BALANCING STABILITY AND PLASTICITY: DYNAMIC SUPERVISION REGULATION FOR MULTI-LABEL CONTINUAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual test-time adaptation (CTTA) aims to adapt a source-trained model to continuously changing unlabeled test distributions without accessing source data during deployment. Existing CTTA methods are predominantly developed for single-label classification, where predictions compete through softmax. In multi-label recognition, however, each category is independently modeled with a sigmoid output, and online adaptation simultaneously involves multiple positive, negative, and uncertain label predictions. As a result, persistent pseudo-label supervision can repeatedly reinforce erroneous label decisions, whereas soft consistency alone, although stable, may remain overly conservative under challenging distribution shifts. In this work, we study continual test-time adaptation for source-trained multi-label classifiers and propose *Dynamic Supervision Regulation* (DSR) to balance adaptation stability and plasticity. DSR maintains teacher–student soft consistency throughout the target stream as a persistent stability anchor, while selectively injecting tri-state hard supervision as an additional plasticity signal. Specifically, confident teacher predictions are assigned positive or negative pseudo-labels, whereas ambiguous labels are excluded from hard supervision. A temporal Bernoulli student–teacher discrepancy is further used as an online control signal to determine when tri-state hard supervision should be activated, thereby avoiding continuous exposure to potentially noisy discrete pseudo-labels. Experiments on VOC-C, COCO-C, and NUS-WIDE-C demonstrate that DSR provides stable and effective continual adaptation across diverse multi-label distribution shifts while mitigating prediction drift and excessive positive predictions caused by overly aggressive supervision.

1 INTRODUCTION

Deep neural networks often degrade when the test distribution differs from the source distribution. Test-time adaptation (TTA) addresses this issue by adapting a source-pretrained model online using unlabeled target data (Wang et al., 2021; Niu et al., 2022; 2023), while continual test-time adaptation (CTTA) further considers non-stationary target streams in which prediction and adaptation are repeatedly interleaved (Wang et al., 2022; Yuan et al., 2023; Liu et al., 2024). In this setting, erroneous self-supervision can propagate through successive updates and accumulate over time, leading to long-term model drift. Existing CTTA methods mitigate this problem from different perspectives, but are predominantly developed for single-label classification or dense prediction.

Extending CTTA to multi-label recognition is not a simple replacement of softmax with sigmoid. Unlike single-label classification, which is typically organized around a competitive categorical decision, multi-label models make independent class-wise predictions and may simultaneously contain positive, negative, and uncertain labels within the same image. Consequently, each test-time update involves many parallel label-wise decisions, and distribution shift may introduce erroneous pseudo-supervision for multiple classes simultaneously. This challenge is further complicated by the severe positive–negative imbalance and heterogeneous confidence distributions across classes, making aggressive discretization particularly vulnerable to uncertain predictions. Recent multi-label TTA methods such as BEM (Wu et al., 2025) and BCP (Li et al., 2026) have begun to address these

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Motivation for regulating hard supervision in ML-CTTA.** (a) Representative predictions on VOC-C under different supervision strategies. (b) Thresholded positive prediction errors of CoTTA, Always Hard, and DSR on VOC-C, COCO-C, and NUS-WIDE-C.

label-wise adaptation challenges, but mainly consider standard TTA rather than continual streams in which pseudo-supervision errors can be repeatedly reinforced through online updates.

This setting creates a pronounced stability–plasticity tension in multi-label CTTA. Soft teacher–student consistency preserves continuous class-wise confidence and avoids prematurely discretizing uncertain labels, making it a stable supervision signal over long target streams. However, soft consistency can remain overly conservative under challenging distribution shifts and may provide insufficient correction for biased predictions. In contrast, hard pseudo-supervision provides a more explicit corrective signal, but repeated updates on erroneous label-wise targets can progressively amplify prediction errors.

Figure 1 further illustrates this tension. In the representative VOC-C example, Soft-only maintains stable predictions but fails to recover a missed positive label, whereas persistent Hard supervision corrects the missed label while introducing an additional false positive. DSR retains the desired correction without the same degree of prediction drift. At the dataset level, thresholded positive prediction errors are substantially higher on COCO-C and NUS-WIDE-C than on VOC-C. As more label-wise decisions are involved simultaneously, persistent exposure to discrete pseudo-labels becomes increasingly vulnerable to prediction errors under distribution shift.

These observations suggest that reliable ML-CTTA requires not only determining *what* label-wise supervision should be constructed, but also *when* hard pseudo-supervision should be introduced. To address this problem, we propose *Dynamic Supervision Regulation* (DSR). DSR maintains teacher–student soft consistency throughout adaptation as a persistent stability anchor, while introducing tri-state hard supervision as an auxiliary plasticity signal. Confident teacher predictions are converted into positive or negative pseudo-labels, whereas ambiguous predictions are excluded from hard pseudo-supervision. DSR further uses a temporal Bernoulli student–teacher discrepancy as an online control signal to determine when additional Hard supervision should be activated, thereby reducing continuous exposure to potentially noisy discrete pseudo-labels and balancing adaptation stability and plasticity over continually evolving target distributions.

Our contributions are summarized as follows:

- We identify and study the stability–plasticity tension in multi-label continual test-time adaptation, where multiple independent label-wise decisions make repeated self-supervision particularly vulnerable to accumulated pseudo-label errors.
- We propose Dynamic Supervision Regulation (DSR), which combines a persistent soft-consistency anchor with tri-state hard supervision as an auxiliary plasticity signal, and dynamically regulates its exposure using a temporal Bernoulli student–teacher discrepancy.
- We evaluate DSR on VOC-C, COCO-C, and NUS-WIDE-C and provide systematic analyses of supervision design, exposure regulation, corruption-order robustness, routing behavior, and prediction stability, demonstrating stable and effective adaptation across diverse multi-label continual distribution shifts.

2 Related Work

2.1 Test-Time and Continual Test-Time Adaptation

Test-time adaptation (TTA) adapts a source-trained model to unlabeled target data during inference without accessing source samples. Representative methods improve online adaptation through entropy minimization (Wang et al., 2021), reliable-sample selection and source preservation (Niu et al., 2022), robust optimization (Niu et al., 2023), or test-time feature augmentation (Cho et al., 2025). These methods are typically designed for stationary or short-term target distributions.

Continual test-time adaptation (CTTA) instead considers non-stationary target streams and long-term error accumulation. CoTTA (Wang et al., 2022) introduces an EMA teacher and stochastic restoration, while subsequent work explores robust teacher–student learning (Döbler et al., 2023), memory-based or memory-efficient adaptation (Song et al., 2023; Yuan et al., 2023), dynamic sample selection (Wang et al., 2024), and stability–plasticity modeling (Liu et al., 2024; Park et al., 2025). Despite this progress, existing CTTA methods are predominantly developed for single-label classification or dense prediction, where supervision is built around competitive categorical predictions. Multi-label recognition instead requires adaptation over multiple independent positive and negative label decisions.

