To Adapt or Not to Adapt? Selective Adaptation for Vision-Language Models

Siru Jiang[1,2†], Yuwei Liang[2,3†], Jian Liang[2,3⋆], Ran He[2,3], and Tieniu Tan[1,2,4]

[1] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
[2] NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[4] Nanjing University, China
{sirujiang324, liangjian92}@gmail.com

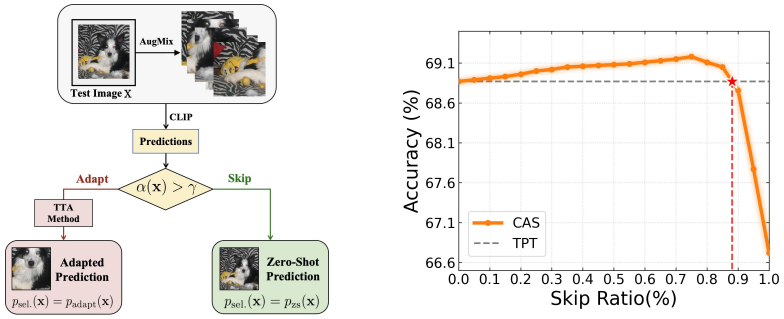
Abstract. Test-time adaptation (TTA) has emerged as a prominent strategy for adapting vision-language models to distribution shifts during inference. We conduct a per-sample analysis of model predictions before and after adaptation, and observe two failure modes in existing TTA methods that echo previous work. Adaptations are frequently negligible, yielding no change in the model's predictions, and more severely, they can be detrimental by flipping previously correct predictions to incorrect ones. This naturally raises a question: *Can we identify and skip such negligible or harmful adaptations?* In this work, we introduce a new problem of **selective adaptation**, which aims to determine whether a given test sample should undergo adaptation or be skipped. To this end, we propose Cross-Augmentation Similarity (CAS), a simple baseline that performs adaptation only when predictions across augmented views exhibit low similarity. Notably, CAS not only preserves but in some cases improves overall accuracy, even when skipping nearly 85% of the adaptation process. We hope other researchers will explore this new direction and surpass the performance of our baseline. Our code is available at `https://github.com/sirujiang/selective-adaptation`.

Keywords: Test-time Adaptation · Selective Adaptation · Vision-Language Models

1 Introduction

Vision-language models (VLMs), such as CLIP [42], ALIGN [26], Flamingo [2], and LLaVA [34], are pretrained on large-scale image–text pairs and have demonstrated strong generalization across diverse vision tasks [57, 69, 72, 73]. Among them, CLIP [42] aligns visual and textual representations in a shared embedding space, enabling impressive zero-shot performance. However, VLMs remain sensitive to distribution shifts and often suffer from performance degradation when the test distribution differs from that of pretraining [29, 33, 44, 65].

⋆ Corresponding author. † These authors are co-first authors.



(a) Selective Adaptation **(b)** Accuracy versus Skip Ratio

Fig. 1: (a) Given a test image x, a score $\alpha(x)$ is computed. Samples with low scores undergo adaptation, while those with high scores are skipped, with the zero-shot prediction used instead. (b) The proposed CAS maintains or slightly improves accuracy across a wide range of skip ratios (the proportion of skipped adaptations) on ImageNet.

In recent years, test-time adaptation (TTA) has emerged as an effective paradigm for mitigating distribution shifts by adapting VLMs with unlabeled test data. In the context of image classification, existing TTA methods mainly focus on improving prediction accuracy [6, 10, 49] or enhancing model calibration [1,46,64]. Despite their effectiveness, they implicitly assume that adaptation is beneficial for all test samples. To better understand this issue, we conduct a preliminary per-sample analysis of the adaptation process under the classic TPT framework [49].

Consistent with observations in prior work [10], we find that a large proportion of predictions remain unchanged before and after adaptation, leading to unnecessary computational overhead. More critically, some originally correct predictions flip to incorrect after adaptation, leading to performance degradation. Notably, these negligible or even harmful adaptations account for more than 90% of all adaptation processes. Motivated by these observations, we propose to identify and skip ineffective adaptations, thereby improving efficiency while preserving accuracy. Unlike previous TTA approaches that enhance efficiency through parameter-free retrieval [7,27] or lightweight parameter optimization [6, 25], our method improves efficiency at the per-sample level by adapting only when necessary.

In this work, we introduce a new problem, termed **selective adaptation**, which formulates a binary detection task to identify whether the adaptation for a given test sample should be performed or skipped. An illustration of this problem is shown in Fig. 1 (a). Similar to out-of-distribution (OOD) detection [17, 20, 22, 35, 36, 38, 62], it relies on a scoring function to identify ineffective adaptations. Specifically, samples with higher scores tend to correspond to negligible or harmful cases and use zero-shot predictions directly, while those with lower scores continue to undergo adaptation. Generally, the goal of selective adaptation is to skip as many ineffective adaptations as possible while maintaining or even improving overall accuracy. To evaluate this problem, we adopt the standard Area Under the ROC Curve (AUC) [8, 11] for detection quality,

and introduce Accuracy Expectation with a Triangular Prior (AEP) to measure expected accuracy under different skip ratios.

Test-time augmentation [10, 31, 45, 49] has been used in many TTA methods [49, 70] to generate multiple views for an input sample. We argue that the prediction similarity between the test sample and its augmented views may be closely correlated with adaptation effectiveness. This motivates our simple baseline Cross-Augmentation Similarity (CAS), which computes a score based on the prediction similarity across multiple augmented views. We compare CAS against random skipping and established OOD detection methods [35, 38] under representative TTA frameworks [6, 10, 48, 49]. Overall, CAS achieves an AUC of around 90%. More importantly, it preserves and even improves the accuracy of full adaptation while skipping 85% of adaptations on ImageNet, its variants, and multiple fine-grained benchmarks. The result of ImageNet is shown in Fig. 1 (b). Beyond classification accuracy, CAS also maintains calibration performance in calibration-oriented TTA methods [46, 64]. Our contributions are summarized as follows:

- We introduce selective adaptation, an underexplored direction to improve TTA efficiency by detecting whether a test sample can benefit from adaptation.
- We provide Cross-Augmentation Similarity (CAS), a simple baseline based on prediction similarity across test-time augmented views.
- Extensive experiments validate that CAS maintains and even improves TTA performance, providing a baseline for future research to advance.

2 Related Work

Test-time adaptation (TTA). TTA aims to mitigate performance degradation caused by distribution shifts by adapting models to unlabeled test data [33, 47]. TTA approaches can be broadly categorized into two paradigms based on how they process test data. Online TTA [53, 55, 59, 66, 67, 74] processes streaming data and updates model parameters by leveraging historical knowledge from previous test samples. In contrast, episodic TTA, such as MEMO [70] and TTT [52], treats each test sample independently, making adaptation more challenging. Throughout this work, we focus exclusively on the episodic paradigm.

With the rise of VLMs [2, 26, 42], increasing attention has been devoted to applying TTA [46, 48, 49] to CLIP [42]. A line of work explores training-based TTA methods [13, 31, 49], which adapt models at test time by optimizing a subset of parameters using unlabeled test samples. The pioneering work TPT [49] adapts the model by optimizing learnable prompts through entropy minimization with confidence selection. Alternatively, another line of work explores training-free paradigms, such as ZERO [10], MTA [68], and TPS [51], which enable fast test-time adaptation without requiring gradient updates. In addition to accuracy improvement, recent studies have also explored other aspects of performance, including calibration [1, 46, 64] and adversarial robustness [48, 56, 60].

Efficiency in TTA. In addition to improving adaptation performance, several studies [6, 27, 40, 75] focus on enhancing efficiency. In episodic TTA, most existing work [6, 25] focuses on making the optimization process more efficient. TTL [25] improves efficiency by optimizing low-rank adapters and STS [6] adapts only a small number of parameters. Instead of refining the optimization algorithm, we propose selective skipping, which identifies and skips ineffective adaptations to improve efficiency without sacrificing performance. Notably, a line of work in online TTA [27, 40, 71, 75] has also explored efficiency improvements. For example, EATA [40] improves efficiency by selecting informative samples for adaptation based on entropy, while other methods reduce computational overhead via key–value cache retrieval [7, 27, 71, 75]. These approaches differ fundamentally from our selective adaptation approach.

Out-of-distribution (OOD) detection and selective classification. OOD detection [17, 20, 38, 61] aims to identify test samples that differ from the training distribution to ensure model reliability. Early work initially introduced MSP [22] for detecting misclassified or OOD samples, followed by Energy [35], MCM [38] and Doctor [17]. We leverage the scoring functions provided by these methods as a comparison strategy and employ AUC to evaluate the effectiveness of our selective adaptation baseline. Selective classification [5, 14, 15, 32], also known as classification with a rejection option, is a machine learning framework that allows a model to abstain from making a prediction when it is uncertain. While selective classification typically abstains from making predictions after inference [16], selective adaptation introduced in this paper instead focuses on rejecting samples before performing adaptation.

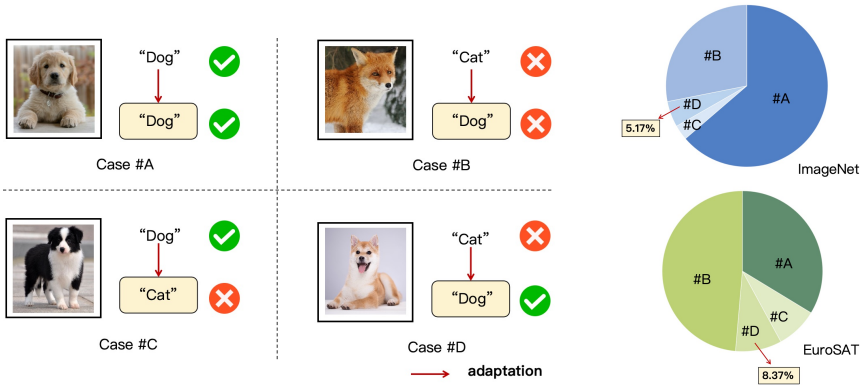
3 Method

3.1 Preliminaries

CLIP [42] is a widely used VLM due to its strong zero-shot generalization capability. It consists of two parts, a visual encoder $f_v(\cdot)$ and a text encoder $f_t(\cdot)$. For a K-class classification task with a label space $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$, let x denote an input image and $y \in \mathcal{Y}$ denote its ground-truth label. The visual encoder extracts visual features from the input image $v = f_v(x)$. For the text encoder, each class label $y_k \in \mathcal{Y}$ is converted into a textual prompt c_k (e.g., using a template such as "a photo of a [class]") and encoded into a textual feature $t_k = f_t(c_k)$. The prediction probability is then computed as

$$p^k(x) = p(y = k \mid x) = \frac{\exp\left(\cos(v, t_k)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos(v, t_j)/\tau\right)}, \tag{1}$$

where τ is a temperature parameter and $\cos(\cdot, \cdot)$ denotes cosine similarity. Given a test image x, $y_{\text{zs}}(x) = \arg\max_k \; p^k_{\text{zs}}(x)$ and $y_{\text{adapt}}(x) = \arg\max_k \; p^k_{\text{adapt}}(x)$, where $p_{\text{zs}}(x)$ and $p_{\text{adapt}}(x)$ denote the zero-shot and adapted probability vectors, and $y_{\text{zs}}(x)$ and $y_{\text{adapt}}(x)$ are their corresponding predicted labels.



(a) Four Adaptation Cases **(b)** Distribution of Cases

Fig. 2: (a) We categorize the adaptation process into four cases (A–D). Checkmarks and crosses indicate prediction correctness, with zero-shot prediction shown above and TTA prediction shown below. The arrow denotes the adaptation process. (b) Proportion of each case on ImageNet and EuroSAT under TPT [49] framework. The percentage of the beneficial-only case is highlighted.

3.2 Problem Formulation: Selective Adaptation

Motivation. Consistent with ZERO [10], we find that a large proportion of adaptations are negligible or even harmful. As illustrated in Fig. 2 (a), the outcomes of adaptation can be categorized into four cases based on prediction changes between the zero-shot and adapted models. Based on their impact on performance, we further group these cases into three types: negligible adaptation, harmful adaptation, and beneficial adaptation.

- **Harmful adaptation.** The zero-shot prediction is correct, but becomes incorrect after adaptation$(y_{\text{zs}}(x) = y, y_{\text{adapt}}(x) \neq y)$. These adaptations will reduce the overall performance.
- **Negligible adaptation.** The zero-shot prediction remains unchanged after adaptation $(y_{\text{zs}}(x) = y,\ y_{\text{adapt}}(x) = y)$ *or* $(y_{\text{zs}}(x) \neq y,\ y_{\text{adapt}}(x) \neq y)$. Adapting these adaptations incurs unnecessary computational overhead.
- **Beneficial adaptation.** The zero-shot prediction is incorrect but is adapted to correct successfully $(y_{\text{zs}}(x) \neq y,\ y_{\text{adapt}}(x) = y)$. These adaptations are the only ones to improve the performance.

Notably, negligible and harmful adaptations dominate the test set in many datasets. Fig. 2 (b) further shows that these cases account for over 90% of test samples on datasets such as ImageNet and EuroSAT, highlighting the importance of identifying and skipping ineffective adaptations.

Formulation. We formulate the selective adaptation problem as a binary detection task. Given a test sample x, we decide whether to perform adaptation or skip it. Following the paradigm used in OOD detection, we define a decision

function $G_\gamma(x)$ based on a scoring function $\alpha(x)$ and a threshold γ:

$$G_\gamma(x) = \begin{cases} \text{Not Adapt} & \text{if } \alpha(x) \geq \gamma \\ \text{Adapt} & \text{otherwise} \end{cases} \tag{2}$$

where γ is designed to control the trade-off between performance and efficiency, and $\alpha(x)$ is defined based on prediction similarity across augmented views. A larger $\alpha(x)$ indicates a higher likelihood of skipping adaptation for the input image x. Given a fixed threshold γ, a unique skip ratio s is determined, where s denotes the proportion of test samples for which adaptation is skipped.

