model input: $\mathbf{T}^{\text{Em}} = [t_{\text{CLS}}^{\text{Em}}, t_1^{\text{Em}}, t_2^{\text{Em}}, \cdots, t_N^{\text{Em}}]$. Inside each encoder block, this token sequence is sequentially processed by the MHSA and the FFN. Specifically, the MHSA module employs fully-connected (FC) layers to transform input tokens into Q, K, and V matrices. The self-attention is then carried out via $\text{Softmax}(\frac{QK^T}{\sqrt{D}})V$, and the resulting output is further projected by another FC layer. Subsequently, the FFN processes the sequence through two FC layers.
Token Aggregation. Token aggregation methods [2, 21, 27, 60] are *training-free*, *plug-and-play* approaches that improve the inference efficiency of ViTs by reducing the number of image tokens. Those methods are orthogonal to the ViT structure and are capable of flexibly changing the compression rate. Let N_l and N_{l+1} denote the sequence lengths of input image tokens for layer l and $l+1$, respectively, and let $\mathbf{x}_l^{\text{In}}$ and $\mathbf{x}_l^{\text{Out}}$ represent the corresponding input and output token sequences for the token aggregation module at layer l. Specifically, token aggregation methods separate the token sequence into the [CLS] token t_{CLS} and the image token sequence $\mathbf{T}_l^{\text{In}} = [t_1^l, t_2^l, \cdots, t_N^l]$. During the aggregation step, existing methods utilize different metrics $g(\cdot)$ to evaluate tokens within $\mathbf{T}_l^{\text{In}}$. For instance, ToMe [2] applies similarity-based merging, while EViT [27] applies attentive-token preservation. Based on those evaluation metrics, an aggregation matrix $P_l = g(\mathbf{T}_l^{\text{In}}, t_{\text{CLS}}) \in \mathbb{R}^{N_{l+1} \times N_l}$ is computed and applied to obtain the aggregated output token sequence $\mathbf{T}_l^{\text{Out}} = P_l \mathbf{T}_l^{\text{In}}$. Finally, the [CLS] token t_{CLS}^l is prepended to $\mathbf{T}_l^{\text{Out}}$, forming the input for subsequent modules in ViT.

3.2 Analysis on Existing Approaches

Experimental evidence in Fig. 1 indicates that straightforwardly combining token aggregation with existing TTA approaches leads to significant performance degradation. Although notable TTA methods use norm-tuning to facilitate adaptation [23,38,39,52], they still fail to compensate for the performance loss caused by token aggregation despite its efficiency. To delve deeper into this issue and ground our analysis in established theory, we draw on prior work in domain adaptation [12, 69] and investigate how information propagates during adaptation. Prior studies [61, 68] have revealed that information loss can be quantified by the mutual information ($I(\cdot;\cdot)$) between input image embeddings and the ground-truth label Y, demonstrating that maximizing mutual information [42]

[5] "Em" denotes "Embedding".

significantly improves prediction performance under unsupervised conditions. Under our notations, this established conclusion can be expressed as:

Proposition 1. *For an image embedding* $\mathbf{x}$ *and its corresponding ground-truth* Y*, a higher value of* $I(\mathbf{x}; Y)$ *indicates better encoding quality of* $\mathbf{x}$*, which is more desirable.*

First, we analyze why naively combining token aggregation with TTA can hurt adaptation. Our analysis is intended to provide a principled design direction rather than a uniqueness proof for every design choice. Let $f_{\theta;g;\mathbf{x}}(\cdot)$ represent the forwarding process with the pre-trained ViT model θ, token aggregation with metric $g(\cdot)$ and input image $\mathbf{x}$. We use $f_{\theta;I;\mathbf{x}}(\cdot)$ as the forwarding process without token aggregation, as the metric $g(\cdot)$ always returns the identity matrix.

Corollary 1. *With token aggregation, the output [CLS] token contains less information about the label* Y*:* $I(f_{\theta;g;\mathbf{x}}(t^{\text{Em}}_{\text{CLS}}); Y) < I(f_{\theta;I;\mathbf{x}}(t^{\text{Em}}_{\text{CLS}}); Y)$.

Since token aggregation induces a many-to-one mapping on tokens: $\mathbf{T}^{\text{Out}}_l = P_l \mathbf{T}^{\text{In}}_l$, and can be viewed as a Markov chain, following the data processing inequality, $I(\mathbf{T}^{\text{Out}}_l; Y) \leq I(\mathbf{T}^{\text{In}}_l; Y)$, where the inequality is strict when aggregation is non-injective. Corollary 1 indicates that simply applying token aggregation leads to a strict degradation in model predictions. To address this, it is essential to compensate for the resulting information loss. While prominent TTA methods [23, 39, 52] fine-tune the normalization layers for adaptation, we show that *norm-tuning alone does not compensate for the information loss due to token aggregation* in ViT models, as LN is deterministic and cannot increase mutual information: $I(\text{LN}(\mathbf{T}^{\text{Out}}_l); Y) \leq I(\mathbf{T}^{\text{Out}}_l; Y)$.

Corollary 2. *In Pre-Norm ViT models, LayerNorm* $(\text{LN}(\cdot))$ *tuning cannot compensate for the mutual information loss caused by token aggregation.*

Corollary 2 holds as long as LayerNorm is deterministic and cannot reconstruct discarded information. Please refer to the appendix for detailed proofs to Corollary 1 and Corollary 2.