2.2 Multi-Label Recognition and Test-Time Adaptation

Multi-label recognition models each category independently with sigmoid outputs, allowing multiple labels to coexist within one image and introducing a strong positive–negative imbalance. Asymmetric Loss (ASL) (Ridnik et al., 2021a) addresses this imbalance and is widely used for multi-label learning. These characteristics make categorical entropy minimization and single-label pseudo-supervision difficult to transfer directly to multi-label adaptation.

Recent work has therefore explored multi-label-specific TTA. BEM (Wu et al., 2025) revisits entropy-based adaptation by jointly considering multiple candidate labels, BCP (Li et al., 2026) incorporates online conditional label priors, and PuRF (Liang et al., 2026) develops purified cache-based adaptation for multi-label vision–language models. However, these methods mainly address standard TTA rather than continually evolving target streams with repeated online updates. In contrast, we study source-trained multi-label classifiers under ML-CTTA and focus on regulating the exposure to discrete label-wise pseudo-supervision to balance long-term stability and plasticity.

3 Method

3.1 Problem Formulation

We study Multi-Label Continual Test-Time Adaptation (ML-CTTA). Let f_{θ_0} denote a multi-label classifier pre-trained on a labeled source dataset $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$, where $y_i^S \in \{0,1\}^C$ denotes the annotation over C classes. At test time, the source data are unavailable and the model receives an unlabeled target stream of mini-batches $\mathcal{B}_t = \{x_i^t\}_{i=1}^B$, whose underlying distribution may evolve over time, i.e., $\mathcal{P}_t^T \neq \mathcal{P}_{t+1}^T$. At each time step t, the model makes predictions on the current target batch and adapts online for subsequent inputs, $\theta_t \rightarrow \theta_{t+1}$, without access to source data, target annotations, the full target stream in advance, or explicit domain boundaries.

For a target sample x_i^t, the current model produces $\mathbf{z}_i^t = f_{\theta_t}(x_i^t) \in \mathbb{R}^C$ and independent class-wise sigmoid predictions $p_{i,c}^t = \sigma(z_{i,c}^t)$, yielding $\mathbf{p}_i^t \in [0,1]^C$. Unlike softmax classification, multiple classes can be simultaneously positive. The objective of ML-CTTA is to continually adapt the model along the evolving unlabeled target stream while maintaining accurate multi-label predictions and limiting the accumulation of erroneous self-supervision.

3.2 Method Overview

We propose *Dynamic Supervision Regulation* (DSR) for multi-label continual test-time adaptation. DSR is built on an EMA teacher–student framework (Tarvainen & Valpola, 2017; Wang et al., 2022) and combines two complementary supervision signals. Given an incoming target batch $\mathcal{B}_t$, the



Figure 2: Overview of DSR. (a) EMA teacher–student continual adaptation with class-wise sigmoid predictions. (b) Temporal student–teacher discrepancy regulates the exposure to hard supervision. (c) Confident predictions are converted into positive, negative, or ignored targets for tri-state hard supervision.

student f_{θ_t} and EMA teacher $f_{\bar{\theta}_t}$ produce class-wise sigmoid predictions $\mathbf{p}^t$ and $\mathbf{q}^t$. *Stable Soft Consistency* uses the teacher's continuous predictions as a persistent stability anchor, while *Tri-State Hard Supervision* converts only confident predictions into positive or negative pseudo-labels and ignores ambiguous labels, providing an additional plasticity signal.

Rather than switching between the two objectives, DSR keeps soft consistency active throughout adaptation and regulates only the exposure to hard supervision. A temporal Bernoulli student–teacher discrepancy is compared against a slowly evolving historical reference; when the current discrepancy becomes sufficiently large, the tri-state hard objective is selectively added to the soft objective. The discrepancy is used solely as an online control signal. Figure 2 summarizes the overall framework; the supervision objectives, temporal regulation, and continual update procedure are detailed in the following subsections.

3.3 Stable Soft Consistency and Tri-State Hard Supervision

We instantiate the stability and plasticity signals of DSR using two complementary multi-label objectives. *Stable Soft Consistency* preserves the continuous teacher probabilities over all classes, while *Tri-State Hard Supervision* converts only confident teacher predictions into explicit positive or negative pseudo-labels and excludes ambiguous labels. Soft consistency remains active throughout the target stream, whereas hard supervision is used only as an auxiliary signal whose exposure is dynamically regulated.

3.3.1 Stable Soft Consistency

At time step t, let $p^t_{i,c}$ and $q^t_{i,c}$ denote the student and EMA-teacher sigmoid predictions for category c of target sample x^t_i, respectively. We retain the continuous teacher prediction as the supervision target and optimize

$$\mathcal{L}^t_{\text{soft}} = \frac{1}{BC} \sum_{i=1}^{B} \sum_{c=1}^{C} \ell_{\text{BCE}} \left(p^t_{i,c}, q^t_{i,c} \right). \tag{1}$$

This preserves the teacher's confidence information without discretizing uncertain label-wise predictions, and we therefore maintain $\mathcal{L}^t_{\text{soft}}$ throughout the continual stream.

3.3.2 TRI-STATE HARD SUPERVISION

To provide more explicit correction on confident predictions, we construct tri-state pseudo-labels from the EMA teacher:

$$\hat{y}_{i,c}^t = \begin{cases} 1, & q_{i,c}^t \geq \tau_{c,t}^+, \\ 0, & q_{i,c}^t \leq \tau^-, \\ \text{ignore}, & \tau^- < q_{i,c}^t < \tau_{c,t}^+, \end{cases} \tag{2}$$

where $\tau_{c,t}^+$ is a class-wise adaptive positive threshold and τ^- is a shared conservative negative threshold. Ambiguous predictions are excluded from hard pseudo-supervision. The positive threshold is estimated from a per-class prediction history with temporal smoothing; full update details are given in Appendix A.5.

Let $\Omega_t = \{(i,c) \mid \hat{y}_{i,c}^t \neq \text{ignore}\}$ denote the valid sample–label pairs. We optimize them using Asymmetric Loss (Ridnik et al., 2021a):

$$\mathcal{L}_{\text{hard}}^t = \frac{1}{|\Omega_t|} \sum_{(i,c)\in\Omega_t} \ell_{\text{ASL}}\left(p_{i,c}^t, \hat{y}_{i,c}^t\right). \tag{3}$$

Compared with continuous soft consistency, this objective provides stronger directional correction but is more sensitive to erroneous pseudo-labels when repeatedly applied. DSR therefore injects it selectively according to the temporal regulation mechanism in Sec. 3.4.

3.4 TEMPORAL DISCREPANCY-BASED SUPERVISION REGULATION

Persistent hard pseudo-supervision may repeatedly reinforce erroneous label decisions over a continual stream. DSR therefore keeps soft consistency active throughout adaptation and regulates only the exposure to additional hard supervision. At time step t, we quantify the instantaneous discrepancy between the online student and the temporally smoothed EMA teacher using the average Bernoulli KL divergence:

$$D_t = \frac{1}{BC} \sum_{i=1}^{B} \sum_{c=1}^{C} D_{\text{KL}}\left(\text{Bern}(q_{i,c}^t) \parallel \text{Bern}(p_{i,c}^t)\right). \tag{4}$$

To evaluate the current discrepancy relative to the recent adaptation state, we maintain a slowly evolving temporal reference:

$$\tau_{t+1} = \alpha_D \tau_t + (1-\alpha_D) D_t, \tag{5}$$

where α_D controls temporal smoothing. Hard supervision is activated when the instantaneous discrepancy sufficiently exceeds this reference:

$$r_t = \mathbb{I}\left[D_t > \lambda_D \tau_t\right], \qquad r_t \in \{0,1\}, \tag{6}$$

where λ_D controls routing sensitivity. When $r_t = 0$, adaptation uses soft consistency alone; when $r_t = 1$, tri-state hard supervision is added to the soft objective, which remains active throughout adaptation.

During a short initial warm-up, r_t is fixed to zero while the temporal reference is initialized. Importantly, D_t is used only as a temporal control signal describing the relative student–teacher prediction state, rather than as a direct estimate of sample reliability or domain-shift magnitude. Initialization details are provided in Appendix A.6.

3.5 OPTIMIZATION OBJECTIVE AND CONTINUAL UPDATE

Combining the supervision objectives in Sec. 3.3 with the temporal regulation mechanism in Sec. 3.4, the final online adaptation objective is

$$\mathcal{L}_{\text{DSR}}^t = \mathcal{L}_{\text{soft}}^t + r_t \lambda_h \mathcal{L}_{\text{hard}}^t, \tag{7}$$

where $r_t \in \{0,1\}$ controls the exposure to hard supervision and λ_h controls the contribution of the hard objective. Thus, soft consistency remains active throughout adaptation, while tri-state hard supervision is introduced only when $r_t = 1$. During warm-up, $r_t = 0$.

At each time step, the student is first optimized using $\mathcal{L}_{\mathrm{DSR}}^{t}$, yielding $\theta_{t+1}^{\mathrm{opt}}$. The EMA teacher is then updated as

$$\bar{\theta}_{t+1} = \alpha_T \bar{\theta}_t + (1 - \alpha_T)\theta_{t+1}^{\mathrm{opt}}, \tag{8}$$

providing a temporally smoothed reference for subsequent target batches. Following CoTTA (Wang et al., 2022), the student subsequently undergoes lightweight stochastic restoration toward the source parameters, while target-domain BN statistics are retained across the continual stream. Additional continual-update details are provided in Appendix A.4.

4 Experiments

4.1 Experimental Setup

Datasets and continual protocol. We evaluate ML-CTTA on PASCAL VOC (Everingham et al., 2010), MS-COCO (Lin et al., 2014), and NUS-WIDE (Chua et al., 2009), constructing VOC-C, COCO-C, and NUS-WIDE-C with the standard 15 common corruptions (Hendrycks & Dietterich, 2019). Unless otherwise specified, all corruptions are evaluated at severity 5 in a fixed continual sequence. The model processes the entire target stream without resetting its parameters or target-domain normalization statistics between corruptions. Target labels are used only for evaluation. The complete corruption order and dataset details are provided in Appendix A.

Baselines. We compare DSR with the frozen source model (*Source*) and representative single-label TTA/CTTA methods including TENT (Wang et al., 2021), CoTTA (Wang et al., 2022), ViDA (Liu et al., 2024), DSS (Wang et al., 2024), FATA-TENT (Cho et al., 2025), and GOLD (Lai et al., 2026). We retain their original implementations as closely as possible while adapting them to independent multi-label prediction. All methods use the same dataset-specific source checkpoint and continual stream; reproduction details are provided in Appendix B. Recent multi-label TTA methods BEM (Wu et al., 2025), BCP (Li et al., 2026), and PuRF (Liang et al., 2026) are CLIP-based and are therefore not directly compared with our source-trained TResNet-M setting.

Evaluation metrics. We use mean average precision (mAP) as the primary metric and report full-stream mAP (*Full*) over the complete continual sequence. Additional diagnostic metrics, including F1, FPR, FNR, and predicted positives per image, are used only for analysis and are detailed in Appendix A.

Implementation details. We use TResNet-M (Ridnik et al., 2021b) source models trained with Asymmetric Loss (ASL) (Ridnik et al., 2021a). DSR performs full-model adaptation with AdamW (Loshchilov & Hutter, 2019) and batch size 32. We use a 20-step soft-consistency warm-up, $\lambda_D = 1.2$, and $\lambda_h = 0.25$; the remaining optimization, EMA, thresholding, restoration, and dataset-specific hyperparameters are provided in Appendix A.

4.2 Main Results

Table 1 compares DSR with representative TTA and CTTA methods under the unified common-15 continual protocol. DSR achieves the best full-stream mAP on all three benchmarks, reaching 82.22%, 68.21%, and 43.44% on VOC-C, COCO-C, and NUS-WIDE-C, respectively. Compared with the frozen source models, this corresponds to gains of 5.87, 2.11, and 6.26 mAP. Among the evaluated adaptive baselines, the next-best full-stream results are GOLD (Lai et al., 2026) with 80.12% on VOC-C, CoTTA (Wang et al., 2022) with 65.31% on COCO-C, and ViDA (Liu et al., 2024) with 40.90% on NUS-WIDE-C, leaving DSR margins of 2.10, 2.90, and 2.54 mAP. The performance gap becomes particularly evident on the more challenging continual streams. TENT (Wang et al., 2021), DSS (Wang et al., 2024), and the recent FATA-TENT baseline (Cho et al., 2025) degrade substantially on COCO-C and NUS-WIDE-C, whereas DSR maintains stable full-stream performance across all three datasets. Together with the per-corruption results, this suggests that successful multi-label continual adaptation requires carefully constructing hard pseudo-supervision and regulating its exposure rather than simply increasing adaptation strength. We examine these factors in Secs. 4.3 and 4.4.

Table 1: Per-corruption and full-stream mAP (%) under the common-15 continual order at severity 5. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Gau.	Shot	Imp.	Def.	Mot.	Zoom	Snow	Frost	Fog	Bri.	Con.	Ela.	Pix.	JPEG	Sat.	Full ↑
(a) VOC-C																
Source	**74.51**	79.29	74.57	72.35	72.87	56.14	72.90	74.16	78.59	88.63	77.48	78.21	**88.80**	87.39	87.51	76.35
TENT	71.64	78.40	72.24	73.73	75.48	68.60	76.00	72.27	82.05	86.83	78.75	83.54	85.84	84.55	85.68	78.63
CoTTA	71.46	77.61	71.42	73.28	74.65	67.50	75.55	71.67	81.86	86.75	77.49	83.05	85.91	84.67	85.53	78.13
ViDA	71.47	78.05	71.56	73.31	75.16	67.86	75.33	71.35	81.79	86.85	78.00	83.40	85.86	84.32	85.57	78.28
DSS	71.51	78.17	72.35	72.27	74.03	66.67	75.37	71.48	81.35	85.64	72.31	81.71	84.29	82.49	82.93	74.68
FATA-TENT	67.17	74.48	68.70	70.29	73.04	65.81	73.33	69.78	79.66	85.72	76.86	82.08	84.71	83.33	84.05	75.98
GOLD	73.99	80.26	75.69	76.14	78.07	73.16	79.06	76.74	82.68	86.51	**81.55**	83.94	85.73	84.88	85.32	80.12
DSR (Ours)	74.37	**80.83**	**76.25**	**77.32**	**80.26**	**74.52**	**80.89**	**78.91**	**85.44**	**89.27**	81.23	**86.41**	88.32	**87.92**	**88.34**	**82.22**
(b) COCO-C																
Source	**66.23**	**67.92**	**66.29**	**67.64**	64.75	41.38	55.81	57.88	71.61	**77.60**	**69.60**	69.69	**78.99**	**76.66**	**75.81**	66.10
TENT	60.75	58.63	38.29	29.04	20.81	8.97	10.28	7.75	13.00	11.85	6.74	7.73	8.22	6.25	6.05	14.04
CoTTA	60.55	63.89	59.26	64.15	65.14	49.93	56.82	56.23	71.19	73.98	64.73	71.72	75.13	72.90	72.12	65.31
ViDA	60.58	63.93	59.38	64.19	65.10	49.77	56.61	56.15	71.25	73.88	65.44	71.80	75.44	72.69	72.06	65.30
DSS	61.25	62.54	54.99	39.83	37.37	17.04	20.24	13.78	9.81	13.23	4.47	7.32	7.68	6.73	6.40	22.23
FATA-TENT	58.48	64.79	63.51	62.81	63.41	51.44	54.56	51.97	61.24	62.18	51.98	55.18	58.13	51.79	48.50	55.87
GOLD	62.77	66.48	64.53	64.52	64.93	**55.69**	60.72	58.97	64.84	66.91	62.92	64.89	67.09	65.42	63.69	61.27
DSR (Ours)	63.05	66.64	62.89	67.13	**68.37**	54.83	**60.80**	**60.47**	**73.54**	75.90	68.60	**74.01**	76.98	74.98	74.36	**68.21**
(c) NUS-WIDE-C																
Source	28.96	30.36	29.98	26.74	31.13	38.78	42.14	44.10	49.28	**60.14**	45.68	29.17	41.68	**52.86**	55.65	37.18
TENT	29.20	28.02	18.06	12.82	12.68	11.65	9.01	6.81	8.40	7.62	5.89	5.96	5.65	5.56	6.03	8.79
CoTTA	29.36	33.12	29.02	24.58	33.58	45.65	43.45	44.03	51.83	57.34	44.79	40.42	40.91	43.35	55.29	40.81
ViDA	29.47	33.27	29.09	25.94	34.31	46.03	43.30	44.05	51.62	57.23	45.28	39.98	41.43	42.56	55.07	40.90
DSS	28.37	23.71	13.53	5.38	6.24	6.51	6.03	4.57	3.56	4.64	2.87	3.13	3.10	3.34	3.74	6.22
FATA-TENT	28.81	34.17	28.35	20.36	23.52	27.97	21.35	17.13	15.85	13.68	6.31	7.36	6.89	5.49	4.66	15.88
GOLD	29.98	**36.00**	**34.15**	27.31	28.92	31.47	28.92	26.74	27.32	29.26	24.83	21.20	21.76	21.64	20.86	18.04
DSR (Ours)	**31.35**	35.69	32.11	**27.90**	**37.49**	**48.61**	**46.00**	**46.67**	**54.60**	59.49	**48.56**	**43.53**	**44.45**	46.00	**57.52**	**43.44**

4.3 Ablation Studies

We organize the ablations around three aspects: (i) the complementary roles of soft consistency and hard pseudo-supervision, (ii) the representation of positive, negative, and uncertain labels in hard targets, and (iii) the regulation of hard-supervision exposure using temporal discrepancy. Unless stated otherwise, all variants use the same source checkpoint, Target BN-EMA (Schneider et al., 2020), optimizer, and continual evaluation protocol, together with the same EMA teacher and stochastic restoration whenever these components are applicable. Strong augmentation is disabled throughout these ablations so that the results isolate the supervision mechanism itself.

Table 2: **Component ablation of DSR.** All variants use the same Target BN-EMA. Checkmarks denote enabled components. Results are full-stream mAP (%).

Variant	Soft	Hard	Dynamic	VOC-C	COCO-C	NUS-WIDE-C
Soft-only	✓			80.806	67.039	41.352
Hard-only		✓		80.493	29.400	13.181
Soft + Always Hard	✓	✓		82.009	68.182	43.078
DSR (Ours)	✓	✓	✓	**82.222**	**68.214**	**43.440**

Component Contributions. Table 2 confirms the complementary roles of the two supervision signals. Soft-only provides a stable adaptation baseline, whereas Hard-only becomes unstable on the more challenging COCO-C and NUS-WIDE-C streams. Combining both signals improves performance, and temporal regulation further provides the best overall results. This indicates that hard supervision is most effective as a dynamically regulated auxiliary signal on top of the persistent Soft objective.

Hard-Target Construction. Table 3 highlights the importance of explicitly modeling positive, negative, and uncertain labels. Positive-only collapses on COCO-C and NUS-WIDE-C, whereas Negative-only recovers much of the performance, supporting the importance of negative supervision

Table 3: **Hard-target construction under Soft + Always Hard.** Example targets use $\mathbf{q} = [0.92, 0.65, 0.42, 0.12]$; 1, $\varnothing$, and 0 denote positive, ignored, and negative targets. Thresholds correspond to the displayed example. Results are full-stream mAP (%).

