# NOT EVERY CORRECTION HELPS: GAIN-GUIDED CONTINUAL TEST-TIME ADAPTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Continual test-time adaptation (CTTA) adapts a source model to an unlabeled test stream whose distribution may change over time. Existing TTA methods often assess prediction reliability using confidence or entropy, which primarily reflect the model's self-certainty for the current sample. In CTTA, accumulated target observations can provide complementary evidence for correcting the source prediction, but this history may become misaligned as the target distribution changes. The key question is therefore not how much the correction differs from the source prediction, but whether and how strongly it should be applied. This paper proposes **G**ain-**A**ware **IN**tervention (**GAIN**), a backpropagation-free CTTA framework guided by a simple principle: *history proposes, gain decides*. **GAIN** maintains compact target statistics to form a correction proposal and a posterior-predictive evaluator that accounts for estimation uncertainty. The resulting source-relative gain estimates the proposal's benefit and determines a sample-specific intervention strength along a continuous path through efficient one-dimensional optimization. Gain-controlled predictions then update the target statistics online, limiting the propagation of unreliable corrections, all without backpropagation, sample storage, or replay. Across five benchmarks, our method achieves strong predictive performance, with favorable accuracy–calibration–efficiency trade-offs in continual adaptation. On ImageNet-C, for example, **GAIN** achieves 61.9% accuracy with near-source calibration. It remains stable under diverse and challenging continual shifts while running $15.9\times$ faster than a representative optimization-based CTTA baseline.

## 1 INTRODUCTION

Continual test-time adaptation (CTTA) (Wang et al., 2022) adapts a source model online to an unlabeled target stream whose distribution evolves over time. A central challenge is deciding when to retain the source prediction and when to correct it. Existing methods assess adaptation reliability from the current sample through entropy-based objectives (Wang et al., 2021; Song et al., 2023; Han et al., 2025a), reliability-aware sample selection (Niu et al., 2022; Lee et al., 2024b; Gong et al., 2023), or region-level confidence modeling (Hu et al., 2025). Such measures can be useful indicators of prediction reliability, yet they primarily characterize the model's *self-certainty*. They do not explicitly capture how the same prediction may be supported differently by the evolving target context. This observation motivates us to look beyond the current output and treat accumulated target observations as additional predictive evidence.

Recent CTTA methods exploit historical target information through feature statistics or distribution estimation (Zhang et al., 2025c; Han et al., 2025b; Zhang et al., 2025a), allowing accumulated target information to guide subsequent predictions. However, this creates a distinct challenge in continual adaptation. As the target distribution evolves, historical statistics may become misaligned with the current domain, while erroneous corrections can accumulate in the target state and affect subsequent adaptation. Consequently, a history-induced correction is not necessarily beneficial simply because it differs from the source prediction. This raises a more fundamental question for CTTA: not whether target history suggests a correction, but whether that correction is worth applying.

Motivated by this perspective, we introduce **G**ain-**A**ware **IN**tervention (**GAIN**), following the principle that *history proposes, gain decides*. **GAIN** uses accumulated target statistics to propose a target-side correction and evaluates its source-relative gain while accounting for uncertainty in the evolving



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.

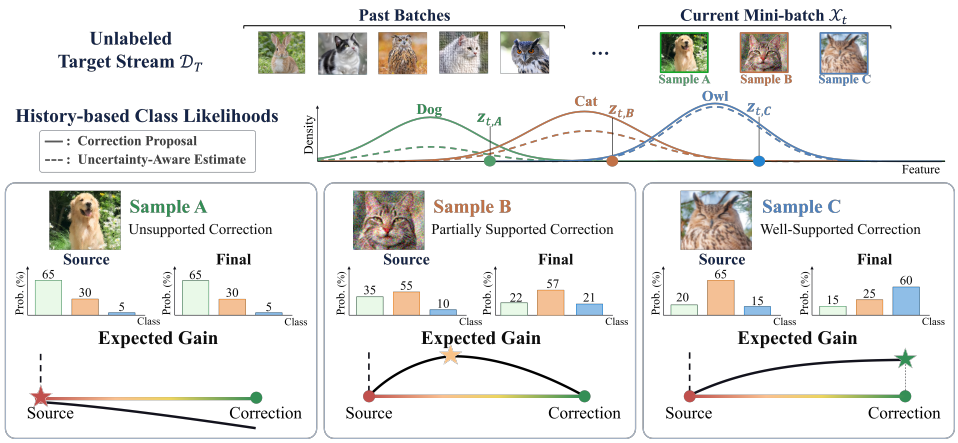



Figure 1: **Gain-Guided Sample-wise Intervention.** Past target batches are summarized by compact statistics to construct a history-based correction proposal (solid) and an uncertainty-aware posterior-predictive estimate (dashed) that evaluates it. The resulting source-relative gain determines a sample-specific strength along the source-to-correction path: unsupported corrections are rejected (Sample A), partially supported corrections are applied conservatively (Sample B), and well-supported corrections are fully adopted (Sample C).

target statistics. The resulting gain determines whether and how strongly the correction should influence the current prediction.

To leverage correction gain for reliable continual adaptation, **GAIN** combines sample-specific intervention with online target-state maintenance. As shown in Fig. 1, the intervention strength is selected along a continuous path from source retention to full target correction through a concave one-dimensional objective with an efficient global solution. The resulting predictions then update the target state causally, limiting the propagation of unsupported history-induced corrections to subsequent adaptation. Throughout adaptation, the source model remains frozen, requiring no backpropagation, parameter updates, or replay of past target samples. Across diverse continual-shift settings, **GAIN** improves accuracy while maintaining strong calibration and long-horizon stability.

- We introduce a *source-relative correction-gain* perspective for CTTA, shifting the focus from the apparent reliability of the current prediction to whether a history-derived correction improves upon source retention.
- **GAIN** analytically marginalizes class-center uncertainty and selects a sample-specific intervention through a concave one-dimensional evidence-path objective, while updating compact target statistics online without sample storage or replay.
- Extensive experiments demonstrate that **GAIN** consistently achieves a strong accuracy–calibration–efficiency trade-off across structured, dynamic, mixed-domain, and long-horizon continual shifts.

## 2 Related Work

**Reliability and Stability in Test-Time Adaptation.** CTTA extends test-time adaptation to evolving unlabeled streams, where repeated self-adaptation can amplify prediction errors and lead to long-term instability (Wang et al., 2022). Prior work improves reliability through entropy-based adaptation and sample filtering or reweighting (Wang et al., 2021; Niu et al., 2022; Lee et al., 2024b; Wang et al., 2024), stabilized or accelerated optimization (Song et al., 2023; Niu et al., 2023; Han et al., 2025a; Duan et al., 2025; Choi et al., 2025), and representation, structural, geometric, or subspace adaptation (Liu et al., 2024a; Yang et al., 2024; Wang et al., 2025; Ni et al., 2025; Liu et al., 2026; Murphy et al., 2026; Lai et al., 2026). CAS (Jiang et al., 2026) uses cross-augmentation similarity to make a binary adapt-or-skip decision when adaptation may be harmful. Beyond this binary decision, **GAIN** evaluates whether a history-induced correction improves over retaining the source prediction and continuously controls its intervention strength through source-relative gain.

**Knowledge Preservation in CTTA.** Under continual shifts, preserving useful knowledge is important for mitigating forgetting and cross-domain interference. Existing methods preserve or reuse information through source-weight restoration or ensembling (Wang et al., 2022; Marsden et al., 2024), sample storage (Yuan et al., 2023), domain-specific modules or experts (Liu et al., 2024b; Lee et al., 2024a; Zhao et al., 2026), and compact prompt or knowledge pools (Niu et al., 2024; Zhang et al., 2025c; Zhou et al., 2025). More recently, DO-ALL (Jang et al., 2026) improves long-term stability by distilling synthetic source anchors for replay during adaptation. In contrast, **GAIN** keeps the source model frozen and maintains only compact target statistics, using gain-guided intervention to regulate how history-induced corrections influence current predictions and subsequent target-state updates without sample storage or replay.

**Distributional Modeling in CTTA.** Distributional modeling has been increasingly explored for continual adaptation. PETAL (Brahma & Rai, 2023) formulates lifelong TTA probabilistically, while BayesTTA (Cui et al., 2025) incrementally models class-conditional distributions under temporal shifts. DOTA (Han et al., 2025b) further estimates evolving test-time feature distributions and derives posterior predictions from accumulated target statistics. Related statistical TTA methods incorporate source-informed priors or analytic inference (Zanella et al., 2025; Zhang et al., 2025a;b). While these approaches improve target-side estimation, accumulated statistics may become inaccurate or stale as the target distribution changes. Rather than directly treating the target estimate as the final prediction, **GAIN** uses it as a correction proposal and evaluates its source-relative gain to determine whether and how strongly it should influence the source prediction, thereby limiting the propagation of unreliable corrections through subsequent adaptation.

# 3 PRELIMINARIES

## 3.1 CONTINUAL TEST-TIME ADAPTATION

**Problem Setup.** Given a source model $f_\theta$ pre-trained on a labeled source domain $\mathcal{D}_S$, continual test-time adaptation (CTTA) considers an unlabeled target stream $\mathcal{D}_T = \{\mathcal{X}_t\}_{t=1}^{T}$ whose distribution may change over time. We decompose the source model as $f_\theta = g_\theta \circ \phi_\theta$, where $\phi_\theta$ and $g_\theta$ denote the feature extractor and classifier head, respectively. At time $t$, the model observes a target mini-batch $\mathcal{X}_t$ with $|\mathcal{X}_t| = B_t$. For each sample $\mathbf{x}_t \in \mathcal{X}_t$, the frozen source model produces

$$\mathbf{s}_t = f_\theta(\mathbf{x}_t) \in \Delta^{K-1}, \qquad \hat{y}_t^s = \arg\max_k s_{t,k}, \tag{1}$$

where $\mathbf{s}_t$ denotes the source predictive distribution and $\hat{y}_t^s$ is the corresponding predicted class. Preceding target observations $\mathcal{H}_t = \bigcup_{\tau<t} \mathcal{X}_\tau$ form the accumulated target context. We keep $f_\theta$ frozen and exploit $\mathcal{H}_t$ for single-pass adaptation without gradient-based model updates.

**Beyond Source Confidence.** Since target labels are unavailable, CTTA commonly relies on source-output proxies such as confidence or entropy to assess prediction reliability (Wang et al., 2021; Niu et al., 2022; Han et al., 2025a). However, these signals reflect only the model's self-certainty and ignore accumulated target context, under which the same source prediction may have different reliability (Appendix A). Historical observations can therefore provide complementary predictive information, but their relevance may diminish as the target distribution changes. We thus ask whether conditioning on $\mathcal{H}_t$ provides useful evidence for the current sample.

## 3.2 ACCUMULATED TARGET CONTEXT AS PREDICTIVE EVIDENCE

**Historical Target Evidence.** Let $q_{t,k}^0 \triangleq P_T(Y_t = k \mid \mathbf{x}_t)$ denote the target posterior probability for class $k$ given the current observation alone, and let $q_{t,k}^H \triangleq P_T(Y_t{=}k \mid \mathbf{x}_t, \mathcal{H}_t)$ denote the corresponding probability additionally conditioned on the accumulated target context. By Bayes' rule,

$$q_{t,k}^H = q_{t,k}^0 \frac{P_T(\mathcal{H}_t \mid Y_t = k, \mathbf{x}_t)}{P_T(\mathcal{H}_t \mid \mathbf{x}_t)}, \tag{2}$$

showing that historical context contributes class-dependent evidence beyond the current observation. In log-probability space, this contribution is

$$\log q_{t,k}^H = \log q_{t,k}^0 + \xi_{t,k}, \qquad \xi_{t,k} \triangleq \log(q_{t,k}^H / q_{t,k}^0) = i(Y_t = k; \mathcal{H}_t \mid \mathbf{x}_t), \tag{3}$$

where $\xi_{t,k}$ is the conditional pointwise mutual information (C-PMI) (Fano, 1966; Ren et al., 2023). Positive and negative values indicate that the accumulated target context provides additional evidence for and against class $k$, respectively. Details are provided in Appendix B.1.

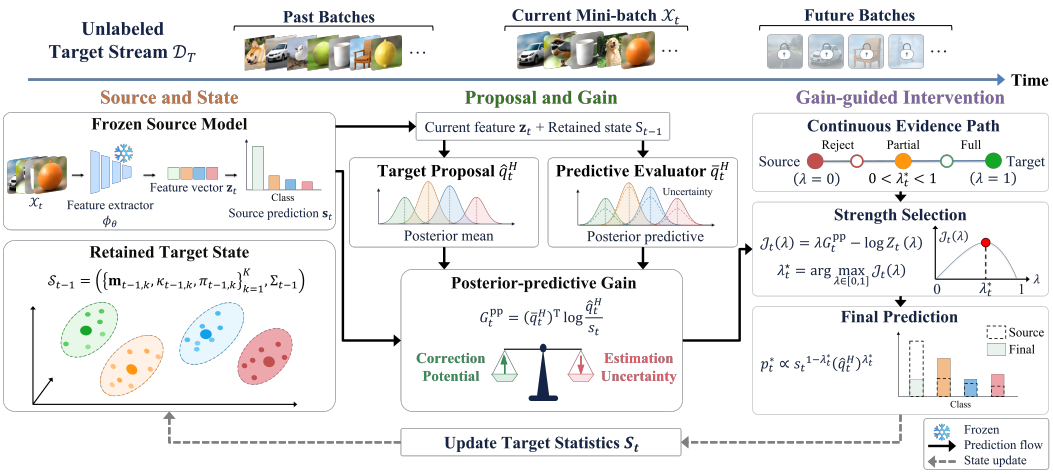



Figure 2: **Overview of GAIN: history proposes, gain decides. (i)** The frozen source model produces the source prediction $\mathbf{s}_t$ and current feature $\mathbf{z}_t$, while the retained target state $\mathcal{S}_{t-1}$ summarizes past batches. **(ii)** Together, $\mathbf{z}_t$ and $\mathcal{S}_{t-1}$ form the target proposal $\hat{\mathbf{q}}_t^H$ and predictive evaluator $\bar{\mathbf{q}}_t^H$. The evaluator accounts for uncertainty to estimate the proposal's source-relative gain $G_t^{\text{pp}}$. **(iii)** This source-relative gain determines the intervention strength $\lambda_t^\star$ along the continuous evidence path from $\mathbf{s}_t$ to $\hat{\mathbf{q}}_t^H$. The final prediction $\mathbf{p}_t^\star$ then updates the target statistics $\mathcal{S}_t$ for future batches.

**Proposition 3.1** (Non-Negative Predictive Value of Target History)**.** *For any predictive distribution $\mathbf{r} \in \Delta^{K-1}$, define the logarithmic risk under the history-conditioned target posterior as $\mathcal{R}_t(\mathbf{r}) \triangleq \mathbb{E}_{Y_t \sim \mathbf{q}_t^H}[-\log r_{Y_t}]$. Conditioning on the accumulated target context $\mathcal{H}_t$ then yields non-negative predictive value under logarithmic loss:*

$$\mathcal{R}_t(\mathbf{q}_t^0) - \mathcal{R}_t(\mathbf{q}_t^H) = D_{\text{KL}}(\mathbf{q}_t^H \| \mathbf{q}_t^0) = \mathbb{E}_{Y_t \sim \mathbf{q}_t^H}[\xi_{t,Y_t}] \geq 0. \tag{4}$$

*Equality holds if and only if $\mathbf{q}_t^H = \mathbf{q}_t^0$.*

Proposition 3.1 establishes that target history is predictively useful in the oracle setting (proof in Appendix B.2). In practice, however, CTTA only has access to an estimate of $\mathbf{q}_t^H$ constructed from finite, unlabeled, and potentially stale observations. Thus, informative history does not guarantee that the resulting estimated correction is beneficial, motivating our source-relative gain formulation.

## 4 GAIN: GAIN-AWARE INTERVENTION

For a current sample $\mathbf{x}_t \in \mathcal{X}_t$, we denote its representation by $\mathbf{z}_t = \phi_\theta(\mathbf{x}_t) \in \mathbb{R}^D$. Conceptually, let $\mathcal{Z}_t^H = \{\mathbf{z}_i^H \mid \mathbf{z}_i^H = \phi_\theta(\mathbf{x}_i^H), \mathbf{x}_i^H \in \mathcal{H}_t\}$ with $N_t^H = |\mathcal{H}_t|$ denote the representations associated with preceding target observations. We introduce $\mathcal{Z}_t^H$ only for notational convenience. In practice, we do not store or replay these historical representations, but maintain their aggregate effect through recursive sufficient statistics.

### 4.1 RELIABLE TARGET-SIDE GAIN ESTIMATION

**From Historical Evidence to Correction Gain.** Preliminaries establish that accumulated target context has non-negative predictive value in the oracle setting. At test time, however, the history-conditioned posterior $\mathbf{q}_t^H$ is unavailable and must be approximated by an estimate $\hat{\mathbf{q}}_t^H$. If the source prediction $\mathbf{s}_t$ is fully replaced by this target estimate, the resulting conditional log-risk reduction is

$$\Delta_t^{\text{full}} \triangleq \mathcal{R}_t(\mathbf{s}_t) - \mathcal{R}_t(\hat{\mathbf{q}}_t^H) = D_{\text{KL}}(\mathbf{q}_t^H \| \mathbf{s}_t) - D_{\text{KL}}(\mathbf{q}_t^H \| \hat{\mathbf{q}}_t^H). \tag{5}$$

Eq. 5 shows that informative target history does not necessarily yield a beneficial correction, since estimation error can offset its correction potential. This motivates two practical requirements: constructing a target-side correction proposal from accumulated history and evaluating whether that proposal improves upon retaining the source prediction.

**Source-Anchored Probabilistic Target Estimation.** As illustrated in Fig. 2, we address both requirements through a probabilistic model of accumulated target evidence. Its posterior mean defines a class-specific target correction proposal, while the posterior-predictive distribution accounts for estimation uncertainty when evaluating the source-relative utility of that proposal. Specifically, for each class $k$, we model target features with a class-conditional Gaussian distribution and place a source-centered prior on its unknown class center $\boldsymbol{\mu}_k$ to stabilize estimation when target evidence is limited or noisy:

$$p_T(\mathbf{z} \mid Y = k, \boldsymbol{\mu}_k) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \Sigma_{t-1}), \qquad \boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{c}_k, \frac{\Sigma_{t-1}}{\kappa_0}), \tag{6}$$

where $\mathbf{c}_k$ is the source-derived class prototype (*i.e.*, the k-th classifier-head weight) and $\kappa_0$ controls the strength of the source-centered prior. We initialize $\Sigma_0 = \mathbf{I}_D$ and update the shared diagonal covariance causally from preceding target observations.

Using the recursively maintained sufficient statistics induced by soft class assignments of preceding target observations, together with the pre-$t$ covariance estimate, we obtain the following fractional Gaussian posterior over the class center:

$$\boldsymbol{\mu}_k \mid \mathcal{H}_t \approx \mathcal{N}(\mathbf{m}_{t-1,k}, \frac{\Sigma_{t-1}}{\kappa_{t-1,k}}), \qquad \kappa_{t-1,k} = \kappa_0 + n_{t-1,k}, \tag{7}$$

where $n_{t-1,k}$ and $\mathbf{m}_{t-1,k}$ denote the retained target support and the source-anchored center. The complete fractional-posterior derivation and recursive updates are given in Appendices C.1 and E.1.