Evaluation metrics. To evaluate the selective skipping strategy comprehensively, we introduce the following metrics:

- **Area under the ROC curve (AUC).** We formulate the identification of ineffective adaptations as a binary detection task. AUC [8,11] measures how well a scoring function distinguishes ineffective adaptations from beneficial ones. An AUC of 0.5 corresponds to random guessing, while higher values indicate stronger discriminative capability.
- **Accuracy expectation with triangular prior (AEP).** Since the optimal skip ratio s^* may vary across deployment scenarios, we propose a metric to evaluate overall performance by computing the expected accuracy under a prior distribution of s. Specifically, we adopt a triangular prior with probability density function $f(s) = 2(1-s)$ for $s \in [0,1]$. As s increases, efficiency gains become more significant, and slight accuracy degradation becomes more acceptable. Therefore, performance is assigned a lower weight at larger skip ratios. The metric is defined as

$$\text{AEP} = \int_0^1 \text{acc}(s) \cdot 2(1-s)\, ds, \tag{3}$$

  where $\text{acc}(s)$ denotes the model accuracy under a skip ratio s. By integrating the skip-accuracy curve, AEP provides a comprehensive evaluation of performance under different computational budgets. Notably, $\text{acc}(s)$ can be replaced with other performance metrics (e.g., ECE [18]) to evaluate different aspects of model behavior.

3.3 Cross-Augmentation Similarity as a Simple Baseline

Test-time augmentation [28,45], derived from data augmentation techniques [63], applies random transformations to test samples during inference. It has been widely adopted in TTA methods such as MEMO [70] and TPT [49]. Specifically, TPT [49] generates $(N-1)$ augmented views for a test image x using AugMix [23]. Let $\{\mathcal{A}_i(x)\}_{i=0}^{N-1}$ denote the augmented views, where $\mathcal{A}_0(x)$ is the original image. A cutoff percentile $\rho \in [0,1]$ is applied over the N augmented views to select high-confidence samples. Views whose prediction entropy is lower than the threshold β are retained to form the high-confidence set S:

$$S = \{i \mid H(p_{\text{zs}}(\mathcal{A}_i(x))) \leq \beta,\ i \in [0, N-1]\}, \tag{4}$$

Fig. 3: Pipeline of the proposed selective adaptation framework. Augmented views and class prompts are encoded to obtain zero-shot predictions. CAS measures cross-view prediction agreement. Samples with low CAS scores undergo adaptation, while those with high scores directly use zero-shot predictions, enabling efficient adaptation.

where β represents the ρ-percentile entropy threshold, and $H(\cdot)$ denotes Shannon entropy. To promote cross-view consistency, TPT optimizes textual prompts by minimizing the entropy of the averaged predictions over the selected set S:

$$L_{\text{TPT}}(x) = H\left(\frac{1}{|S|}\sum_{i\in S} p_{\text{zs}}(\mathcal{A}_i(x))\right), \tag{5}$$

The adapted prediction $p_{\text{adapt}}(x)$ is then obtained from the original view $\mathcal{A}_0(x)$ using the updated model. However, when predictions in S are consistent with $p_{\text{zs}}(\mathcal{A}_0(x))$, they provide little informative supervision for model updates. We therefore argue that the similarity among selected predictions is closely related to the effectiveness of subsequent adaptation. To measure this, we define a prediction consistency score $\alpha_{\text{Con}}(x) = \sum_{i\in S}\mathbb{I}(p_{\text{zs}}(\mathcal{A}_i(x)) = p_{\text{zs}}(\mathcal{A}_0(x)))$, which measures how many augmented views produce the same prediction as the original view. Higher consistency suggests that the prediction is already stable across different views, implying limited benefit from further adaptation.

Hard prediction consistency can be improved by cross-augmentation similarity. Samples may share the same predicted label while exhibiting different probability distributions. In such cases, adaptation may still be beneficial and should therefore not be skipped. To address this issue, we introduce CAS, a scoring function for selective adaptation that jointly considers prediction consistency and distribution similarity:

$$\alpha_{\text{CAS}}(x) = \sum_{i\in S} \widetilde{\cos}(p_{\text{zs}}(\mathcal{A}_i(x)), p_{\text{zs}}(\mathcal{A}_0(x))) \cdot \mathbb{I}(y_{\text{zs}}(\mathcal{A}_i(x)) = y_{\text{zs}}(\mathcal{A}_0(x))). \tag{6}$$

For each augmented view $\mathcal{A}_i(x)$, a reweighting factor is computed via cosine similarity and normalized across selected views. Specifically, letting p_i and p_0 denote $p_{\text{zs}}(\mathcal{A}_i(x))$ and $p_{\text{zs}}(\mathcal{A}_0(x))$ respectively, the normalized similarity is given by

Table 1: Performance comparison of different selection strategies under various TTA methods with ViT-B/16. We report AUC and AEP on ImageNet and its variants. The best results are highlighted in **bold**.

Method	Strategy	ImageNet		ImageNet-A		ImageNet-V		ImageNet-R		ImageNet-K		Avg.	
		AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑
TPT [49]	Random	50.77	68.19	51.08	52.54	51.38	62.62	48.54	75.97	49.46	47.23	50.25	61.31
	Energy [35]	57.27	68.41	52.50	52.70	57.70	62.95	64.31	76.64	54.04	47.41	57.16	61.62
	MCM [38]	63.82	68.50	53.73	52.73	61.44	62.95	71.09	76.77	54.33	47.36	60.88	61.66
	CAS	**91.10**	**69.00**	**88.62**	**54.70**	**89.69**	**63.45**	**93.45**	**77.14**	**84.24**	**47.93**	**89.42**	**62.44**
R-TPT [47]	Random	50.63	68.50	50.67	54.55	51.71	63.09	48.27	75.87	49.85	47.14	50.23	61.83
	Energy [35]	58.17	68.78	52.22	54.82	58.69	63.45	64.03	76.56	54.11	47.35	57.44	62.19
	MCM [38]	65.35	68.90	53.33	54.75	63.44	63.46	71.37	76.68	55.17	47.27	61.73	62.21
	CAS	**91.54**	**69.45**	**89.77**	**57.76**	**89.55**	**64.03**	**93.52**	**77.13**	**84.86**	**48.02**	**89.85**	**63.28**
STS [6]	Random	50.23	68.17	50.84	57.21	51.41	63.17	48.30	75.94	50.11	47.45	50.18	62.39
	Energy [35]	58.21	68.41	53.69	57.78	59.03	63.58	64.71	76.78	53.78	47.55	57.88	62.82
	MCM [38]	65.62	68.45	55.68	57.75	64.29	63.55	71.67	76.78	54.17	47.38	62.29	62.78
	CAS	**94.51**	**68.89**	**91.02**	**61.23**	**93.03**	**64.14**	**94.96**	**77.11**	**88.10**	**48.14**	**92.32**	**63.90**
ZERO [10]	Random	51.08	68.57	50.71	55.92	51.42	63.22	48.59	76.10	49.80	47.69	50.32	62.30
	Energy [35]	57.04	68.72	51.89	56.29	57.66	63.53	63.79	76.78	54.09	47.83	56.89	62.63
	MCM [38]	64.22	68.83	52.92	56.16	62.71	63.57	70.64	76.86	54.82	47.74	61.06	62.63
	CAS	**94.58**	**69.36**	**91.71**	**59.58**	**93.35**	**64.22**	**95.05**	**77.27**	**88.41**	**48.50**	**92.62**	**63.79**

$\widetilde{\cos}(p_i, p_0) = \frac{\cos(p_i,p_0)}{\sum_{j\in S}\cos(p_j,p_0)}$. As illustrated in Fig. 3, augmentations with higher consistency with the original prediction are assigned larger weights, thereby contributing more to the final score. The pseudo-code is provided in the Appendix.

4 Experiment

4.1 Experimental Setup

Datasets. To comprehensively evaluate the efficiency and performance of CAS, we conduct experiments across a range of benchmarks, including ImageNet and its variants, as well as fine-grained datasets. We first use ImageNet [9] and its four variants: ImageNet-A (natural adversarial examples) [24], ImageNet-V (re-collected images) [43], ImageNet-R (artistic renditions) [21], and ImageNet-K (sketch-style images with domain shifts) [54]. We further evaluate our method on fine-grained datasets to assess cross-domain generalization, including Flowers102 [39], DTD [4], Pets [41], UCF101 [50], Caltech101 [12], Aircraft [37], EuroSAT [19], Cars [30], Food101 [3], and SUN397 [58]. In the episodic TTA setting, no training data is available at test time, and all experiments are conducted strictly in a zero-shot manner.

Baselines. To validate the generalizability of our method, we integrate CAS into four representative TTA methods: TPT [49], ZERO [10], R-TPT [48], and STS [6]. We compare CAS with three sample selection strategies (Random Skipping, Energy [35], and MCM [38]) to demonstrate its effectiveness. Random skipping serves as a lower-bound baseline for comparison. As representative OOD detection approaches, Energy [35] utilizes a logit-based energy score, while

Table 2: Performance comparison of different selection strategies under various TTA methods with ViT-B/16. We report AUC and AEP on fine-grained datasets. The best results are highlighted in **bold**.

Method	Strategy	Metric	Flow.	DTD	Pets	UCF	Cal.	Air.	Euro.	Cars	Food	SUN	Avg.
TPT [49]	Random	AUC	49.04	49.06	55.59	49.33	46.03	53.72	49.55	49.94	50.63	50.12	50.30
		AEP	68.23	45.98	**87.60**	67.08	94.12	**23.89**	42.55	66.28	84.35	64.48	64.46
	Energy [35]	AUC	63.48	53.52	69.46	60.01	63.46	46.97	75.99	54.58	69.82	59.56	61.69
		AEP	68.42	46.26	87.36	67.62	**94.39**	23.36	**43.97**	66.28	84.54	64.95	64.72
	MCM [38]	AUC	68.47	69.48	75.21	65.47	73.69	47.83	70.17	63.01	79.74	65.92	67.90
		AEP	68.53	46.78	87.30	67.64	94.33	23.09	43.38	66.47	84.60	65.08	64.72
	CAS	AUC	**89.64**	**86.47**	**95.61**	**89.67**	**94.42**	**65.06**	**82.57**	**89.45**	**96.50**	**90.34**	**87.97**
		AEP	**68.65**	**47.25**	87.30	**68.09**	94.22	23.86	43.19	**66.68**	**84.68**	**65.58**	**64.95**
R-TPT [47]	Random	AUC	47.42	50.76	53.97	49.31	49.28	55.72	49.08	50.93	50.50	50.10	50.71
		AEP	67.91	45.53	**87.38**	66.66	94.01	**24.42**	36.84	66.47	84.06	64.58	63.79
	Energy [35]	AUC	62.55	56.79	68.78	59.34	62.93	46.64	69.58	53.17	70.36	59.60	60.97
		AEP	68.14	45.72	87.02	67.15	**94.26**	23.67	38.26	66.29	84.20	65.05	63.98
	MCM [38]	AUC	66.63	69.75	73.97	65.86	73.75	48.17	65.63	63.70	80.36	66.48	67.43
		AEP	68.25	46.10	86.94	67.12	94.14	23.55	**37.49**	66.54	84.19	65.17	63.95
	CAS	AUC	**89.14**	**87.20**	**97.44**	**89.31**	**95.35**	**64.73**	**78.31**	**90.12**	**96.47**	**90.59**	**87.87**
		AEP	**68.48**	**46.66**	86.92	**67.52**	94.01	24.20	37.26	**66.88**	**84.26**	**65.71**	**64.19**
STS [6]	Random	AUC	49.91	49.81	52.74	48.41	51.65	54.23	50.33	51.31	50.16	50.18	50.87
		AEP	**66.37**	45.40	**87.14**	65.93	93.87	**24.59**	39.41	66.82	**83.34**	64.12	63.70
	Energy [35]	AUC	60.70	58.16	68.22	57.45	63.41	47.38	69.11	53.17	70.39	59.50	60.75
		AEP	66.10	45.79	86.71	66.41	**94.05**	24.32	**39.86**	66.63	83.31	64.47	63.77
	MCM [38]	AUC	64.91	69.78	74.22	65.83	76.92	48.47	66.03	63.23	79.69	66.54	67.56
		AEP	66.05	45.96	86.60	66.32	93.93	24.18	39.57	66.91	83.17	64.54	63.72
	CAS	AUC	**93.07**	**89.97**	**97.68**	**93.24**	**98.94**	**76.44**	**81.97**	**92.29**	**96.85**	**93.70**	**91.42**
		AEP	66.03	**46.18**	86.49	**66.59**	93.63	24.56	39.26	**67.20**	83.17	**64.88**	**63.80**
ZERO [10]	Random	AUC	46.49	50.30	54.13	47.44	49.82	56.01	49.12	50.85	50.87	49.98	50.50
		AEP	66.80	44.96	**87.68**	65.76	93.93	**24.90**	38.73	66.89	83.71	64.59	63.80
	Energy [35]	AUC	56.72	56.22	67.86	57.99	66.72	47.31	69.19	53.62	69.98	59.01	60.46
		AEP	66.71	45.16	87.44	66.26	**94.23**	24.36	**39.17**	66.71	**83.76**	65.02	63.88
	MCM [38]	AUC	62.91	67.47	73.80	65.89	79.23	48.10	64.73	62.39	79.73	65.93	67.02
		AEP	66.84	45.26	87.35	66.19	94.11	24.41	38.59	66.96	83.70	65.15	63.86
	CAS	AUC	**91.95**	**90.93**	**97.58**	**93.37**	**98.29**	**76.71**	**81.48**	**91.60**	**96.87**	**93.75**	**91.25**
		AEP	**66.99**	**45.46**	87.30	**66.44**	93.90	24.81	38.69	**67.32**	83.73	**65.58**	**64.02**

MCM [38] measures confidence by evaluating the alignment between visual features and textual concepts.