Hard target	Threshold	Example target	VOC-C	COCO-C	NUS-WIDE-C
Positive-only	$\tau^+_{c,t} = 0.60$	$[1, 1, \varnothing, \varnothing]$	80.607	30.075	7.289
Negative-only	$\tau^- = 0.30$	$[\varnothing, \varnothing, \varnothing, 0]$	82.007	67.493	41.414
Binary	$\tau = 0.50$	$[1, 1, 0, 0]$	81.964	67.298	33.044
Fixed Tri-State	$(\tau^+, \tau^-) = (0.60, 0.30)$	$[1, 1, \varnothing, 0]$	81.926	67.925	41.205
Adaptive Tri-State	$(\tau^+_{c,t}, \tau^-) = (0.60, 0.30)$	$[1, 1, \varnothing, 0]$	**82.009**	**68.182**	**43.078**

for effective adaptation. However, Negative-only remains below Adaptive Tri-State on COCO-C and NUS-WIDE-C, suggesting that positive supervision provides additional benefit. Binary supervision also underperforms Adaptive Tri-State, particularly on NUS-WIDE-C, supporting the selective exclusion of ambiguous predictions from hard targets. Tri-State supervision combines positive and negative correction while retaining an Ignore region, and adaptive class-wise positive thresholds consistently improve over their fixed-threshold counterpart across all three datasets.

Table 4: **Exposure regulation across datasets.** Full-stream mAP (%).

Strategy	VOC-C	COCO-C	NUS-WIDE-C
Always Hard	82.009	68.182	43.078
Uniform Weak	80.813	67.159	42.390
Random-K	80.964 ± 0.013	67.208 ± 0.007	42.344 ± 0.016
DSR	**82.222**	**68.214**	**43.440**

Exposure Regulation. Table 4 compares hard-supervision exposure strategies across VOC-C, COCO-C, and NUS-WIDE-C. To isolate exposure timing, Random-K samples without replacement K post-warm-up batches, where K equals the number of Hard-active batches produced by DSR, and uses the same hard-loss weight $\lambda_h = 0.25$. Uniform Weak instead activates Hard on all N post-warm-up batches with $\lambda_h^{\text{weak}} = 0.25K/N$, matching only the nominal aggregate Hard-loss weight, not the realized gradient magnitude or supervision strength. For VOC-C, COCO-C, and NUS-WIDE-C, (K, N) is $(317, 2302)$, $(4177, 18967)$, and $(8348, 39308)$, yielding $\lambda_h^{\text{weak}} \approx 0.0344$, 0.0551, and 0.0531, respectively. DSR achieves the highest full-stream mAP on all three datasets and outperforms both matched controls, indicating that the timing of Hard supervision contributes beyond its overall weight or frequency. Nevertheless, Always Hard remains competitive, suggesting that temporal routing refines the contribution of the Hard branch rather than accounting for DSR's entire performance advantage.

4.4 Analysis

Robustness to Corruption Order. Across 10 shuffled corruption orders, DSR maintains consistently low variance and strong performance on all three benchmarks (Table 5), while several continually optimized baselines remain unstable on the more challenging multi-label streams. This indicates that DSR is not tied to a particular corruption ordering.

Table 5: Average full-stream mAP (%) over 10 shuffled corruption orders at severity 5.

Dataset	Source	TENT	CoTTA	**DSR**
VOC-C	76.347	78.708 ± 0.026	78.416 ± 0.105	$\mathbf{82.716 \pm 0.200}$
COCO-C	66.101	15.819 ± 0.937	65.416 ± 0.102	$\mathbf{68.362 \pm 0.070}$
NUS-WIDE-C	37.121	11.769 ± 1.194	40.872 ± 0.106	$\mathbf{43.619 \pm 0.052}$

Temporal Routing Behavior. Figure 3(a) visualizes the discrepancy-based routing decisions on VOC-C. Hard supervision is activated on 317 of 2,322 batches (13.65%), with activations concentrated around defocus blur and contrast. The trajectory illustrates that Hard is introduced selectively when the current student–teacher discrepancy exceeds its evolving temporal reference, while Soft supervision remains active throughout the stream.

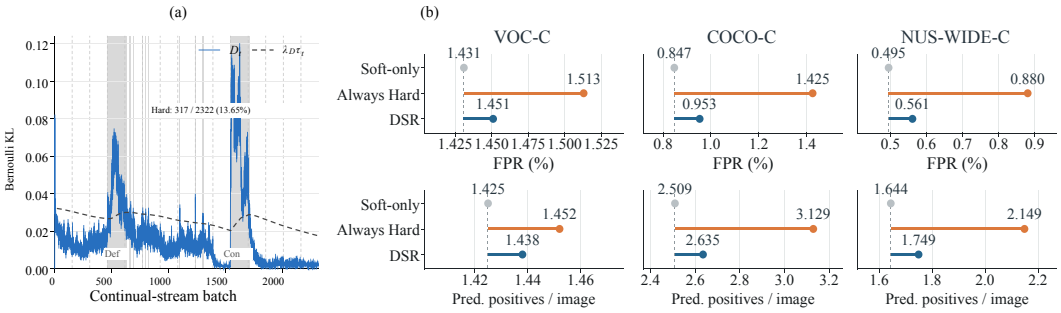

Figure 3: **Temporal routing and prediction behavior.** (a) Bernoulli student–teacher discrepancy and temporal routing threshold on VOC-C; shading marks Hard-active batches. (b) Full-stream false-positive rate (FPR) and predicted positives per image on VOC-C, COCO-C, and NUS-WIDE-C. Gray, orange, and blue indicate Soft-only, Always Hard, and DSR; dashed vertical lines mark Soft-only. Horizontal scales differ across plots. The panels use separate archived evaluations and are not configuration-matched.

Prediction Behavior Across Datasets. Figure 3(b) shows that Always Hard produces more predicted positive labels and a higher FPR than DSR on all three datasets. DSR remains closer to Soft-only on both measures, with smaller differences on VOC-C than on COCO-C and NUS-WIDE-C. This qualitative pattern suggests that regulating Hard exposure can limit excessive positive-label expansion while retaining the benefit of additional supervision, complementing the full-stream mAP results in Table 2.

Cross-Dataset Routing Patterns. Figure 4 further shows that this selectivity is not tied to a fixed set of corruption types. Hard activations are highly concentrated on VOC-C, whereas COCO-C and NUS-WIDE-C exhibit broader routing patterns across the corruption stream. The distinct behaviors across datasets suggest that DSR does not follow a predefined update schedule, but adapts the exposure to Hard supervision according to the evolving target stream.

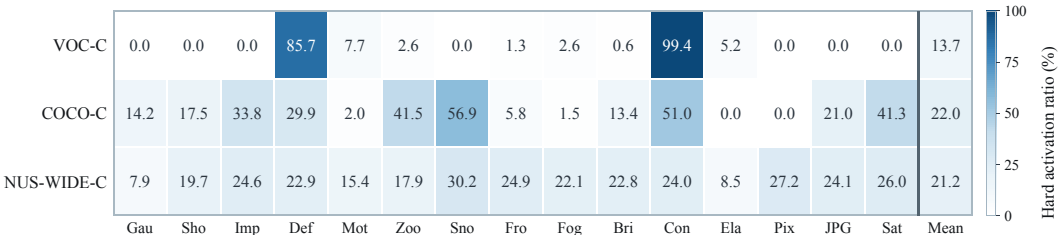

Figure 4: **Cross-dataset routing patterns of DSR.** Hard activation ratios (%) across individual corruption domains and the overall continual stream.

5 Conclusion

We study multi-label continual test-time adaptation, where label-wise pseudo-supervision can accumulate errors over a changing target stream. Dynamic Supervision Regulation (DSR) retains teacher–student soft consistency while selectively adding tri-state hard supervision. A temporal Bernoulli student–teacher discrepancy determines when the Hard objective is activated, limiting continuous exposure to potentially erroneous discrete targets. On VOC-C, COCO-C, and NUS-WIDE-C, DSR achieves the highest full-stream mAP among the evaluated methods. Ablations highlight the complementary roles of Soft and Hard supervision and the value of regulating Hard exposure; prediction analyses further suggest reduced positive-label expansion. These results support selective supervision as a way to balance stability and plasticity in multi-label continual adaptation.

AI USE STATEMENT

For writing and literature-related tasks, we used generative AI tools to assist with literature retrieval and summarization, language polishing, and drafting and revising portions of the manuscript. These uses included identifying relevant studies, summarizing existing literature, improving language and paragraph organization, and assisting with the preparation and revision of selected passages. The authors take responsibility for the final adopted content, the accuracy of references, and the claims and conclusions presented in the paper.

ETHICS STATEMENT

This study investigates multi-label continual test-time adaptation using existing VOC, COCO, and NUS-WIDE image benchmarks and their synthetically corrupted variants. It does not involve new data collection from human participants. Existing images may contain people and reflect sampling, geographic, cultural, or annotation biases in the original datasets. Our evaluation does not measure demographic fairness or establish privacy protection. Moreover, online self-supervision can reinforce incorrect predictions under distribution shifts. Improvements on these benchmarks therefore do not guarantee reliable or equitable behavior in real-world or high-stakes applications. Any deployment would require application-specific validation, monitoring of adaptation behavior, and appropriate safeguards against harmful prediction errors.

REPRODUCIBILITY STATEMENT

Section 3 describes the supervision objectives, temporal routing mechanism, and continual update procedure, while Section 4 specifies the experimental setting and evaluation protocol. Appendix A provides the dataset splits and filtering rules, corruption order, evaluation metrics, hyperparameters, threshold updates, and initialization details. Appendix B describes the multi-label adaptations of the evaluated baselines, and Appendices C and D provide additional ablations and supervision analyses. We have also prepared a compact code package containing the core DSR implementation, three dataset-specific YAML configurations, dependency versions, and a README describing the method interface and input conventions. The package focuses on the adaptation method; datasets, source checkpoints, data loaders, and the evaluation runner must be supplied separately.

REFERENCES

Younggeol Cho, Youngrae Kim, Junho Yoon, Seunghoon Hong, and Dongman Lee. Feature augmentation based test-time adaptation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 6838–6847, 2025.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, 2009.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7704–7714, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3866–3875, 2026.

Qiru Li, Ao Zhou, Zhiwei Jiang, Zifeng Cheng, Cong Wang, Yafeng Yin, and Qing Gu. Multi-label test-time adaptation with bayesian conditional priors. In *Forty-Third International Conference on Machine Learning*, 2026.

Yiwen Liang, Hui Chen, Yizhe Xiong, Mengyao Lyu, Yuhan Cao, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Towards purified multi-label test-time adaptation of vision-language models. In *European Conference on Computer Vision*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. ViDA: Homeostatic visual domain adapter for continual test time adaptation. In *International Conference on Learning Representations*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15319–15328, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Hyewon Park, Hyejin Park, Jueun Ko, and Dongbo Min. Hybrid-tta: Continual test-time adaptation via dynamic domain shift detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2877–2886, 2025.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021a.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben-Baruch, Gilad Sharir, and Itamar Friedman. TResNet: High performance GPU-dedicated architecture. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1400–1409, 2021b.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11920–11929, 2023.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

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*, pp. 7201–7211, 2022.

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. Continual test-time domain adaptation via dynamic sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1701–1710, 2024.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, 2020.

Xiangyu Wu, Feng Yu, Yang Yang, Qing-Guo Chen, and Jianfeng Lu. Multi-label test-time adaptation with bound entropy minimization. In *International Conference on Learning Representations*, 2025.

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*, pp. 15922–15932, 2023.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Additional Experimental and Implementation Details

A.1 Datasets and Continual Protocol

The target test images are taken from the PASCAL VOC 2007 test split (Everingham et al., 2010), the COCO 2014 validation split (Lin et al., 2014), and a fixed subset of the official NUS-WIDE test split (Chua et al., 2009). Each corruption domain therefore contains 4,952 VOC images, 40,504 COCO images, or 83,898 NUS-WIDE images. For NUS-WIDE, we adopt only the test-sample filtering rule used by ML-BEM (Wu et al., 2025): among the 107,859 images in the official test split, we retain images having at least one positive annotation over the 81 concept classes. This removes 23,961 images with all-zero concept labels and leaves 83,898 target images. This adoption is limited to the sample-filtering rule and does not reproduce ML-BEM's experimental protocol or continual-stream ordering. Separately, our source training and validation splits are formed within the official train split after applying the corresponding 81-concept filtering. For NUS-WIDE-C, the target images are further arranged using a fixed, label-independent random permutation with seed 0, and the same permutation is reused in all 15 corruption domains.

We construct VOC-C, COCO-C, and NUS-WIDE-C using the standard set of 15 common image corruptions (Hendrycks & Dietterich, 2019) covering noise, blur, weather, and digital distortions. Unless otherwise specified, all corruptions are evaluated at severity level 5 and follow the fixed order: Gaussian noise, shot noise, impulse noise, defocus blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate, JPEG compression, and saturation.

The label-based filtering above is performed offline only to define fixed dataset splits and the evaluation set. Target annotations are never used for online prediction, Hard routing, pseudo-label construction, or model updates; they are accessed only for evaluation.

For each dataset, the source model is trained on the corresponding clean source split and subsequently adapted online to the unlabeled corrupted test stream. The complete sequence is processed continually without resetting model parameters or target-domain normalization statistics between corruption domains.

A.2 Evaluation Metrics

We use mean average precision (mAP) as the primary evaluation metric. Full-stream mAP (*Full*) is computed from predictions pooled over the complete continual sequence, rather than by averaging per-corruption mAP. For DSR, predictions are obtained from the EMA teacher after updating target BN statistics on the current batch, but before the student gradient update and teacher-parameter EMA update. For diagnostic analyses, we additionally report class-wise F1 (CF1), overall F1 (OF1), false-positive rate (FPR), false-negative rate (FNR), and predicted positives per image (PredPos/Img). Binary metrics use a prediction threshold of 0.5. The thresholded positive prediction error in Figure 1 instead uses a common threshold of 0.7 for all compared methods; this threshold is specific to that diagnostic analysis.

Table 6: Implementation details and hyperparameters of DSR.