Using the posterior mean geometry, we define

$$d_{t,k} = (\mathbf{z}_t - \mathbf{m}_{t-1,k})^\top \Sigma_{t-1}^{-1} (\mathbf{z}_t - \mathbf{m}_{t-1,k}), \qquad \hat{\ell}^H_{t,k} = \log \pi_{t-1,k} - \frac{1}{2} d_{t,k}, \tag{8}$$

and obtain the posterior-mean target proposal:

$$\hat{q}^H_{t,k} = \frac{\pi_{t-1,k} \exp(-d_{t,k}/2)}{\sum_j \pi_{t-1,j} \exp(-d_{t,j}/2)}. \tag{9}$$

The distribution $\hat{\mathbf{q}}^H_t = [\hat{q}^H_{t,1}, \ldots, \hat{q}^H_{t,K}]^\top$ specifies the target-side correction proposed by the posterior-mean geometry, where $\pi_{t-1,k}$ is the pre-$t$ target class prior. When $\mathcal{H}_t = \varnothing$, we set $\hat{\mathbf{q}}^H_t = \mathbf{s}_t$.

**Posterior-Predictive Gain Evaluation.** The proposal $\hat{\mathbf{q}}^H_t$ is constructed from the posterior-mean target geometry and therefore does not account for the remaining uncertainty in the estimated class centers. To incorporate this uncertainty, we analytically marginalize the latent class centers. From Eq. 7 and the class-conditional observation model in Eq. 6, the posterior-predictive likelihood for class $k$ is $p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t) = \mathcal{N}(\mathbf{z}_t; \mathbf{m}_{t-1,k}, h_{t,k}\Sigma_{t-1})$, $h_{t,k} \triangleq 1 + \kappa^{-1}_{t-1,k}$. Thus, classes with less precisely estimated centers induce broader posterior-predictive distributions. Normalizing these predictive likelihoods gives the posterior-predictive evaluator

$$\bar{q}^H_{t,k} \propto \pi_{t-1,k} h_{t,k}^{-D/2} \exp(-\frac{d_{t,k}}{2h_{t,k}}), \qquad h_{t,k} = 1 + \kappa^{-1}_{t-1,k}. \tag{10}$$

Here, $\hat{\mathbf{q}}^H_t$ and $\bar{\mathbf{q}}^H_t$ serve distinct roles: the former specifies the correction proposed by the estimated target geometry, whereas the latter evaluates that correction after accounting for class-center uncertainty under the posterior-predictive working model. Define the source-relative correction evidence $\hat{\xi}^s_{t,k} \triangleq \log \frac{\hat{q}^H_{t,k}}{s_{t,k}}$. Then the oracle gain in Eq. 5 can be written as $\Delta^{\mathrm{full}}_t = (\mathbf{q}^H_t)^\top \hat{\boldsymbol{\xi}}^s_t$. Since $\mathbf{q}^H_t$ is unavailable at test time, we evaluate the same correction under the posterior-predictive distribution:

$$G^{\mathrm{pp}}_t \triangleq (\bar{\mathbf{q}}^H_t)^\top \hat{\boldsymbol{\xi}}^s_t = D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t \| \mathbf{s}_t\right) - D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t \| \hat{\mathbf{q}}^H_t\right). \tag{11}$$

The first KL term captures the potential benefit of correcting the source prediction, while the second KL measures the mismatch between the proposed correction and its posterior-predictive evaluation. Thus, source–target disagreement alone does not justify correction; the proposed correction must also remain supported after accounting for uncertainty in the estimated target geometry. Accordingly, $G^{\mathrm{pp}}_t$ is the expected gain of the proposed correction under the posterior-predictive working model, rather than a lower bound on the unknown oracle gain $\Delta^{\mathrm{full}}_t$. See Appendix C.2 for details.

### 4.2 Posterior-Predictive Evidence Intervention

The posterior-predictive gain in Eq. 11 quantifies how strongly the proposed correction is supported after accounting for uncertainty in the target geometry. We now translate this quantity into the extent of intervention on the frozen source prediction.

**Continuous Evidence Path.** Rather than directly replacing $\mathbf{s}_t$ with the target estimate, we continuously scale the source-relative correction evidence $\hat{\boldsymbol{\xi}}_t^s$ by an intervention coefficient $\lambda \in [0,1]$:

$$p_{t,k}^{(\lambda)} = \frac{s_{t,k}\exp(\lambda \hat{\xi}_{t,k}^s)}{Z_t(\lambda)} = \frac{s_{t,k}^{1-\lambda}(\hat{q}_{t,k}^H)^\lambda}{\sum_{j=1}^K s_{t,j}^{1-\lambda}(\hat{q}_{t,j}^H)^\lambda}, \tag{12}$$

where $Z_t(\lambda)=\sum_j s_{t,j}\exp(\lambda\hat{\xi}_{t,j}^s)$. The two endpoints satisfy $\mathbf{p}_t^{(0)}=\mathbf{s}_t$ and $\mathbf{p}_t^{(1)}=\hat{\mathbf{q}}_t^H$. Thus, $\lambda$ controls the amount of target-side correction introduced relative to the source prediction. Importantly, for the current prediction, the posterior-predictive distribution $\bar{q}_t^H$ serves as an evaluator rather than as an additional replacement prediction. Define its conditional logarithmic risk as $\bar{\mathcal{R}}_t(\mathbf{p})\triangleq-\sum_k \bar{q}_{t,k}^H \log p_k$. Then, the reduction in posterior-predictive risk relative to the source prediction is

$$\begin{aligned}\mathcal{J}_t(\lambda) &\triangleq \bar{\mathcal{R}}_t(\mathbf{s}_t) - \bar{\mathcal{R}}_t(\mathbf{p}_t^{(\lambda)}) \\ &= \lambda(\bar{\mathbf{q}}_t^H)^\top \hat{\boldsymbol{\xi}}_t^s - \log Z_t(\lambda) \\ &= \lambda G_t^{\mathrm{pp}} - \log Z_t(\lambda), \qquad \lambda \in [0,1].\end{aligned} \tag{13}$$

Hence, intervention is determined by the gain of the same target correction after accounting for uncertainty in the estimated target geometry, while the normalization term follows exactly from the source-relative evidence path.

**Theorem 4.1** (Globally Optimal Posterior-Predictive Intervention)**.** *For the evidence path $\mathbf{p}_t^{(\lambda)}$ in Eq. 12, the posterior-predictive objective $\mathcal{J}_t(\lambda)$ is concave over $\lambda \in [0,1]$. Consequently, it admits a globally optimal intervention coefficient $\lambda_t^\star$, yielding the final adapted prediction*

$$p_{t,k}^\star = p_{t,k}^{(\lambda_t^\star)} \propto s_{t,k}^{1-\lambda_t^\star}\left(\hat{q}_{t,k}^H\right)^{\lambda_t^\star}, \qquad \lambda_t^\star = \arg\max_{\lambda\in[0,1]} \mathcal{J}_t(\lambda). \tag{14}$$

The proof and the efficient one-dimensional solution for $\lambda_t^\star$ are provided in Appendix D.

**Continual Target Update.** In CTTA, historical statistics can become mismatched with the current distribution, while unreliable corrections may accumulate through subsequent state updates. **GAIN** mitigates this propagation by updating the target state only from the gain-controlled predictions. Specifically, we define the reliability-weighted assignment $\omega_{t,b,k} = \zeta_{t,b} p_{t,b,k}^\star$, where $\zeta_{t,b} = s_{t,b,\hat{y}_{t,b}^\star}$ measures frozen-source support for the adapted prediction $\hat{y}_{t,b}^\star = \arg\max_k p_{t,b,k}^\star$. The reliability-weighted class support $\kappa_{t,k}$ and predictive class mass $\hat{\kappa}_{t,k}$ are then accumulated as:

$$\kappa_{t,k} = \kappa_{t-1,k} + \sum_{b=1}^{B_t} \omega_{t,b,k}, \qquad \hat{\kappa}_{t,k} = \hat{\kappa}_{t-1,k} + \sum_{b=1}^{B_t} p_{t,b,k}^\star, \tag{15}$$

with $\kappa_{0,k} = \hat{\kappa}_{0,k} = \kappa_0$. While $\hat{\kappa}_{t,k}$ reflects how frequently class $k$ is predicted, $\kappa_{t,k}$ measures how strongly these assignments are supported. To further prevent frequently predicted classes from being progressively reinforced, we define the historical class prior by combining reliability-normalized support with inverse-support balancing:

$$\pi_{t,k} \propto \frac{\kappa_{t,k}}{\hat{\kappa}_{t,k}} \cdot \frac{1}{\hat{\kappa}_{t,k}} = \frac{\kappa_{t,k}}{\hat{\kappa}_{t,k}^2}. \tag{16}$$

The same reliability-weighted evidence recursively updates the class centers and shared covariance, yielding $\mathcal{S}_t = \left(\{\mathbf{m}_{t,k}, \kappa_{t,k}, \pi_{t,k}\}_{k=1}^K, \Sigma_t\right)$. The updated state is used only from time $t+1$ onward, limiting the repeated reinforcement of unreliable history-induced corrections without storing or replaying previous target samples. Full recursive updates are provided in Appendix E.1.

# 5 Experiments

## 5.1 Experimental Setup

**Datasets and Metrics.** We evaluate **GAIN** on ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-3DCC (Kar et al., 2022), ImageNet-R (Hendrycks et al., 2021), ImageNet-V2 (Recht et al., 2019), and ImageNet-Sketch (Wang et al., 2019) to assess robustness under diverse distribution shifts. For ImageNet-C and ImageNet-3DCC, we use corruption severity 5 unless otherwise specified and perform continual adaptation without reset across the stream. We consider four complementary stream settings: continual structured change (CSC) (Wang et al., 2022), continual dynamic change (CDC) (Zhang et al., 2025c), mixed-domain shift (MDS) (Niu et al., 2023; Hu et al., 2025), and long-horizon adaptation (LHA) (Liu et al., 2024b) over 10 repeated corruption cycles. We report top-1 accuracy (Acc.) and expected calibration error (ECE) (Naeini et al., 2015).

Table 1: **CSC results on ImageNet-C.** Accuracy (Acc., %) and expected calibration error (ECE, %) with ViT-Base at severity level 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only.

| Method | BP-free | Metric | Noise | | | Blur | | | | Weather | | | | Digital | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Gauss. | Shot | Impu. | Defo. | Glas. | Moti. | Zoom | Snow | Fros. | Fog | Brig. | Cont. | Elas. | Pix. | JPEG | |
| Source | – | Acc. ↑ | 47.0 | 48.2 | 47.9 | 31.5 | 21.2 | 41.5 | 36.7 | 50.1 | 45.8 | 42.3 | 73.6 | 8.6 | 42.5 | 62.0 | 63.8 | 44.2 |
| | | ECE ↓ | 3.6 | 4.1 | 3.7 | 4.3 | 5.4 | 3.9 | 9.1 | 2.3 | 4.9 | 17.4 | 3.1 | 3.9 | 9.0 | 3.3 | 2.7 | 5.4 |
| Tent (ICLR 2021) | ✗ | Acc. ↑ | 47.8 | 51.1 | 50.8 | 34.2 | 27.0 | 45.5 | 41.6 | 56.0 | 52.3 | 49.7 | 76.1 | 27.2 | 44.3 | 65.6 | 66.1 | 49.0 |
| | | ECE ↓ | 5.8 | 8.3 | 10.5 | 12.1 | 18.0 | 13.9 | 18.2 | 12.4 | 14.6 | 13.3 | 7.0 | 19.1 | 20.2 | 10.3 | 8.7 | 12.8 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 47.1 | 48.4 | 48.6 | 31.7 | 21.9 | 42.9 | 38.0 | 51.8 | 47.3 | 44.7 | 74.1 | 10.0 | 43.6 | 63.6 | 64.8 | 45.2 |
| | | ECE ↓ | **4.0** | 5.3 | 5.6 | **3.1** | 8.4 | 7.2 | 13.0 | 6.6 | 10.9 | 11.6 | 4.7 | **2.7** | 15.7 | 8.3 | 5.5 | 7.5 |
| SAR (ICLR 2023) | ✗ | Acc. ↑ | 54.2 | 54.1 | 52.3 | 47.7 | 36.3 | 53.8 | 49.1 | 59.7 | 57.6 | 58.2 | 75.6 | 46.6 | 46.4 | 61.6 | 63.4 | 54.4 |
| | | ECE ↓ | 4.8 | 5.7 | 7.6 | 4.9 | 12.9 | 9.6 | 13.5 | 9.8 | 10.2 | 9.7 | 5.1 | 14.8 | 14.0 | 7.0 | 6.5 | 9.1 |
| ROID (WACV 2024) | ✗ | Acc. ↑ | 56.3 | 62.3 | 60.7 | 51.0 | 50.6 | **59.2** | 54.8 | 63.9 | 62.2 | 64.0 | 78.7 | 50.7 | 58.8 | 69.5 | 70.0 | 60.8 |
| | | ECE ↓ | 56.4 | 61.9 | 60.5 | 51.3 | 51.3 | 58.9 | 54.6 | 63.4 | 62.3 | 64.5 | 78.3 | 49.1 | 58.8 | 69.9 | 69.5 | 60.7 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 52.3 | 57.5 | 57.1 | 47.8 | 43.1 | 54.5 | 51.1 | 61.1 | 57.3 | 59.3 | 75.7 | 47.2 | 50.9 | 66.5 | 66.9 | 56.6 |
| | | ECE ↓ | 6.8 | 11.0 | 13.7 | 10.8 | 20.4 | 14.7 | 19.5 | 13.9 | 16.5 | 15.3 | 8.1 | 24.0 | 23.3 | 11.2 | 11.0 | 14.7 |
| DeYO (ICLR 2024) | ✗ | Acc. ↑ | 52.8 | 60.6 | 59.8 | 45.3 | 46.2 | 57.5 | 50.9 | 60.7 | 60.0 | 60.2 | 77.2 | 51.1 | 55.1 | 67.7 | 69.5 | 58.3 |
| | | ECE ↓ | 5.7 | 6.3 | 8.3 | 6.9 | 13.3 | 10.1 | 15.5 | 11.0 | 11.1 | 10.7 | 5.5 | 14.6 | 13.2 | 8.3 | 7.8 | 9.9 |
| AEA (ICLR 2025) | ✗ | Acc. ↑ | 54.1 | 55.1 | 55.6 | **55.5** | **51.9** | 58.6 | 49.3 | 11.4 | 25.7 | 69.2 | 73.7 | **64.5** | 59.1 | 70.1 | 66.9 | 54.7 |
| | | ECE ↓ | 18.6 | 19.1 | 20.0 | 23.1 | 25.0 | 22.1 | 26.6 | 23.3 | 22.2 | 19.4 | 15.8 | 22.0 | 27.9 | 20.4 | 21.1 | 21.8 |
| ReCAP (ICML 2025) | ✗ | Acc. ↑ | 37.9 | 47.8 | 52.9 | 49.1 | 50.9 | 56.3 | 53.1 | 58.4 | 61.5 | 65.8 | 76.3 | 62.8 | 57.9 | 67.3 | 67.2 | 57.7 |
| | | ECE ↓ | 9.1 | 9.4 | 9.8 | 7.9 | 11.1 | 9.2 | 12.8 | 10.4 | 9.5 | 8.5 | 5.0 | 10.8 | 12.6 | 8.1 | 7.8 | 9.5 |
| REM (ICML 2025) | ✗ | Acc. ↑ | 56.5 | 61.9 | 60.8 | 46.8 | 51.0 | 56.5 | **57.2** | 62.5 | 64.8 | 64.6 | 76.8 | 53.2 | 58.4 | 71.1 | 69.8 | 60.8 |
| | | ECE ↓ | 5.5 | 6.0 | 7.2 | 10.4 | 13.1 | 10.8 | 11.9 | 8.6 | **7.0** | 8.6 | 5.2 | 11.0 | 11.5 | 6.3 | **4.9** | 8.5 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 57.8 | 61.3 | 60.7 | 52.8 | 48.6 | 52.3 | 53.1 | 60.7 | 63.1 | 62.6 | 78.0 | 55.6 | 54.9 | 69.1 | 70.4 | 60.1 |
| | | ECE ↓ | 7.4 | 8.1 | 8.0 | 6.4 | **4.7** | 7.2 | 6.6 | 10.9 | 9.6 | 8.3 | 9.7 | 7.7 | 8.4 | 10.8 | 10.0 | 8.2 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 51.2 | 56.3 | 55.6 | 50.6 | 50.4 | 52.7 | 55.8 | 62.5 | 60.6 | 57.9 | 74.8 | 50.0 | **60.7** | 64.5 | 63.5 | 57.8 |
| | | ECE ↓ | 8.7 | 8.6 | 8.6 | 8.8 | 11.0 | 10.6 | 8.2 | 7.2 | 7.9 | 9.5 | **4.0** | 13.1 | 8.4 | 7.3 | 6.6 | 8.6 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 57.2 | 57.6 | 59.1 | 48.9 | 38.0 | 55.2 | 47.4 | 63.6 | 65.1 | 68.5 | 78.5 | 33.1 | 47.5 | 68.5 | 69.9 | 57.2 |
| | | ECE ↓ | 35.5 | 36.8 | 36.3 | 44.1 | 55.5 | 40.5 | 47.9 | 33.7 | 31.7 | 27.6 | 20.2 | 57.8 | 48.5 | 29.3 | 28.0 | 38.2 |
| FreqCTTA (AAAI 2026) | ✗ | Acc. ↑ | 52.3 | 54.9 | 57.8 | 53.4 | 50.3 | 57.2 | 53.5 | 65.0 | 62.0 | 64.8 | 78.4 | 48.3 | 56.5 | **73.1** | 69.0 | 59.8 |
| | | ECE ↓ | 4.6 | 7.0 | 7.8 | 7.5 | 9.7 | 9.0 | 13.1 | 8.9 | 9.5 | 11.0 | 5.5 | 14.7 | 14.2 | 7.8 | 8.1 | 9.2 |
| NEO (ICLR 2026) | ✓ | Acc. ↑ | 56.7 | 57.1 | 57.4 | 46.7 | 35.6 | 52.8 | 45.5 | 62.7 | 63.7 | 68.6 | 78.0 | 36.4 | 45.5 | 66.9 | 67.1 | 56.0 |
| | | ECE ↓ | 10.6 | 7.1 | 9.7 | 6.4 | 5.7 | **3.6** | **4.9** | 5.0 | 20.7 | 51.0 | 8.9 | 23.9 | **5.9** | **5.8** | 6.8 | 11.7 |
| GOLD (CVPR 2026) | ✗ | Acc. ↑ | **59.9** | **64.4** | **64.3** | 42.7 | 44.6 | 56.4 | 48.2 | 64.7 | 64.1 | 63.8 | 77.9 | 30.6 | 55.2 | 68.8 | 70.2 | 58.4 |
| | | ECE ↓ | 27.9 | 25.6 | 26.1 | 40.3 | 42.3 | 33.1 | 40.8 | 27.5 | 27.5 | 25.7 | 17.1 | 48.7 | 35.4 | 24.3 | 23.6 | 31.1 |
| GAIN (Ours) | ✓ | Acc. ↑ | 57.7 | 59.3 | 60.0 | 53.2 | 44.1 | 58.3 | 53.2 | **65.6** | **67.0** | **72.6** | **78.7** | 61.9 | 55.6 | 69.7 | **71.3** | **61.9** |
| | | ECE ↓ | 4.6 | **4.8** | **5.3** | 5.6 | 5.1 | 5.8 | 5.5 | **4.9** | 9.2 | **5.4** | 5.3 | 9.0 | 7.0 | 6.6 | 5.9 | **6.0** |

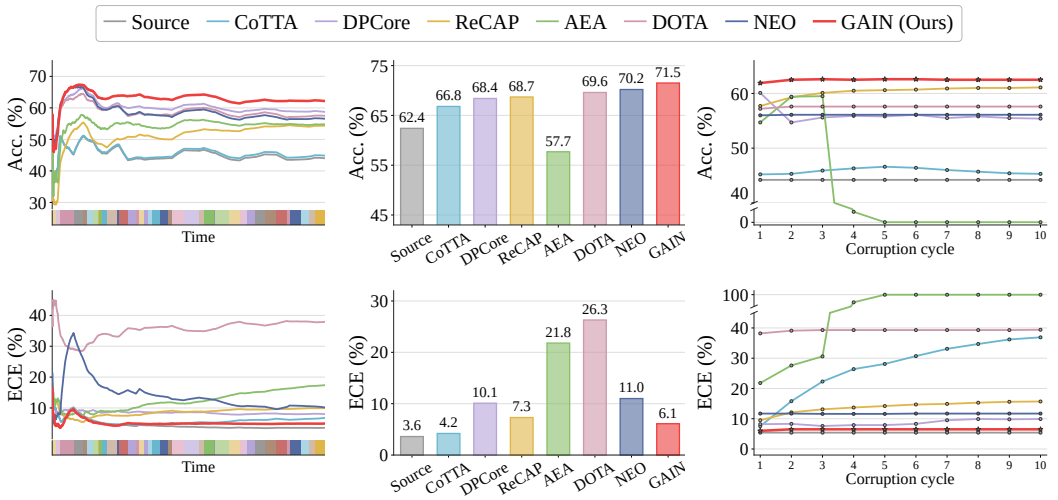
**Implementation Details.** All experiments use an ImageNet-pretrained ViT-B/16 with a batch size of 64 on a single NVIDIA RTX A6000 GPU. The source model remains frozen, while **GAIN** updates only compact target statistics without backpropagation or parameter updates. We set $\kappa_0 = 3$ and use the same hyperparameters across CSC, CDC, MDS, and long-horizon evaluation. Unless otherwise specified, main-paper experiments are conducted on ImageNet-C. ECE is computed from the final outputs of each method's official implementation. Further details are provided in Appendix F.

### 5.2 Main Results on ImageNet-C

**CSC Scenario.** Continual structured change (CSC) introduces abrupt domain transitions while carrying historical information across corruptions, making error accumulation a key challenge. As shown in Table 1, **GAIN** achieves 61.9% accuracy with 6.0% ECE, improving accuracy over the frozen source by 17.7 points with only a 0.6-point increase in ECE. In contrast, ROID reaches 60.8% accuracy with 60.7% ECE, while the BP-free DOTA obtains 57.2% accuracy with 38.2% ECE. These results show that **GAIN** enables accurate and well-calibrated adaptation without backpropagation, supporting gain-guided source-relative intervention.

**CDC Scenario.** Continual dynamic change (CDC) further challenges adaptation through recurring corruptions with irregular durations and frequencies, making historical target statistics less consistently aligned with the current distribution. As shown in Fig. 3a, **GAIN** maintains high accuracy and low calibration error throughout the dynamic stream, achieving 61.8% mean accuracy and 6.2% ECE. In contrast, several baselines exhibit either accuracy degradation or substantial miscalibration. These results show that gain-guided intervention remains reliable under irregular distribution changes by evaluating the source-relative utility of history-induced corrections.

**MDS Scenario.** Mixed-domain shift interleaves samples from heterogeneous corruption domains, making historical target statistics less specific to the current sample. As shown in Fig. 3b, **GAIN**




(a) CDC Scenario (b) MDS Scenario (c) LHA Scenario

Figure 3: **Accuracy and calibration under three challenging CTTA scenarios on ImageNet-C.** **(a)** CDC evaluates recurring corruptions with irregular durations. **(b)** MDS interleaves samples from multiple corruption domains, with results averaged over severity levels 1–5. **(c)** LHA evaluates error accumulation and long-term stability over 10 repeated corruption cycles.

achieves the highest accuracy averaged across severity levels of 71.5% with a low ECE of 6.1%, outperforming the BP-free NEO and DOTA in the accuracy–calibration trade-off. These results show that gain-guided intervention remains effective under heterogeneous target shifts.

**LHA Scenario.** The long-horizon setting evaluates adaptation stability under repeated exposure, where small errors may accumulate over time. As shown in Fig. 3c, **GAIN** remains stable across all 10 rounds: accuracy increases from 61.9% in R1 to 62.5–62.6% thereafter, while ECE stays around 6.5%. In contrast, CoTTA suffers severe calibration drift, DPCore exhibits noticeable accuracy degradation, ReCAP becomes increasingly miscalibrated, and AEA eventually collapses. DOTA remains persistently miscalibrated, while NEO is stable but substantially less accurate. These results demonstrate that **GAIN** maintains stable accuracy and calibration over long horizons by limiting the propagation of unreliable history-induced corrections.

### 5.3 Experiments on ImageNet-3DCC

To evaluate robustness under more realistic shifts, we further consider ImageNet-3DCC (Kar et al., 2022), which covers diverse geometry- and imaging-related corruptions and provides a complementary test beyond conventional 2D corruptions. As shown in Fig. 4, **GAIN** achieves the highest classification accuracy while maintaining near-source inference speed, reaching 18.5× the inference speed of CoTTA, with low calibration error under CSC. In contrast, REM and DPCore achieve competitive accuracy at substantially lower inference speeds, while DOTA remains less accurate and more poorly calibrated. These results show that gain-guided intervention preserves a strong accuracy–calibration–efficiency trade-off beyond ImageNet-C under more diverse corruption shifts. Additional experiments and analyses are provided in Appendix G.

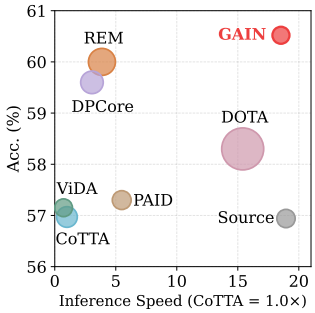



Figure 4: **Efficiency–accuracy trade-off on ImageNet-3DCC.** Inference speed is normalized to CoTTA; bubble size indicates ECE.

### 5.4 Ablation Studies and Further Analysis

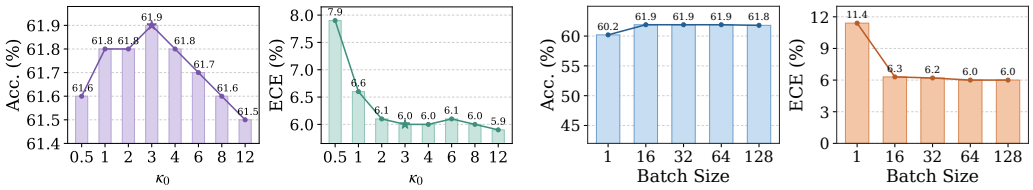
**Ablation Studies.** Table 2 compares different intervention rules under the same target-side estimation framework. Source retention does not exploit target evidence, whereas full correction substantially improves accuracy but leads to poor calibration. Fixed, entropy-based, and disagreement-based interventions partially alleviate this trade-off, but remain inferior to **GAIN**. In contrast, **GAIN** determines the intervention strength from the source-relative posterior-predictive gain, achieving the best overall performance with 61.9% accuracy, 6.0% ECE, and 1.9 NLL. These results reinforce our principle: history proposes, while gain decides whether and how strongly to intervene.

Table 2: **Ablation of intervention rules under CSC.** Signal denotes the criterion for setting the intervention strength $\lambda_t$.

| Intervention Rule | Signal | $\lambda_t$ | Acc. ↑ | ECE ↓ | NLL ↓ |
|---|---|---|---|---|---|
| Source Retention | – | 0 | 44.2 | 5.4 | 3.0 |
| Full Correction | – | 1 | 58.6 | 11.2 | 2.5 |
| Fixed Intervention | – | 0.5 | 59.7 | 8.9 | 2.1 |
| Entropy-based | $H(\mathbf{s}_t)$ | $\lambda_t^{\text{ent}}$ | 58.9 | 8.4 | 2.1 |
| Disagreement-based | $\text{JS}(\mathbf{s}_t, \hat{\mathbf{q}}_t^H)$ | $\lambda_t^{\text{dis}}$ | 58.9 | 9.2 | 2.1 |
| **GAIN** (Ours) | $G_t^{\text{pp}}$ | $\lambda_t^\star$ | **61.9** | 6.0 | **1.9** |

Table 3: **Efficiency analysis on ImageNet-C.** BP/FP: backward/forward propagation counts. Speed is normalized to CoTTA (1.0×); higher is faster.

| Method | #BP | #FP | Param.(M) ↓ | Mem.(GB) ↓ | Speed ↑ | Acc. ↑ | ECE ↓ |
|---|---|---|---|---|---|---|---|
| CoTTA | 1 | 5.1 | 86.42 | 23.01 | 1.0× | 45.2 | 7.5 |
| AEA | 1 | 1 | 0.04 | 6.05 | 6.3× | 54.7 | 21.8 |
| REM | 1 | 3 | 0.03 | 26.77 | 3.3× | 60.8 | 8.5 |
| DPCore | 7.9 | 9.9 | 1.03 | 8.59 | 0.6× | 60.1 | 8.2 |
| PAID | 1 | 1 | 0.81 | 11.63 | 4.6× | 57.8 | 8.6 |
| DOTA | 0 | 1 | 0 | 9.54 | 12.7× | 57.2 | 38.2 |
| **GAIN** (Ours) | **0** | **1** | **0** | **0.81** | **15.9**× | **61.9** | **6.0** |




(a) Source-centered Prior Strength $\kappa_0$. (b) Test-Time Batch Size

Figure 5: **Sensitivity analysis.** We study the sensitivity of **GAIN** to (a) the source-centered prior strength $\kappa_0$ and (b) the test-time batch size. For each setting, we report both accuracy and ECE.

**Sensitivity to $\kappa_0$.** We study the sensitivity of **GAIN** to the prior strength $\kappa_0$, which controls the influence of the source prior on target-side estimation. As shown in Fig. 5a, increasing $\kappa_0$ from 0.5 to 2 improves both accuracy and calibration, with accuracy rising from 61.6% to 61.8% and ECE decreasing from 7.9% to 6.1%. Performance remains stable for $\kappa_0 \in [2, 4]$, with the highest accuracy of 61.9% achieved at $\kappa_0 = 3$. Larger values slightly improve ECE but gradually reduce accuracy as the source prior becomes more dominant. We therefore set $\kappa_0 = 3$ by default, which offers a favorable accuracy–calibration trade-off without careful tuning.

**Effect of Test-Time Batch Size.** We evaluate the effect of test-time batch size on adaptation performance. As shown in Fig. 5b, single-sample updates yield less reliable target statistics, with 60.2% accuracy and 11.4% ECE. Increasing the batch size to 16 improves accuracy to 61.9% and sharply reduces ECE to 6.3%. Beyond 16 samples, performance largely saturates: accuracy remains within 61.8–61.9%, while ECE only gradually decreases to 6.0% at a batch size of 128. This indicates that **GAIN** does not require large test-time batches. For a fair comparison, we use a default test-time batch size of 64 in all main experiments.

**Computational Efficiency.** Table 3 compares the computational efficiency of different CTTA methods on ImageNet-C. **GAIN** requires only a single forward pass, without backpropagation or trainable parameter updates, and uses only 0.81 GB of memory. With computational speed normalized to CoTTA (1.0×), **GAIN** achieves the highest relative speed of 15.9×, while also attaining the best accuracy of 61.9% with only 6.0% ECE. Compared with the BP-free DOTA, **GAIN** is also faster while improving accuracy by 4.7 points and reducing ECE from 38.2% to 6.0%. These results demonstrate that **GAIN** achieves a favorable accuracy–calibration–efficiency trade-off with a lightweight, forward-only adaptation pipeline.

## 6 Conclusion

We presented **GAIN**, a gain-guided framework for continual test-time adaptation following the principle that *history proposes, gain decides*. Rather than directly trusting history-derived corrections, **GAIN** evaluates their source-relative utility with a posterior-predictive evaluator and adaptively controls the intervention strength along a continuous evidence path. Combined with causal target-statistic updates, **GAIN** limits the propagation of unreliable corrections while enabling efficient adaptation without backpropagation or replay. Across structured, dynamic, mixed-domain, and long-horizon shifts, **GAIN** achieves strong accuracy, calibration, and stability, highlighting the effectiveness of gain-guided intervention for reliable continual adaptation.

**Limitations and future work.** Our study follows the standard closed-set CTTA setting, where the source and target domains share the same label space. Extending gain-guided intervention to open-set adaptation and broader prediction tasks presents a promising direction for future work.

## REPRODUCIBILITY STATEMENT

All experiments are implemented in PyTorch 2.7.1 and conducted on a single NVIDIA RTX A6000 GPU. Detailed experimental settings and hyperparameter configurations are provided in the Appendix. The source code will be released publicly upon publication.

## AI USE STATEMENT

During the preparation of this manuscript, large language models (LLMs) were used as assistive tools for language editing, improving clarity, and manuscript revision. We also used feedback from Google's Paper Assistant Tool (PAT), provided through ICLR, to support manuscript revision. All suggestions were critically reviewed by the authors, and all methodological decisions, experiments, analyses, and conclusions were independently verified by the authors. The authors take full responsibility for the content of this paper.

## REFERENCES

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):1103–1130, 2016. 15

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Wonjeong Choi, Do-Yeon Kim, Jungwuk Park, Jungmoon Lee, Younghyun Park, Dong-Jun Han, and Jaekyun Moon. Adaptive energy alignment for accelerating test-time adaptation. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2025. 2

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991. 18

Shuang Cui, Jinglin Xu, Yi Li, Xiongxin Tang, Jiangmeng Li, Jiahuan Zhou, Fanjiang Xu, Fuchun Sun, and Hui Xiong. Bayestta: Continual-temporal test-time adaptation for vision-language models via gaussian discriminant analysis. *arXiv preprint arXiv:2507.08607*, 2025. 3

Dexin Duan, Rui Xu, Peilin Liu, and Fei Wen. Lifelong test-time adaptation via online learning in tracked low-dimensional subspace. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 2

Robert Mario Fano. *Transmission of Information: A statistical theory of communications*. MIT Press, 1966. 3, 14

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2023. 1

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2025a. 1, 2, 3, 23

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025b. 1, 3

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2019. 6, 23

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 24

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2025. 1, 6, 25

Hyun-Kurl Jang, Jihun Kim, Hyeokjun Kweon, and Kuk-Jin Yoon. Distill once, adapt life-long: Exploring dataset distillation for continual test-time adaptation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2026. 3

Siru Jiang, Yuwei Liang, Jian Liang, Ran He, and Tieniu Tan. To adapt or not to adapt? selective adaptation for vision-language models. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2026. 2

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 8, 24

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. Becotta: Input-dependent online blending of experts for continual test-time adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2024a. 3

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. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024b. 1, 2

Chang Liu, Ruotong Zhao, Li Gao, and Yupei Zhang. Sateen: Learning structural alignment for continual test-time adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2026. 2

Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024a. 2

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 *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024b. 3, 6, 23, 25

Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, 2024. 3

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo—no-optimization test-time adaptation through latent re-centering. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2026. 2

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, 2015. 6, 24

Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2022. 1, 2, 3

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 *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2023. 2, 6, 25

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2024. 3

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2019. 6, 24

Liliang Ren, Mankeerat Sidhu, Qi Zeng, Revanth Gangi Reddy, Heng Ji, and ChengXiang Zhai. C-pmi: Conditional pointwise mutual information for turn-level dialogue evaluation. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp. 80–85, 2023. 3, 14

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2021. 1, 2, 3

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2019. 6, 24

Kunyu Wang, Xueyang Fu, Yuanfei Bao, Chengjie Ge, Chengzhi Cao, Wei Zhai, and Zheng-Jun Zha. Paid: Pairwise angular-invariant decomposition for continual test-time adaptation. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 2

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 6, 23, 25

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 *WACV*, 2024. 2

Xu Yang, Xuan Chen, Moqi Li, Kun Wei, and Cheng Deng. A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Maxime Zanella, Clément Fuchs, Christophe De Vleeschouwer, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025a. 1, 3

Yufei Zhang, Yicheng Xu, Hongxin Wei, Zhiping Lin, Xiaofeng Zou, Cen Chen, and Huiping Zhuang. Analytic continual test-time adaptation for multi-modality corruption. In *Proc. of ACM International Conference on Multimedia (MM)*, 2025b. 3

Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. Dpcore: Dynamic prompt coreset for continual test-time adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2025c. 1, 3, 6, 23, 25

Jianchao Zhao, Chenhao Ding, SongLin Dong, Jiangyang Li, Qiang Wang, Yuhang He, and Yihong Gong. Shared & domain self-adaptive experts with frequency-aware discrimination for continual test-time adaptation. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, 2026. 3

Jiahuan Zhou, Chao Zhu, Zhenyu Cui, Zichen Liu, Xu Zou, and Gang Hua. Class-aware domain knowledge fusion and fission for continual test-time adaptation. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025. 3

# Appendix

This Appendix provides additional theoretical derivations, implementation details, and experimental results supporting our method. The contents are organized as follows:

- Appendix A: Why source-only reliability is insufficient under continual target shift;
- Appendix B: Accumulated target context as additional predictive evidence;
- Appendix C: Posterior-predictive target-side gain estimation;
- Appendix D: Posterior-predictive evidence intervention;
- Appendix E: Causal continual target-statistic updates and algorithmic implementation;
- Appendix F: Experimental setup and implementation details;
- Appendix G: Additional experimental results and analyses.

## A Source-Only Reliability Is Insufficient

The main paper argues that source confidence alone is generally insufficient for deciding whether the current source prediction should be modified. We formalize this observation below.

Let

$$\hat{y}_t^s = \arg\max_k s_{t,k}, \qquad C_t^s = \mathbb{I}[Y_t = \hat{y}_t^s] \tag{17}$$

denote the source prediction and its correctness indicator.

**Proposition A.1** (Insufficiency of source-only reliability)**.** *Suppose there exist a source predictive distribution* $\mathbf{s}$ *and two target histories* $h$ *and* $h'$ *with positive probability such that*

$$P_T(C_t^s = 1 \mid \mathbf{s}_t = \mathbf{s}, \mathcal{H}_t = h) \neq P_T(C_t^s = 1 \mid \mathbf{s}_t = \mathbf{s}, \mathcal{H}_t = h'). \tag{18}$$

*Then no function depending only on* $\mathbf{s}_t$ *can recover the conditional correctness probability under both histories.*

**Proof.** Assume that there exists a function $g$ such that $g(\mathbf{s}_t) = P_T(C_t^s = 1 \mid \mathbf{s}_t, \mathcal{H}_t)$ for every admissible history. For the same source distribution $\mathbf{s}$ in Eq. 18, this would require simultaneously

$$g(\mathbf{s}) = P_T(C_t^s = 1 \mid \mathbf{s}_t = \mathbf{s}, \mathcal{H}_t = h) = a, \tag{19}$$

$$g(\mathbf{s}) = P_T(C_t^s = 1 \mid \mathbf{s}_t = \mathbf{s}, \mathcal{H}_t = h') = b \tag{20}$$

with $a \neq b$. This is a contradiction. □

This result does not imply that confidence is uninformative. Rather, it shows that confidence is not a sufficient statistic for adaptation reliability when the correctness of an identical source prediction depends on the evolving target context.

## B Accumulated Target Context as Predictive Evidence

This section provides the derivations underlying the C-PMI interpretation in Eq. 3, the non-negative oracle predictive value in Eq. 26, and the practical correction gain in Eq. 5.

### B.1 Conditional Pointwise Mutual Information Representation

For notational clarity, consider a fixed realization $X_t = \mathbf{x}_t$ and $\mathcal{H}_t = h$, and define

$$q_k^0 \triangleq P_T(Y_t = k \mid X_t = \mathbf{x}_t), \qquad q_k^H \triangleq P_T(Y_t = k \mid X_t = \mathbf{x}_t, \mathcal{H}_t = h). \tag{21}$$

Conditioning additionally on the maintained target context and applying Bayes' rule gives

$$\begin{aligned} q_k^H &= \frac{P_T(\mathcal{H}_t = h \mid Y_t = k, X_t = \mathbf{x}_t) P_T(Y_t = k \mid X_t = \mathbf{x}_t)}{P_T(\mathcal{H}_t = h \mid X_t = \mathbf{x}_t)} \\ &= q_k^0 \frac{P_T(\mathcal{H}_t = h \mid Y_t = k, X_t = \mathbf{x}_t)}{P_T(\mathcal{H}_t = h \mid X_t = \mathbf{x}_t)}. \end{aligned} \tag{22}$$

Thus, historical conditioning reweights the current-sample posterior by a class-dependent evidence term. The conditional pointwise mutual information (C-PMI) (Fano, 1966; Ren et al., 2023) associated with the realization is

$$\begin{aligned} i(Y_t = k; \mathcal{H}_t = h \mid X_t = \mathbf{x}_t) &\triangleq \log \frac{P_T(Y_t = k \mid \mathcal{H}_t = h, X_t = \mathbf{x}_t)}{P_T(Y_t = k \mid X_t = \mathbf{x}_t)} \\ &= \log \frac{P_T(Y_t = k, \mathcal{H}_t = h \mid X_t = \mathbf{x}_t)}{P_T(Y_t = k \mid X_t = \mathbf{x}_t) P_T(\mathcal{H}_t = h \mid X_t = \mathbf{x}_t)} \\ &= \log \frac{P_T(\mathcal{H}_t = h \mid Y_t = k, X_t = \mathbf{x}_t)}{P_T(\mathcal{H}_t = h \mid X_t = \mathbf{x}_t)} \\ &= \log \frac{q_k^H}{q_k^0}. \end{aligned} \tag{23}$$

We therefore obtain

$$\xi_{t,k} = \log \frac{q_{t,k}^H}{q_{t,k}^0} = i(Y_t = k; \mathcal{H}_t \mid \mathbf{x}_t). \tag{24}$$

Equivalently,

$$\log q_{t,k}^H = \log q_{t,k}^0 + \xi_{t,k}. \tag{25}$$

Hence, the class-wise log-posterior correction induced by accumulated target context is exactly a conditional pointwise mutual-information quantity, rather than an ad hoc calibration score.

### B.2 Non-Negative Oracle Predictive Value

**Proposition 3.1** [Non-Negative Predictive Value of Target History] Conditioning on the accumulated target context $\mathcal{H}_t$ yields non-negative oracle predictive value under logarithmic loss:

$$\mathcal{R}_t(\mathbf{q}_t^0) - \mathcal{R}_t(\mathbf{q}_t^H) = D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \mathbf{q}_t^0\big) = \mathbb{E}_{Y_t \sim \mathbf{q}_t^H}\big[\xi_{t,Y_t}\big] \geq 0. \tag{26}$$

The inequality is strict whenever $\mathbf{q}_t^H \neq \mathbf{q}_t^0$.

**Proof of Proposition 3.1** For any predictive distribution $\mathbf{r} \in \Delta^{K-1}$, define the conditional logarithmic risk under the oracle history-conditioned target posterior as

$$\mathcal{R}_t(\mathbf{r}) \triangleq \mathbb{E}_{Y_t \sim \mathbf{q}_t^H}\left[-\log r_{Y_t}\right] = -\sum_{k=1}^{K} q_{t,k}^H \log r_k. \tag{27}$$

The predictive value of conditioning on the maintained target context is therefore

$$\begin{aligned} \mathcal{R}_t(\mathbf{q}_t^0) - \mathcal{R}_t(\mathbf{q}_t^H) &= -\sum_{k=1}^{K} q_{t,k}^H \log q_{t,k}^0 + \sum_{k=1}^{K} q_{t,k}^H \log q_{t,k}^H \\ &= \sum_{k=1}^{K} q_{t,k}^H \log \frac{q_{t,k}^H}{q_{t,k}^0} \\ &= D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \mathbf{q}_t^0\big) \\ &= \mathbb{E}_{Y_t \sim \mathbf{q}_t^H}\left[\xi_{t,Y_t}\right] \geq 0, \end{aligned} \tag{28}$$

where the last equality follows from the C-PMI $\xi_{t,k} = \log(q_{t,k}^H / q_{t,k}^0)$, and the inequality follows from the non-negativity of KL divergence. Equality holds if and only if $\mathbf{q}_t^H = \mathbf{q}_t^0$. □

This result characterizes the value of conditioning on history under the oracle target posterior; it does not guarantee that an estimate constructed from accumulated unlabeled statistics improves upon the source prediction. The latter depends on the estimation gap, as shown in the practical correction-gain decomposition below.

### B.3 Practical Correction Gain

The oracle result above assumes access to the true history-conditioned posterior $\mathbf{q}_t^H$. In practice, CTTA can only infer an estimate $\hat{\mathbf{q}}_t^H$ from finite, unlabeled target observations accumulated so far. Consider fully replacing the source prediction $\mathbf{s}_t$ with $\hat{\mathbf{q}}_t^H$. The resulting reduction in conditional logarithmic risk is

$$\begin{aligned}\Delta_t^{\text{full}} &\triangleq \mathcal{R}_t(\mathbf{s}_t) - \mathcal{R}_t(\hat{\mathbf{q}}_t^H) = \sum_{k=1}^{K} q_{t,k}^H \log \frac{\hat{q}_{t,k}^H}{s_{t,k}} \\ &= \sum_{k=1}^{K} q_{t,k}^H \left( \log \frac{q_{t,k}^H}{s_{t,k}} - \log \frac{q_{t,k}^H}{\hat{q}_{t,k}^H} \right) \\ &= \underbrace{D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \mathbf{s}_t\big)}_{\text{correction potential}} - \underbrace{D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \hat{\mathbf{q}}_t^H\big)}_{\text{estimation gap}}. \end{aligned} \tag{29}$$

Unlike the non-negative oracle value, the practical correction gain $\Delta_t^{\text{full}}$ is not guaranteed to be positive. The first term measures how far the source prediction lies from the oracle history-conditioned posterior, whereas the second measures the remaining gap between its finite-sample estimate and the oracle. Consequently, full correction is beneficial if and only if

$$D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \hat{\mathbf{q}}_t^H\big) < D_{\mathrm{KL}}\big(\mathbf{q}_t^H \| \mathbf{s}_t\big). \tag{30}$$

This limitation is particularly relevant to CTTA, where continuously evolving target distributions can render accumulated statistics noisy, biased, or stale. Consequently, methods that directly reuse historical target information for calibration may propagate estimation errors and lead to unstable or degraded adaptation.

## C Reliable Target-Side Gain Estimation

This section provides the derivations underlying the target-side estimator introduced in Sec. 4.1. We first derive the source-anchored working posterior over the unknown target class centers, and then marginalize the remaining class-center uncertainty to construct a posterior-predictive evaluator of the proposed correction.

### C.1 Source-Anchored Probabilistic Target Estimation

Before processing the current sample at time $t$, let $\mathcal{Z}_t^H = \{\mathbf{z}_i^H\}_{i=1}^{N_t^H}$ denote the historical target representations from $\mathcal{H}_t$, where $\mathbf{z}_i^H = \phi_\theta(\mathbf{x}_i^H) \in \mathbb{R}^D$. Throughout the following derivation, we condition on the soft responsibilities induced by preceding target predictions and the pre-t covariance estimate $\Sigma_{t-1}$. Thus, the uncertainty derived below characterizes uncertainty in the target class center conditional on the current historical statistical state.

For each class $k$, we assume the class-conditional Gaussian model $p_T\left(\mathbf{z}_i^H \,|\, Y_i = k, \boldsymbol{\mu}_k\right) = \mathcal{N}\left(\mathbf{z}_i^H; \boldsymbol{\mu}_k, \Sigma_{t-1}\right)$, and place a source-centered prior on the unknown target class center

$$p(\boldsymbol{\mu}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{c}_k, \frac{\Sigma_{t-1}}{\kappa_0}\right), \tag{31}$$

where $\mathbf{c}_k$ denotes the source-derived class prototype and $\kappa_0$ controls the strength of the source-centered prior. At a fixed time $t$, $\Sigma_{t-1}$ is shared across classes, while its estimate is updated causally as new target observations become available.

Since historical target labels are unavailable, each historical representation $\mathbf{z}_i^H$ contributes to class $k$ through a soft responsibility $\omega_{i,k} \in [0,1]$ in Eq. 75, determined by the reliability-weighted prediction when the observation is incorporated into the target state. These responsibilities are used only to derive the corresponding sufficient statistics and need not be stored explicitly. Conditional on the soft responsibilities, we define the class-$k$ fractional likelihood (Bissiri et al., 2016):

$$\widetilde{\mathcal{L}}_{t,k}(\boldsymbol{\mu}_k) \triangleq \prod_{i=1}^{N_t^H} \mathcal{N}\left(\mathbf{z}_i^H; \boldsymbol{\mu}_k, \Sigma_{t-1}\right)^{\omega_{i,k}}. \tag{32}$$

Combining the fractional likelihood with the source-centered prior gives

$$\widetilde{p}\left(\boldsymbol{\mu}_k \mid \mathcal{Z}_t^H, \{\omega_{i,k}\}_{i=1}^{N_t^H}, \Sigma_{t-1}\right) \propto p(\boldsymbol{\mu}_k)\widetilde{\mathcal{L}}_{t,k}(\boldsymbol{\mu}_k). \tag{33}$$

For brevity, we suppress the fixed conditioning on the soft responsibilities and $\Sigma_{t-1}$ below. Taking the negative logarithm of Eq. 33 and multiplying by 2, we obtain the following expression, where $\doteq$ denotes equality up to additive terms independent of $\boldsymbol{\mu}_k$:

$$\begin{aligned}
-2\log\widetilde{p}\left(\boldsymbol{\mu}_k \mid \mathcal{Z}_t^H\right) &\doteq -2\log p(\boldsymbol{\mu}_k) - 2\log\widetilde{\mathcal{L}}_{t,k}(\boldsymbol{\mu}_k) \\
&\doteq \kappa_0(\boldsymbol{\mu}_k - \mathbf{c}_k)^\top \Sigma_{t-1}^{-1}(\boldsymbol{\mu}_k - \mathbf{c}_k) + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}(\mathbf{z}_i^H - \boldsymbol{\mu}_k)^\top \Sigma_{t-1}^{-1}(\mathbf{z}_i^H - \boldsymbol{\mu}_k) \\
&= \kappa_0\left[\boldsymbol{\mu}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k - 2\mathbf{c}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k + \mathbf{c}_k^\top \Sigma_{t-1}^{-1}\mathbf{c}_k\right] \\
&\quad + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}\left[(\mathbf{z}_i^H)^\top \Sigma_{t-1}^{-1}\mathbf{z}_i^H - 2(\mathbf{z}_i^H)^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k\right] \\
&= \left(\kappa_0 + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}\right)\boldsymbol{\mu}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k - 2\left(\kappa_0\mathbf{c}_k + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}\mathbf{z}_i^H\right)^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k \\
&\quad + \underbrace{\kappa_0\mathbf{c}_k^\top \Sigma_{t-1}^{-1}\mathbf{c}_k + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}(\mathbf{z}_i^H)^\top \Sigma_{t-1}^{-1}\mathbf{z}_i^H}_{\text{independent of } \boldsymbol{\mu}_k} \\
&\doteq \left(\kappa_0 + \sum_{i=1}^{N_t^H} \omega_{i,k}\right)\boldsymbol{\mu}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k - 2\left(\kappa_0\mathbf{c}_k + \sum_{i=1}^{N_t^H} \omega_{i,k}\mathbf{z}_i^H\right)^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k.
\end{aligned} \tag{34}$$

Define the posterior precision parameter and the corresponding source-anchored posterior-mean center as

$$\kappa_{t-1,k} = \kappa_0 + \sum\nolimits_{i=1}^{N_t^H} \omega_{i,k}, \qquad \mathbf{m}_{t-1,k} = \frac{\kappa_0\mathbf{c}_k + \sum_{i=1}^{N_t^H} \omega_{i,k}\mathbf{z}_i^H}{\kappa_{t-1,k}}. \tag{35}$$

Substituting Eq. 35 into Eq. 34 and completing the square gives

$$\begin{aligned}
-2\log\widetilde{p}\left(\boldsymbol{\mu}_k \mid \mathcal{Z}_t^H\right) &\doteq \kappa_{t-1,k}\boldsymbol{\mu}_k^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k - 2\kappa_{t-1,k}\mathbf{m}_{t-1,k}^\top \Sigma_{t-1}^{-1}\boldsymbol{\mu}_k \\
&\doteq \kappa_{t-1,k}\left(\boldsymbol{\mu}_k - \mathbf{m}_{t-1,k}\right)^\top \Sigma_{t-1}^{-1}\left(\boldsymbol{\mu}_k - \mathbf{m}_{t-1,k}\right).
\end{aligned} \tag{36}$$

Therefore, conditional on the effective responsibilities and the pre-$t$ covariance estimate, the induced fractional posterior is

$$\widetilde{p}\left(\boldsymbol{\mu}_k \mid \mathcal{Z}_t^H, \{\omega_{i,k}\}_{i=1}^{N_t^H}, \Sigma_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}_{t-1,k}, \frac{\Sigma_{t-1}}{\kappa_{t-1,k}}\right). \tag{37}$$

Hence, the posterior mean $\mathbf{m}_{t-1,k}$ provides the history-conditioned estimate of the target class center, while $\Sigma_{t-1}/\kappa_{t-1,k}$ characterizes the corresponding center-estimation uncertainty. Here, $\kappa_{t-1,k}$ measures the accumulated reliability-weighted support for class $k$, so stronger accumulated support yields a more concentrated estimate of its target center. The soft responsibilities are introduced only for derivation. In practice, their aggregate effect is maintained through recursive sufficient statistics in Appendix E.1, without storing or replaying historical representations.

**Posterior-Mean Target Proposal.** Given the class-center posterior in Eq. 37, we first construct a target-side proposal using its posterior mean $\mathbf{m}_{t-1,k}$. Specifically, plugging $\mathbf{m}_{t-1,k}$ into the class-conditional Gaussian model gives

$$\hat{p}_T\left(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t\right) \triangleq \mathcal{N}\left(\mathbf{z}_t; \mathbf{m}_{t-1,k}, \Sigma_{t-1}\right). \tag{38}$$

Applying Bayes' rule with the pre-$t$ class prior $\pi_{t-1,k}$ yields

$$\hat{q}_{t,k}^H = \frac{\pi_{t-1,k}\hat{p}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t)}{\sum_{j=1}^K \pi_{t-1,j}\hat{p}_T(\mathbf{z}_t \mid Y_t = j, \mathcal{H}_t)}. \tag{39}$$

Expanding the Gaussian density,

$$\hat{p}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_{t-1}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}_t - \mathbf{m}_{t-1,k})^\top \Sigma_{t-1}^{-1}(\mathbf{z}_t - \mathbf{m}_{t-1,k})\right]. \tag{40}$$

Define the corresponding squared Mahalanobis distance

$$d_{t,k} \triangleq (\mathbf{z}_t - \mathbf{m}_{t-1,k})^\top \Sigma_{t-1}^{-1} (\mathbf{z}_t - \mathbf{m}_{t-1,k}). \tag{41}$$

Since $\Sigma_{t-1}$ is shared across classes, the Gaussian normalization factor $(2\pi)^{-D/2}|\Sigma_{t-1}|^{-1/2}$ is independent of $k$ and therefore cancels in Eq. 39. Hence,

$$\hat{q}^H_{t,k} = \frac{\pi_{t-1,k} \exp(-d_{t,k}/2)}{\sum_{j=1}^K \pi_{t-1,j} \exp(-d_{t,j}/2)} \propto \pi_{t-1,k} \exp\left(-\frac{1}{2} d_{t,k}\right). \tag{42}$$

Equivalently, the class-wise discriminant score can be written as

$$\hat{\ell}^H_{t,k} = \log \pi_{t-1,k} - \frac{1}{2} d_{t,k}, \qquad \hat{\mathbf{q}}^H_t = \mathrm{softmax}(\hat{\boldsymbol{\ell}}^H_t). \tag{43}$$

Thus, $\hat{\mathbf{q}}^H_t$ is the target-side correction proposed by the posterior-mean target geometry. Importantly, this is a plug-in estimate: it uses the posterior mean $\mathbf{m}_{t-1,k}$ but does not yet account for the remaining class-center uncertainty $\Sigma_{t-1}/\kappa_{t-1,k}$. We marginalize this uncertainty next to construct the posterior-predictive evaluator.

### C.2 Posterior-Predictive Gain Evaluation

The posterior-mean proposal $\hat{\mathbf{q}}^H_t$ above is constructed by plugging the posterior mean $\mathbf{m}_{t-1,k}$ into the class-conditional model. We next marginalize the remaining uncertainty in the target class centers and use the resulting posterior-predictive distribution to evaluate the same proposed correction.

**Posterior-Predictive Target Distribution.** Recall from Eq. 37 that

$$\boldsymbol{\mu}_k \mid \mathcal{H}_t \sim \mathcal{N}\left(\mathbf{m}_{t-1,k}, \frac{\Sigma_{t-1}}{\kappa_{t-1,k}}\right), \tag{44}$$

while the class-conditional representation model is

$$\mathbf{z}_t \mid Y_t = k, \boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_{t-1}). \tag{45}$$

Marginalizing the latent class center gives

$$\begin{aligned} p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t) &= \int p_T(\mathbf{z}_t \mid Y_t = k, \boldsymbol{\mu}_k)\, \widetilde{p}(\boldsymbol{\mu}_k \mid \mathcal{H}_t)\, d\boldsymbol{\mu}_k \\ &= \mathcal{N}\left(\mathbf{z}_t; \mathbf{m}_{t-1,k}, \Sigma_{t-1} + \frac{\Sigma_{t-1}}{\kappa_{t-1,k}}\right) \\ &= \mathcal{N}(\mathbf{z}_t; \mathbf{m}_{t-1,k}, h_{t,k}\Sigma_{t-1}), \end{aligned} \tag{46}$$

where

$$h_{t,k} \triangleq 1 + \kappa^{-1}_{t-1,k}. \tag{47}$$

Hence, the posterior-predictive covariance accounts for both the within-class representation variability and the remaining uncertainty in the estimated class center.

Applying Bayes' rule with the pre-$t$ class prior $\pi_{t-1,k}$ gives

$$\bar{q}^H_{t,k} = \frac{\pi_{t-1,k} p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t)}{\sum_{j=1}^K \pi_{t-1,j} p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = j, \mathcal{H}_t)}. \tag{48}$$

Expanding Eq. 46,

$$p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t) = \frac{\exp\left[-\frac{1}{2}(\mathbf{z}_t - \mathbf{m}_{t-1,k})^\top (h_{t,k}\Sigma_{t-1})^{-1}(\mathbf{z}_t - \mathbf{m}_{t-1,k})\right]}{(2\pi)^{D/2}|h_{t,k}\Sigma_{t-1}|^{1/2}}. \tag{49}$$

Using

$$|h_{t,k}\Sigma_{t-1}| = h^D_{t,k}|\Sigma_{t-1}|, \qquad (h_{t,k}\Sigma_{t-1})^{-1} = h^{-1}_{t,k}\Sigma^{-1}_{t-1}, \tag{50}$$

together with the Mahalanobis distance $d_{t,k}$ in Eq. 41, we obtain

$$p^{\mathrm{pp}}_T(\mathbf{z}_t \mid Y_t = k, \mathcal{H}_t) = C_t\, h^{-D/2}_{t,k} \exp\left(-\frac{d_{t,k}}{2h_{t,k}}\right), \tag{51}$$

where $C_t = (2\pi)^{-D/2}|\Sigma_{t-1}|^{-1/2}$ is independent of $k$ because $\Sigma_{t-1}$ is shared across classes. The common factor therefore cancels during class normalization, yielding

$$\bar{q}^H_{t,k} = \frac{\pi_{t-1,k} h_{t,k}^{-D/2} \exp\left(-\frac{d_{t,k}}{2h_{t,k}}\right)}{\sum_{j=1}^{K} \pi_{t-1,j} h_{t,j}^{-D/2} \exp\left(-\frac{d_{t,j}}{2h_{t,j}}\right)}. \tag{52}$$

Equivalently, $\bar{\ell}^H_{t,k} = \log \pi_{t-1,k} - \frac{D}{2}\log h_{t,k} - \frac{d_{t,k}}{2h_{t,k}}$, and $\bar{\mathbf{q}}^H_t = \mathrm{softmax}(\bar{\boldsymbol{\ell}}^H_t)$.

Compared with the posterior-mean proposal, $h_{t,k}$ attenuates the distance penalty for uncertain class centers, while $(D/2)\log h_{t,k}$ accounts for the corresponding increase in predictive volume. The latter cannot, in general, be removed when $\kappa_{t-1,k}$ varies across classes.

**Posterior-Predictive Correction Gain.** Recall the source-relative correction evidence $\hat{\xi}^s_{t,k} \triangleq \log \frac{\hat{q}^H_{t,k}}{s_{t,k}}$. The oracle practical gain in Eq. 29 can be written as $\Delta^{\mathrm{full}}_t = (\mathbf{q}^H_t)^\top \hat{\boldsymbol{\xi}}^s_t$, where the unknown $\mathbf{q}^H_t$ evaluates the correction proposed by $\hat{\mathbf{q}}^H_t$. Since $\mathbf{q}^H_t$ is unavailable at test time, we instead evaluate the same correction under the posterior-predictive target distribution:

$$G^{\mathrm{pp}}_t \triangleq (\bar{\mathbf{q}}^H_t)^\top \hat{\boldsymbol{\xi}}^s_t = \sum_{k=1}^{K} \bar{q}^H_{t,k} \log \frac{\hat{q}^H_{t,k}}{s_{t,k}}. \tag{53}$$

Equivalently, defining the posterior-predictive logarithmic risk

$$\bar{\mathcal{R}}_t(\mathbf{p}) \triangleq -\sum_{k=1}^{K} \bar{q}^H_{t,k} \log p_k, \tag{54}$$

we have $G^{\mathrm{pp}}_t = \bar{\mathcal{R}}_t(\mathbf{s}_t) - \bar{\mathcal{R}}_t(\hat{\mathbf{q}}^H_t)$. Moreover,

$$\begin{aligned} G^{\mathrm{pp}}_t &= \sum_{k=1}^{K} \bar{q}^H_{t,k} \left[\log \frac{\bar{q}^H_{t,k}}{s_{t,k}} - \log \frac{\bar{q}^H_{t,k}}{\hat{q}^H_{t,k}}\right] \\ &= D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t \| \mathbf{s}_t\right) - D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t \| \hat{\mathbf{q}}^H_t\right). \end{aligned} \tag{55}$$

The first term measures the potential benefit of correcting the source prediction, whereas the second measures the mismatch between the proposed correction and its posterior-predictive evaluation. Hence, source–target disagreement alone does not imply a reliable correction; the correction must remain supported after accounting for uncertainty in the estimated target geometry.

**Relation to the Oracle Gain.** The posterior-predictive gain evaluates the same source-relative correction as the oracle practical gain, but replaces the unknown history-conditioned target posterior $\mathbf{q}^H_t$ with the posterior-predictive evaluator $\bar{\mathbf{q}}^H_t$. Their difference is therefore

$$\Delta^{\mathrm{full}}_t - G^{\mathrm{pp}}_t = \left(\mathbf{q}^H_t - \bar{\mathbf{q}}^H_t\right)^\top \hat{\boldsymbol{\xi}}^s_t. \tag{56}$$

Let $\mathrm{span}(\hat{\boldsymbol{\xi}}^s_t) \triangleq \max_k \hat{\xi}^s_{t,k} - \min_k \hat{\xi}^s_{t,k}$. Since both $\mathbf{q}^H_t$ and $\bar{\mathbf{q}}^H_t$ are probability distributions, their difference sums to zero. Hence,

$$\left|\Delta^{\mathrm{full}}_t - G^{\mathrm{pp}}_t\right| \le \mathrm{TV}\left(\mathbf{q}^H_t, \bar{\mathbf{q}}^H_t\right) \mathrm{span}(\hat{\boldsymbol{\xi}}^s_t), \tag{57}$$

where $\mathrm{TV}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1$. By Pinsker's inequality (Cover et al., 1991), this further implies

$$\left|\Delta^{\mathrm{full}}_t - G^{\mathrm{pp}}_t\right| \le \mathrm{span}(\hat{\boldsymbol{\xi}}^s_t) \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathbf{q}^H_t \| \bar{\mathbf{q}}^H_t\right)}. \tag{58}$$

Thus, the discrepancy between the posterior-predictive and oracle gains is controlled jointly by the accuracy of the posterior-predictive evaluator and the magnitude of the proposed correction. In particular, $G^{\mathrm{pp}}_t$ approaches the oracle practical gain whenever $\bar{\mathbf{q}}^H_t$ approaches $\mathbf{q}^H_t$, while aggressive source-relative corrections amplify errors in the target evaluator. For numerical stability, probabilities entering log-ratios are lower bounded by $\epsilon_p$ before renormalization.

# D Posterior-Predictive Evidence Intervention

This section provides the derivations underlying the posterior-predictive evidence intervention introduced in Sec. 4.2. We first derive the continuous source-to-target evidence path and its exact posterior-predictive risk reduction. We then establish the concavity of the resulting objective, characterize its boundary behavior, and derive an efficient global solution for the intervention strength.

## D.1 Continuous Evidence Path and Predictive Objective

**Continuous Evidence Path.** To control the extent of target-side intervention, we construct a continuous evidence path by scaling the source-relative evidence with $\lambda \in [0,1]$. Recall that $\hat{\xi}^s_{t,k} = \log(\hat{q}^H_{t,k}/s_{t,k})$. Scaling this evidence and normalizing across classes gives

$$p^{(\lambda)}_{t,k} = \frac{s_{t,k}\exp(\lambda\hat{\xi}^s_{t,k})}{Z_t(\lambda)} = \frac{s^{1-\lambda}_{t,k}(\hat{q}^H_{t,k})^\lambda}{\sum_{j=1}^K s^{1-\lambda}_{t,j}(\hat{q}^H_{t,j})^\lambda}, \tag{59}$$

where $Z_t(\lambda)=\sum_{j=1}^K s_{t,j}\exp(\lambda\hat{\xi}^s_{t,j})$ is the normalization factor. The two endpoints satisfy $\mathbf{p}^{(0)}_t=\mathbf{s}_t$ and $\mathbf{p}^{(1)}_t=\hat{\mathbf{q}}^H_t$. Hence, $\lambda$ continuously controls the amount of target-side evidence introduced into the source prediction.

**Posterior-Predictive Path-Wise Gain.** As established in Sec. 4.1, the posterior-predictive distribution $\bar{\mathbf{q}}^H_t$ evaluates the correction proposed by $\hat{\mathbf{q}}^H_t$ after accounting for uncertainty in the estimated target geometry. Define the corresponding conditional logarithmic risk as

$$\bar{\mathcal{R}}_t(\mathbf{p}) \triangleq -\sum_{k=1}^K \bar{q}^H_{t,k}\log p_k. \tag{60}$$

From Eq. 59, $\log\frac{p^{(\lambda)}_{t,k}}{s_{t,k}} = \lambda\hat{\xi}^s_{t,k} - \log Z_t(\lambda)$. Therefore, the posterior-predictive risk reduction relative to the source prediction is

$$\begin{aligned}\mathcal{J}_t(\lambda) &\triangleq \bar{\mathcal{R}}_t(\mathbf{s}_t) - \bar{\mathcal{R}}_t\big(\mathbf{p}^{(\lambda)}_t\big) = \sum_{k=1}^K \bar{q}^H_{t,k}\log\frac{p^{(\lambda)}_{t,k}}{s_{t,k}} \\ &= \lambda(\bar{\mathbf{q}}^H_t)^\top\hat{\boldsymbol{\xi}}^s_t - \log Z_t(\lambda) \\ &= \lambda G^{\mathrm{pp}}_t - \log Z_t(\lambda), \qquad \lambda\in[0,1].\end{aligned} \tag{61}$$

Here, $G^{\mathrm{pp}}_t$ is the posterior-predictive correction gain introduced in Sec. 4.1. Thus, Eq. 61 gives the exact reduction in posterior-predictive logarithmic risk along the evidence path. The same proposed correction $\hat{\mathbf{q}}^H_t$ determines the path, whereas $\bar{\mathbf{q}}^H_t$ determines how strongly that correction is supported under the uncertainty-aware target working model.

Equivalently, since $\bar{\mathcal{R}}_t(\mathbf{p}) = H(\bar{\mathbf{q}}^H_t) + D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t\|\mathbf{p}\right)$, maximizing $\mathcal{J}_t(\lambda)$ is equivalent to finding the point on the source–target evidence path that minimizes $D_{\mathrm{KL}}\left(\bar{\mathbf{q}}^H_t\|\mathbf{p}^{(\lambda)}_t\right)$. Hence, the intervention coefficient can also be interpreted as the forward-KL projection of the posterior-predictive evaluator onto the continuous source–target evidence path.

## D.2 Globally Optimal Posterior-Predictive Intervention

We now establish the global solution of Eq. 61 and prove Theorem 4.1.

**Concavity and Global Optimality.** Since $G^{\mathrm{pp}}_t$ is independent of $\lambda$, differentiating the log-normalizer gives

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log Z_t(\lambda) = \frac{\sum_{k=1}^K s_{t,k}\exp(\lambda\hat{\xi}^s_{t,k})\hat{\xi}^s_{t,k}}{Z_t(\lambda)} = \sum\nolimits_{k=1}^K p^{(\lambda)}_{t,k}\hat{\xi}^s_{t,k} = \mathbb{E}_{k\sim\mathbf{p}^{(\lambda)}_t}\left[\hat{\xi}^s_{t,k}\right]. \tag{62}$$

Therefore, $\mathcal{J}'_t(\lambda) = G^{\mathrm{pp}}_t - \mathbb{E}_{k\sim\mathbf{p}^{(\lambda)}_t}\left[\hat{\xi}^s_{t,k}\right]$. Differentiating once more yields

$$\begin{aligned}\mathcal{J}''_t(\lambda) &= -\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\log Z_t(\lambda) = -\left\{\mathbb{E}_{k\sim\mathbf{p}^{(\lambda)}_t}\left[(\hat{\xi}^s_{t,k})^2\right] - \left(\mathbb{E}_{k\sim\mathbf{p}^{(\lambda)}_t}\left[\hat{\xi}^s_{t,k}\right]\right)^2\right\} \\ &= -\operatorname{Var}_{k\sim\mathbf{p}^{(\lambda)}_t}\left[\hat{\xi}^s_{t,k}\right] \le 0.\end{aligned} \tag{63}$$

Hence, $\mathcal{J}_t(\lambda)$ is concave over $\lambda \in [0,1]$. If $\hat{\mathbf{q}}_t^H \neq \mathbf{s}_t$, the source-relative evidence is not constant across classes and the variance is strictly positive, yielding strict concavity. Therefore, the maximizer is unique except in the degenerate case $\hat{\mathbf{q}}_t^H = \mathbf{s}_t$.

**Boundary Behavior.** The derivatives at the two endpoints further characterize when intervention is suppressed or fully applied. At $\lambda = 0$, $\mathbf{p}_t^{(0)} = \mathbf{s}_t$, and therefore

$$\begin{aligned}\mathcal{J}_t'(0) &= G_t^{\mathrm{pp}} - \sum\nolimits_{k=1}^{K} s_{t,k} \log \frac{\hat{q}_{t,k}^H}{s_{t,k}} \\ &= G_t^{\mathrm{pp}} + D_{\mathrm{KL}}\left(\mathbf{s}_t \| \hat{\mathbf{q}}_t^H\right). \end{aligned} \tag{64}$$

Since $\mathcal{J}_t$ is concave, $\mathcal{J}_t'(0) \leq 0$ implies that the objective is non-increasing from the source endpoint, and hence

$$G_t^{\mathrm{pp}} \leq -D_{\mathrm{KL}}\left(\mathbf{s}_t \| \hat{\mathbf{q}}_t^H\right) \quad \Longrightarrow \quad \lambda_t^\star = 0. \tag{65}$$

Thus, the posterior-predictive evaluator assigns sufficiently negative gain to the proposed correction that even an infinitesimal target-side intervention is rejected.

At $\lambda = 1$, $\mathbf{p}_t^{(1)} = \hat{\mathbf{q}}_t^H$, giving

$$\begin{aligned}\mathcal{J}_t'(1) &= G_t^{\mathrm{pp}} - \sum\nolimits_{k=1}^{K} \hat{q}_{t,k}^H \log \frac{\hat{q}_{t,k}^H}{s_{t,k}} \\ &= G_t^{\mathrm{pp}} - D_{\mathrm{KL}}\left(\hat{\mathbf{q}}_t^H \| \mathbf{s}_t\right). \end{aligned} \tag{66}$$

Since the objective remains non-decreasing up to the target endpoint,

$$G_t^{\mathrm{pp}} \geq D_{\mathrm{KL}}\left(\hat{\mathbf{q}}_t^H \| \mathbf{s}_t\right) \quad \Longrightarrow \quad \lambda_t^\star = 1. \tag{67}$$

In this regime, the posterior-predictive evaluator sufficiently supports the full target-side correction.

Finally, when

$$\mathcal{J}_t'(0) > 0 \qquad \text{and} \qquad \mathcal{J}_t'(1) < 0, \tag{68}$$

strict concavity guarantees a unique interior optimum $\lambda_t^\star \in (0,1)$ satisfying $\mathbb{E}_{k \sim \mathbf{p}_t^{(\lambda_t^\star)}}\left[\hat{\xi}_{t,k}^s\right] = G_t^{\mathrm{pp}}$. Thus, the optimal partial intervention is the point along the evidence path at which the path-wise expected source-relative evidence matches the gain supported by the posterior-predictive evaluator.

**Efficient Global Solution.** Combining the three regimes gives

$$\lambda_t^\star = \begin{cases} 0, & G_t^{\mathrm{pp}} \leq -D_{\mathrm{KL}}\left(\mathbf{s}_t \| \hat{\mathbf{q}}_t^H\right), \\ 1, & G_t^{\mathrm{pp}} \geq D_{\mathrm{KL}}\left(\hat{\mathbf{q}}_t^H \| \mathbf{s}_t\right), \\ \text{the unique root of } \mathcal{J}_t'(\lambda) = 0, & \text{otherwise.} \end{cases} \tag{69}$$

Equivalently, the interior case corresponds to $-D_{\mathrm{KL}}\left(\mathbf{s}_t \| \hat{\mathbf{q}}_t^H\right) < G_t^{\mathrm{pp}} < D_{\mathrm{KL}}\left(\hat{\mathbf{q}}_t^H \| \mathbf{s}_t\right)$.

Since $\mathcal{J}_t'(\lambda)$ is monotone non-increasing, the interior root can be found efficiently by bisection. Reaching a tolerance $\epsilon_\lambda$ requires $O(\log(1/\epsilon_\lambda))$ iterations, with $O(K)$ computation per iteration.

Substituting the optimal intervention strength into the evidence path gives

$$p_{t,k}^\star = p_{t,k}^{(\lambda_t^\star)} = \frac{s_{t,k}^{1-\lambda_t^\star} (\hat{q}_{t,k}^H)^{\lambda_t^\star}}{\sum_{j=1}^{K} s_{t,j}^{1-\lambda_t^\star} (\hat{q}_{t,j}^H)^{\lambda_t^\star}} \propto s_{t,k}^{1-\lambda_t^\star} \left(\hat{q}_{t,k}^H\right)^{\lambda_t^\star}. \tag{70}$$

If $\hat{\mathbf{q}}_t^H = \mathbf{s}_t$, the evidence path collapses to a single prediction and $\mathcal{J}_t(\lambda) \equiv 0$. In this degenerate case, we set $\lambda_t^\star = 0$ as a conservative tie-breaking rule.

This completes the proof of Theorem 4.1.

### D.3 Scope of Optimality and Relation to Target Risk

**Optimality under the Working Evaluator.** Fix the current representation and the pre-$t$ statistical state, so that $\mathbf{s}_t$, $\hat{\mathbf{q}}_t^H$, and $\bar{\mathbf{q}}_t^H$ remain fixed while optimizing $\lambda$. Assume $s_{t,k}, \hat{q}_{t,k}^H > 0$ for every class.

Theorem 4.1 establishes global optimality with respect to the posterior-predictive logarithmic risk $\bar{\mathcal{R}}_t(\mathbf{p}) = -\sum_k \bar{q}^H_{t,k} \log p_k$ on the prescribed evidence path. In particular,

$$\lambda^\star_t \in \arg\min_{\lambda\in[0,1]} \bar{\mathcal{R}}_t\big(\mathbf{p}^{(\lambda)}_t\big). \tag{71}$$

Since both the source prediction and the target proposal belong to this path, the selected prediction satisfies

$$\bar{\mathcal{R}}_t(\mathbf{p}^\star_t) \le \min\left\{\bar{\mathcal{R}}_t(\mathbf{s}_t), \bar{\mathcal{R}}_t(\hat{\mathbf{q}}^H_t)\right\}. \tag{72}$$

Indeed, optimality implies $\mathcal{J}_t(\lambda^\star_t) \ge \max\{\mathcal{J}_t(0), \mathcal{J}_t(1)\} = \max\{0, G^{\mathrm{pp}}_t\}$. This is optimality over the fixed evidence path, not over all predictive distributions.

### D.4 Why Use the Posterior-Predictive Distribution as an Evaluator?

A natural alternative is to directly use the posterior-predictive distribution $\bar{\mathbf{q}}^H_t$ as the adapted prediction. However, its role in **GAIN** is deliberately different. The posterior-mean distribution $\hat{\mathbf{q}}^H_t$ specifies the correction suggested by the estimated target geometry, whereas $\bar{\mathbf{q}}^H_t$ marginalizes class-center uncertainty and is used to evaluate whether this correction remains beneficial relative to the frozen source prediction.

This separation is important because using the same distribution both to propose and evaluate a correction leads to a degenerate intervention. Consider an arbitrary target distribution $\mathbf{r}_t$ and the evidence path

$$p^{(\lambda)}_{t,k} = \frac{s^{1-\lambda}_{t,k} r^{\lambda}_{t,k}}{\sum_j s^{1-\lambda}_{t,j} r^{\lambda}_{t,j}}. \tag{73}$$

If $\mathbf{r}_t$ is also used as the evaluator, the corresponding objective is $J^r_t(\lambda) = \mathbb{E}_{\mathbf{r}_t}\left[\log \frac{\mathbf{p}^{(\lambda)}_t}{\mathbf{s}_t}\right]$. At $\lambda = 1$, $\mathbf{p}^{(1)}_t = \mathbf{r}_t$ and $\frac{\partial J^r_t(\lambda)}{\partial\lambda}\Big|_{\lambda=1} = D_{\mathrm{KL}}(\mathbf{r}_t\|\mathbf{s}_t) - \mathbb{E}_{\mathbf{r}_t}\left[\log \frac{\mathbf{r}_t}{\mathbf{s}_t}\right] = 0$.

Since $J^r_t(\lambda)$ is concave, its optimum is attained at $\lambda^\star = 1$ (except for the degenerate case $\mathbf{r}_t = \mathbf{s}_t$). Therefore, self-evaluation simply reduces to directly adopting the target distribution and provides no meaningful mechanism for deciding whether the proposed correction should be applied.

**GAIN** instead separates the two roles: $\hat{\mathbf{q}}^H_t$ proposes the correction, and $\bar{\mathbf{q}}^H_t$ evaluates its source-relative utility. This asymmetric proposal–evaluation design enables $\lambda^\star_t$ to reject, partially apply, or fully accept the history-derived correction rather than automatically trusting the target estimate.

**Direct posterior-predictive prediction.** To empirically examine whether the posterior-predictive evaluator should instead be used directly as the prediction, we compare **GAIN** with a variant that sets $\mathbf{p}^\star_t = \bar{\mathbf{q}}^H_t$. As shown in Table 7, directly predicting with $\bar{\mathbf{q}}^H_t$ is inferior to using it as an evaluator of $\hat{\mathbf{q}}^H_t$. This confirms that accounting for target-statistic uncertainty is most effective for assessing the utility of a proposed correction rather than indiscriminately replacing the source prediction.

## E Continual Target Update and Algorithm

### E.1 Continual Target Statistics

Our target-side estimator is maintained through compact sufficient statistics rather than explicit replay of preceding target representations. At time step $t$, the state constructed from $\mathcal{H}_t = \bigcup_{\tau<t} \mathcal{X}_\tau$ is used to estimate $\hat{\mathbf{q}}^H_t$ in Sec. 4.1. After the current mini-batch is predicted, the statistics are updated once and carried forward to the next time step. This section details the resulting causal recursion.

**Causal Update Protocol.** Let $\mathcal{X}_t = \{\mathbf{x}_{t,b}\}_{b=1}^{B_t}$ denote the target mini-batch at time step $t$, with frozen representations $\mathbf{z}_{t,b} = \phi_\theta(\mathbf{x}_{t,b})$. The predictive state available before processing $\mathcal{X}_t$ is

$$\mathcal{S}_{t-1} = \left(\{\mathbf{m}_{t-1,k}, \kappa_{t-1,k}, \pi_{t-1,k}\}_{k=1}^K, \Sigma_{t-1}\right). \tag{74}$$

This state remains fixed while predicting all samples in $\mathcal{X}_t$. After obtaining the gain-controlled predictions $\mathbf{p}^\star_{t,b}$, we assign each sample to class $k$ with reliability-weighted responsibility

$$\omega_{t,b,k} = \zeta_{t,b} p^\star_{t,b,k}, \qquad \text{where } \zeta_{t,b} = s_{t,b,\hat{y}^\star_{t,b}}, \hat{y}^\star_{t,b} = \arg\max_k p^\star_{t,b,k}. \tag{75}$$

The resulting weights satisfy $\sum_{k=1}^{K} \omega_{t,b,k} = \zeta_{t,b}$. Here, $\mathbf{p}^{\star}_{t,b}$ determines the class-wise allocation, while $\zeta_{t,b}$ measures how strongly the frozen source model supports the adapted prediction. Hence, corrections weakly supported by the source contribute less to the accumulated target statistics.

All predictions in $\mathcal{X}_t$ are computed before any state update. Therefore, each prediction depends only on preceding target observations and the current sample, while the current mini-batch can affect only subsequent adaptation. This batch-causal predict-then-update protocol avoids within-batch feedback and limits the propagation of unreliable corrections through the continual target state.

**Cold-Start Initialization.** When $\mathcal{H}_1{=}\varnothing$, no historical target evidence is available. As specified in Sec. 4.1, we set $\hat{\mathbf{q}}_1^H = \mathbf{s}_1$, which yields the conservative choice $\lambda_1^{\star}{=}0$ and $\mathbf{p}^{\star}_{1,b}{=}\mathbf{s}_{1,b}$. We initialize

$$n_{0,k} = 0, \qquad \mathbf{U}_{0,k} = \mathbf{0}, \qquad \mathbf{V}_{0,k} = \mathbf{0}, \qquad Q_{0,k} = 0, \tag{76}$$

where $n_{t,k}$ denotes the accumulated reliability-weighted class support, $\mathbf{U}_{t,k}$ and $\mathbf{V}_{t,k}$ are the corresponding weighted first- and second-moment statistics, and $Q_{t,k}$ accumulates squared weights for covariance estimation. Together with $\mathbf{m}_{0,k} = \mathbf{c}_k$, $\kappa_{0,k} = \kappa_0$, and $\Sigma_0 = \mathbf{I}_D$, these quantities initialize the target state. The source predictions of $\mathcal{X}_1$ then provide the initial reliability-weighted assignments for subsequent target-state updates.

**Effective Class Support and Center Update.** For class $k$, the current mini-batch contributes reliability-weighted soft support $\Delta n_{t,k} = \sum_{b=1}^{B_t} \omega_{t,b,k}$. We accumulate this support as

$$n_{t,k} = n_{t-1,k} + \Delta n_{t,k}, \qquad \kappa_{t,k} = \kappa_0 + n_{t,k}, \tag{77}$$

where $\kappa_0$ controls the strength of the source-centered prior and $\kappa_{t,k}$ denotes the resulting effective class support. We further maintain the responsibility-weighted first moment

$$\mathbf{U}_{t,k} = \mathbf{U}_{t-1,k} + \sum\nolimits_{b=1}^{B_t} \omega_{t,b,k}\mathbf{z}_{t,b}, \tag{78}$$

from which the source-anchored target center is recovered as

$$\mathbf{m}_{t,k} = \frac{\kappa_0\mathbf{c}_k + \mathbf{U}_{t,k}}{\kappa_0 + n_{t,k}} = \frac{\kappa_0\mathbf{c}_k + \mathbf{U}_{t,k}}{\kappa_{t,k}}. \tag{79}$$

Thus, the target center is updated from accumulated reliability-weighted evidence, while the source prior stabilizes the estimate when target support is limited.

**Effective-Support Shared Covariance Update.** We instantiate the shared covariance in Sec. 4.1 with a diagonal estimator. In addition to the first moment, we maintain the responsibility-weighted second moment

$$\mathbf{V}_{t,k} = \mathbf{V}_{t-1,k} + \sum\nolimits_{b=1}^{B_t} \omega_{t,b,k}\mathbf{z}_{t,b}^{\odot 2}, \qquad \mathbf{V}_{0,k} = \mathbf{0}. \tag{80}$$

For $n_{t,k}{>}0$, the corresponding target empirical center is $\bar{\mathbf{z}}_{t,k}{=}\frac{\mathbf{U}_{t,k}}{n_{t,k}}$, yielding the within-class scatter

$$\mathbf{R}_{t,k} = \mathbf{V}_{t,k} - \mathbf{U}_{t,k}^{\odot 2}/n_{t,k}. \tag{81}$$

Unlike the source-anchored center $\mathbf{m}_{t,k}$ used for classification, $\bar{\mathbf{z}}_{t,k}$ is estimated solely from target evidence, isolating within-class target dispersion from source–target center shift.

Raw assignment mass does not directly quantify the statistical support available for covariance estimation. We therefore additionally maintain the squared effective-weight mass

$$Q_{t,k} = Q_{t-1,k} + \sum\nolimits_{b=1}^{B_t} \omega_{t,b,k}^2, \qquad Q_{0,k} = 0, \tag{82}$$

and define the weighted residual degrees of freedom

$$\nu_{t,k} = n_{t,k} - \frac{Q_{t,k}}{n_{t,k}}, \qquad n_{t,k} > 0. \tag{83}$$

For $n$ unit-weight hard assignments, $\nu_{t,k} = n - 1$. In particular, a singleton gives $\nu_{t,k} = 0$ and therefore does not spuriously increase the statistical support of the covariance estimate.

Pooling the within-class scatter across classes gives

$$\nu_t = \sum\nolimits_{k:n_{t,k}>0} \nu_{t,k}, \qquad \hat{\mathbf{v}}_t^T = \frac{\sum_{k:n_{t,k}>0} \mathbf{R}_{t,k}}{\nu_t}, \qquad \nu_t > 0. \tag{84}$$

We initialize the shared covariance isotropically with $\mathbf{v}_0 = \mathbf{1}$, equivalently $\Sigma_0 = \mathbf{I}_D$, and shrink the empirical target variance toward this initialization:

$$\mathbf{v}_t = \begin{cases} \dfrac{\kappa_0 \mathbf{v}_0 + \nu_t \hat{\mathbf{v}}_t^T}{\kappa_0 + \nu_t}, & \nu_t > 0, \\ \mathbf{v}_0, & \nu_t = 0. \end{cases} \tag{85}$$

The resulting shared covariance is

$$\Sigma_t = \mathrm{Diag}\left(\max\{\mathbf{v}_t, \epsilon_\Sigma \mathbf{1}\}\right), \tag{86}$$

where the maximum is applied element-wise and $\epsilon_\Sigma$ is a fixed numerical variance floor. We reuse $\kappa_0$ as the shrinkage pseudo-support to avoid introducing an additional tuning parameter. The effective degrees of freedom prevent weak or singleton class support from prematurely overriding the isotropic initialization.

**Reliability-Balanced Historical Class Prior.** As defined in Eq. 77, $n_{t,k}$ denotes the retained target support accumulated from the reliability-weighted class weights $\omega_{t,b,k} = \zeta_{t,b} p^\star_{t,b,k}$, providing the effective support used in the Gaussian target-state update. For the historical class prior, we use this reliability-weighted support together with the corresponding predictive class mass obtained from the same adapted predictions. Specifically, we maintain the accumulated predictive class mass $\hat{\kappa}_{t,k} = \hat{n}_{t,k} + \kappa_0$ with $\hat{n}_{t,k} = \hat{n}_{t-1,k} + \sum_{b=1}^{B_t} p^\star_{t,b,k}$ . We then construct the historical class prior by combining reliability-normalized support with inverse-support balancing:

$$\pi_{t,k} \propto \underbrace{\frac{n_{t,k} + \kappa_0}{\hat{n}_{t,k} + \kappa_0}}_{\text{reliability-normalized support}} \cdot \underbrace{\frac{1}{\hat{n}_{t,k} + \kappa_0}}_{\text{class balancing}} = \frac{\kappa_{t,k}}{\hat{\kappa}_{t,k}^2}, \tag{87}$$

where the proportionality is normalized across classes. The first factor measures how strongly the accumulated predictive mass for class $k$ is supported by the reliability-weighted target statistics, while the second prevents frequently assigned classes from dominating the historical prior. At cold start, $n_{0,k} = \hat{n}_{0,k} = 0$ for all classes, so both factors are class-independent and the normalized prior reduces to $\pi_{0,k} = 1/K$. The prior is updated only after the current mini-batch has been predicted; hence, $\pi_{t-1,k}$ is used for samples in $\mathcal{X}_t$, while the updated $\pi_{t,k}$ is carried forward to time $t+1$.

**Carried-Forward State and Efficiency.** After processing $\mathcal{X}_t$, the updated predictive state

$$\mathcal{S}_t = \left(\{\mathbf{m}_{t,k}, \kappa_{t,k}, \pi_{t,k}\}_{k=1}^K, \Sigma_t\right) \tag{88}$$

is used only from time step $t+1$ onward. The quantities $\{n_{t,k}, \hat{n}_{t,k}, \mathbf{U}_{t,k}, \mathbf{V}_{t,k}, Q_{t,k}\}_{k=1}^K$ are auxiliary sufficient statistics used only for recursive state updates. Consequently, the method neither stores preceding target representations nor revisits earlier samples. With a diagonal shared covariance, the maintained statistics require $O(KD)$ memory independent of stream length, and all updates consist only of responsibility-weighted vector operations without backpropagation or replay.

### E.2 Overall Algorithm

Algorithm 1 summarizes the causal implementation of **GAIN**. The pre-$t$ target state is fixed while predicting the entire mini-batch, and all sufficient statistics are updated only after the corresponding predictions have been obtained. The detailed recursions are given in Sec. E.1.

## F Detailed Experimental Setup

**Datasets.** We conduct our main continual adaptation experiments on ImageNet-C (Hendrycks & Dietterich, 2019), which contains 15 corruption types at five severity levels: Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate, and JPEG compression. Following established CTTA protocols (Wang et al., 2022; Zhang et al., 2025c; Han et al., 2025a), we evaluate all corruptions at severity level 5 unless otherwise specified. For each corruption, we use a fixed set of 5,000 images following the RobustBench-based evaluation protocol (Zhang et al., 2025c; Liu et al., 2024b), yielding 75,000 samples per complete corruption cycle. The same per-corruption samples are used across

**Algorithm 1** **G**ain-**A**ware **IN**tervention (GAIN)

**Input:** Unlabeled target stream $\{\mathcal{X}_t\}_{t=1}^{T}$; frozen source model $f_\theta$ with feature extractor $\phi_\theta$; source prototypes $\{\mathbf{c}_k\}_{k=1}^{K}$; source-centered prior strength $\kappa_0$.
**Output:** Adapted predictions $\{\mathbf{p}^\star_{t,b}\}$.

**Initialize:** $\mathbf{m}_{0,k}=\mathbf{c}_k$, $\kappa_{0,k}=\kappa_0$, $\hat{\kappa}_{0,k}=\kappa_0$, $\pi_{0,k}=1/K$, and $\mathbf{\Sigma}_0=\mathbf{I}_D$.
**for** $t=1,\dots,T$ **do**
  Receive $\mathcal{X}_t=\{\mathbf{x}_{t,b}\}_{b=1}^{B_t}$ and freeze the pre-$t$ state $\mathcal{S}_{t-1}$ in Eq. 74.
  Compute features $\mathbf{z}_{t,b}=\phi_\theta(\mathbf{x}_{t,b})$ and source predictions $\mathbf{s}_{t,b}$ using Eq. 1.
  **if** $t=1$ **then**
    $\hat{q}^H_{t,b}\leftarrow s_{t,b}$, $\bar{q}^H_{t,b}\leftarrow s_{t,b}$, $\lambda^\star_{t,b}\leftarrow 0$, $p^\star_{t,b}\leftarrow s_{t,b}$.
  **else**
    **for** $b=1,\dots,B_t$ **do**
      Construct the target proposal $\hat{\mathbf{q}}^H_{t,b}$ from $d_{t,b,k}$ and $\hat{\ell}^H_{t,b,k}$ using Eq. 8 and Eq. 9.
      Form source-relative evidence $\hat{\xi}^s_{t,b,k}\leftarrow\log(\hat{q}^H_{t,b,k}/s_{t,b,k})$.
      Construct $\bar{\mathbf{q}}^H_{t,b}$ and evaluate $G^{\mathrm{pp}}_{t,b}$ using Eq. 10 and Eq. 11.
      Determine $\lambda^\star_{t,b}$ by Eq. 69.
      Obtain $\mathbf{p}^\star_{t,b}$ from Eq. 70.
    **end for**
  **end if**
  Obtain $\omega_{t,b,k}$ by Eq. 75.
  Update $\{n_{t,k},\kappa_{t,k},\mathbf{m}_{t,k}\}$ using Eqs. 77–79.
  Update $\Sigma_t$ using Eqs. 80–86.
  Update $\{\hat{\kappa}_{t,k},\pi_{t,k}\}$ using Eq. 87 and carry $\mathcal{S}_t$ in Eq. 88 to time $t+1$.
**end for**
**return** $\{\mathbf{p}^\star_{t,b}\}$.

all stream settings, which differ only in their temporal organization and repetition. Target labels are never accessed during adaptation and are used solely for evaluation.

To further evaluate the robustness and generalizability of **GAIN** beyond ImageNet-C, we include additional evaluations on both CTTA and standard TTA settings. For continual adaptation, we consider ImageNet-3DCC (Kar et al., 2022), which contains 12 corruption types at five severity levels and introduces geometry-aware transformations that produce more realistic distribution shifts. We further extend the evaluation beyond CTTA to standard TTA on ImageNet-R (Hendrycks et al., 2021), ImageNet-V2 (Recht et al., 2019), and ImageNet-Sketch (Wang et al., 2019), assessing whether gain-guided intervention generalizes across diverse forms of domain shift.

**Metrics.** We report top-1 accuracy (Acc.) and expected calibration error (ECE) (Naeini et al., 2015) to evaluate predictive performance and confidence calibration. For each target sample $i$, let $\mathbf{p}_i$ denote the predicted distribution, $\hat{y}_i=\arg\max_k p_{i,k}$ the predicted label, and $c_i=\max_k p_{i,k}$ the prediction confidence. Top-1 accuracy is

$$\mathrm{Acc}=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i=y_i]. \tag{89}$$

ECE partitions predictions into $M=20$ confidence bins $\{\mathcal{B}_m\}_{m=1}^{M}$ and measures the discrepancy between empirical accuracy and mean confidence:

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|\mathcal{B}_m|}{N}\left|\mathrm{acc}(\mathcal{B}_m)-\mathrm{conf}(\mathcal{B}_m)\right|, \tag{90}$$

where $\mathrm{acc}(\mathcal{B}_m)=|\mathcal{B}_m|^{-1}\sum_{i\in\mathcal{B}_m}\mathbb{I}[\hat{y}_i=y_i]$ and $\mathrm{conf}(\mathcal{B}_m)=|\mathcal{B}_m|^{-1}\sum_{i\in\mathcal{B}_m}c_i$. For CSC and CDC, ECE is computed separately for each corruption and then averaged over corruption types, with the same protocol used within each LHA cycle. For MDS, ECE is computed over the pooled mixed stream at each severity level. We report both metrics in percentage points, with higher Acc. and lower ECE indicating better performance. For ablation studies, we additionally report negative log-likelihood (NLL), $\mathrm{NLL}=-N^{-1}\sum_i\log p_{i,y_i}$, where lower is better.

Table 4: **CDC results on ImageNet-C.** Accuracy (Acc., %) and expected calibration error (ECE, %) with ViT-Base at severity level 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only. Our results are averaged over five runs.

| Method | BP-free | Metric | Noise | | | Blur | | | | Weather | | | | Digital | | | | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Gauss. | Shot | Impu. | Defo. | Glas. | Moti. | Zoom | Snow | Fros. | Fog | Brig. | Cont. | Elas. | Pix. | JPEG | |
| Source | – | Acc. ↑ | 47.0 | 48.2 | 47.9 | 31.5 | 21.2 | 41.5 | 36.7 | 50.1 | 45.8 | 42.3 | 73.6 | 8.6 | 42.5 | 62.0 | 63.8 | 44.2 |
| | | ECE ↓ | 3.6 | 4.1 | 3.7 | 4.3 | 5.4 | 3.9 | 9.1 | 2.3 | 4.9 | 17.4 | 3.1 | 3.9 | 9.0 | 3.3 | 2.7 | 5.4 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 46.1 | 47.4 | 47.5 | 33.3 | 22.3 | 42.7 | 38.6 | 50.3 | 46.1 | 44.2 | 74.0 | 6.6 | 43.3 | 62.4 | 64.6 | 44.6 |
| | | ECE ↓ | 7.0 | 6.9 | 10.3 | **4.1** | 12.1 | 8.4 | 17.3 | 5.5 | 8.9 | 10.0 | 5.2 | **2.2** | 12.8 | 5.7 | **4.1** | 8.0 |
| SAR (ICLR 2023) | ✗ | Acc. ↑ | 53.7 | 57.5 | 56.5 | 51.4 | 45.7 | 56.8 | 48.1 | 61.2 | 57.9 | 53.7 | 77.3 | 41.6 | 50.8 | 66.7 | 66.5 | 56.4 |
| | | ECE ↓ | 5.2 | 8.4 | 8.8 | 10.4 | 11.2 | 9.0 | 13.4 | 8.9 | 9.7 | **7.2** | **4.4** | 11.6 | 10.1 | **5.2** | 4.9 | 8.6 |
| ROID (WACV 2024) | ✗ | Acc. ↑ | 56.5 | 58.8 | 56.5 | 50.9 | 47.4 | 54.5 | 52.2 | 62.5 | 59.3 | 60.9 | 78.2 | 43.9 | 56.3 | 67.0 | 67.7 | 58.2 |
| | | ECE ↓ | 56.4 | 58.7 | 56.4 | 50.8 | 47.3 | 54.4 | 52.1 | 62.4 | 59.2 | 60.8 | 78.1 | 43.8 | 56.2 | 66.9 | 67.6 | 58.1 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 53.7 | 56.6 | 55.3 | 51.4 | 43.7 | 54.4 | 50.2 | 61.4 | 57.1 | 59.5 | 76.1 | 40.0 | 49.1 | 67.7 | 67.2 | 56.2 |
| | | ECE ↓ | 11.7 | 10.6 | 15.2 | 15.8 | 19.7 | 11.9 | 20.2 | 10.4 | 13.1 | 14.3 | 7.8 | 22.1 | 18.0 | 7.1 | 6.7 | 13.6 |
| DeYO (ICLR 2024) | ✗ | Acc. ↑ | 55.8 | 58.8 | 57.0 | 50.9 | 47.9 | 54.7 | 48.6 | 60.8 | 59.1 | 61.4 | 77.3 | 41.9 | 53.4 | 67.4 | 68.6 | 57.6 |
| | | ECE ↓ | 9.1 | 7.8 | 10.8 | 12.8 | 13.7 | 10.1 | 16.5 | 9.7 | 9.8 | 11.4 | 6.1 | 13.8 | 12.1 | 6.2 | 5.3 | 10.3 |
| AEA (ICLR 2025) | ✗ | Acc. ↑ | 47.9 | 47.7 | 51.9 | 48.5 | 47.6 | 49.7 | 49.8 | 47.0 | 55.2 | 65.1 | 75.0 | 37.4 | 49.1 | 64.4 | 64.5 | 53.4 |
| | | ECE ↓ | 22.4 | 20.5 | 25.2 | 27.8 | 26.1 | 24.7 | 60.6 | 26.0 | 21.8 | 21.0 | 16.6 | 34.4 | 28.2 | 18.9 | 19.6 | 26.2 |
| ReCAP (ICML 2025) | ✗ | Acc. ↑ | 40.8 | 42.5 | 46.0 | 51.5 | 48.2 | 53.2 | 50.0 | 55.0 | 57.9 | 63.7 | 75.7 | 57.3 | 49.8 | 62.6 | 64.6 | 54.6 |
| | | ECE ↓ | 11.7 | 10.1 | 13.0 | 10.9 | 12.5 | 9.6 | 15.2 | 9.5 | 8.8 | 9.3 | 5.3 | 8.6 | 12.3 | 6.0 | 6.0 | 9.9 |
| REM (ICML 2025) | ✗ | Acc. ↑ | 55.8 | 58.6 | 56.9 | 50.9 | **50.8** | 53.0 | **54.4** | 61.8 | 63.0 | 63.8 | 77.1 | 46.9 | 57.7 | **69.9** | 69.3 | 59.3 |
| | | ECE ↓ | 7.6 | 6.9 | 8.2 | 12.2 | 12.4 | 10.2 | 12.2 | 8.5 | **7.4** | 8.6 | 4.9 | 11.9 | 10.6 | 5.3 | 4.6 | 8.8 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 55.8 | 58.0 | 57.3 | 47.2 | 46.0 | 53.1 | 52.4 | 61.9 | 64.0 | 60.0 | 76.9 | 48.8 | 53.4 | 69.7 | **70.6** | 58.3 |
| | | ECE ↓ | 10.3 | 10.7 | 10.4 | 7.8 | 6.5 | 10.0 | 8.7 | 12.5 | 12.9 | 10.2 | 12.6 | 8.7 | 9.3 | 12.9 | 12.9 | 10.4 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 49.8 | 54.8 | 49.9 | 45.1 | 48.5 | 49.4 | 51.0 | 61.5 | 59.9 | 55.8 | 71.1 | 44.7 | **57.9** | 65.5 | 64.6 | 55.3 |
| | | ECE ↓ | 10.9 | 9.1 | 12.0 | 13.3 | 12.2 | 11.9 | 11.7 | 6.6 | 8.3 | 10.7 | 4.6 | 14.2 | 8.1 | 5.4 | 5.4 | 9.6 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 59.1 | 59.2 | 60.1 | 50.2 | 38.1 | 56.1 | 46.9 | 64.1 | 64.8 | 66.3 | 78.3 | 32.5 | 47.4 | 68.5 | 69.4 | 57.4 |
| | | ECE ↓ | 37.0 | 36.9 | 36.7 | 44.9 | 55.9 | 40.4 | 47.9 | 33.0 | 31.5 | 25.9 | 20.3 | 56.8 | 48.0 | 28.6 | 27.3 | 38.1 |
| NEO (ICLR 2026) | ✓ | Acc. ↑ | 56.2 | 56.5 | 56.9 | 46.9 | 36.0 | 52.6 | 45.4 | 62.9 | 63.9 | 68.9 | 78.2 | 36.4 | 45.8 | 67.0 | 67.1 | 56.1 |
| | | ECE ↓ | 10.0 | 6.7 | 9.1 | 6.8 | **5.3** | **4.0** | 4.9 | 5.4 | 20.8 | 51.5 | 9.0 | 23.8 | 5.9 | 5.7 | 6.7 | 11.7 |
| GAIN (Ours) | ✓ | Acc. ↑ | **59.9** ±0.73 | **59.7** ±0.76 | **60.8** ±1.19 | **54.4** ±0.64 | 43.9 ±0.98 | **58.7** ±0.39 | 51.6 ±0.91 | **66.1** ±0.46 | **67.0** ±0.19 | **72.2** ±0.46 | **78.4** ±0.25 | **61.7** ±0.61 | 53.1 ±2.64 | 69.2 ±0.64 | 70.3 ±0.49 | **61.8** ±0.28 |
| | | ECE ↓ | **5.2** ±0.52 | **5.7** ±0.70 | **5.9** ±1.12 | 5.9 ±0.35 | 5.6 ±0.31 | 5.4 ±0.38 | **4.8** ±0.52 | **5.2** ±0.36 | 8.8 ±0.52 | 7.3 ±3.06 | 5.1 ±0.75 | 10.9 ±0.95 | **5.8** ±0.40 | 5.9 ±1.04 | 4.9 ±0.46 | **6.2** ±0.23 |

**Considered Settings.** We evaluate **GAIN** under four complementary forms of continual distribution shift. ***Continual Structured Change (CSC)*** follows the conventional CTTA protocol (Wang et al., 2022), where the model encounters all 15 ImageNet-C corruptions sequentially at severity level 5, with 5,000 consecutive samples per corruption and no reset across domain transitions. We use the standard corruption order adopted in prior CTTA work. ***Continual Dynamic Change (CDC)*** follows DPCore (Zhang et al., 2025c) and reorganizes the same corruption domains into a less structured stream, where domains recur with non-uniform durations and frequencies. We use the released CDC construction with a Dirichlet concentration parameter of $\delta = 1$. ***Mixed-Domain Shift (MDS)*** follows the mixed-domain evaluation protocol used in Wild TTA (Niu et al., 2023; Hu et al., 2025), where samples from multiple corruption domains are interleaved within the same test stream rather than appearing in locally homogeneous domain segments. This setting introduces concurrent domain heterogeneity and tests whether adaptation remains reliable when accumulated target statistics reflect a mixture of shifts. Finally, ***Long-Horizon Adaptation (LHA)*** evaluates stability under repeated exposure to previously observed shifts. Following prior repeating-domain protocols (Liu et al., 2024b; Zhang et al., 2025c), we repeat the complete 15-corruption stream for 10 rounds (R1–R10) without resetting either the model or the target state. No corruption identity or domain boundary is provided to **GAIN** in any setting.

**Implementation Details.** We use an ImageNet-pretrained ViT-B/16 as the source model and set the test-time mini-batch size to 64 for all main experiments. The feature representation $\mathbf{z} = \phi_\theta(\mathbf{x})$ and source prediction $\mathbf{s} = f_\theta(\mathbf{x})$ are obtained from the frozen source network defined in Section 3.1; neither the backbone nor the classifier is updated during adaptation. **GAIN** maintains only the recursive target statistics described in Section E.1. We instantiate the shared covariance with a diagonal estimator, initialize $\Sigma_0 = \mathbf{I}_D$, and set the source-centered prior strength to $\kappa_0 = 3$. The intervention coefficient $\lambda^\star_{t,b}$ is determined independently for each test sample by the endpoint conditions in Eq. 69, with the interior case solved by one-dimensional bisection rather than treated as a tuned mixing coefficient. Unless otherwise specified, all hyperparameters and numerical tolerances are fixed across CSC, CDC, MDS, and long-horizon evaluation. For consistency, we compute ECE from the final outputs returned by each method's official implementation. Experiments are implemented in PyTorch and conducted on a single NVIDIA RTX A6000 GPU.

Table 5: **MDS results on ImageNet-C.** Accuracy (Acc.,%) and expected calibration error (ECE,%) across corruption severity levels 5–1. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only. Our results are averaged over five runs.

| Method | BP-free | Metric | Level 5 | Level 4 | Level 3 | Level 2 | Level 1 | **Avg.** |
|---|---|---|---|---|---|---|---|---|
| Source | – | Acc. ↑ | 44.2 | 55.4 | 68.3 | 69.2 | 74.8 | 62.4 |
| | | ECE ↓ | 3.4 | 4.3 | 4.2 | 3.3 | 2.8 | 3.6 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 50.3 | 62.0 | 69.4 | 73.9 | 78.2 | 66.8 |
| | | ECE ↓ | **5.4** | **4.6** | 3.9 | 3.9 | 3.4 | 4.2 |
| SAR (ICLR 2023) | ✗ | Acc. ↑ | 56.1 | 64.9 | 70.9 | 74.0 | 78.3 | 68.8 |
| | | ECE ↓ | 8.5 | 5.9 | 4.4 | 3.8 | 3.2 | 5.2 |
| ROID (WACV 2024) | ✗ | Acc. ↑ | 56.9 | 65.5 | 71.2 | 74.4 | 78.5 | 69.3 |
| | | ECE ↓ | 56.8 | 65.4 | 71.1 | 74.3 | 78.4 | 69.2 |
| DeYO (ICLR 2024) | ✗ | Acc. ↑ | 55.2 | 64.0 | 69.8 | 72.8 | 77.4 | 67.8 |
| | | ECE ↓ | 10.8 | 7.6 | 5.9 | 5.1 | 4.0 | 6.7 |
| AEA (ICLR 2025) | ✗ | Acc. ↑ | 52.4 | 60.2 | 67.8 | 52.2 | 55.8 | 57.7 |
| | | ECE ↓ | 19.4 | 16.3 | 12.1 | 33.5 | 27.6 | 21.8 |
| ReCAP (ICML 2025) | ✗ | Acc. ↑ | 56.7 | 65.0 | 70.6 | 73.4 | 77.7 | 68.7 |
| | | ECE ↓ | 11.0 | 8.2 | 6.5 | 5.9 | 4.8 | 7.3 |
| REM (ICML 2025) | ✗ | Acc. ↑ | **59.5** | 66.8 | 72.1 | 74.8 | 78.6 | 70.4 |
| | | ECE ↓ | 8.6 | 6.6 | 5.3 | 4.9 | 4.2 | 5.9 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 54.3 | 64.5 | 69.8 | 75.0 | 78.6 | 68.4 |
| | | ECE ↓ | 8.1 | 10.3 | 10.3 | 11.3 | 10.7 | 10.1 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 53.0 | 62.2 | 68.9 | 72.9 | 77.8 | 67.0 |
| | | ECE ↓ | 7.0 | 4.7 | **3.5** | **2.8** | **2.1** | **4.0** |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 55.6 | 65.9 | 71.8 | 75.3 | 79.2 | 69.6 |
| | | ECE ↓ | 37.9 | 29.4 | 24.4 | 21.5 | 18.3 | 26.3 |
| NEO (ICLR 2026) | ✓ | Acc. ↑ | 56.7 | 66.7 | 72.4 | 75.8 | 79.6 | 70.2 |
| | | ECE ↓ | 9.4 | 12.0 | 11.4 | 11.1 | 11.1 | 11.0 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | 59.0 ± 0.07 | **68.2** ± 0.03 | **73.6** ± 0.04 | **76.6** ± 0.06 | **80.1** ± 0.03 | **71.5** ± 0.02 |
| | | ECE ↓ | 8.1 ± 0.07 | 6.6 ± 0.15 | 5.6 ± 0.03 | 5.1 ± 0.06 | 5.2 ± 0.04 | 6.1 ± 0.04 |

Table 6: **LHA results on ImageNet-C.** Accuracy (Acc., %) and expected calibration error (ECE, %) over 10 repeated corruption cycles with ViT-Base at severity 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only.

| Method | BP-free | Metric | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | – | Acc. ↑ | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 | 44.2 |
| | | ECE ↓ | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 | 5.4 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 45.2 | 45.3 | 45.9 | 46.3 | 46.6 | 46.4 | 46.0 | 45.7 | 45.4 | 45.3 | 45.8 |
| | | ECE ↓ | 7.5 | 15.8 | 22.3 | 26.4 | 28.1 | 30.7 | 33.1 | 34.7 | 36.2 | 36.9 | 27.2 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 56.6 | 57.3 | 57.3 | 57.4 | 57.5 | 57.5 | 57.3 | 54.1 | 54.8 | 54.9 | 56.5 |
| | | ECE ↓ | 14.7 | 21.0 | 24.1 | 26.3 | 28.0 | 29.3 | 29.9 | 31.1 | 31.9 | 32.9 | 26.9 |
| AEA (ICLR 2025) | ✗ | Acc. ↑ | 54.7 | 59.4 | 59.5 | 8.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 18.2 |
| | | ECE ↓ | 21.8 | 27.6 | 30.6 | 90.3 | 99.9 | 99.9 | 99.9 | 99.9 | 99.9 | 99.9 | 77.0 |
| ReCAP (ICML 2025) | ✗ | Acc. ↑ | 57.7 | 59.3 | 60.1 | 60.5 | 60.6 | 60.7 | 60.9 | 61.0 | 61.0 | 61.1 | 60.3 |
| | | ECE ↓ | 9.5 | 12.1 | 13.1 | 13.7 | 14.2 | 14.7 | 14.9 | 15.3 | 15.6 | 15.7 | 13.9 |
| REM (ICML 2025) | ✗ | Acc. ↑ | 60.8 | 61.3 | 61.4 | 62.0 | 62.1 | 62.2 | 61.9 | 61.9 | 61.9 | 61.8 | 61.7 |
| | | ECE ↓ | 8.5 | 10.3 | 11.1 | 11.4 | 11.8 | 12.1 | 12.7 | 13.0 | 13.3 | 13.7 | 11.8 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 60.1 | 54.7 | 55.6 | 55.9 | 55.8 | 56.1 | 55.5 | 55.8 | 55.5 | 55.4 | 56.0 |
| | | ECE ↓ | 8.2 | 8.3 | 7.6 | 7.9 | 7.9 | 8.3 | 9.5 | 9.9 | 9.8 | 9.9 | 8.7 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 57.8 | 53.9 | 50.9 | 48.6 | 46.7 | 45.2 | 43.6 | 42.3 | 41.3 | 40.2 | 47.0 |
| | | ECE ↓ | 8.6 | 10.5 | 11.2 | 12.1 | 13.0 | 13.9 | 14.7 | 15.7 | 16.2 | 16.9 | 13.3 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 57.2 | 57.6 | 57.6 | 57.6 | 57.6 | 57.6 | 57.6 | 57.6 | 57.6 | 57.6 | 57.5 |
| | | ECE ↓ | 38.2 | 39.1 | 39.3 | 39.3 | 39.3 | 39.3 | 39.3 | 39.3 | 39.3 | 39.4 | 39.2 |
| NEO (ICLR 2026) | ✓ | Acc. ↑ | 56.0 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 | 56.1 |
| | | ECE ↓ | 11.7 | 11.7 | 11.6 | 11.6 | 11.6 | 11.7 | 11.7 | 11.7 | 11.7 | 11.7 | 11.7 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | **61.9** | **62.6** | **62.6** | **62.6** | **62.6** | **62.6** | **62.5** | **62.5** | **62.5** | **62.5** | **62.5** |
| | | ECE ↓ | **6.0** | **6.5** | **6.5** | **6.5** | **6.5** | **6.5** | **6.5** | **6.5** | **6.5** | **6.5** | **6.4** |

# G Additional Experimental Results

## G.1 Detailed Experimental Results on ImageNet-C

**Continual Dynamic Change (CDC).** CDC introduces irregular and recurring domain shifts, making accumulated target statistics more susceptible to staleness. As shown in Table 4, **GAIN** achieves the highest accuracy of 61.8% with 6.2% ECE, outperforming the strongest competing baseline REM by 2.5 accuracy points while reducing ECE by 2.6 points. Several methods degrade from CSC to CDC, including DPCore, ReCAP, and PAID, while ROID and DOTA exhibit severe miscalibration. These results show that **GAIN** remains reliable under dynamic shifts by evaluating the source-relative utility of history-induced corrections before intervention.

**Mixed-Domain Shift (MDS).** Table 5 reports results across corruption severity levels. **GAIN** achieves the highest average accuracy of 71.5%, outperforming REM (70.4%) and NEO (70.2%), and ranks first from Levels 4 to 1. It also maintains a low average ECE of 6.1%, substantially below other high-accuracy BP-free methods such as NEO (11.0%) and DOTA (26.3%). Overall, **GAIN** maintains a favorable accuracy–calibration trade-off under heterogeneous mixed-domain shifts.

Table 7: **Ablation of the proposal–evaluator roles under CSC.** We compare direct target-side prediction, reversed proposal–evaluation roles, and the asymmetric design used by GAIN.

| Variant | Proposal | Gain evaluation | Intervention | Acc. ↑ | ECE ↓ | NLL ↓ |
|---|---|---|---|---|---|---|
| Proposal as Prediction | $\hat{\mathbf{q}}^H$ | – | – | 58.6 | 11.2 | 2.5 |
| Evaluator as Prediction | $\bar{\mathbf{q}}^H$ | – | – | 58.9 | 12.8 | 2.4 |
| Swapped Roles | $\bar{\mathbf{q}}^H$ | $\hat{\mathbf{q}}^H$ | $\tilde{\lambda}^\star$ | 61.7 | 8.0 | 2.0 |
| **GAIN** (Ours) | $\hat{\mathbf{q}}^H$ | $\bar{\mathbf{q}}^H$ | $\lambda^\star$ | **61.9** | **6.0** | **1.9** |

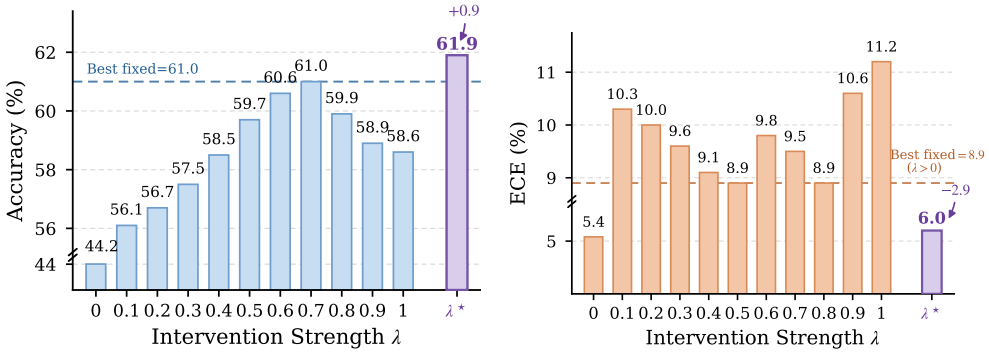



Figure 6: **Fixed vs. adaptive intervention strength.** We compare fixed $\lambda \in [0, 1]$ with the proposed sample-wise adaptive $\lambda^\star$. No single nonzero fixed intervention matches the accuracy–calibration trade-off of **GAIN**: the best fixed accuracy is 61.0%, while the lowest nonzero fixed ECE is 8.9%. The adaptive $\lambda^\star$ achieves 61.9% accuracy with 6.0% ECE.

**Long-Horizon Adaptation (LHA).** The long-horizon setting evaluates stability over ten repeated corruption cycles without resetting the adaptation state. As shown in Table 6, **GAIN** remains stable throughout all rounds, with accuracy increasing from 61.9% to 62.5–62.6% and ECE remaining around 6.5%. In contrast, DPCore shows a marked accuracy decline and PAID progressively deteriorates, while REM and ReCAP become increasingly miscalibrated despite competitive accuracy. These results show that gain-guided intervention limits the reinforcement of unreliable history-induced corrections and maintains stable adaptation over long horizons.

### G.2 Further Ablation Study and Analysis

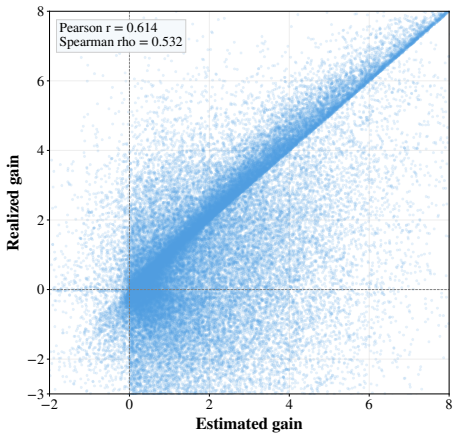
**Proposal–Evaluator Roles.** Table 7 examines the asymmetric roles of the posterior-mean proposal $\hat{\mathbf{q}}_t^H$ and posterior-predictive evaluator $\bar{\mathbf{q}}_t^H$. Directly using either distribution as the prediction performs substantially worse than **GAIN**, showing that improved target-side estimation alone does not guarantee a reliable correction. In particular, using $\bar{\mathbf{q}}_t^H$ directly yields 58.9% accuracy and 12.8% ECE, supporting its role as an uncertainty-aware evaluator rather than a replacement prediction. Reversing the proposal and evaluator retains competitive accuracy with 61.7% but degrades ECE from 6.0% to 8.0%. These results support the intended asymmetry of **GAIN**: $\hat{\mathbf{q}}_t^H$ specifies the correction, while $\bar{\mathbf{q}}_t^H$ evaluates its source-relative utility.

**Fixed vs. Adaptive Intervention.** Figure 6 further compares **GAIN** with fixed intervention strengths $\lambda \in [0, 1]$. Increasing $\lambda$ initially improves accuracy by incorporating more target-side evidence, but aggressive correction eventually degrades both accuracy and calibration. No single nonzero fixed value achieves the same trade-off as the sample-wise adaptive intervention: the best fixed accuracy reaches 61.0%, while the lowest nonzero fixed ECE remains 8.9%. In contrast, **GAIN** achieves 61.9% accuracy with 6.0% ECE, showing that correction strength should adapt to the estimated source-relative gain rather than remain fixed across samples.

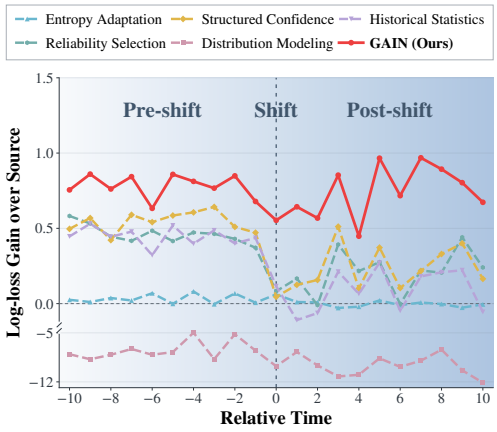
**Continual Target-state Update.** Table 8 first examines the reliability-weighted responsibility used for target-state updates in Eq. 75. Removing source support $\zeta$ consistently degrades all assignment variants, indicating that indiscriminately accumulating current predictions can amplify unreliable evidence over time. In particular, weighting the gain-controlled prediction by $\zeta$ improves accuracy from 59.9% to 61.9% and reduces ECE from 11.9% to 6.0%. This supports our update $\omega_{t,b,k} = \zeta_{t,b} p^\star_{t,b,k}$: predictions weakly supported by the frozen source contribute less to future target statistics. Moreover, using the intervened prediction $\mathbf{p}_t^\star$ outperforms updating with either the source prediction or the unfiltered target proposal, showing that gain-guided intervention also provides more reliable evidence for subsequent adaptation. We further ablate the reliability-balanced

Table 8: **Ablation of continual target-state updates under CSC.** RB denotes the reliability-balanced historical prior.

| Variant | Assignment | Source Support $\zeta$ | Class Prior | CSC | | |
|---|---|---|---|---|---|---|
| | | | | Acc.↑ | ECE↓ | NLL↓ |
| *State Assignment* | | | | | | |
| Source Assignment | $\mathbf{s}_t$ | – | RB | 58.5 | 13.7 | 2.2 |
| | | ✓ | RB | 60.8 | 7.2 | 2.0 |
| Proposal Assignment | $\hat{\mathbf{q}}_t^H$ | – | RB | 59.6 | 12.2 | 2.2 |
| | | ✓ | RB | 61.6 | 6.2 | 2.0 |
| **GAIN** (Ours) | $\mathbf{p}_t^\star$ | – | RB | 59.9 | 11.9 | 2.1 |
| | | ✓ | RB | **61.9** | **6.0** | **1.9** |
| *Historical Class Prior* | | | | | | |
| Uniform Prior | $\mathbf{p}_t^\star$ | ✓ | $1/K$ | 59.9 | 6.5 | 2.0 |
| Reliability Only | $\mathbf{p}_t^\star$ | ✓ | $\kappa_{t,k}/\hat{\kappa}_{t,k}$ | 59.6 | 6.8 | 2.0 |
| Balancing Only | $\mathbf{p}_t^\star$ | ✓ | $1/\hat{\kappa}_{t,k}$ | 60.5 | **5.5** | 2.0 |
| **GAIN** (Ours) | $\mathbf{p}_t^\star$ | ✓ | RB | **61.9** | 6.0 | **1.9** |




(a) Estimated vs. Realized Correction Gain




(b) Correction Gain Over Time

Figure 7: **Correction gain analysis.** (a) The estimated posterior-predictive gain is positively associated with the realized source-relative correction gain across test samples. (b) Not every correction helps: the source-relative benefit of adaptation varies throughout the stream, and target-driven corrections may yield limited or even negative gain.

historical class prior in Eq. 87, which combines reliability-normalized support with inverse-support balancing. Reliability-normalized support favors classes whose accumulated predictions are better supported, while inverse-support balancing prevents frequently predicted classes from progressively dominating the prior. The full prior achieves the best accuracy and NLL while maintaining low ECE, reaching 61.9% accuracy, 1.9 NLL, and 6.0% ECE. Together, the reliability-weighted state update and reliability-balanced historical prior play complementary roles in limiting the reinforcement of unreliable predictions and class bias, thereby mitigating error accumulation and maintaining a stable continual target state.

**Does Estimated Gain Reflect Correction Utility?** To assess whether the posterior-predictive gain reflects correction utility, we compare $G_t^{\mathrm{pp}}$ with the realized gain $\log(\hat{q}_{t,y_t}^H / s_{t,y_t})$, measured using ground-truth labels $y_t$ only for retrospective evaluation. This quantifies the full proposal's log-loss improvement over the source prediction, whose conditional expectation corresponds to Eq. 5. As shown in Fig. 7a, the estimated and realized gains are positively associated across samples (Pearson $r = 0.614$, Spearman $\rho = 0.532$). Despite finite, unlabeled observations and non-stationary target shifts, $G_t^{\mathrm{pp}}$ therefore meaningfully tracks and ranks the utility of a proposed correction. Unlike a confidence score, $G_t^{\mathrm{pp}}$ estimates the expected benefit of a specific correction relative to retaining the source prediction. Fig. 7b further shows why this utility must be evaluated continually. Around distribution shifts, correction benefit changes substantially, with several alternative signals yielding limited or even negative gain, whereas **GAIN** remains consistently positive across the transition. Together, these results directly support our principle: history proposes, while gain determines whether and how strongly to intervene.

Table 9: **CSC results on ImageNet-3DCC.** Accuracy (Acc., %) and expected calibration error (ECE, %) with ViT-Base at severity 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only.

| Method | BP-free | Metric | Depth of field | | Noise | | | Lighting | Weather | Video | | | Camera motion | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Near foc. | Far foc. | Color quant. | ISO | Low light | Flash | Fog 3D | Bit err. | H.265 abr. | H.265 crf | XY-mot. | Z-mot. | |
| Source | – | Acc. ↑ | 71.1 | 62.6 | 55.9 | 62.1 | 61.7 | 45.2 | 44.6 | 36.0 | 71.8 | 77.4 | 45.6 | 49.3 | 56.9 |
| | | ECE ↓ | 7.0 | 5.0 | 3.9 | 14.9 | 8.2 | 4.3 | 5.5 | 8.6 | 9.7 | 9.7 | 4.1 | 4.9 | 7.2 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 70.8 | 62.4 | 57.1 | 61.1 | 62.2 | 45.4 | 44.9 | 35.7 | 72.0 | 77.3 | 46.1 | 49.6 | 57.0 |
| | | ECE ↓ | 6.1 | 4.8 | **3.8** | 8.1 | **2.7** | 8.0 | **4.8** | 18.3 | **3.3** | 3.7 | 15.0 | 18.5 | 8.1 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 70.8 | 62.3 | 57.2 | 60.9 | 62.2 | 45.3 | 44.7 | 35.9 | 72.3 | 77.5 | 46.5 | 50.2 | 57.1 |
| | | ECE ↓ | 6.6 | 4.8 | 4.1 | 12.4 | 7.2 | **4.3** | 4.8 | 10.5 | 7.3 | 6.9 | **4.9** | **7.1** | **6.7** |
| REM (ICML 2025) | ✗ | Acc. ↑ | **74.9** | **68.0** | 61.8 | 65.6 | **70.8** | **49.7** | **49.5** | 31.2 | 63.9 | 77.3 | 49.0 | **57.6** | 60.0 |
| | | ECE ↓ | 3.4 | **4.5** | 6.1 | **5.0** | 4.8 | 10.7 | 8.6 | 36.6 | 15.3 | 6.2 | 16.8 | 14.7 | 11.1 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 74.5 | 67.7 | **62.1** | 63.0 | 69.3 | 48.1 | 41.2 | 34.4 | 74.0 | 78.3 | 48.8 | 53.6 | 59.6 |
| | | ECE ↓ | 12.6 | 10.9 | 10.4 | 11.5 | 11.6 | 6.7 | 4.8 | **5.9** | 12.7 | 10.5 | 5.9 | 7.6 | 9.2 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 73.3 | 65.4 | 58.4 | 56.2 | 65.2 | 47.7 | 46.5 | 33.5 | 68.8 | 74.1 | 43.6 | 54.6 | 57.3 |
| | | ECE ↓ | **2.9** | 4.8 | 6.7 | 8.1 | 6.1 | 9.6 | 9.5 | 14.7 | 4.1 | **3.3** | 13.2 | 8.4 | 7.6 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 71.1 | 63.2 | 57.2 | 63.3 | 63.9 | 46.6 | 45.6 | 36.3 | 74.1 | 78.8 | 48.2 | 51.0 | 58.3 |
| | | ECE ↓ | 24.4 | 32.0 | 37.7 | 32.4 | 32.5 | 48.0 | 48.8 | 58.3 | 24.1 | 19.9 | 47.8 | 45.5 | 37.6 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | 71.8 | 64.7 | 59.5 | **65.9** | 66.5 | 48.2 | 46.8 | **40.3** | **74.8** | **78.8** | **52.4** | 56.5 | **60.5** |
| | | ECE ↓ | 3.1 | 7.2 | 7.8 | 6.4 | 6.5 | 8.4 | 6.9 | 6.2 | 6.8 | 7.0 | 8.2 | 8.3 | 6.9 |

Table 10: **CDC results on ImageNet-3DCC.** Accuracy (Acc.,%) and expected calibration error (ECE,%) with ViT-Base at severity 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only. Our results are averaged over five runs.

| Method | BP-free | Metric | Depth of field | | Noise | | | Lighting | Weather | Video | | | Camera motion | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Near foc. | Far foc. | Color quant. | ISO | Low light | Flash | Fog 3D | Bit err. | H.265 abr. | H.265 crf | XY-mot. | Z-mot. | |
| Source | – | Acc. ↑ | 71.1 | 62.6 | 55.9 | 62.1 | 61.7 | 45.2 | 44.6 | 36.0 | 71.8 | 77.4 | 45.6 | 49.3 | 56.9 |
| | | ECE ↓ | 7.0 | 5.0 | 3.9 | 14.9 | 8.2 | 4.3 | 5.5 | 8.6 | 9.7 | 9.7 | 4.1 | 4.9 | 7.2 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 71.1 | 62.4 | 57.1 | 61.9 | 62.4 | 45.0 | 45.2 | 35.7 | 71.9 | 77.4 | 45.5 | 49.6 | 57.1 |
| | | ECE ↓ | **3.0** | 5.7 | 6.7 | **3.4** | **3.7** | 14.3 | 4.8 | 15.9 | **3.1** | 5.0 | 9.1 | 10.1 | 7.1 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 71.1 | 62.5 | 57.2 | 61.0 | 62.2 | 45.3 | 44.7 | 36.0 | 72.1 | 77.4 | 45.9 | 49.9 | 57.1 |
| | | ECE ↓ | 5.6 | 4.7 | **3.9** | 10.5 | 7.5 | **5.2** | **4.3** | 10.4 | 7.8 | 8.4 | **4.4** | **5.7** | **6.5** |
| REM (ICML 2025) | ✗ | Acc. ↑ | **74.3** | **65.2** | 59.2 | 63.8 | **68.0** | 42.7 | 43.7 | 33.3 | 73.3 | 76.6 | 52.1 | 54.0 | 58.9 |
| | | ECE ↓ | 5.0 | **3.8** | 9.9 | 7.4 | 6.8 | 31.2 | 30.7 | 28.4 | 7.0 | 6.0 | 14.7 | 15.2 | 13.8 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 72.7 | 63.8 | **60.2** | 60.2 | 66.7 | **48.4** | 41.6 | 33.6 | **74.7** | **78.9** | 49.6 | 51.9 | 58.5 |
| | | ECE ↓ | 12.1 | 10.3 | 11.6 | 12.5 | 11.6 | 6.2 | 4.8 | **7.2** | 12.3 | 12.3 | 6.5 | 8.2 | 9.6 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 70.2 | 61.2 | 54.9 | 54.2 | 64.3 | 45.2 | 45.4 | 34.1 | 71.4 | 77.0 | 46.8 | 53.7 | 56.5 |
| | | ECE ↓ | 4.4 | 6.6 | 9.3 | 9.2 | 6.1 | 12.1 | 10.8 | 16.2 | 3.5 | **2.8** | 12.4 | 9.5 | 8.6 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 72.2 | 63.9 | 57.8 | 63.7 | 63.6 | 46.6 | 45.4 | 36.4 | 73.8 | 77.8 | 47.6 | 50.2 | 58.3 |
| | | ECE ↓ | 24.8 | 32.9 | 38.6 | 32.6 | 32.0 | 48.4 | 49.7 | 57.1 | 24.1 | 19.7 | 47.0 | 44.3 | 37.6 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | 72.9 ±0.26 | 64.9 ±0.86 | 59.2 ±1.39 | **65.6** ±0.44 | 66.2 ±0.33 | 47.4 ±0.77 | **46.5** ±0.38 | **39.9** ±0.27 | 73.8 ±0.57 | 78.3 ±0.29 | **52.2** ±0.38 | **54.4** ±0.70 | **60.1** ±0.09 |
| | | ECE ↓ | 6.7 ±0.51 | 6.9 ±1.84 | 7.5 ±0.93 | 6.7 ±0.52 | 6.3 ±0.28 | 8.0 ±1.00 | 7.5 ±0.25 | 7.4 ±0.81 | 5.7 ±0.97 | 6.7 ±0.51 | 7.6 ±0.13 | 8.1 ±0.42 | 7.1 ±0.10 |

### G.3 More Results on ImageNet-3DCC

Tables 9–12 further evaluate **GAIN** on ImageNet-3DCC under CSC, CDC, MDS, and LHA. Compared with ImageNet-C, ImageNet-3DCC introduces more diverse shifts involving depth of field, lighting and weather, video compression, and camera motion, providing a complementary test of adaptation under heterogeneous distribution changes. **GAIN** achieves the highest average accuracy under CSC (60.5%), CDC (60.1%), and MDS (69.6%), while maintaining competitive calibration. The long-horizon setting further stresses error accumulation over 10 repeated corruption cycles without reset. **GAIN** remains stable throughout the stream, maintaining 60.5–60.8% accuracy with an average ECE of 8.2%, whereas several baselines exhibit substantial accuracy degradation or calibration drift. These results indicate that gain-guided intervention generalizes beyond ImageNet-C to more diverse corruption mechanisms, dynamic and mixed shifts, and prolonged continual adaptation.

### G.4 Evaluation under Test-Time Adaptation

Beyond continual test-time adaptation, we further evaluate **GAIN** under standard test-time adaptation (TTA) to examine its generalization to non-continual domain shifts. Following prior TTA evaluation, we consider shifts from ImageNet to ImageNet-R, ImageNet-V2, and ImageNet-Sketch, and compare against representative adaptation methods. Table 13 reports the top-1 accuracy on each target domain and their average. **GAIN** achieves the highest mean accuracy of 63.4%, with the best performance on ImageNet-V2 and ImageNet-Sketch, demonstrating that the proposed gain-guided intervention remains effective beyond continual adaptation.

Table 11: **MDS results on ImageNet-3DCC.** Accuracy (Acc., %) and expected calibration error (ECE, %) across severity levels 5–1. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only. Our results are averaged over five runs.

| Method | BP-free | Metric | Level 5 | Level 4 | Level 3 | Level 2 | