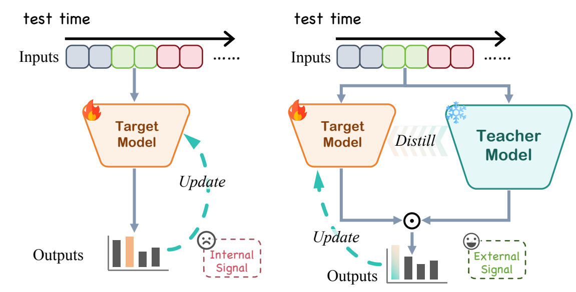
Test-Time Distillation for Continual Model Adaptation

Xiao Chen[1]*, Jiazhen Huang[1]*, Zhiming Liu[1],
Qinting Jiang[1], Fanding Huang[1], Jingyan Jiang[2]†, Zhi Wang[1]†
[1] Shenzhen International Graduate School, Tsinghua University [2] Shenzhen Technology University

Abstract

*Deep neural networks often suffer performance degradation upon deployment due to distribution shifts. Continual Test-Time Adaptation (CTTA) aims to address this issue in an unsupervised manner. However, existing methods that rely on self-supervision are prone to an inherent self-referential feedback loop that amplifies initial prediction errors, leading to model drift. We revisit this limitation and propose Test-Time Distillation (TTD), which reframes adaptation as a distillation process guided by a frozen Vision-Language Model (VLM) as an external signal. While promising, we find that direct distillation is fraught with two pitfalls: (1) the Generalist Trap, where the VLM's broad but non-specialized knowledge leads to suboptimal performance on specific tasks and shifts; and (2) the Entropy Bias, where naive model fusion techniques based on entropy fail due to the disparate calibration of heterogeneous models. These pitfalls highlight the need to build a robust supervisory signal and leverage it to guide the target model toward stable adaptation. Hence, we present **CoDiRe**, a **Co**ntinual **Di**stillation and **Re**ctification framework for TTD. CoDiRe first constructs a robust blended teacher by dynamically fusing the predictions of the VLM and the target model. Critically, it circumvents the Entropy Bias by leveraging Maximum Softmax Probability (MSP) as a more reliable confidence metric for weighting each model's expertise. Then applies an Optimal Transport-based rectification to further align predictions with the blended teacher, enabling continuous and stable adaptation. Extensive experiments show that CoDiRe outperforms state-of-the-art baselines, exceeding CoTTA by 10.55% with only 48% of its time cost on ImageNet-C. Project page is publicly available at https://github.com/walawalagoose/TTD.*



(a) Test-Time Adaptaion. (b) Test-Time Distillation.

Figure 1. **Comparison between TTA and TTD.** (a) TTA updates the source-pretrained target model solely based on the internal signals with a self-supervised loss. (b) TTD introduces a VLM as a teacher model to provide external signals during inference.

1. Introduction

Deep neural networks (DNNs) [52] frequently encounter deployment environments that deviate from their training distributions, leading to degraded performance and reliability. Test-Time Adaptation (TTA) [23, 48, 55, 76] therefore aims to align models to the target distribution on the fly, without access to source data or offline fine-tuning. Among these methods, Continual Test-Time Adaptation (CTTA) [56] enables sequential adaptation across evolving distribution shifts. The prevailing CTTA approach, from CoTTA [56] to more recent continual variants [3, 37, 66], is rooted in self-supervision. These methods rely exclusively on the model's own predictions as an *internal signal* to generate learning targets, typically through self-distillation.

However, we argue that this reliance on a self-referential signal constitutes a limitation. Under a significant domain shift, the model's initial predictions are inherently noisy and unreliable. Employing such outputs as supervisory signals creates a risky feedback loop: initial errors might be amplified rather than corrected, leading to gradual drift away from the optimum. The core problem of CTTA is therefore not merely adapting, but adapting reliably without reinforcing its own biases. This motivates our central question: *Can we leverage external knowledge to construct a stable and robust anchor signal that breaks this error accumulation cycle and guides the model towards a reliable optimization direction?*

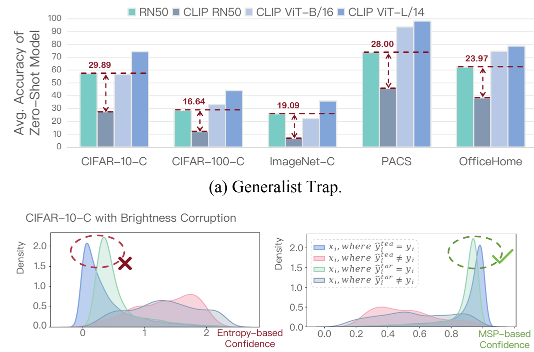
To identify such an external anchor, we leverage the vast open-world knowledge encapsulated by modern Vision-Language Models (VLMs) [7, 21, 67], with CLIP [46] as a representative exemplar. Pre-trained on web-scale image-text pairs, CLIP develops a rich semantic understanding that is orthogonal to the inductive biases of any single, task-

*Equal contribution.
†Co-corresponding authors. jiangjingyan@sztu.edu.cn; wangzhi@sz.tsinghua.edu.cn.

specific classifier. Therefore, CLIP is independent of the source training set and, crucially, is not influenced by the target domain shift affecting the target model. This makes CLIP an ideal candidate to serve as a signal source for adaptation. Therefore, we introduce a new paradigm, **Test-Time Distillation (TTD)**, which reframes adaptation as a distillation process guided by a frozen VLM teacher, shifting from self-referential correction to external guidance.

Although CLIP provides an appealing source of external supervision for stabilizing TTA, the central challenge lies in transforming its broad, open-world knowledge into a dependable and task-relevant signal for the target model. Although direct distillation from CLIP appears promising in principle, it often suffers from two critical pitfalls in practice: (1) *Generalist Trap*: While CLIP possesses broad semantic knowledge, its generalist nature makes it vulnerable under domain shifts and thus not always effective. As shown in Figure 2(a), CLIP consistently underperforms supervised classifiers with the same backbone on task-specific benchmarks, and struggles when facing certain shifts like common corruptions. Notably, scaling does not help much to bridge this gap. On ImageNet-C, even CLIP ViT-L/14 (304M parameters, more than $10\times$ larger than RN50 with 23M) only achieves about 10% performance gain over RN50 source-pretrained on ImageNet, and still trails the source-pretrained ViT-B/16 (as illustrated in Section 5.2). Recent studies [72, 74] have further revealed that scaling CLIP size is insufficient for reliable out-of-distribution (OOD) generalization. This suggests that neither CLIP alone nor the target model alone offers a sufficiently reliable supervisory signal. To escape this pitfall, we aim to apply certain model fusion techniques to integrate the knowledge from both models to form a more robust signal. (2) *Entropy Bias*: Determining a reliable fusion technique introduces a new challenge. Recent works have adopted entropy-based confidence [4, 25, 43, 51] as a proxy for model expertise and thus as weights for model merging, feature fusion, or ensembling. However, as illustrated in Figure 2(b), we find that heterogeneous models exhibit distinct entropy landscapes due to architectural, calibration, and training differences, making entropy inherently biased toward models with globally lower entropy and leading to skewed distillation targets. Hence, our core objective has shifted to: *how to build a robust supervisory signal from an effective fusion technique, and leverage it as teacher to guide the target model toward stable adaptation?*