3.3 NAVIA

Corollary 2 implies that simply applying token aggregation to existing TTA methods introduces an information loss that cannot be compensated by subsequent deterministic operations. To retain useful information and neutralize the negative effects of token aggregation, optimization should occur *prior to applying token aggregation matrix and within input tokens*. Motivated by recent advances in prompt tuning [19, 43, 48], we address this issue at the input stage, especially during the construction of input token embeddings within each encoder block. However, in resource-constrained deployment environments, prepending randomly initialized learnable prompts requires intensive tuning to locate optimal embeddings [37], or even requires training an additional processing module to generate effective prompts [30]. As the complexity of self-attention is

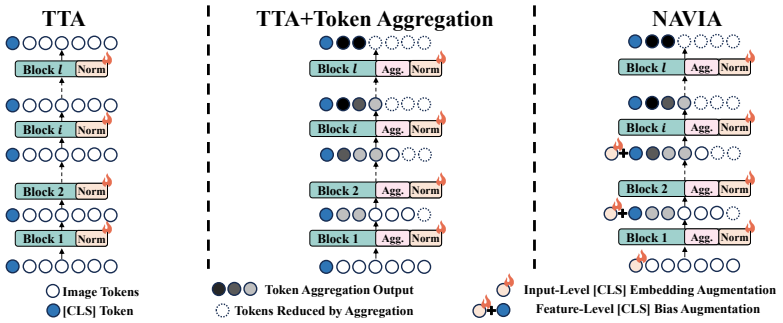


Fig. 2: Pipeline comparison between standard TTA methods, ETTA via token aggregation, and our proposed NAVIA. "Agg." refers to token aggregation, and "Norm" refers to LayerNorms in ViTs that are tunable during TTA. We employ input-level [CLS] embedding augmentation and feature-level [CLS] bias augmentation to neutralize negative effects caused by token aggregation.

quadratically proportional to sequence length, such an approach also introduces significant computational overhead during inference due to inserting extra tokens. Consequently, the constraints of ETTA make it preferable to directly tune existing prompts, and the [CLS] token offers a promising alternative.

Input-Level [CLS] Embedding Augmentation. Prior studies [45, 71] have shown that the [CLS] token can effectively encode domain-specific information beneficial for cross-domain learning. Inspired by those findings, we propose to leverage the [CLS] token to *encode domain-specific compensation information specifically targeted at addressing the information loss induced by token aggregation*. We refer it as **information augmentation**, serving as a complementary mechanism to norm-tuning. We theoretically demonstrate that optimizing the [CLS] token via entropy minimization increases the upper bound of learnable information:

Theorem 1. *Let δ represent a learned bias vector for the* [CLS] *embedding optimized through entropy minimization. Optimizing δ yields increased mutual information between network input and ground-truth label, formally:*

$$I([t_{\text{CLS}}^{\text{Em}} + \delta, P_0\mathbf{T}_{\text{img}}^{\text{Em}}]; Y) > I([t_{\text{CLS}}^{\text{Em}}, P_0\mathbf{T}_{\text{img}}^{\text{Em}}]; Y) \tag{1}$$

Please refer to the supplementary material for detailed proof to Theorem 1.

Theorem 1 motivates placing a lightweight learnable information carrier before aggregation, where the carrier can still influence the compressed token sequence, therefore potentially leading δ to close the information gap caused by introducing the token aggregation operation P_0. Here we leverage [CLS] embedding optimization to augment its information. Specifically, for each image batch $\mathbf{x}$ during the TTA process, we directly update the augmentation vector δ using gradients derived from the optimization objective $\mathcal{L}(\theta; \mathbf{x})$:

$$\widetilde{\delta} = \delta - \eta_\delta \nabla_\delta \mathcal{L}(\theta; \mathbf{x}), \tag{2}$$

where η_δ denotes the learning rate for δ, and ∇_δ the gradient operation with respect to δ. Through optimization, input-level [CLS] embedding augmentation compensates for the mutual information, increasing the upper bound of learnable information for ETTA.

Feature-Level [CLS] Bias Augmentation. Simply learning an augmentation vector δ for the [CLS] embedding has shown effectiveness in compensating for information loss (see Sec. 4 for details). However, since the augmentation vector only operates on input-level, it has limited flexibility to handle the information loss on image features in deeper layers. As token aggregation progressively reduces the total number of tokens, information loss also accumulates gradually throughout the forward pass. Consequently, despite the input-level augmentation, it is still necessary to develop a feature-level augmentation approach to progressively compensate for such information loss.

Building upon the former input-level [CLS] embedding augmentation, a straightforward approach is to learn an augmentation vector δ_l for the [CLS] token at every layer l, further enabling each δ_l to compensate for the information loss at its corresponding layer. However, this approach significantly increases the number of trainable parameters to Ld, where L is the total number of layers in the ViT model. Prior TTA research [54] indicates that increasing trainable parameters complicates the optimization process and adds computational overhead. In the ETTA scenario where computation budget is strictly limited, training a large number of parameters can be unstable and inefficient.

To achieve an optimal balance between efficiency and performance, we adopt the design of only augmenting [CLS] biases in *shallow* layers of the backbone. This design follows the following observations.

Observations. (1) From the token feature perspective, the aggregation matrix at layer l $P_l = g(\mathbf{T}_l^{\text{In}}, t_{\text{CLS}})$ depends on the [CLS] token. Augmenting [CLS] token before aggregation influences the aggregation choice in following layers. (2) From the information perspective, layerwise token aggregation is a lossy information compression process. Applying biases in shallow layers acts as a pre-emphasis on the useful information that is about to be compressed, helping to preserve discriminative signals for deeper layers. (3) From the model structure perspective, ViT adopts the residual structure, in which an update applied to [CLS] in a shallow block is directly reused by more subsequent blocks through the residual connections, whereas a late update has fewer chances to take effect.