Metrics. To comprehensively evaluate the effectiveness of selective adaptation, we introduce two complementary metrics that assess both detection quality and overall performance. AUC measures the selection strategy's ability to distinguish between beneficial and ineffective adaptations, independent of the decision threshold. Meanwhile, AEP evaluates the strategy's performance across varying skip ratios. It computes the expected accuracy under a predefined prior distribution of skip ratios (e.g., a triangular prior), thereby reflecting the overall efficiency–accuracy trade-off of the selection strategy.

Table 3: Performance comparison of different selection strategies across various TTA methods with ViT-B/16. We report AEP, ECE expectation with triangular prior (EEP), and AUC on ImageNet and its variants.

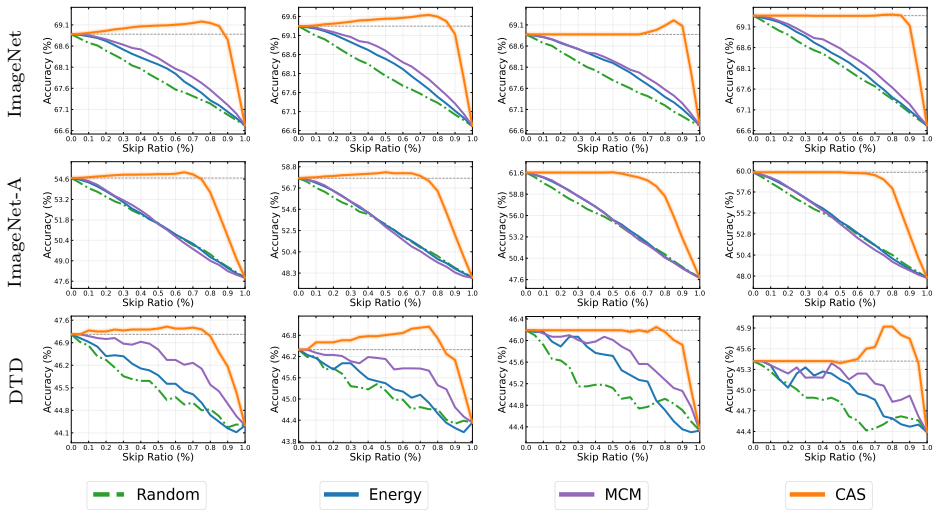
Method	Strategy	ImageNet			ImageNet-A			ImageNet-V			ImageNet-R			ImageNet-K			Avg.		
		AEP↑	EEP↓	AUC↑	AEP↑	EEP↓	AUC↑	AEP↑	EEP↓	AUC↑	AEP↑	EEP↓	AUC↑	AEP↑	EEP↓	AUC↑	AEP↑	EEP↓	AUC↑
TPT [49]	Random	68.19	7.40	50.77	52.54	12.62	51.08	62.62	8.52	51.38	75.97	**3.25**	48.54	47.23	12.07	49.46	61.31	8.77	50.25
	Energy [35]	68.41	7.68	57.27	52.70	13.23	52.50	62.95	9.01	57.70	76.64	3.27	64.31	47.41	12.12	54.04	61.62	9.06	57.16
	MCM [38]	68.50	7.91	63.82	52.73	13.66	53.73	62.95	9.15	61.44	76.77	3.54	71.09	47.36	12.17	54.33	61.66	9.29	60.88
	CAS	**69.00**	**7.33**	**91.10**	**54.70**	**12.02**	**88.62**	**63.45**	**8.54**	**89.69**	**77.14**	3.89	**93.45**	**47.93**	**11.24**	**84.24**	**62.44**	**8.60**	**89.42**
C-TPT [64]	Random	67.88	**3.88**	50.83	50.14	7.95	50.04	62.03	**5.02**	51.21	75.20	2.05	49.29	47.03	8.34	50.57	60.46	5.45	50.39
	Energy [35]	68.14	4.26	58.26	50.66	8.20	55.21	62.31	5.42	60.21	75.63	**1.79**	64.36	47.15	8.61	54.40	60.78	5.66	58.49
	MCM [38]	68.21	4.31	64.98	50.63	8.56	56.72	62.28	5.48	63.78	75.73	1.80	71.42	47.16	8.62	55.57	60.80	5.75	62.49
	CAS	**68.59**	4.01	**86.50**	**51.74**	**7.53**	**84.29**	**62.68**	5.05	**86.40**	**76.02**	2.91	**90.22**	**47.60**	**7.67**	**80.95**	**61.33**	**5.43**	**85.67**
O-TPT [46]	Random	67.18	**1.95**	51.35	48.11	7.22	51.65	61.32	3.02	52.08	74.02	3.87	50.15	46.51	5.36	50.96	59.43	**4.28**	51.24
	Energy [35]	67.28	2.05	57.35	48.26	7.05	55.41	61.35	**2.97**	57.80	74.02	**3.81**	62.41	46.49	5.59	52.00	59.48	4.29	56.99
	MCM [38]	67.35	2.03	63.50	48.22	7.18	57.28	61.32	**2.97**	61.37	74.10	3.83	69.81	46.50	5.64	53.12	59.50	4.33	61.02
	CAS	**67.70**	2.25	**76.05**	**49.29**	**6.57**	**78.18**	**61.71**	3.16	**75.91**	**74.54**	4.24	**83.31**	**46.93**	**5.27**	**69.70**	**60.03**	4.30	**76.63**

Implementation details. We use CLIP-ViT-B/16 [42] as the backbone model and follow the standard TPT setting. The text prompt is initialized with the template "a photo of a". For each image, we generate $N = 64$ augmented views via AugMix [23]. The confidence threshold ρ is set to 0.1 and kept fixed across all datasets. Experiments are conducted with multiple random seeds. All baseline results are reproduced following the previous benchmark [47].

4.2 Results

Performance on ImageNet and its variants. We evaluate the performance of CAS on ImageNet and its variants. As demonstrated in Table 1, CAS consistently outperforms all competing selection strategies across diverse TTA methods [6, 10, 48, 49] with respect to both AEP and AUC. Specifically, when integrated into the ZERO [10], CAS achieves state-of-the-art performance on ImageNet, attaining an AEP of 69.36% and an AUC of 94.58%. The latter represents an outperformance of 30.36% over the second-best baseline, MCM [38]. On ImageNet-A, CAS improves the AEP of R-TPT [48] to 57.76%, validating its superior ability in filtering out ineffective adaptations. Across the evaluated datasets, CAS maintains an AUC around 90%, peaking at 92.62% when combined with ZERO [10]. These results show that CAS effectively distinguishes beneficial from harmful adaptations, achieving a good balance between efficiency and accuracy under distribution shifts.

Performance on fine-grained datasets. We evaluate CAS on fine-grained benchmarks, with results summarized in Table 2. Across all evaluated strategies, CAS consistently delivers the highest average performance. Under the TPT framework [49], CAS achieves a peak average AUC of 87.97% and an AEP of 64.95%. It outperforms MCM [38] and Energy [35] by margins of 20.07% and 26.28% in AUC, respectively, demonstrating a superior ability to identify ineffective adaptations. Specifically, when integrated into ZERO [10], CAS demonstrates highly competitive results, achieving an AUC of 98.29% on Caltech101 and 96.87% on Food101.