Parameter	VOC-C	COCO-C	NUS-WIDE-C
Backbone	TResNet-M	TResNet-M	TResNet-M
Number of classes	20	80	81
Optimizer	AdamW	AdamW	AdamW
Batch size	32	32	32
Learning rate	3×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	0	0	0
Updated parameters	Full model	Full model	Full model
Teacher EMA momentum α_T	0.999	0.999	0.999
Target BN new-batch factor	0.1	0.1	0.1
Stochastic restoration probability p_r	0.01	0.01	0.01
Warm-up batches	20	20	20
Discrepancy EMA coefficient α_D	0.999	0.999	0.999
Routing coefficient λ_D	1.2	1.2	1.2
Hard-loss weight λ_h	0.25	0.25	0.25
Negative pseudo-label threshold τ^-	0.30	0.30	0.30
Positive percentile ρ	95	95	95
Positive-threshold range	$[0.60, 0.95]$	$[0.60, 0.95]$	$[0.60, 0.95]$
Per-class prediction queue size	4096	4096	4096
Positive-threshold EMA coefficient β	0.9	0.9	0.9
Positive-threshold initialization τ_0^+	0.8	0.8	0.8
ASL γ_{pos}	0	0	0
ASL γ_{neg}	2	2	2
ASL clipping parameter	0.05	0.05	0.05

A.3 Implementation Details

We use TResNet-M (Ridnik et al., 2021b) as the source architecture for all datasets, with source models trained using Asymmetric Loss (ASL) (Ridnik et al., 2021a). All methods are initialized from the same dataset-specific TResNet-M source checkpoint. The source checkpoints are calibrated using the source-side Global Balance procedure, with the resulting bias merged into the classifier before test-time adaptation.

DSR performs full-model adaptation using AdamW (Loshchilov & Hutter, 2019) with batch size 32. All trainable student parameters, including BN affine weights and biases, are optimized jointly. BN running statistics are updated separately with a new-batch factor of 0.1 following target-statistics adaptation practice (Schneider et al., 2020). The EMA teacher momentum is 0.999. Stochastic restoration is applied with probability 0.01 only to trainable student weight and bias elements; teacher parameters and BN running statistics are not stochastically restored. We use a 20-batch Soft Consistency warm-up and no additional test-time image augmentation for the main DSR method. The learning rate is 3×10^{-4} for VOC-C and 1×10^{-4} for COCO-C and NUS-WIDE-C. Table 6 summarizes the remaining implementation details.

A.4 Continual Update Details

At each time step t, the student is first optimized using $\mathcal{L}^t_{\mathrm{DSR}}$, producing $\theta^{\mathrm{opt}}_{t+1}$. The EMA teacher is updated from this optimized student state before stochastic restoration, such that the teacher preserves the temporally smoothed sequence of adapted model states.

To constrain long-term parameter drift, we adopt stochastic restoration following CoTTA (Wang et al., 2022). After updating the EMA teacher, an element-wise restoration mask M_t is sampled over trainable student weight and bias elements, and a small subset is restored toward the source-pretrained parameters θ_0:

$$\theta_{t+1} = M_t \odot \theta_0 + (1 - M_t) \odot \theta^{\mathrm{opt}}_{t+1}, \qquad M_t \sim \mathrm{Bernoulli}(p_r), \tag{9}$$

where p_r denotes the restoration probability and $\odot$ represents element-wise multiplication. Restoration is not applied to teacher parameters or BN running statistics; the EMA teacher retains the update obtained from $\theta^{\mathrm{opt}}_{t+1}$.

Target-domain batch-normalization running statistics are continuously updated throughout the target stream and retained across successive corruption domains rather than restored to their source values. Consequently, the continual adaptation state consists of accumulated target-domain normalization statistics, a temporally smoothed EMA teacher, and an online student regularized by stochastic restoration.

TCA stabilizes continual adaptation by preserving inter-class centroid topology and encouraging intra-class compactness (Ni et al., 2025). DSR instead regulates supervision exposure, retaining Soft consistency throughout adaptation and selectively activating the additional Hard objective.

A.5 Adaptive Positive Threshold

Different categories may exhibit substantially different confidence distributions under domain shift. We therefore estimate the positive pseudo-label threshold separately for each class using a running prediction history queue $Q_{c,t}$ with capacity 4096. At each time step, the current batch of clean teacher predictions is first appended to the corresponding per-class queues. We then compute the 95th percentile and clip this candidate threshold before applying temporal EMA:

$$\widetilde{\tau}^{+}_{c,t} = \mathrm{clip}\left(\mathrm{Percentile}_{95}(Q_{c,t}), 0.60, 0.95\right), \tag{10}$$

$$\tau^{+}_{c,t} = \beta \tau^{+}_{c,t-1} + (1 - \beta)\widetilde{\tau}^{+}_{c,t}, \qquad \beta = 0.9. \tag{11}$$

The positive threshold is initialized as $\tau^{+}_{c,0} = 0.8$ and remains fixed at 0.8 during the first 20 warm-up batches. The shared negative threshold is fixed to $\tau^{-} = 0.3$. This asymmetric design allows the positive threshold to adapt to class-specific confidence statistics while retaining a stable negative decision boundary. Class-dependent confidence thresholding has also been explored in semi-supervised learning to account for different class learning states (Zhang et al., 2021). In DSR, thresholds are estimated from running teacher-probability histories for label-wise target selection, while temporal discrepancy separately controls batch-level Hard exposure.

A.6 Regulation Initialization

At the beginning of the target stream, the historical discrepancy reference is not yet sufficiently established. We therefore use a short Soft Consistency warm-up during which r_t is fixed to zero. During this warm-up, the observed student–teacher discrepancies are accumulated, and their mean is used to initialize the temporal reference. After warm-up, D_t and τ_t are updated online throughout the continual stream and r_t is determined according to Eq. 6.

Unless otherwise specified, we use a 20-step warm-up with $\alpha_D = 0.999$ and $\lambda_D = 1.2$.

B Baseline Implementation Details

All baselines use the same dataset-specific TResNet-M source checkpoint, 224×224 inputs, and batch size 32. They are evaluated on the same common-15 continual stream at severity 5 using a

single continual pass. For methods originally designed for single-label prediction, we use independent sigmoid/Bernoulli predictions and adapt their objectives to multi-label outputs while preserving their core adaptation mechanisms as closely as possible.

Table 7: Implementation summary of the evaluated baselines under the multi-label continual test-time adaptation setting.