Based on the observations above, formally, we select a layer subset $\mathbf{L}_{\text{aug}} \subset \{0, 1, \ldots, L-1\}$ under a fixed parameter budget constraint $|\mathbf{L}_{\text{aug}}| = L_{\text{bgt}}$. We choose the first L_{bgt} blocks as the selected subset:

$$\mathbf{L}_{\text{aug}} = \{0, 1, \ldots, L_{\text{bgt}} - 1\}, \tag{3}$$

Adopting this solution, we specifically update the augmentation vector δ_l for each $l \in \mathbf{L}_{\text{aug}}$ using the gradient derived from the optimization objective $\mathcal{L}(\theta; \mathbf{x})$:

$$\widetilde{\delta_l} = \delta_l - \eta_{\delta_l} \nabla_{\delta_l} \mathcal{L}(\theta; \mathbf{x}), \tag{4}$$

where η_{δ_l} is the learning rate applied across selected layers. Empirically, with our feature-level [CLS] bias augmentation, we could encourage the [CLS] feature

to follow a similar separation trend as that in conducting adaptation without token aggregation. Through visualization in Sec. 4.3, we could observe that while plain token aggregation yields noticeably more overlap in intermediate layers, our approach progressively separates different visual concepts. Overall, feature-level [CLS] bias augmentation leverages shallow layer learnable biases to progressively compensate for information layerwise, further boosting performance.

Overall TTA Framework. The proposed information augmentation strategies are implemented with the TTA framework, constituting our ETTA method, NAVIA. For the overall learning objective, we follow recent established approaches [9, 37, 41], employing entropy minimization and feature statistics discrepancy loss to reduce prediction uncertainty and feature shifting. Specifically, to calculate the above discrepancy loss, following [37, 41], we first sample a small set (e.g., 64) of unlabeled images from the source domain to compute the mean feature $\{\mu_l^S\}_{l=1}^L$ and mean variance $\{\sigma_l^S\}_{l=1}^L$ of each layer l. Note that such source-statistics term is a one-time pre-deployment caching step rather than a deployment-time requirement for source data. During TTA, we calculate the corresponding mean $\{\mu_l(\mathbf{T}_l)\}_{l=1}^L$ and variance $\{\sigma_l(\mathbf{T}_l)\}_{l=1}^L$ of target domain samples and align these with the source statistics. Formally, the loss is defined as:

$$\mathcal{L}_{TTA} = \sum_{c \in \mathcal{C}} -\hat{y}_c \log \hat{y}_c + \lambda \sum_{l=1}^{L} (\parallel \mu_l(\mathbf{T}_l) - \mu_l^S \parallel^2 + \parallel \sigma_l(\mathbf{T}_l) - \sigma_l^S \parallel^2), \tag{5}$$

where $\mathcal{C}$ denotes the set of categories, $\hat{y}_c$ is the probability prediction for category c, and λ is a trade-off parameter. During the TTA process, each incoming test batch is processed by a single forward pass before performing back-propagation. Importantly, the predictions for each batch are obtained *before* back-propagation to ensure computational efficiency, with only one forward pass per batch during TTA. For ETTA, NAVIA applies ToMe [2] during the forward pass, and leverages information augmentation of the [CLS] embedding and shallow-layer [CLS] biases during back-propagation. Following previous TTA methods, NAVIA also adopts standard norm-tuning for basic distribution alignment.

4 Experiments

4.1 Experimental Setting

Datasets&Models. Following [37], we evaluate our approach on four commonly used OOD benchmarks: ImageNet-C [16], ImageNet-R [15], ImageNet-V2 [40], and ImageNet-S [53] with ViT-B/16 and ViT-L/16.

Baselines. We compare NAVIA against: (1) *Online TTA methods*, including Tent [52], LAME [3], CoTTA [54], SAR [39], FOA (including the BP-free FOA_{CMA} variant and the FOA_{BP} variant) [37] , and DeYO [23]. All tested TTA methods with BP *employ LayerNorm tuning*. (2) *Token aggregation methods*, including ToMe [2], EViT [27], Tofu [21], and TCA [56], perform training-free token reduction within the ViTs, reducing r tokens in each layer. For fair comparison, they

Table 1: Performance (%) comparison with state-of-the-art methods on ImageNet-C (severity level 5). Subscripts represent the number of tokens reduced in each layer through aggregation. "Time (s)" denotes wall-clock time for processing a single image batch. **Bold** represents the *best token aggregation result*. †: BP-free methods.

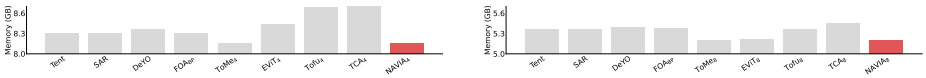
	Noise			Blur				Weather				Digital						
Method	Gauss.	Shot	Imp.	Def.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Cont.	Elas.	Pix.	JPEG	Avg.	GFLOPs	Time (s)
Backbone: ViT-L/16																		
NoAdapt	62.5	62.0	63.3	52.9	45.3	60.7	55.2	66.0	62.3	62.6	79.9	40.1	56.2	74.3	72.8	61.1	55.60	0.50
Tent	62.9	67.4	68.5	64.8	62.1	67.9	64.7	71.2	68.5	67.8	79.0	61.1	68.5	78.1	74.7	68.5	55.60	0.51
LAME†	62.3	61.8	63.1	52.7	45.1	60.5	55.0	65.8	61.9	62.0	79.8	39.9	55.8	74.2	72.6	60.9	55.60	0.50
CoTTA	59.0	60.8	60.2	48.5	53.7	49.7	54.2	56.0	62.5	60.2	69.9	45.6	63.5	66.6	65.2	58.4	111.20	5.01
SAR	67.2	68.3	68.2	63.4	60.4	67.7	64.2	70.4	69.3	22.3	13.5	57.4	67.0	77.5	74.7	60.8	111.20	1.08
DeYO	4.5	65.3	53.8	55.7	48.7	32.9	56.0	47.3	64.5	40.3	4.3	7.0	71.4	45.9	76.3	44.9	111.20	0.76
FOA_{CMA}†	65.3	64.4	65.2	62.4	56.7	63.8	59.3	71.5	72.3	73.7	82.0	63.7	67.1	76.9	76.5	68.2	1588.64	7.73
FOA_{BP}	68.5	70.1	67.0	62.3	64.4	66.0	66.5	72.1	70.8	65.1	82.6	65.6	71.9	78.7	77.2	69.9	55.60	0.59
ToMe_2	68.6	68.8	68.5	60.7	64.1	60.1	66.7	73.0	70.4	59.9	82.1	**64.0**	72.6	78.3	76.2	68.9	50.15	0.51
ToMe_4	62.8	68.7	67.3	61.5	61.6	59.5	58.2	61.7	69.0	57.1	81.9	57.1	68.2	76.4	75.8	65.8	**42.98**	0.44
EViT_2	64.7	65.3	67.0	61.2	64.3	64.8	65.2	67.2	63.2	63.4	81.4	53.2	69.4	77.9	69.3	66.5	53.75	0.56
EViT_4	59.4	38.2	61.7	30.5	60.4	62.6	50.8	60.7	16.9	54.2	79.0	41.2	60.6	72.1	67.3	54.4	46.55	0.49
Tofu_2	65.0	69.1	67.0	62.6	64.4	66.0	64.4	71.2	70.2	62.5	81.3	56.6	72.3	77.9	71.1	68.1	50.15	0.52
Tofu_4	63.7	63.2	65.1	60.8	62.7	58.1	55.2	64.4	51.0	52.1	79.6	56.4	64.8	74.6	70.4	62.8	**42.98**	0.44
TCA_2	58.1	70.1	69.8	60.2	53.9	62.9	63.2	74.4	68.9	70.5	82.3	45.4	74.5	60.7	76.1	66.1	52.76	1.89
TCA_4	64.6	68.6	66.9	43.1	53.0	62.6	62.0	63.9	67.3	71.6	81.9	50.5	72.0	76.9	75.3	65.3	44.93	0.92
NAVIA_2	**70.0**	**71.4**	**70.1**	**64.4**	**67.1**	**69.7**	**67.7**	**74.9**	**72.4**	**73.8**	**82.6**	62.2	**74.7**	**79.5**	**78.3**	**71.9**	50.15	0.51
NAVIA_4	67.4	69.5	69.0	63.2	65.6	68.7	67.3	74.6	70.6	71.7	82.1	59.5	74.3	78.9	77.5	70.7	**42.98**	0.44
Backbone: ViT-B/16																		
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	15.71	0.32
Tent	60.3	61.6	61.8	59.2	56.5	63.5	59.2	54.3	64.5	2.3	79.1	67.4	61.5	72.5	70.6	59.6	15.71	0.32
LAME†	56.5	56.5	57.2	46.4	34.7	52.7	44.2	58.4	61.5	63.1	77.4	24.7	44.6	66.6	67.2	54.1	15.71	0.32
CoTTA	63.6	63.8	64.1	55.5	51.1	63.6	55.5	70.0	69.4	71.5	78.5	9.7	64.5	73.4	71.2	61.7	31.42	1.30
SAR	59.2	60.5	60.7	57.5	55.6	61.8	57.6	65.9	63.5	69.1	78.7	45.7	62.4	71.9	70.3	62.7	31.42	0.63
DeYO	62.4	64.0	63.9	61.0	60.7	66.4	62.9	70.9	69.6	73.7	80.5	67.2	69.9	75.7	73.7	68.2	31.42	0.45
FOA_{CMA}†	61.5	63.2	63.3	59.3	56.7	61.4	57.7	69.4	69.6	73.4	81.1	67.7	62.7	73.9	73.0	66.3	446.49	4.19
FOA_{BP}	64.4	66.3	66.3	63.0	63.7	68.6	67.5	73.7	71.4	76.3	81.6	70.2	73.2	77.4	75.0	70.6	15.71	0.39
ToMe_4	63.3	65.3	65.1	62.3	63.1	68.0	67.0	73.2	70.8	75.6	81.3	68.4	73.0	77.0	74.9	69.9	14.29	0.32
ToMe_8	62.0	63.5	63.4	61.0	62.2	67.1	66.0	72.5	69.8	74.8	81.0	67.0	72.0	76.3	74.3	68.9	**12.22**	0.30
EViT_4	62.7	64.4	64.1	61.8	62.5	67.4	65.6	72.2	70.4	73.8	80.9	68.5	71.8	76.2	73.6	69.1	14.80	0.37
EViT_8	59.9	61.8	61.6	60.2	60.8	65.6	63.8	70.4	68.2	72.3	79.8	66.2	70.5	74.6	71.6	67.2	12.73	0.34
Tofu_4	63.5	64.9	65.1	61.9	62.7	67.7	66.6	73.0	70.6	75.7	81.1	68.6	72.6	76.8	74.6	69.7	14.29	0.32
Tofu_8	62.2	63.9	63.7	60.9	61.8	66.8	65.4	72.1	69.6	74.4	80.7	67.3	71.8	76.2	74.0	68.7	**12.22**	0.30
TCA_4	64.3	66.0	66.0	62.1	63.3	68.1	67.2	73.5	71.2	71.8	81.4	65.3	73.1	77.1	74.9	69.7	14.81	1.42
TCA_8	63.4	65.1	65.1	60.9	62.4	67.0	65.9	72.8	70.4	69.9	81.1	64.6	72.1	76.5	74.3	68.8	12.68	1.23
NAVIA_4	**64.2**	**65.9**	**65.7**	**63.5**	**64.4**	**69.0**	**67.4**	**74.0**	**72.0**	**76.9**	**81.4**	**69.8**	**73.8**	**77.4**	**75.5**	**70.7**	14.29	0.32
NAVIA_8	62.7	64.4	64.2	62.1	63.5	67.9	66.5	73.4	71.1	76.4	81.0	68.8	73.1	76.7	74.9	69.8	**12.22**	0.30

are implemented on the best-performing FOA_{BP}. Following [60], experiments are conducted at two compression levels: moderate (r=2 for ViT-L/16 and r=4 for ViT-B/16) and high (r=4 for ViT-L/16 and r=8 for ViT-B/16).

Implementation Details. Following [37], we set the learning rate to 5e-3, and the trade-off parameter λ as 30. We set the batch size as 32 for ViT-L/16 and 64 for ViT-B/16. We use SGD for both LayerNorm tuning and information augmentation. For token aggregation, we adopt ToMe [2] for NAVIA, given its superior performance.

Table 2: Performance (%) comparison with state-of-the-art methods on ImageNet-R/V2/Sketch. Subscripts represent the number of tokens reduced in each layer through aggregation. **Bold** represents the *best token aggregation result*. †: BP-free methods.

Method	R	V2	S	Avg.	Method	R	V2	S	Avg.	Method	R	V2	S	Avg.
Backbone: ViT-L/16														
NoAdapt	63.1	76.0	51.8	63.6	$\text{FOA}_{\text{CMA}}{}^{\dagger}$	66.2	76.2	52.2	64.9	Tofu_2	69.4	75.0	52.6	65.7
TENT	63.8	76.2	55.5	65.2	FOA_{BP}	71.3	75.4	54.6	67.1	Tofu_4	51.5	70.5	49.4	57.1
CoTTA	70.4	66.6	51.1	62.7	ToMe_2	70.8	73.3	53.3	65.8	TCA_2	70.9	73.8	54.2	66.3
SAR	68.1	76.2	55.3	66.5	ToMe_4	66.7	71.5	51.1	63.1	TCA_4	69.7	74.0	53.1	65.6
DEYO	71.4	67.9	51.8	63.7	EViT_2	52.6	74.2	53.7	60.2	NAVIA_2	**72.0**	**76.4**	**57.7**	**68.7**
LAME†	64.1	76.0	51.6	63.9	EViT_4	42.4	73.4	46.8	54.2	NAVIA_4	71.8	75.9	57.0	68.2
Backbone: ViT-B/16														
NoAdapt	59.5	75.4	44.9	59.9	$\text{FOA}_{\text{CMA}}{}^{\dagger}$	63.8	75.4	49.9	63.0	Tofu_4	67.3	75.3	52.1	64.9
Tent	63.9	75.4	49.9	63.0	FOA_{BP}	67.7	75.4	52.2	65.1	Tofu_8	66.8	74.9	51.6	64.4
CoTTA	63.5	75.4	50.0	62.9	ToMe_4	67.6	75.3	52.2	65.0	TCA_4	67.7	73.5	51.9	64.4
SAR	63.3	75.1	48.7	62.4	ToMe_8	66.8	74.6	52.0	64.5	TCA_8	67.0	73.0	51.7	63.9
DeYO	66.1	74.8	52.2	64.4	EViT_4	66.0	74.8	51.3	64.0	NAVIA_4	**68.8**	**75.5**	**53.6**	**66.0**
LAME†	59.0	75.2	44.4	59.6	EViT_8	64.4	73.8	49.9	62.7	NAVIA_8	68.4	75.2	53.4	65.6



(a) ViT-L/16 **(b)** ViT-B/16

Fig. 3: Peak memory usage comparison for back propagation methods.

4.2 Main Results

As shown in Tab. 1 and Tab. 2, NAVIA consistently outperforms all token aggregation baselines across different settings, backbones, and compression rates, achieving the best token aggregation performance in 35 of 36 domains. Notably, on the ViT-L/16 backbone, NAVIA even significantly outperforms the TTA state-of-the-art, FOA, by a maximum of 3.1% (on ImageNet-Sketch). Existing token aggregation baselines, however, yield significant accuracy drops when applied to a large ViT-L/16 backbone. For example, under high compression rates, EViT incurs a 15.5% drop on ImageNet-C. This validates Corollary 2 that LayerNorm tuning cannot compensate for the performance drop brought by token aggregation. When paired with our [CLS] augmentation methods, NAVIA successfully regains the performance, well demonstrating its effectiveness and superiority, and empirically supporting the MI motivation (Theorem 1) as well.

Computational Efficiency. NAVIA improves the ETTA trade-off by being *faster, lighter, and more accurate* than prior TTA methods that incur heavy multi-pass overhead. As shown in Tab. 1, on the larger ViT-L/16 backbone, even the highly compressed NAVIA_4 outperforms the TTA state-of-the-art, FOA_{BP}, with 23% less GFLOPs and 25% wall-clock speedup. When compared with its compute-intensive variant FOA_{CMA}, NAVIA_4 is over 17 times faster while delivering 2.5% accuracy improvement. As illustrated in Fig. 1, NAVIA yields the best Pareto frontier over all compared baselines, with the best accuracy among

Table 3: Ablation study results. [CLS]-E: Input-level [CLS] embedding augmentation. [CLS]-S: Feature-level [CLS] bias augmentation. **Bold** represents the best result.

r	[CLS]-E	[CLS]-S	C	R	V2	Sketch	Avg.
4	×	×	69.9	67.6	75.3	52.2	66.3
	√	×	70.3	68.2	75.2	52.6	66.6
	×	√	70.4	68.0	**75.5**	53.4	66.8
	√	√	**70.7**	**68.8**	**75.5**	**53.6**	**67.2**
8	×	×	68.9	66.8	74.6	52.0	65.6
	√	×	69.3	67.5	75.0	52.9	66.2
	×	√	69.5	67.3	74.7	53.0	66.1
	√	√	**69.8**	**68.4**	**75.2**	**53.4**	**66.7**

Table 4: Analysis of augmenting [CLS] bias on different layers and different token tuning approaches for [CLS] embedding augmentation. Wea.: Weather. Dig.: Digital. **Bold** represents the best result.

Method	Noise	Blur	Wea.	Dig.	Avg.
Layer Selection					
Shallow:{0,1,2,3,4,5}	**63.7**	**65.0**	**75.5**	**73.4**	**69.8**
Deep:{6,7,8,9,10,11}	63.5	64.6	74.4	72.8	69.2
Uniform:{0,2,4,6,8,10}	63.4	64.5	74.6	72.8	69.2
Token Tuning Approach					
Init. [CLS] Embed.	60.2	60.8	72.0	69.3	65.9
Add New Tokens	63.6	64.7	74.7	73.0	69.4
Tune Trained [CLS]	**63.7**	**65.0**	**75.5**	**73.4**	**69.8**

the most efficient methods. Furthermore, as shown in Fig. 3, the memory footprint of NAVIA remains comparable to the most efficient baselines, indicating that the augmentation approaches do not come with extra memory burden.