(a) TPT [49] (b) R-TPT [48] (c) STS [6] (d) ZERO [10]

Fig. 4: Accuracy (%) versus skip ratio (%) under different skipping strategies across four TTA methods on ImageNet, ImageNet-A, and DTD. CAS consistently maintains higher accuracy even when skipping a large proportion of samples. Compared with other strategies, CAS demonstrates a superior efficiency–accuracy trade-off.

4.3 Impact on Calibration

While our main results focus on TTA methods [6,10,48,49] that aim to improve classification accuracy, several prior works emphasize calibration, including C-TPT [64] and O-TPT [46]. To evaluate the generalization of our approach, we further incorporate CAS into these calibration-oriented methods. We additionally report the ECE expectation with a triangular prior (EEP), which is computed in the same manner as AEP in Eq. (3). The results are summarized in Table 3. Across all methods, CAS consistently achieves the highest AEP and AUC, while maintaining lower EEP. Under TPT [49], CAS maintains the lowest average EEP of 8.60% across ImageNet and its variants. This indicates that selective skipping guided by CAS does not amplify overconfidence or destabilize prediction margins. For calibration-oriented methods such as C-TPT [64] and O-TPT [46], CAS continues to generalize effectively. In C-TPT [64], CAS achieves the highest average AEP of 61.33% and AUC of 85.67% while preserving competitive calibration performance with an EEP of 5.43%. Similarly, under O-TPT [46], CAS yields the best AEP of 60.03% and AUC of 76.63%, while the other three show comparable performance in EEP. Overall, these results suggest that our method is not limited to accuracy-oriented TTA methods. It also generalizes effectively to calibration-oriented methods, consistently improving robustness and reliability without sacrificing calibration performance.

Table 4: Ablation study of CAS across various TTA methods on both ImageNet and its variants and fine-grained benchmarks, evaluated using AUC and AEP.

Dataset	Strategy	TPT [49]		R-TPT [47]		STS [6]		ZERO [10]	
		AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑
ImageNet & its variants	MCM [38]	60.88	61.66	61.73	62.21	62.29	62.78	61.06	62.63
	Similarity	82.84	62.31	85.50	63.17	88.41	63.84	87.78	63.73
	Consistency	89.14	62.44	**89.89**	**63.28**	**92.65**	**63.90**	92.59	**63.79**
	CAS	**89.42**	**62.44**	89.85	**63.28**	92.32	**63.90**	**92.62**	**63.79**
Fine-grained	MCM [38]	81.08	64.70	80.70	63.95	79.54	63.70	79.60	63.86
	Similarity	81.04	64.81	81.44	63.93	83.61	63.60	82.91	63.79
	Consistency	87.65	64.93	87.62	64.13	91.27	**63.84**	90.82	63.97
	CAS	**87.97**	**64.95**	**87.87**	**64.19**	**91.42**	63.80	**91.25**	**64.02**

Table 5: Accuracy (%) of TTA methods on ImageNet with ViT-B/16 at an 85% skip ratio. Total inference time is reported in hours (h). CLIP* denotes CLIP with 64 augmentations.

Metric	CLIP*		TPT [49]		R-TPT [48]		STS [6]		ZERO [10]	
	Base	CAS	Base	CAS	Base	CAS	Base	CAS	Base	CAS
Acc. (%)	66.72	–	68.88	**69.03**	69.36	**69.43**	68.83	**69.20**	69.28	**69.32**
Time (h)	1.65	–	7.76	**2.55**	6.73	**2.40**	2.03	**1.68**	5.42	**4.96**
Speedup	–		**3.04**×		**2.81**×		**1.21**×		**1.09**×	

4.4 In-depth Analysis

Trade-off between efficiency and accuracy. To evaluate the trade-off between efficiency and performance, we show the accuracy–skip ratio curves for four TTA methods [6, 10, 48, 49], varying the skip ratio s from 0 to 1 with a step size of 0.05. The results demonstrate that CAS consistently maintains or improves accuracy compared to the full-adaptation baseline ($s = 0$), even when skipping up to 85% of samples. This suggests that standard TTA may over-adapt certain samples, while CAS selectively identifies and bypasses harmful adaptations, thereby mitigating performance degradation. Notably, R-TPT [48] achieves an accuracy above 69.60% when skipping nearly 80% of samples on the ImageNet dataset, while STS [6] reaches 69.10% with a skip ratio of 90%. In contrast, selection strategies based on Energy [35] and MCM [38] exhibit steep accuracy declines as the skip ratio increases. In general, Figure 4 shows that CAS can effectively improve computational overhead without sacrificing accuracy, serving as a good baseline for the selective adaptation problem.

Ablation study. CAS consists of augmentation prediction consistency and similarity reweighting, which correspond to the consistency and similarity components, respectively. To analyze their individual contributions, we evaluate each component independently as a scoring function. The results are reported in Table 4. Both consistency and similarity serve as effective skipping strategies compared to the classical MCM criterion [38], consistently yielding an improvement

Table 6: Performance comparison of different AEP metric functions under TPT [49] framework. We report AUC and AEP on ImageNet and its variants.

Method	Strategy	ImageNet	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-K	Avg.
$f(s) = 1$	Random	67.83	51.41	62.23	75.43	46.93	60.77
	Energy [35]	68.03	51.49	62.53	76.13	47.13	61.06
	MCM [38]	68.16	51.45	62.52	76.26	47.07	61.09
	CAS	**68.91**	**53.99**	**63.32**	**76.96**	**47.80**	**62.19**
$f(s) = 2s$	Random	67.47	50.27	61.85	74.88	46.63	60.22
	Energy [35]	67.66	50.28	62.12	75.62	46.84	60.50
	MCM [38]	67.82	50.17	62.10	75.76	46.78	60.53
	CAS	**68.81**	**53.28**	**63.19**	**76.78**	**47.66**	**61.94**

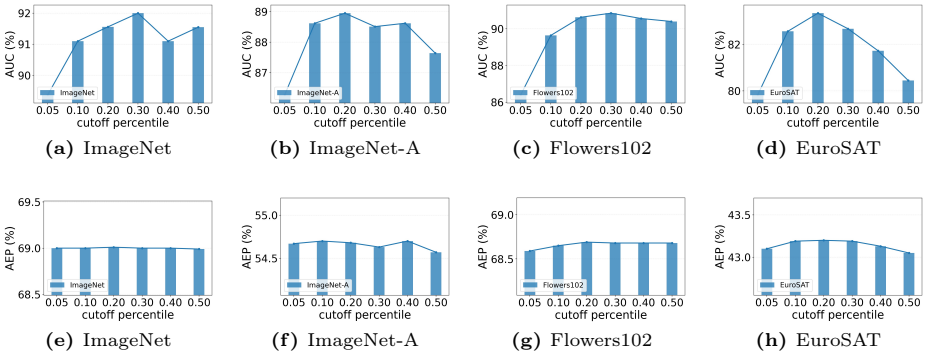
Table 7: Performance using resized crops/horizontal flip [10] as data augmentation strategy under ZERO [10] with ViT-B/16. We report AUC and AEP on ImageNet and its variants.

Method	Strategy	ImageNet		ImageNet-A		ImageNet-V		ImageNet-R		ImageNet-K		Avg.	
		AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑	AUC↑	AEP↑
ZERO [10]	Random	51.18	68.50	50.66	55.91	51.33	63.18	48.78	76.15	49.99	47.70	50.39	62.29
	Energy [35]	57.12	68.67	51.98	56.30	57.93	63.51	63.68	76.82	53.94	47.81	56.93	62.62
	MCM [38]	64.30	68.77	52.96	56.15	62.89	63.52	70.56	76.93	54.80	47.72	61.10	62.62
	CAS	**94.59**	**69.27**	**91.89**	**59.54**	**93.33**	**64.16**	**95.08**	**77.34**	**88.37**	**48.49**	**92.65**	**63.76**

of about 30% in AUC on ImageNet and its variants. However, the integrated CAS baseline achieves the most stable and competitive performance overall, attaining the best or near-best AUC and AEP across all methods. While consistency yields slightly higher gains than CAS in a few isolated cases, CAS demonstrates more stable improvements, particularly on fine-grained datasets. Notably, CAS improves the AUC from 90.82% to 91.25% under ZERO [10], further validating the effectiveness of combining both components.

Computation cost. To validate the efficiency of our baseline, we measure the total inference time of different TTA methods [6,10,48,49] with and without CAS, as shown in Table 5. Specifically, we report the total inference time on ImageNet using ViT-B/16. By integrating CAS, the overall adaptation time of TPT [49] decreases from 7.76 to 2.55 hours. This delivers a 3.04× speedup over full adaptation while marginally improving accuracy from 68.88% to 69.03%. For training-free TTA methods such as ZERO [10] and computationally efficient methods like STS [6], CAS further reduces computational overhead by approximately 20% and 10%, respectively, while simultaneously improving accuracy. These results demonstrate that CAS effectively accelerates the overall adaptation process without sacrificing TTA performance.

Different AEP measuring functions. We choose a decreasing linear function because lower skip ratios may be more preferred in real-world deployments as they sacrifice less performance. And $f(s) = 2(1 - s)$ was explicitly chosen because its integral evaluates exactly to 1. To verify that our method is not



(a) ImageNet **(b)** ImageNet-A **(c)** Flowers102 **(d)** EuroSAT

(e) ImageNet **(f)** ImageNet-A **(g)** Flowers102 **(h)** EuroSAT

Fig. 5: Sensitivity analysis of CAS under different cutoff percentile ρ ranging from 0.05 to 0.50 under TPT with ViT-B/16. (a)-(d) demonstrate the AUC performance on ImageNet, ImageNet-A, Flowers102, and EuroSAT, respectively, while (e)-(h) show the AEP results for the same datasets.

dependent on a specific AEP metric function, we further evaluate CAS with two alternative AEP functions, including $f(s) = 1$ and $f(s) = 2s$. As shown in Table 6, CAS consistently achieves the best performance under both alternative settings. Specifically, CAS obtains the highest average AEP of 62.19% with $f(s) = 1$ and 61.94% with $f(s) = 2s$, outperforming other strategies. These results indicate that CAS remains robust across different AEP measuring functions.

Different data augmentations. To examine whether CAS depends on a specific augmentation strategy, we replace the AugMix [23] augmentation with the random resized crops / horizontal flips used in ZERO [10]. As shown in Table 7, CAS still achieves the best performance on ImageNet and its variants. Specifically, CAS obtains the highest average AUC of 92.65% and average AEP of 63.76%, outperforming other strategies by a clear margin, indicating that CAS remains effective under different data augmentation settings.

Sensitivity to cutoff percentile ρ. To evaluate the sensitivity of CAS to cutoff percentile ρ, we vary it from 0.05 ($|S| = 3$) to 0.50 ($|S| = 32$) across four methods [6,10,48,49] . As illustrated in Figure 5, the AUC remains robust across different ρ. Initially, increasing the number of selected augmentations improves performance. For example, on ImageNet, increasing ρ by 20% improves AUC by about 1%. However, when ρ becomes too large, additional views with high entropy are introduced, which may harm performance and lead to slight degradation. Similar trends are observed on ImageNet-A and fine-grained datasets. In contrast, AEP varies only marginally across different ρ, indicating that larger ratios bring limited overall benefit. Notably, on EuroSAT, AEP even declines as ρ increases, suggesting that excessive augmentations may negatively affect performance. Based on these results, we set the cutoff percentile ρ to 0.1 as the default setting, as this configuration maintains high detection quality while minimizing computational overhead.

4.5 Case Study

For correct-to-correct samples, predictions are stable across augmentations, indicating that the model has learned robust and invariant representations. Conversely, wrong-to-wrong samples yield consistently incorrect predictions, suggesting stable but biased representations that augmentation alone cannot rectify. Meanwhile, wrong-to-correct samples lie near decision boundaries, where augmentations provide consistent corrective signals. In contrast, correct-to-wrong samples are overly sensitive: perturbations disrupt originally correct cues, leading to performance degradation. Overall, stable samples offer limited adaptation gains, boundary samples benefit the most from adaptation, and highly sensitive samples risk degradation from improper updates.

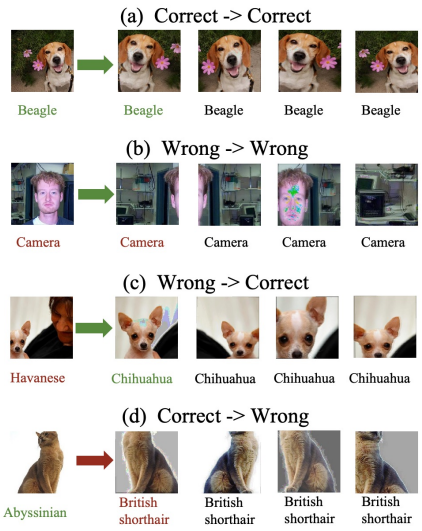


Fig. 6: Visualization of augmented views and their corresponding predictions.

5 Conclusion

While existing TTA methods generally prioritize overall performance gains, this paper shifts the focus toward adaptation efficiency at the per-sample level. Our main contribution is the introduction of a new selective adaptation problem, which aims to determine whether a given test sample should undergo adaptation or be skipped. We also introduce CAS as a simple baseline that maintains performance with reduced computational overhead, while improving end performance serves as an added benefit. We hope this work inspires the community to further investigate this problem and build upon our baseline. Additionally, we encourage the exploration of new directions, such as extending selective adaptation to tasks beyond image classification.

Acknowledgements

We thank Dr. Lijun Sheng for his critical discussions, and the anonymous reviewers for their constructive comments and helpful suggestions that improved this paper. This work was funded by the National Natural Science Foundation of China under Grants 62276256 and U2441251, Beijing Natural Science Foundation Z260008, and National Key Research and Development Program of China 2026ZD1500301.

References

1. Ahamed, S.A., Thanthrige, U.S.K.P.M., Rodrigo, R., Khan, M.H.: A-TPT: Angular diversity calibration properties for test-time prompt tuning of vision-language models. In: ICLR (2026)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS. pp. 23716–23736 (2022)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: ECCV. pp. 446–461 (2014)
4. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR. pp. 3606–3613 (2014)
5. Cortes, C., DeSalvo, G., Mohri, M.: Learning with rejection. In: ALT. pp. 67–82 (2016)
6. Dafnis, K.M., Metaxas, D.N.: Test-time spectrum-aware latent steering for zero-shot generalization in vision-language models. In: NeurIPS. pp. 151169–151194 (2025)
7. Dastmalchi, H., An, A., et al.: Etta: Efficient test-time adaptation for vision-language models through dynamic embedding updates. In: BMVC (2025)
8. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: ICML. pp. 233–240 (2006)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
10. Farina, M., Franchi, G., Iacca, G., Mancini, M., Ricci, E.: Frustratingly easy test-time adaptation of vision-language models. In: NeurIPS. pp. 129062–129093 (2024)
11. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**(8), 861–874 (2006)
12. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPRW. pp. 178–178 (2004)
13. Feng, C.M., Yu, K., Liu, Y., Khan, S., Zuo, W.: Diverse data augmentation with diffusions for effective test-time prompt tuning. In: ICCV. pp. 2704–2714 (2023)
14. Fisch, A., Jaakkola, T.S., Barzilay, R.: Calibrated selective classification. TMLR (2022)
15. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: NeurIPS. p. 4885–4894 (2017)
16. Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: ICML. pp. 2151–2159 (2019)
17. Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., Piantanida, P.: Doctor: A simple method for detecting misclassification errors. In: NeurIPS. pp. 5669–5681 (2021)
18. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330 (2017)
19. Helber, P., Bischke, B., Dengel, A., Borth, D.: Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In: IGARSS. pp. 204–207 (2018)
20. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML. pp. 8759–8773 (2022)

21. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: ICCV. pp. 8340–8349 (2021)
22. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
23. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: ICLR (2020)
24. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: CVPR. pp. 15262–15271 (2021)
25. Imam, R., Gani, H., Huzaifa, M., Nandakumar, K.: Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In: WACV. pp. 5449–5459 (2025)
26. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916 (2021)
27. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: CVPR. pp. 14162–14171 (2024)
28. Kim, I., Kim, Y., Kim, S.: Learning loss for test-time augmentation. In: NeurIPS. pp. 4163–4174 (2020)
29. Kim, Y., Cho, D., Han, K., Panda, P., Hong, S.: Domain adaptation without source data. IEEE Transactions on Artificial Intelligence **2**(6), 508–518 (2021)
30. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCVW. pp. 554–561 (2013)
31. Li, X., Zhang, D., Du, Z., Zhu, L., Chen, Z., Li, J.: Pataug: Augmentation of augmentation for test-time adaptation. In: ACM MM. pp. 5080–5089 (2025)
32. Liang, H., Peng, L., Sun, J.: Selective classification under distribution shifts. TMLR (2024)
33. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. IJCV **133**(1), 31–64 (2025)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. pp. 34892–34916 (2023)
35. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS. pp. 21464–21475 (2020)
36. Lu, S., Wang, Y., Sheng, L., He, L., Zheng, A., Liang, J.: Out-of-distribution detection: A task-oriented survey of recent advances. ACM Computing Surveys **58**(2), 1–39 (2025)
37. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
38. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: NeurIPS. pp. 35087–35102 (2022)
39. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP. pp. 722–729 (2008)
40. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: ICML. pp. 16888–16905 (2022)
41. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505 (2012)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)

43. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICML. pp. 5389–5400 (2019)
44. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV. pp. 213–226 (2010)
45. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: ICCV. pp. 1214–1223 (2021)
46. Sharifdeen, A., Munir, M.A., Baliah, S., Khan, S., Khan, M.H.: O-tpt: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In: CVPR. pp. 19942–19951 (2025)
47. Sheng, L., Liang, J., He, R., Wang, Z., Tan, T.: The illusion of progress? a critical look at test-time adaptation for vision-language models. In: NeurIPS (2025)
48. Sheng, L., Liang, J., Wang, Z., He, R.: R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In: CVPR. pp. 29958–29967 (2025)
49. 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. In: NeurIPS. pp. 14274–14289 (2022)
50. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
51. Sui, E., Wang, X., Yeung-Levy, S.: Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In: WACV. pp. 825–835 (2025)
52. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: ICML. pp. 9229–9248 (2020)
53. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
54. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: NeurIPS. pp. 10506–10518 (2019)
55. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022)
56. Wang, X., Chen, K., Zhang, J., Chen, J., Ma, X.: Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In: CVPR. pp. 19910–19920 (2025)
57. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: CVPR. pp. 10704–10713 (2023)
58. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. pp. 3485–3492 (2010)
59. Xiao, Z., Yan, S., Hong, J., Cai, J., Jiang, X., Hu, Y., Shen, J., Wang, C., Snoek, C.G.M.: Dynaprompt: Dynamic test-time prompt tuning. In: ICLR (2025)
60. Xing, S., Zhao, Z., Sebe, N.: Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In: CVPR. pp. 15172–15182 (2025)
61. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. IJCV **132**(12), 5635–5662 (2024)
62. Yang, P., Liang, J., Cao, J., He, R.: Auto: Adaptive outlier optimization for online test-time ood detection. arXiv preprint arXiv:2303.12267 (2023)
63. Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. In: NeurIPS. vol. 32 (2019)

64. Yoon, H.S., Yoon, E., Tee, J.T.J., Hasegawa-Johnson, M.A., Li, Y., Yoo, C.D.: C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In: ICLR (2024)
65. Yu, Y., Sheng, L., He, R., Liang, J.: Benchmarking test-time adaptation against distribution shifts in image classification. arXiv preprint arXiv:2307.03133 (2023)
66. Yu, Y., Sheng, L., He, R., Liang, J.: Stamp: Outlier-aware test-time adaptation with stable memory replay. In: ECCV. pp. 375–392 (2024)
67. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: CVPR. pp. 15922–15932 (2023)
68. Zanella, M., Ben Ayed, I.: On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In: CVPR. pp. 23783–23793 (2024)
69. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE TPAMI **46**(8), 5625–5644 (2024)
70. Zhang, M.M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. In: NeurIPS (2022)
71. Zhang, T., Wang, J., Guo, H., Dai, T., Chen, B., Xia, S.T.: Boostadapter: Improving vision-language test-time adaptation via regional bootstrapping. In: NeurIPS. pp. 67795–67825 (2024)
72. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
73. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)
74. 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: CVPR. pp. 29999–30009 (2025)
75. Zhou, S., Yin, M., Sun, L., Yang, S., Xie, D., Zhu, J.: Training-free test-time adaptation via shape and style guidance for vision-language models. In: NeurIPS. pp. 152968–152982 (2025)