Method	Multi-label adaptation	Updated parameters	Key components
TENT	Factorized Bernoulli entropy over independent sigmoid predictions	BN affine parameters only	Entropy minimization
CoTTA	Element-wise BCE against detached EMA-teacher sigmoid probabilities	Full student model	EMA teacher; stochastic restoration; confidence-gated augmentation
ViDA	Symmetric Bernoulli teacher–student consistency	Base model and ViDA adapters	Dual-rank adapters; EMA teacher
DSS	Sigmoid pseudo-supervision with sample selection and complementary negative learning	Full student model	EMA teacher; dynamic sample selection
FATA-TENT	Bernoulli entropy with multi-hot BCE for feature-augmented samples	Normalization affine parameters before layer4	Feature augmentation
GOLD	Bernoulli teacher–student consistency with prototype and contrastive objectives	BN affine parameters and GOLD-specific adaptation modules	Source-side preparation; source prototypes; golden subspace

TENT (Wang et al., 2021). We replace categorical entropy with factorized Bernoulli entropy over independent sigmoid outputs. Only BN affine parameters are optimized, while the remaining backbone and classifier parameters are frozen.

CoTTA (Wang et al., 2022). Softmax teacher targets are replaced by independent sigmoid probabilities, and the student minimizes element-wise BCE against detached EMA-teacher outputs. We retain the EMA teacher, stochastic restoration, confidence-gated augmentation, and full-model student adaptation.

ViDA (Liu et al., 2024). The categorical consistency objective is replaced by symmetric Bernoulli teacher–student consistency. ViDA adapters are inserted into selected TResNet blocks, and both the base model and adapters are adapted under EMA-teacher supervision.

DSS (Wang et al., 2024). Dynamic sample selection is retained under independent sigmoid predictions. Selected samples provide multi-label pseudo-supervision with complementary negative learning, and the full student model is adapted.

FATA-TENT (Cho et al., 2025). Categorical entropy and single-label pseudo-targets are replaced by Bernoulli entropy and multi-hot sigmoid targets, respectively. The feature augmentation mechanism is retained, while optimization is restricted to the corresponding normalization affine parameters before layer4.

GOLD (Lai et al., 2026). For each dataset, the corresponding source-side preparation is performed before deployment; source images and labels are used only during this stage, and no source data are accessed during the target stream. Source prototypes and golden-subspace-related states are prepared in advance. At test time, adaptation is restricted to the GOLD-specific modules and BN affine parameters, using the Bernoulli teacher–student and prototype-based contrastive objectives.

Recent multi-label TTA methods such as BEM (Wu et al., 2025), BCP (Li et al., 2026), and PuRF (Liang et al., 2026) are developed primarily for vision–language or CLIP-style settings,

whereas our study considers source-trained TResNet-M discriminative classifiers. Because their backbones, source initialization, and adaptation interfaces differ substantially from our setting, we do not include them in the direct quantitative comparison.

C Additional Ablation Studies

C.1 Effect of Adaptation Scope

We compare four update scopes on VOC-C under the standard Full15 continual protocol at severity 5 without resetting between corruption domains. Restricted scopes produce similar results around 80.2–80.3 mAP, whereas full-model adaptation reaches 82.222 mAP. This supports the use of full-model adaptation in the main experiments.

Table 8: Effect of adaptation scope on VOC-C. Results are EMA-teacher full-stream mAP (%) under the standard Full15 continual protocol.

Update scope	mAP (%)
Norm only	80.231
Head only	80.196
Layer4 only	80.305
Full model	**82.222**

C.2 Sensitivity to the Routing Coefficient

Performance remains stable across the tested values, with the default $\lambda_D = 1.2$ attaining the highest mAP. Varying λ_D also changes the frequency of Hard supervision activation, but DSR is not highly sensitive within this tested range on VOC-C.

Table 9: Sensitivity to the routing coefficient λ_D on VOC-C. All other DSR settings are kept unchanged.

λ_D	Hard-active batches	Activation ratio (%)	mAP (%)
1.0	480 / 2322	20.672	81.920
1.2	**317 / 2322**	**13.652**	**82.222**
1.4	319 / 2322	13.738	81.813

D Understanding Tri-State Hard Supervision

D.1 Why Negative Supervision Matters

Across the 8,348 Hard-active batches of the NUS-WIDE-C main run, 95.8455% of sample–class pairs fall in the Negative region, while only 0.8737% fall in the Positive region (Table 10). This dominance makes negative supervision essential for constraining uncontrolled positive expansion. Negative-label dominance is also a recognized challenge in supervised multi-label learning (Wu et al., 2020); in continual adaptation, it additionally interacts with noisy hard targets. Although Negative pseudo-label accuracy reaches 99.2258%, the remaining 0.7742% corresponds to 160,554 true-positive labels incorrectly assigned as Negative, comparable to the 160,421 correctly supervised positives in the Positive region. Negative supervision is therefore necessary, but its scale amplifies even a small error rate, motivating conservative assignment together with an Ignore region.

Table 10: Ground-truth composition of the tri-state regions on NUS-WIDE-C over Hard-active batches. GT ratios are computed within each region.

Region	Share (%)	GT = 1 (%)	GT = 0 (%)
Positive	0.8737	84.8618	15.1382
Negative	95.8455	0.7742	99.2258
Ignore	3.2808	45.2690	54.7310

D.2 Why Ignore Ambiguous Predictions

The Ignore region is highly mixed, containing 45.2690% positive and 54.7310% negative ground-truth labels. It is therefore unsuitable for direct binary assignment. As Table 11 shows, binarizing these pairs at 0.5 would make 31.0776% incorrect hard targets, including 18.3202% true positives forced to Negative. Such an incorrect Negative target creates an explicit wrong-direction hard update for the corresponding class.

Table 11: Forced binary assignment within the Ignore region on NUS-WIDE-C. All ratios are computed over the entire Ignore region; the last two rows partition the incorrect assignments.

Outcome	Count	Ratio (%)
Correct	489,234	68.9224
Wrong	220,599	31.0776
GT = 1 → Negative	130,043	18.3202
GT = 0 → Positive	90,556	12.7574

Confidence-based pseudo-label selection has been studied in semi-supervised learning (Sohn et al., 2020). Here, selection operates on individual sample–class pairs and applies only to the additional Hard objective. Importantly, Ignore does not remove these predictions from adaptation: they remain supervised by the persistent Soft Consistency objective, while only the unreliable discrete Hard target is omitted.

D.3 Why ASL Instead of BCE

Tri-State and Asymmetric Loss (ASL) (Ridnik et al., 2021a) address complementary issues: Tri-State controls pseudo-label reliability, whereas ASL addresses the positive–negative imbalance among the retained hard targets. With 95.8455% of Hard-active pairs in the Negative region, symmetric treatment by standard binary cross-entropy can allow abundant negatives to dominate the aggregate hard objective. Reducing the contribution of well-classified examples is also the central motivation of focal loss (Lin et al., 2017). We use ASL with $\gamma_{\mathrm{pos}} = 0$, $\gamma_{\mathrm{neg}} = 2$, and clipping parameter 0.05 to reduce the contribution of easy negatives while retaining informative negative correction. Thus, ASL complements reliability-based Tri-State selection without changing which predictions receive discrete Hard supervision.