4.3 Experimental Analysis

We present experimental analysis results on the ViT-B/16. Please refer to the appendix for further analysis results.

Ablation Studies for Components. We do controlled experiments to analyze the contribution of each component in NAVIA: input-level [CLS] embedding augmentation and feature-level [CLS] bias augmentation. As shown in Tab. 3, both the two components independently improve the model's adaptation performance. Notably, employing both strategies yields the best results across all experimental settings, underscoring their individual effectiveness and mutual complementarity.

Visualizations. We conduct a qualitative analysis to illustrate how NAVIA affects feature representations. Specifically, we select 8 representative classes from ImageNet-R, categorized into two "super-classes": aquatic animals (goldfish, great white shark, hammerhead, stingray) and birds (hen, ostrich, goldfinch, junco). We apply t-SNE [36] to the [CLS] features extracted from layers 3, 6, and 11 of ViT. As shown in Fig. 4, NAVIA progressively separates features between the two super-classes from shallow to deep layers, while effectively preserving the intra-class distributions. Conversely, with ToMe, the [CLS] features from different super-classes in layers 3 and 6 exhibit significant overlap, indicating limited discriminability. Even in layer 11, ToMe seems to misclassify a class to the other super-classes. When compared to adaptation without token aggregation ("Baseline"), NAVIA encourages the [CLS] feature to follow a similar separation trend as that in conducting adaptation without token aggregation, progressively separating different visual concepts as [CLS] is being processed through the forward

pass. These results underscore that our information augmentation guides the ViT to better capture class-specific semantics under distribution shifts, showing that information is compensated layerwise inside ViT.

Design Choices for NAVIA. We study the two design factors of [CLS] bias augmentation: the selection of layers for tuning and the choice for token augmentation. (1) *Layer Selection.* We investigate layer subsets where [CLS] biases are augmented. First, as shown in Tab. 4 (top), compared to other two selections (i.e., deep layers, and uniformly selected layers), augmenting [CLS] biases exclusively in shallow layers consistently yields the best results. Second, considering the number of shallow layers, as shown in Fig. 5a, tuning 4 to 6 shallow layers is optimal. Extending augmentation to more than 6 layers leads to performance saturation or slight degradation, likely due to suboptimal optimization. This can be attributed to the limited number of iterations available during the TTA process. Introducing an excessive number of tunable parameters hinders the convergence of all bias parameters. As a result, the optimization process becomes less effective, which ultimately impairs performance. Consequently, we select the optimal number of layers from the set {4,5,6} based on minimizing the error computed using Eq. (5) on a small set of gathered target images. (2) *Token Selection.* We then compare our proposed pre-trained [CLS] tuning strategy against alternative token augmentation methods. As shown in Tab. 4 (bottom), tuning newly added tokens leads to lower performance, likely due to the difficulty of tuning from scratch with limited TTA steps. Tuning re-initialized [CLS] notably decreases performance, highlighting the importance of preserving pre-trained embeddings that inherently encode class-relevant information.

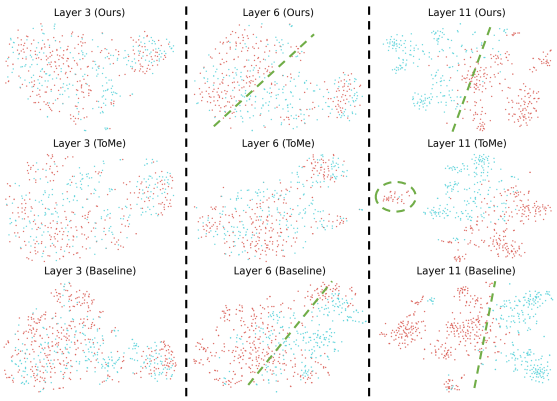


Fig. 4: t-SNE analysis. Colors represent different super-classes (aquatic animals in red, birds in blue).

Convergence Speed Analysis. To analyze the convergence efficiency of NAVIA, we compare it against existing token aggregation baselines following the evaluation protocol established by [9]. Specifically, we perform TTA optimization over a fixed number of iterations and subsequently evaluate the optimized model's average accuracy on a set of 10,000 images reserved from the original ImageNet-R dataset. As illustrated in Fig. 5b, NAVIA consistently achieves the fastest convergence rate and highest final accuracy among all compared methods. Notably, NAVIA demonstrates a rapid accuracy improvement in the initial adaptation

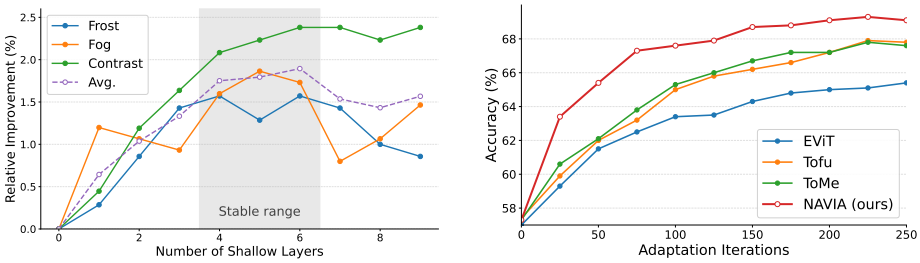


Fig. 5: (a) Relative acc. (%) improvement results for different number of shallow layers with augmented [CLS] biases. (b) Convergence speed comparison on ImageNet-R.

Table 5: Ablation results for **total** trainable parameters on ImageNet-C using ViT-B/16. **Bold** denotes the best result. Note that for ViT-B/16, standard LayerNorm tuning updates a total of 38.4K parameters.

Method	Configuration	Params	Noise	Blur	Weather	Digital	ImageNet-C
Prompt Tuning	+1 token / layer	43.0K	62.9	64.0	74.4	72.2	68.7
Prompt Tuning	+2 tokens / layer	47.6K	63.1	64.0	74.7	72.5	68.9
Prompt Tuning	+3 tokens / layer	52.2K	63.1	64.0	74.9	72.4	69.0
LoRA	$r = 1$	66.0K	63.0	64.0	71.1	72.2	67.9
LoRA	$r = 2$	93.7K	62.9	63.8	74.2	72.4	68.7
LoRA	$r = 3$	121.3K	62.9	63.9	74.4	72.3	68.8
NAVIA (ours)	–	43.7K	**63.7**	**65.0**	**75.5**	**73.4**	**69.8**

iterations, indicating that our proposed information augmentation strategies effectively mitigate the early-stage information loss incurred by token aggregation.

Further Ablation on Trainable Parameters. NAVIA introduces additional trainable parameters compared to plainly attaching token aggregation during TTA. To investigate whether performance gains are primarily driven by the number of additional trainable parameters, we scale up parameter capacity with standard adaptation strategies: prompt tuning [19] and LoRA [18]. For prompt tuning, we insert different number of learnable tokens in the layers where we conduct [CLS] bias augmentation. For LoRA, we attach trainable LoRA modules with different bottleneck rank to QKV-projection in the same subset of layers. As shown in Tab. 5, while increased prompt tokens or higher-rank LoRA moderately affect performance, their performance quickly saturates and neither surpasses our NAVIA, despite requiring substantially larger parameter budgets. These results indicate that performance improvements are not merely a consequence of parameter scaling, but instead arise from the effectiveness of the adaptation mechanism.

5 Conclusion

In this work, we introduced a novel challenge termed Efficient Test-Time Adaptation (ETTA) to directly reduce the backbone forward cost in existing TTA frameworks. While token aggregation methods reduce computational overhead, our analysis reveals that they inherently induce information loss, which cannot be adequately compensated by the commonly used norm-tuning approach in TTA, leading to significant performance degradation. To address this, we proposed to Neutralize Token Aggregation via Information Augmentation (NAVIA), compensating for information loss on both the input-level and the feature-level. Theoretical analysis shows that NAVIA increases the upper bound for learnable information under token aggregation. Extensive experiments validated that NAVIA achieves superior predictive accuracy while reducing computational costs, making test-time adaptation methods more practical for resource-constrained environments.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 62525103, 62271281, 62441235) and Beijing Natural Science Foundation (L247026).

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9163–9171 (2019)
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: ICLR (2023)
3. Boudiaf, M., Mueller, R., Ben Ayed, I., Bertinetto, L.: Parameter-free online test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8344–8353 (2022)
4. Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M.W., Keutzer, K.: Zeroq: A novel zero shot quantization framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13169–13178 (2020)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020)
6. Chen, M., Shao, W., Xu, P., Lin, M., Zhang, K., Chao, F., Ji, R., Qiao, Y., Luo, P.: Diffrate : Differentiable compression rate for efficient vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17164–17174 (October 2023)
7. Chen, T., Cheng, Y., Gan, Z., Yuan, L., Zhang, L., Wang, Z.: Chasing sparsity in vision transformers: An end-to-end exploration. Advances in Neural Information Processing Systems **34**, 19974–19988 (2021)

8. Courbariaux, M., Bengio, Y., David, J.P.: Binaryconnect: Training deep neural networks with binary weights during propagations. Advances in neural information processing systems **28** (2015)
9. Deng, Q., Niu, S., Zhang, R., Chen, Y., Zeng, R., Chen, J., Hu, X.: Learning to generate gradients for test-time adaptation via test-time training layers. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
11. Farina, M., Franchi, G., Iacca, G., Mancini, M., Ricci, E.: Frustratingly easy test-time adaptation of vision-language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
12. Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., Pavlovic, V.: Unsupervised multi-target domain adaptation: An information theoretic approach. IEEE Transactions on Image Processing **29**, 3993–4002 (2020). https://doi.org/10.1109/TIP.2019.2963389, publisher Copyright: © 2020 IEEE.
13. Habib, G., Saleem, T.J., Lall, B.: Knowledge distillation in vision transformers: A critical review (2024), https://arxiv.org/abs/2302.02108
14. Hao, T., Chen, H., Guo, Y., Ding, G.: Consolidator: Mergable adapter with group connections for visual adaptation. In: The Eleventh International Conference on Learning Representations (2023)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 8320–8329. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00823, https://doi.org/10.1109/ICCV48922.2021.00823
16. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
19. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
20. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14162–14171 (2024)
21. Kim, M., Gao, S., Hsu, Y.C., Shen, Y., Jin, H.: Token fusion: Bridging the gap between token pruning and token merging. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1383–1392 (2024)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)

23. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In: The Twelfth International Conference on Learning Representations (2024)
24. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
25. Liang, Y., Chen, H., Lin, Z., Liu, P., Wang, J., Wang, G., Li, L., Zhao, S., Han, J., Ding, G.: Dar-prompt: Dynamic regulation in prompt tuning for multi-label zero-shot learning. IEEE Transactions on Image Processing **34**, 7697–7711 (2025). https://doi.org/10.1109/TIP.2025.3626157
26. Liang, Y., Chen, H., Xiong, Y., Zhou, Z., Lyu, M., Lin, Z., Niu, S., Zhao, S., Han, J., Ding, G.: Advancing reliable test-time adaptation of vision-language models under visual variations (2025), https://arxiv.org/abs/2507.09500
27. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=BjyvwnXXVn_
28. Lin, X., Zhao, C., Pan, W.: Towards accurate binary convolutional neural network. Advances in neural information processing systems **30** (2017)
29. Lin, Z., Courbariaux, M., Memisevic, R., Bengio, Y.: Neural networks with few multiplications. arXiv preprint arXiv:1510.03009 (2015)
30. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI Open **5**, 208–215 (2024)
31. Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., Gao, W.: Post-training quantization for vision transformer. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 28092–28103. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf
32. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 97–105. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/long15.html
33. Lv, M., Su, Z., Pan, L., Xiong, Y., Lin, Z., Chen, H., Zhou, W., Han, J., Ding, G., Luo, C., et al.: Dsmoe: Matrix-partitioned experts with dynamic routing for computation-efficient dense llms. arXiv preprint arXiv:2502.12455 (2025)
34. Lyu, M., Hao, T., Xu, X., Chen, H., Lin, Z., Han, J., Ding, G.: Learn from the learnt: Source-free active domain adaptation via contrastive sampling and visual persistence. In: European Conference on Computer Vision. pp. 228–246. Springer (2024)
35. Ma, X., Fang, G., Wang, X.: Llm-pruner: On the structural pruning of large language models. In: Advances in Neural Information Processing Systems (2023)
36. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
37. Niu, S., Miao, C., Chen, G., Wu, P., Zhao, P.: Test-time model adaptation with only forward passes. In: Proceedings of the 41st International Conference on Machine Learning. pp. 38298–38315 (2024)
38. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)

39. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: The Eleventh International Conference on Learning Representations (2023)
40. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 5389–5400. PMLR (2019), http://proceedings.mlr.press/v97/recht19a.html
41. Samadh, J.H.A., Gani, H., Hussein, N.H., Khattak, M.U., Naseer, M., Khan, F., Khan, S.: Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
42. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
43. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)
44. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)
45. Tang, Y., Chen, S., Kan, Z., Zhang, Y., Guo, Q., He, Z.: Learning visual conditioning tokens to correct domain shift for fully test-time adaptation. IEEE Transactions on Multimedia (2024)
46. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
47. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021), https://arxiv.org/abs/2012.12877
48. Tu, C.H., Mai, Z., Chao, W.L.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7725–7735 (2023)
49. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
50. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Lsnet: See large, focus small. In: Proceedings of the computer vision and pattern recognition conference. pp. 9718–9729 (2025)
51. Wang, A., Chen, H., Lin, Z., Zhao, S., Han, J., Ding, G.: Cait: Triple-win compression toward high accuracy, fast inference, and favorable transferability for vits. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(2), 1373–1389 (2026). https://doi.org/10.1109/TPAMI.2025.3616854
52. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)
53. Wang, H., Ge, S., Lipton, Z.C., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 10506–10518 (2019), https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html

54. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (2022)
55. Wang, Z., Luo, H., WANG, P., Ding, F., Wang, F., Li, H.: Vtc-lfc: Vision transformer compression with low-frequency components. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 13974–13988. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/5a8177df23bdcc15a02a6739f5b9dd4a-Paper-Conference.pdf
56. Wang, Z., Gong, D., Wang, S., Huang, Z., Luo, Y.: Is less more? exploring token condensation as training-free test-time adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 144–154 (2025)
57. Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., Huang, Z.: In search of lost online test-time adaptation: A survey. International Journal of Computer Vision pp. 1–34 (2024)
58. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(5), 1–46 (2020)
59. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: SmoothQuant: Accurate and efficient post-training quantization for large language models. In: Proceedings of the 40th International Conference on Machine Learning (2023)
60. Xiong, Y., Chen, H., Hao, T., Lin, Z., Han, J., Zhang, Y., Wang, G., Bao, Y., Ding, G.: Pyra: Parallel yielding re-activation for training-inference efficient task adaptation. In: European Conference on Computer Vision. pp. 455–473. Springer (2024)
61. Xiong, Y., Chen, H., Lin, Z., Zhao, S., Ding, G.: Confidence-based visual dispersal for few-shot unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11621–11631 (2023)
62. Xiong, Y., Huang, W., Ye, X., Chen, H., Lin, Z., Lian, H., Su, Z., Han, J., Ding, G.: Uniattn: Reducing inference costs via softmax unification for post-training llms. arXiv preprint arXiv:2502.00439 (2025)
63. Xu, X., Chen, H., Lin, Z., Han, J., Gong, L., Wang, G., Bao, Y., Ding, G.: Tad: A plug-and-play task-aware decoding method to better adapt llms on downstream tasks. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 6587–6596 (2024)
64. Xu, X., Chen, H., Lyu, M., Zhao, S., Xiong, Y., Lin, Z., Han, J., Ding, G.: Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1571–1590 (2025)
65. Yang, F., Chen, H., He, Y., Zhao, S., Zhang, C., Ni, K., Ding, G.: Geometry-guided domain generalization for monocular 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6467–6476 (2024)
66. Yang, H.Y., Chen, H., Wang, A., Chen, K., Lin, Z., Tang, Y., Gao, P., Quan, Y., Han, J., Ding, G.: Promptable anomaly segmentation with sam through self-perception tuning (2024), https://arxiv.org/abs/2411.17217
67. Yu, S., Chen, T., Shen, J., Yuan, H., Tan, J., Yang, S., Liu, J., Wang, Z.: Unified visual transformer compression. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=9jsZiUgkCZP
68. Yue, X., Zheng, Z., Zhang, S., Gao, Y., Darrell, T., Keutzer, K., Sangiovanni-Vincentelli, A.: Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)

69. Zhao, H., Ma, C., Chen, Q., Deng, Z.H.: Domain adaptation via maximizing surrogate mutual information. arXiv preprint arXiv:2110.12184 (2021)
70. Zhou, L., Ye, M., Li, S., Li, N., Zhu, X., Deng, L., Liu, H., Lei, Z.: Bayesian test-time adaptation for vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29999–30009 (2025)
71. Zou, Y., Yi, S., Li, Y., Li, R.: A closer look at the CLS token for cross-domain few-shot learning. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=qIkYlfDZaI