To this end, we propose **CoDiRe**, a **Co**ntinual **Di**stillation and **Re**ctification framework for CTTA, which combines a frozen CLIP and the target model to construct a robust teacher signal for TTD. CoDiRe comprises two components: Distillation and Rectification. In Distillation, we interpolate the logits of CLIP and the target model to ensemble a robust blended teacher as a robust supervisory signal. It sidesteps Entropy Bias by utilizing confidence score based on Maximum Softmax Probability (MSP), which proves to be a more robust arbiter between heterogeneous models. This allows it to intelligently fuse the target model and VLM predictions into a high-quality signal. In Rectification, we further refine the target model by deriving a rectification matrix using optimal transport (OT) [8, 54] to align predictive mass with the target label geometry under constraints imposed by the blended teacher. We validate CoDiRe across diverse CTTA benchmarks and shift scenarios, demonstrating that it not only surpasses various TTA baselines based on target model, but also outperforms several variants based on CLIP. Our contributions are summarized as follows:



(a) Generalist Trap.

(b) Entropy Bias.

Figure 2. **Two pitfalls of Test-Time Distillation.** (a) *Generalist Trap.* Under domain shift, CLIP underperforms a supervised classifier with the same backbone. (b) *Entropy Bias.* Heterogeneous models exhibit inconsistent entropy distributions, inducing bias in entropy-based confidence.

- **New Paradigm**: We propose Test-Time Distillation, *first* introducing a VLM as a distillation teacher into CTTA.
- **Empirical Findings**: We identify two practical pitfalls under TTD paradigm: Generalist Trap and Entropy Bias.
- **Novel Methodology**: We develop a TTD method CoDiRe, which constructs a more reliable blended teacher and mitigates the impacts of the aforementioned pitfalls.
- **Extensive Experiments**: We demonstrate substantial and robust gains of our CoDiRe over state-of-the-art baselines across CTTA benchmarks, surpassing CoTTA by 10.55% while requiring only 48% of its time cost on ImageNet-C.

2. Related Works

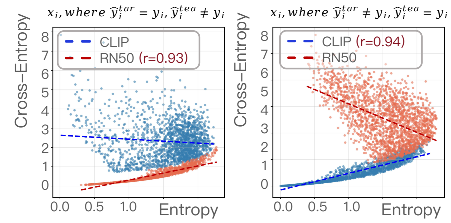
Test-Time Adaptation. TTA mitigates performance degradation on OOD data by fine-tuning a pretrained model during inference using mini-batches in an online manner [40, 64]. Tent [55] first introduced the concept of Fully Test-Time Adaptation, proposing entropy-based self-supervised optimization, which was subsequently adopted by follow-up work [5, 6, 11, 28, 41, 42, 68]. As the field has evolved, TTA has been extended to diverse modalities and tasks, including depth completion [44], action recognition [59], point cloud understanding [22], and time-series anomaly detection [26]. Recent works [18, 24, 49, 71] further extend TTA to VLMs for improved OOD generalization. Building on this foundation, CoTTA [56] introduced CTTA to handle continuously shifting test distributions, mitigating catastrophic forgetting via self-distillation with the source model as the teacher. Subsequent methods [3, 66] further advanced this line of work; however, whether employing entropy-based self-supervised optimization or self-distillation using the source model as the teacher, the supervisory signal is inherently internal, thereby imposing an intrinsic performance ceiling. Instead, we propose the TTD paradigm, which employs CLIP as a teacher to form a reliable and robust external signal.

Vision-Language Models (VLMs). VLMs learn joint visual-textual representations that enable open-vocabulary understanding and zero-shot transfer across diverse tasks. Early methods, such as bottom-up top-down attention [2], BAN [27], and MCAN [65], made notable progress on vision-language benchmarks, while recent models including BLIP [30], MiniGPT4 [77], and BERT-based architectures [38] further advanced multimodal reasoning. Among them, CLIP [46] popularized contrastive pretraining and has shown strong performance across 3D [70], video [35], and depth understanding [69]. In this work, we adopt CLIP as the teacher model, given its broad recognition and widespread use. However, CLIP can underperform on OOD data [24, 49], and its performance can lag behind that of a supervised classifier with the same backbone–a phenomenon we term the Generalist Trap. This motivates our attempt at constructing a more reliable teacher.

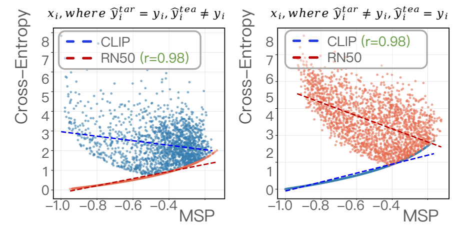
Knowledge Distillation. Knowledge distillation trains lightweight student models under the supervision of large pre-trained teachers, achieving notable success across tasks such as visual recognition [19, 33, 61] and multimodal representation learning [12, 32, 75]. To enable effective transfer, diverse paradigms have emerged, including feature imitation [58, 63], relational distillation [60, 62], and prompt distillation [34, 57]. CLIPPING [45] introduces a hierarchical alignment strategy that promotes student-centric adaptation for efficient assimilation of teacher knowledge while CLIP-KD [63] transfers knowledge by minimizing unimodal feature discrepancies between student and teacher. However, these works primarily focus on training a compact network via offline distillation. In contrast, we propose TTD, wherein the target model self-evolves within the test stream under guidance from CLIP, integrating rich world knowledge while preserving its own distributional sensitivity.

3. Pitfalls in Test-Time Distillation

We introduce TTD as a new paradigm, with CLIP as an external signal to construct the distillation teacher objective. Nevertheless, we identify two pitfalls in practice: the Generalist Trap precludes directly using CLIP as the distillation target, motivating the construction of a more reliable blended teacher; and the Entropy Bias invalidates entropy-based confidence as a cross-model expertise measure for fusion, thereby prompting the exploration for alternative solutions as weights to fuse the two models.



(a) Correlations between Entropy and Cross-Entropy.



(b) Correlations between MSP and Cross-Entropy.

Figure 3. **Correlations between different proxies and cross-entropy.** MSP-based confidence shows a substantially stronger correlation than entropy-based confidence, indicating that MSP is a more reliable proxy for cross-entropy.

Pitfall 1: Generalist Trap. TTD is proposed to address the limitations of self-distillation in existing CTTA methods [3, 56, 66], where reliance on the internal signals to generate learning targets constrains performance. A naive method is to directly adopt the output of CLIP as the teacher signal, but empirical evidence indicates this does not consistently improve the target model due to the Generalist Trap:

> ***Generalist Trap.*** *Under domain shift, CLIP underperforms a supervised classifier with the same backbone.*

As shown in Figure 2(a), on canonical domain generalization (PACS and OfficeHome) and synthetic corruption (CIFAR-10-C and ImageNet-C) benchmarks, CLIP RN50 consistently underperforms an RN50 classifier trained on the source domain. Moreover, CLIP particularly struggles on certain shifts like common corruptions (e.g., noise, blur). Even with a stronger backbone, CLIP ViT-B/16 remains inferior to RN50 on the two corruption datasets.

To remedy this pitfall, two avenues exist: upgrading the teacher model or constructing a more reliable distillation target. However, existing studies [72, 74] demonstrate that: 1) Proprietary and public VLMs such as LLaVA [36], BLIP-2 [31], and GPT-4-Turbo [1], despite often employing CLIP [46] as the vision encoder and possessing substantially more parameters, significantly underperform CLIP on standard image classification tasks; 2) Scaling up CLIP on OOD datasets does not bring remarkable benefit and sometimes can even hurt. Therefore, we pursue the latter path of constructing a more reliable distillation target called blended teacher. Inspired by Ensemble Learning [9, 14], we aim to simply fuse the output logits of CLIP and the target model to mitigate performance disparities.

Pitfall 2: Entropy Bias. The blended teacher is expected to provide a stronger supervisory signal by integrating the knowledge from the two models. Beyond merely averaging, an intuitive idea is to assign higher weights to the model that performs better on the current data. This requires a metric to measure each model's expertise. The cross-entropy (CE) between predictions and the ground truth naturally becomes an oracle choice, since a lower CE indicates higher prediction accuracy. However, it is impossible in the unsupervised TTA setting as labels are unavailable. A common workaround is to utilize some kinds of confidence measure as proxies for CE, as is widely done with entropy in ensembling [51], feature fusion [4, 78] and model merging [43]. However, our experiments reveal an Entropy Bias in TTD:

> ***Entropy Bias.*** *Heterogeneous models exhibit inconsistent entropy distributions, inducing bias in entropy-based confidence.*

As illustrated in Figure 2(b), the peak of CLIP's entropy distribution shows a clear deviation from that of the target model, causing entropy-based confidence to skew the fusion results toward models with globally lower entropy. Tuning the temperature parameter can adjust CLIP's logit sharpness, yet we avoid learning or hand-crafting a task-specific temperature for every scenario. Instead, we seek a more reliable metric in this setting with two properties: 1) effectiveness (strong correlation with ground-truth model expertise); and 2) comparability (distributional consistency across heterogeneous models without pronounced shifts).

Through experiments, we identify Maximum Softmax Probability (MSP) as a superior proxy for CE. In Figure 2(b), MSP-based confidence substantially mitigates inter-model confidence bias, providing a more faithful confidence metric than entropy. Furthermore, in Figure 3, we demonstrate the correlation plots between (MSP, CE) and (entropy, CE), focusing on cases where the target model and CLIP predictions disagree. The results show that under prediction conflicts, MSP-based confidence maintains a significantly higher correlation with cross-entropy than entropy. Consequently, an MSP-driven blended teacher better balances model predictions, yielding a more robust distillation target. We further provide a theoretical analysis motivated by an empirical MSP-accuracy binning experiment in the Supplementary Materials to show the effectiveness of this choice.

4. Methodology

The overview of our proposed CoDiRe is depicted in Figure 4. CoDiRe comprises two core components: Distillation (Section 4.2), which constructs a more reliable blended teacher for distillation; and Rectification (Section 4.3), which further refines the target model under constraints imposed by the blended teacher. In addition, we incorporate a widely adopted entropy loss and conduct a systematic study of reset mechanisms in the CTTA setting (Section 4.4).

4.1. Preliminaries

Given a pretrained classifier $f(\cdot)$ with parameters θ_0 trained on source data $(\mathcal{X}^S, \mathcal{Y}^S)$, we seek to enhance the performance of this target model during inference for a continually evolving target domain in an online manner, aided by a teacher model $\mathcal{F}(\cdot)$, such as CLIP. Unlabeled data $\mathcal{X}^T$, comprising K classes from the target domain, arrive sequentially and asynchronously, and the model has access only to the data available at the current time step. At time step t, the model receives a mini-batch of unlabeled test samples $\{x \mid x \in \mathcal{X}_t^T\}$, and the target model $f(\cdot)$ must adapt its parameters for future inputs, i.e., $\theta_t \rightarrow \theta_{t+1}$. The distribution $\mathcal{X}_t^T$ evolves continually over time. For a test sample x_i, the model produces logits $\mathbf{z}_i$, with post-softmax probabilities $p_i = \sigma(\mathbf{z}_i)$, where $\sigma(\cdot)$ denotes the softmax operator. Specifically, for x_i, we denote $\mathbf{z}_i^{\text{tar}} = f(x_i)$ and $\mathbf{z}_i^{\text{tea}} = \mathcal{F}(x_i)$. We form the blended teacher via a linear interpolation with weight λ_i:

$$\mathbf{z}_i^{\text{bt}} = \lambda_i \cdot \mathbf{z}_i^{\text{tea}} + (1 - \lambda_i) \cdot \mathbf{z}_i^{\text{tar}}, \tag{1}$$

and adopt p_i^{bt} as the inference prediction output, where $p_i^{\text{bt}} = \sigma(\mathbf{z}_i^{\text{bt}})$. To ensure balance and scale consistency, both logits are independently normalized by subtracting its LogSumExp (LSE) term and thus operate in the log-probability space.

4.2. Distillation

Section 3 shows that directly distilling from CLIP is suboptimal, and entropy-based interpolation of the target model and CLIP predictions is likewise ineffective. Empirical evidence indicates that MSP is a superior proxy for CE. To this end, we construct the blended teacher via MSP-based confidence and use the resulting distribution as the distillation target. The MSP-based confidence weight λ_i is computed as:

$$\lambda_i = \frac{\exp(\max(p_i^{\text{tea}}))}{\exp(\max(p_i^{\text{tea}})) + \exp(\max(p_i^{\text{tar}}))}. \tag{2}$$

We also explore other weighting schemes such as naive averaging and entropy-based alternatives, detailed in Section 5.4. The MSP-based blended teacher serves as a superior distillation target compared to CLIP, effectively mitigating the impacts of the pitfalls of the Generalist Trap and Entropy Bias in TTD. Intuitively, higher-confidence blended teachers

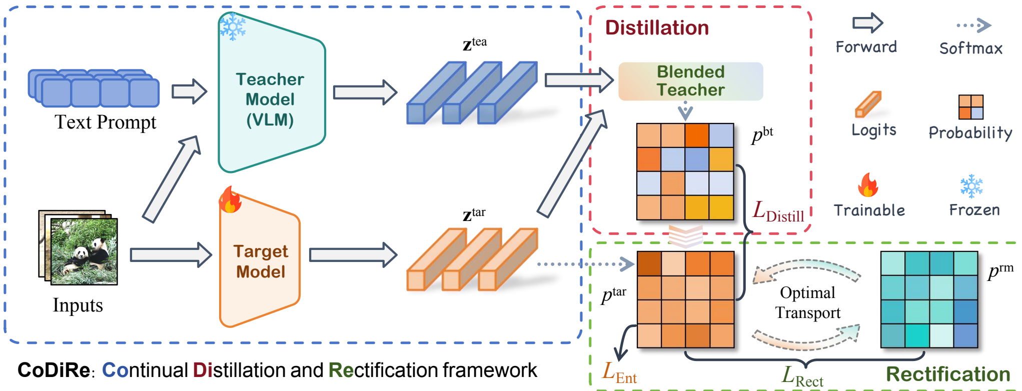


Figure 4. **Overview of our proposed CoDiRe**. CoDiRe introduces a VLM to provide external signals and consists of two components-Distillation and Rectification-by constructing a reliable blended teacher to mitigate the effects of Generalist Trap and Entropy Bias.

should also exert greater influence. We therefore weight the distillation loss by the teacher's confidence. The distillation loss $\mathcal{L}_{\text{Distill}}$ is then computed as:

$$\mathcal{L}_{\text{Distill}}(x_i) = -\max(p_i^{\text{bt}}) \sum_{c=1}^{K} p_{ic}^{\text{bt}} \log p_{ic}^{\text{tar}}. \tag{3}$$

4.3. Rectification

To further refine the target model using the blended teacher, we construct a rectification matrix. By imposing marginal constraints on this matrix, we exert global control over within-batch class assignments, thereby preventing collapse or severe class imbalance (e.g., all samples being confidently mapped to a single class). Concretely, we introduce an optimal transport (OT) [8, 50, 54] step that reconciles the target model's predictions within each mini-batch, yielding a rectified matrix p_i^{rm}. Using p_i^{rm} as refined supervision improves the target model's robustness at the mini-batch level. The rectification preserves fidelity to the original similarities p_i^{tar} while enforcing global margin constraints, resulting in smoothed, distribution-aligned soft scores. Formally, we define the following OT problem:

$$\max_{\mathcal{P}} \operatorname{tr}(\mathbf{P}^{\text{rm}\top}\mathbf{P}^{\text{tar}}), \tag{4}$$

where $\mathbf{P}^{\text{tar}} = (p_1^{\text{tar}}, p_2^{\text{tar}}, \ldots, p_N^{\text{tar}}) \in \mathbb{R}^{K\times N}$ with N the batch size. The transport plan $\mathbf{P}^{\text{rm}} \in \mathbb{R}^{K\times N}$ is treated as the rectified distribution, i.e., $p_i^{\text{rm}} = \mathbf{P}_i^{\text{rm}}$. The plan must satisfy the marginal constraints:

$$\mathcal{P} = \{\mathbf{P}^{\text{rm}} \mid \mathbf{P}^{\text{rm}}\mathbf{1}_N = \mathbf{m},\ \mathbf{P}^{\text{rm}\top}\mathbf{1}_K = \mathbf{u}_N\}, \tag{5}$$

where $\mathbf{m}$ denotes the mini-batch label marginal. We instantiate $\mathbf{m}$ via pseudo-label voting over the three predictions, namely p_i^{bt}, p_i^{tar}, and p_i^{tea}, which serves as a reliable proxy for the true label distribution. We relax the problem and solve it efficiently with the Sinkhorn algorithm [8], which converges to $\mathbf{P}^{\text{rm}}$ within a few iterations.

Finally, we refine the target model using mutual information between p_i^{tar} and p_i^{rm}:

$$\mathcal{L}_{\text{Rect}}(x_i) = -\text{MI}(p_i^{\text{tar}}; p_i^{\text{rm}}), \tag{6}$$

where $\text{MI}(\cdot;\cdot)$ denotes the mutual information loss [20].

4.4. Overall Procedure of CoDiRe

In Section 4.2 and 4.3, we introduced $\mathcal{L}_{\text{Distill}}$ and $\mathcal{L}_{\text{Rect}}$. As most TTA works do [5, 28, 41], we further incorporate a widely adopted entropy loss to enable the target model to update with the distribution of the current data stream:

$$\mathcal{L}_{\text{Ent}}(x_i) = \frac{\mathcal{E}_i}{\exp(\mathcal{E}_i - \tau_{\text{Ent}})}, \tag{7}$$

where $\mathcal{E}_i = -\sum_{c=1}^{K} p_{ic}^{\text{tar}} \log p_{ic}^{\text{tar}}$, and τ_{ENT} controls sensitivity to entropy. Therefore, the final loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Ent}} + \mathcal{L}_{\text{Distill}} + \mathcal{L}_{\text{Rect}}. \tag{8}$$

Beyond the distillation mechanisms, we further address catastrophic forgetting in CTTA by endowing the target model with distribution-awareness. Prior work has explored reset heuristics [37, 56]; for example, CoTTA randomly resets a subset of parameters. In contrast, we neither reset all parameters nor resort to random resets. Instead, we design a distribution-aware layer-wise reset mechanism. Let the target model have parameters θ_t at step t. Specifically, we define an anchor θ^{anchor}, initialized as θ_0 and updated every s steps. We then introduce two displacement vectors [47]:

$$\delta_t = \theta_t - \theta_{t-1}, \tag{9}$$
$$\delta_t^{\text{anchor}} = \theta_{t-1} - \theta^{\text{anchor}}. \tag{10}$$

A divergence between δ_t and δ_t^{anchor} indicates a domain change. We quantify this via cosine similarity:

$$\gamma = \cos(\delta_t, \delta_t^{\text{anchor}}) = \frac{\langle \delta_t, \delta_t^{\text{anchor}} \rangle}{\|\delta_t\| \cdot \|\delta_t^{\text{anchor}}\|}. \tag{11}$$

When $\gamma < \gamma_0$, we regard this as a domain switch, with γ_0 serving as the reset threshold. We observe that deeper

Table 1. **Comparison results under corruption scenarios on CIFAR-10-C.** Classification accuracy of the standard CIFAR-10 → CIFAR-10-C online continual test-time adaptation task while continually adapting to different corruptions at the highest severity 5. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

CIFAR-10-C RN50	Venue	Noise *gauss.*	*shot*	*impul.*	Blur *defoc.*	*glass*	*motion*	*zoom*	Weather *snow*	*frost*	*fog*	*brit.*	Digital *contr.*	*elastic*	*pixel*	*jpeg*	Avg.
		Time →															
Source	-	33.69	40.07	35.74	53.11	49.86	67.04	56.85	73.89	62.59	64.66	89.00	43.57	74.37	41.96	74.43	57.39
BN Adapt	NIPS'20	62.43	64.33	57.30	84.48	59.46	82.02	83.25	76.87	74.30	79.29	87.24	83.33	71.72	73.41	71.10	$74.03_{\pm 0.05}$
Tent	ICLR'21	64.70	70.66	66.39	<u>86.66</u>	66.20	<u>85.09</u>	86.37	82.10	81.57	<u>83.29</u>	90.15	85.03	78.90	81.40	79.19	$79.18_{\pm 0.11}$
MEMO	NIPS'22	50.07	56.89	52.72	70.75	58.65	76.79	71.30	79.74	74.47	74.08	<u>90.72</u>	65.89	77.26	49.76	77.27	$68.42_{\pm 0.03}$
EATA	ICML'22	62.36	64.34	57.31	84.49	59.46	82.02	83.25	76.87	74.31	79.29	87.24	83.33	71.71	73.43	71.11	$74.03_{\pm 0.05}$
SAR	ICLR'23	63.73	68.82	63.33	85.65	64.34	83.20	85.35	79.91	79.09	81.42	88.75	83.93	77.00	79.40	77.75	$77.44_{\pm 0.07}$
DeYO	ICLR'24	<u>67.70</u>	<u>76.27</u>	70.98	86.48	<u>70.04</u>	84.59	<u>87.10</u>	<u>82.25</u>	82.44	82.72	89.49	<u>86.28</u>	<u>79.19</u>	<u>82.69</u>	<u>80.78</u>	<u>80.60</u>$_{\pm 0.07}$
CoTTA	CVPR'22	61.57	63.26	61.05	84.52	53.21	81.93	83.28	76.82	74.23	79.25	87.21	83.30	71.59	73.32	70.97	$73.70_{\pm 0.03}$
NOTE	NIPS'22	55.11	68.63	55.80	34.33	56.29	69.34	78.37	72.35	78.81	64.58	86.97	73.04	63.46	50.32	70.17	$65.17_{\pm 0.07}$
RoTTA	CVPR'23	54.09	53.76	50.24	76.77	53.15	76.76	80.39	71.82	61.30	73.48	84.82	43.55	64.42	72.48	73.94	$66.06_{\pm 0.50}$
SANTA	TMLR'23	65.21	70.75	65.18	85.21	65.67	82.88	84.44	80.31	79.41	81.36	89.13	83.75	77.08	80.27	78.60	$77.95_{\pm 0.28}$
ViDA	ICLR'24	62.53	64.71	57.95	84.67	60.57	82.42	83.85	77.79	75.77	80.17	87.89	83.50	73.56	75.39	73.38	$74.94_{\pm 0.08}$
CLIP	ICML'21	65.39	66.64	<u>76.02</u>	75.75	48.18	78.16	79.08	81.92	<u>84.68</u>	76.14	90.40	80.47	64.74	76.93	71.32	74.39
Ours	-	**79.26**	**84.45**	**84.36**	**89.68**	**74.79**	**89.50**	**91.30**	**90.02**	**91.14**	**88.78**	**95.18**	**91.54**	**83.91**	**88.68**	**86.14**	$\mathbf{87.25}_{\pm 0.06}$

layers tend to capture domain-specific activation statistics and are thus more susceptible to shift-induced drift, whereas shallower layers encode domain-invariant structural cues (e.g., shapes and edges). Accordingly, we selectively reset only the deep, domain-specific last $\alpha\%$ layers. Detailed sensitivity analyses for the reset step size s, switch threshold γ_0, reset ratio α, and a more comprehensive discussion of the reset mechanism are provided in Section 5.4.

5. Experiments

5.1. Experimental Setup

Our experiments are designed to address the following key questions: **RQ1:** How does CoDiRe perform in comparison to existing methods across various real-world scenarios? **RQ2:** What is the contribution of each component of CoDiRe to its overall performance? **RQ3:** How do these components and hyper-parameters work and what are their advantages over other alternatives or baselines?

To address the above questions, we evaluate CoDiRe in two distinct real-world scenarios: corruptions and domain generalizations. As discussed before, all experiments are conducted under the standard CTTA protocols. In the corruption scenarios, we conduct evaluations on CIFAR-10-C and ImageNet-C [17], across 15 different corruption types sequentially at the highest severity level 5. In the domain generalization scenarios, we use two widely adopted datasets: OfficeHome [53] and PACS [29]. Here, we treat one domain as the source domain and concatenate the remaining domains sequentially for continual adaptation. Across both settings, we compare CoDiRe against a comprehensive set of baselines, including: 1) TTA methods: BN Adapt [48], Tent [55], MEMO [68], EATA [41], SAR [42], and DeYO [28]; 2) CTTA methods: CoTTA [56], NOTE [15], RoTTA [66], SANTA [3], ViDA [37], and DPCore [73].

Note that our method's use of CLIP is similar to another popular line of work, namely VLM-TTA, which aims to adapt CLIP itself on OOD data during test time. Though the objective and task differ from ours, we include these methods for a comprehensive and fair comparison: 3) VLM-TTA methods: TPT [49], TDA [24], BoostAdapter [71], and ZERO [13]. Since there are no existing baselines under our proposed TTD paradigm, we also design several 4) TTD methods: Naive Ensemble (NE), BN Adapt w. NE, Tent w. NE, and Distill. CLIP, as detailed in Section 5.2.

In implementation, we consistently adopt CLIP ViT-L/14 as the VLM teacher $\mathcal{F}(\cdot)$. For the target model $f(\cdot)$, we employ a ViT-B/16 [10] backbone architecture on ImageNet-C, and ResNet-50 [16] on the other three datasets following prior work [28]. For more information about experimental details, please refer to the Supplementary Materials.

5.2. Main Results

Comparisons on Corruptions Scenarios. We first report the comparisons on the two corruption datasets CIFAR-10-C and ImageNet-C, as shown in Table 1 and 2, respectively. As noted in Section 3, under domain shift, CLIP underperforms a supervised classifier with the same backbone. Even CLIP ViT-L/14 lags behind a smaller ViT-B/16 trained on ImageNet in terms of performance on ImageNet-C.

In contrast, CoDiRe, by constructing a more reliable blended teacher as the distillation target, achieves the best overall performance across both datasets. On CIFAR-10-C and ImageNet-C, it not only surpasses CLIP's zero-shot capability but also outperforms state-of-the-art CTTA methods, achieving 87.25% and 60.69%, respectively. Notably, CoDiRe keeps CLIP frozen, without requiring access to its parameters, gradients, or architectural details. It can even operate via a simple VLM API, rendering it highly practical for real-world deployment.

Comparisons on Domain Generalization Scenarios. We next evaluate CoDiRe in domain generalization scenarios on OfficeHome and PACS, as shown in Table 3. Unlike corruption scenarios where images are severely degraded,

Table 2. **Comparison results under corruption scenarios on ImageNet-C.** Classification accuracy of the standard ImageNet → ImageNet-C online continual test-time adaptation task while continually adapting to different corruptions at the highest severity 5. The best and second-best results are highlighted in **bold** and underlined, respectively.

ImageNet-C ViT-B/16	Venue	Noise			Blur				Weather				Digital			Avg.
		gauss.	*shot*	*impul.*	*defoc.*	*glass*	*motion*	*zoom*	*snow*	*frost*	*fog*	*brit.*	*contr.*	*elastic*	*pixel*	*jpeg*	
		Time ——————————————————————————————————>															
Source	-	35.06	33.68	36.92	32.46	23.04	36.78	30.56	21.40	27.54	52.84	62.60	50.26	31.68	53.58	57.28	39.05
Tent	ICLR'21	42.31	48.90	52.67	42.41	36.65	50.99	44.75	48.97	54.74	65.37	74.75	**61.08**	45.43	64.92	66.70	53.38$_{\pm0.10}$
MEMO	NIPS'22	41.24	40.80	42.56	32.78	29.32	44.86	37.74	30.24	33.82	53.86	69.34	56.34	33.32	62.32	59.94	44.57$_{\pm0.18}$
EATA	ICML'22	47.93	55.20	57.23	49.21	49.27	55.49	53.18	59.45	62.47	65.27	76.71	57.82	**57.09**	67.63	67.07	58.73$_{\pm0.19}$
SAR	ICLR'23	41.81	48.97	53.16	43.69	38.15	51.29	45.08	48.23	54.93	64.56	75.48	59.76	45.85	64.37	66.54	53.46$_{\pm0.08}$
DeYO	ICLR'24	50.18	56.75	57.95	48.67	49.84	54.75	48.15	58.55	60.65	63.97	75.77	56.45	56.45	66.69	67.55	58.16$_{\pm0.24}$
CoTTA	CVPR'22	54.37	53.67	54.89	50.02	32.02	52.47	45.30	59.83	62.41	64.37	77.76	34.84	45.33	66.67	69.53	54.90$_{\pm0.06}$
NOTE	NIPS'22	54.05	53.45	54.11	50.15	33.22	52.25	45.08	60.27	62.49	65.63	**77.77**	36.30	44.95	67.80	68.97	55.10$_{\pm0.22}$
RoTTA	CVPR'23	53.98	54.02	55.31	49.19	35.31	54.69	48.45	62.87	**65.08**	64.54	**77.77**	38.08	50.17	68.35	70.29	56.54$_{\pm0.40}$
SANTA	TMLR'23	**55.99**	**59.27**	**59.69**	51.08	40.74	56.63	50.87	62.55	64.69	**68.07**	77.48	59.13	50.53	68.29	69.36	59.62$_{\pm0.62}$
ViDA	ICLR'24	54.55	56.05	57.65	50.18	34.78	54.69	46.92	61.55	63.25	64.61	77.70	36.79	49.81	69.26	**70.49**	56.55$_{\pm0.61}$
DPCore	ICML'25	51.15	53.25	53.25	40.06	41.36	53.95	47.21	58.91	55.35	53.97	75.37	51.49	51.39	67.35	63.54	54.51$_{\pm1.02}$
CLIP	ICML'21	22.94	23.20	24.06	31.50	19.80	35.84	33.58	45.00	39.34	47.26	62.88	34.54	25.74	50.68	42.02	35.89
Ours	-	54.43	57.64	57.43	**53.09**	**50.11**	**57.51**	**55.54**	**62.91**	62.76	66.73	77.57	58.65	56.03	**70.02**	69.91	**60.69**$_{\pm0.19}$

Table 3. **Comparison results under domain generalization scenarios on OfficeHome and PACS.** Classification accuracy of the standard OfficeHome-Art/PACS-Art → the rest domains online continual test-time adaptation task while continually adapting to different domains. The best and second-best results are highlighted in **bold** and underlined, respectively.

Source	Target	Source —	BN Adapt NIPS'20	Tent ICLR'21	MEMO NIPS'22	EATA ICML'22	SAR ICLR'23	DeYO ICLR'24	CoTTA CVPR'22	NOTE NIPS'22	RoTTA CVPR'23	SANTA TMLR'23	ViDA ICLR'24	CLIP ICML'21	Ours -
	Clipart	47.93	46.97	47.36	48.24	47.51	47.29	48.17	40.84	49.47	48.94	48.37	48.19	67.67	**70.98**
OfficeHome	*Product*	65.78	61.05	61.25	65.01	61.28	61.12	60.27	44.01	64.59	63.37	62.70	62.53	84.73	**85.39**
Art	*Real*	73.24	70.13	69.76	72.76	69.83	69.80	67.95	60.76	71.10	71.99	71.18	71.29	83.70	**84.77**
	Avg.	62.32	59.39$_{\pm0.29}$	59.46$_{\pm0.18}$	62.00$_{\pm0.04}$	59.54$_{\pm0.18}$	59.40$_{\pm0.22}$	58.80$_{\pm0.13}$	48.54$_{\pm0.13}$	61.72$_{\pm0.03}$	61.43$_{\pm0.14}$	60.75$_{\pm0.17}$	60.67$_{\pm0.10}$	78.70	**80.38**$_{\pm0.04}$
	Cartoon	66.04	74.80	74.97	69.78	74.80	74.93	75.68	75.41	68.87	73.04	74.97	74.79	99.53	**99.73**
PACS	*Photo*	97.84	96.83	96.89	98.20	96.83	96.79	97.09	97.07	97.45	93.73	96.89	96.81	**99.88**	99.84
Art	*Sketch*	57.32	69.32	70.73	60.84	69.32	69.72	71.67	74.15	65.65	71.80	71.43	69.48	95.24	**95.41**
	Avg.	73.73	80.32$_{\pm0.18}$	80.86$_{\pm0.11}$	76.27$_{\pm0.08}$	80.32$_{\pm0.18}$	80.48$_{\pm0.15}$	81.48$_{\pm0.25}$	82.21$_{\pm0.16}$	77.32$_{\pm0.11}$	79.52$_{\pm0.10}$	81.10$_{\pm0.15}$	80.36$_{\pm0.18}$	98.22	**98.33**$_{\pm0.01}$

Table 4. **Comparison results with VLM-TTA and TTD baselines.** Average classification accuracy of the standard CTTA task on CIFAR-10-C, ImageNet-C, OfficeHome and PACS, with the same setting as Table 1, Table 2 and Table 3. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	CIFAR-10-C	ImageNet-C	OfficeHome	PACS	Avg.
CLIP	ICML' 21	74.39	35.89	78.70	98.22	71.80
TPT	NIPS' 22	73.52$_{\pm0.02}$	36.24$_{\pm0.05}$	79.07$_{\pm0.01}$	98.32$_{\pm0.01}$	71.79
TDA	CVPR' 24	75.61$_{\pm0.15}$	37.55$_{\pm0.32}$	79.75$_{\pm0.21}$	98.13$_{\pm0.07}$	72.76
BoostAdapter	NIPS' 24	75.80$_{\pm0.17}$	38.14$_{\pm0.17}$	80.25$_{\pm0.12}$	98.18$_{\pm0.06}$	73.09
ZERO	NIPS' 24	77.89$_{\pm0.04}$	17.43$_{\pm0.17}$	79.46$_{\pm0.10}$	97.66$_{\pm0.05}$	68.11
Naive Ensemble	-	76.90$_{\pm0.00}$	47.95$_{\pm0.27}$	79.89$_{\pm0.00}$	96.29$_{\pm0.00}$	75.26
BN Adapt w. NE	-	84.56$_{\pm0.02}$	47.95$_{\pm0.27}$	80.23$_{\pm0.08}$	97.55$_{\pm0.06}$	77.57
Tent w. NE	-	86.29$_{\pm0.03}$	56.58$_{\pm0.24}$	80.25$_{\pm0.08}$	97.57$_{\pm0.08}$	80.17
Distil. CLIP	-	77.52$_{\pm0.05}$	48.29$_{\pm0.13}$	61.02$_{\pm0.12}$	78.34$_{\pm0.35}$	66.29
Ours	-	**87.25**$_{\pm0.06}$	**60.69**$_{\pm0.19}$	**80.38**$_{\pm0.07}$	**98.33**$_{\pm0.06}$	**81.66**

shifts here arise from stylistic or contextual variations across domains (e.g., cartoons and sketches). Thus, CLIP exhibits robust performance in such scenarios, which greatly benefits from large-scale pre-training on diverse web imagery. Consistent with this, only a zero-shot CLIP could significantly outperform all the TTA/CTTA baselines. Nonetheless, CoDiRe still establishes excellent performance across both datasets. These results highlight CoDiRe's ability to balance the complementary strengths of the task-specific knowledge from the target model and the open-world knowledge from CLIP. It remains effective across diverse shift types, from low-level corruptions to high-level domain variation.

Comparisons with VLM-TTA and TTD baselines. A substantial body of work focuses on directly enhancing CLIP's generalization ability during inference, collectively referred to as VLM-TTA. However, recent evidence [39] indicates that these methods perform poorly on corruption benchmarks such as CIFAR-10-C and ImageNet-C. In contrast, CoDiRe, equipped with the target model, still achieves state-of-the-art performance on both corruption and domain generalization datasets, as shown in Table 4. This demonstrates that CoDiRe not only improves the target model, but also benefits CLIP itself at test-time.

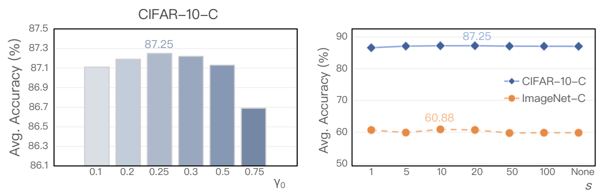
Moreover, we evaluate several naive TTD baselines. A straightforward average interpolation yields a simple prediction p_i^{ens}, where $p_i^{\text{ens}} = \sigma(\mathbf{z}_i^{\text{esn}})$, $\mathbf{z}_i^{\text{esn}} = \frac{1}{2}(\mathbf{z}_i^{\text{tar}} + \mathbf{z}_i^{\text{tea}})$, which we term Naive Ensemble (NE). We further compare BN Adapt and Tent with NE, and evaluate the performance when directly using CLIP logits as the distillation target. The results in Table 4 show that CoDiRe surpasses these naive approaches by constructing a more effective blended teacher, thereby achieving the best performance. The poor results of Distill. CLIP further validates our Generalist Trap pitfall.

5.3. Ablation Study

We conduct an ablation study to assess the contributions of the three loss terms and the distribution-aware reset mechanism. Table 5 reports the results on two corruption benchmarks. Notably, even without any gradient backpropagation, the blended teacher already outperforms both the target model and CLIP. Building on this, incorporating each component yields additional gains: $\mathcal{L}_{\text{Ent}}$ enhances prediction confidence, $\mathcal{L}_{\text{Distill}}$ distills the rich knowledge from blended teacher, $\mathcal{L}_{\text{Rect}}$ further rectifies the target model, and our reset

Table 5. **Ablation study on two corruption datasets.** Average classification accuracy of the standard CTTA task on CIFAR-10-C and ImageNet-C, with the same setting as Table 1 and Table 2. The best and second-best results are highlighted in **bold** and <u>underlined</u>.

	Components				CIFAR-10-C	ImageNet-C	Avg.
	$\mathcal{L}_{\text{Distill}}$	$\mathcal{L}_{\text{Rect}}$	$\mathcal{L}_{\text{Ent}}$	reset			
(1) BT					$84.46_{\pm 0.02}$	$48.05_{\pm 0.15}$	66.26
(2)		✓	✓	✓	$86.71_{\pm 0.02}$	$59.30_{\pm 0.05}$	73.01
(3)	✓			✓	$87.02_{\pm 0.07}$	$59.82_{\pm 0.21}$	73.42
(4)	✓	✓		✓	<u>87.12</u>$_{\pm 0.06}$	$60.30_{\pm 0.14}$	73.71
(5)	✓		✓	✓	$87.11_{\pm 0.08}$	$60.15_{\pm 0.12}$	73.63
(6)	✓	✓	✓		$86.71_{\pm 0.08}$	<u>60.41</u>$_{\pm 0.86}$	73.56
(7) Ours	✓	✓	✓	✓	**87.25**$_{\pm 0.06}$	**60.69**$_{\pm 0.19}$	**73.97**



(a) Reset Threshold. (b) Step Size for Anchor Update.

Figure 5. **Hyper-parameters of Reset Mechanisms.** (a) Hyperparameter study of the reset threshold γ_0 on CIFAR-10-C. (b) Hyperparameter study of the step size s for updating the anchor θ^{anchor} on CIFAR-10-C and ImageNet-C.

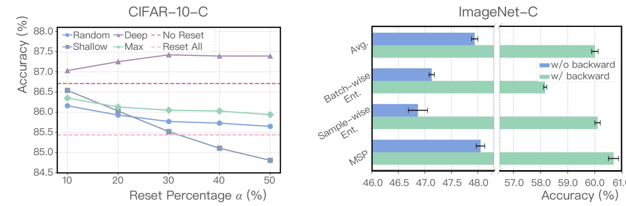
mechanism alleviates catastrophic forgetting during CTTA.

5.4. Discussions

Hyper-parameters. We conduct a sensitivity experiment on the hyper-parameters γ_0 and s introduced in our reset mechanisms. As illustrated in Figure 5, CoDiRe exhibits strong robustness to both hyper-parameters: performance remains stable across wide intervals, with only minor fluctuations at extreme values.

Reset Strategies. We also investigate various reset strategies for the target model, including resetting shallow layers, deep layers, randomly selected layers, and layers with maximum drift, along with no reset and full reset as baselines. As shown in Figure 6(a), selectively resetting deep layers consistently delivers the strongest performance across a wide range of reset percentages α. This corroborates our prior observation that deep layers are more prone to accumulating detrimental domain-specific drift due to their role in modeling higher-level activation statistics and semantic features. In contrast, most alternatives underperform even compared with no reset, as they disrupt beneficial adaptation or interfere with learning of domain-invariant features. Resetting based on maximum shifts is also suboptimal, likely due to heterogeneous learning dynamics and feature distributions across layers. Overall, these results underscore the importance of strategic, layer-aware resetting for robust CTTA.

Fusion Weights. We explore different fusion schemes λ in Equation 2 to combine the logits of the two models. Beyond simple averaging, we adopt dynamic weighting based



(a) Reset strategies. (b) Interpolation weight.

Figure 6. **Discussions of CoDiRe.** (a) Effect of reset strategies under varying percentages α. (b) Effect of different fusion weights.

on batch-wise and sample-wise entropy, alongside our proposed MSP-based approach. Figure 6(b) reports results on ImageNet-C. Direct averaging is a surprisingly strong and straightforward baseline; however, our dynamic MSP-based weighting is overall more robust than fixed averaging. In contrast, entropy-based weights exhibit limited effectiveness and can sometimes cause severe performance degradations, consistent with our Entropy Bias pitfall.

Table 6. **Efficiency analysis on ImageNet-C.** All methods are evaluated on a single NVIDIA Tesla V100 GPU.

Method	**Source**	**Tent**	**CoTTA**	**CLIP**	**Ours**
Memory (GiB) ↓	1.68	6.62	18.41	4.24	8.18
Testing Time (min) ↓	4.23	13.43	36.62	10.14	17.86
Avg Acc. ↑	39.05	53.38	54.90	35.89	**60.69**

Efficiency Analysis. Introducing CLIP as the teacher model inevitably incurs additional overhead. To quantify this, we report the time and memory consumption on ImageNet-C in Table 6. Since CoDiRe keeps CLIP frozen, its extra memory overhead is less than only 15% of Tent's; and the computational cost amounts to only one additional CLIP forward pass, increasing overall latency by merely about 30%. In contrast, CoTTA retains the source, teacher and student model simultaneously, and applies computationally heavy augmentations to maintain robustness, which harms both space and time efficiency. These results also suggest that TTD is a promising, efficient research paradigm.

6. Conclusion

In this work, we proposed Test-Time Distillation (TTD), a new paradigm for Continual Test-Time Adaptation (CTTA) that utilizes a frozen Vision-Language Model as an external signal to mitigate the error accumulation common in self-supervised methods. From comprehensive empirical studies, we identify two critical pitfalls in the practice of TTD: Generalist Trap and Entropy Bias. Our proposed method, CoDiRe, addresses these pitfalls by first constructing a robust blended teacher using a more reliable confidence metric based on Maximum Softmax Probability for distillation. It then further refines the target model's predictions with an Optimal Transport-based rectification step guided by the blended teacher. Extensive experiments show that CoDiRe achieves SoTA performance on standard CTTA benchmarks.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 92467204 and 62472249), the Shenzhen Science and Technology Program (Grant No. KJZD20240903102300001 and JCYJ20250604145014018), and the Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202413). We thank the anonymous reviewers for their efforts, which have helped improve the quality of this paper.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[3] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. SANTA: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023. 1, 3, 6

[4] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7038–7046, 2023. 2, 4

[5] Xiao Chen, Qihui Zhang, and Yan Wang. Cliff: Leveraging ambiguous samples for enhanced test-time adaptation. In *27th European Conference on Artificial Intelligence*, pages 642–649, 2024. 2, 5

[6] Xiao Chen, Zhongjing Du, Jiazhen Huang, Xu Jiang, Li Lu, Jingyan Jiang, and Zhi Wang. Neural collapse in test-time adaptation. *arXiv preprint arXiv:2512.10421*, 2025. 2

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 1

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2, 5

[9] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020. 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[11] Xiao Fan, Jingyan Jiang, Zhaoru Chen, Fanding Huang, Xiao Chen, Qinting Jiang, Bowen Zhang, Xing Tang, and Zhi Wang. Moetta: Test-time adaptation under mixed distribution shifts with moe-layernorm. *arXiv preprint arXiv:2511.13760*, 2025. 2

[12] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438, 2021. 3

[13] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024. 6

[14] Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 4

[15] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 6

[18] Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9772–9781, 2025. 2

[19] Libo Huang, Yan Zeng, Chuanguang Yang, Zhulin An, Boyu Diao, and Yongjun Xu. etag: Class-incremental learning via embedding distillation and task-oriented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12591–12599, 2024. 3

[20] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019. 5

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[22] Jincen Jiang, Qianyu Zhou, Yuhang Li, Xinkui Zhao, Meili Wang, Lizhuang Ma, Jian Chang, Jian Zhang, Xuequan Lu, et al. Pcotta: Continual test-time adaptation for multi-task point cloud understanding. *Advances in Neural Information Processing Systems*, 37:96229–96253, 2024. 2

[23] Qinting Jiang, Chuyang Ye, Dongyan Wei, Bingli Wang, Yuan Xue, Jingyan Jiang, and Zhi Wang. Feature-based instance

neighbor discovery: Advanced stable test-time adaptation in dynamic world. *arXiv preprint arXiv:2506.06782*, 2025. 1

[24] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 6

[25] Shakib Khan, Fariba Dambandkhameneh, Nazim Shaikh, Yao Nie, Raghavan Venugopal, and Xiao Li. Slidemamba: Entropy-based adaptive fusion of gnn and mamba for enhanced representation learning in digital pathology, 2025. 2

[26] Dongmin Kim, Sunghyun Park, and Jaegul Choo. When model meets new normals: test-time adaptation for unsupervised time-series anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13113–13121, 2024. 2

[27] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018. 3

[28] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024. 2, 5, 6

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization, 2017. 6

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[32] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468, 2024. 3

[33] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1512, 2023. 3

[34] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 3

[35] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 3

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[37] Jiaming Liu, Senqiao Yang, Peidong Jia, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 1, 5, 6

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[39] Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1569–1579, 2025. 7

[40] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019. 2

[41] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 2, 5, 6

[42] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 2, 6

[43] Changdae Oh, Yixuan Li, Kyungwoo Song, Sangdoo Yun, and Dongyoon Han. Dawin: Training-free dynamic weight interpolation for robust adaptation. *arXiv preprint arXiv:2410.03782*, 2024. 2, 4

[44] Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20519–20529, 2024. 2

[45] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992, 2023. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3

[47] Sabyasachi Sahoo, Mostafa ElAraby, Jonas Ngnawe, Yann Batiste Pequignot, Frédéric Precioso, and Christian Gagné. A layer selection approach to test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20237–20245, 2025. 5

[48] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Removing covariate shift improves robustness against common corruptions. *CoRR*, abs/2006.16971, 2020. 1, 6

[49] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2, 3, 6

[50] Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19931–19941, 2025. 5

[51] Cody Simons, Dripta S. Raychaudhuri, Sk Miraj Ahmed, Suya You, Konstantinos Karydis, and Amit K. Roy-Chowdhury. Summit: Source-free adaptation of uni-modal models to multi-modal targets, 2023. 2, 4

[52] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017. 1

[53] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 6

[54] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2008. 2, 5

[55] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 6

[56] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 3, 5, 6

[57] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *European Conference on Computer Vision*, pages 304–321. Springer, 2024. 3

[58] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980, 2023. 3

[59] Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Modality-collaborative test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26732–26741, 2024. 2

[60] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3045–3053, 2022. 3

[61] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, pages 534–551. Springer, 2022. 3

[62] Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023. 3

[63] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 3

[64] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[65] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 3

[66] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 1, 3, 6

[67] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[68] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022. 2, 6

[69] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 3

[70] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21769–21780, 2023. 3

[71] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. Boostadapter: Improving test-time adaptation via regional bootstrapping. *arXiv preprint arXiv:2410.15430*, 2024. 2, 6

[72] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024. 2, 3

[73] Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. Dpcore: Dynamic prompt coreset for continual test-time adaptation. *arXiv preprint arXiv:2406.10737*, 2024. 6

[74] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024. 2, 3

[75] Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. Unims: A unified framework for multimodal summarization with knowledge distillation. In

Proceedings of the AAAI conference on artificial intelligence, pages 11757–11764, 2022. 3

[76] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. *arXiv preprint arXiv:2306.03536*, 2023. 1

[77] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[78] Kaifeng Zou, Xiaoyi Feng, Peng Wang, Tao Huang, Zizhou Huang, Zhang Haihang, Yuntao Zou, and Dagang Li. Zero-shot subject-centric generation for creative application using entropy fusion, 2025. 4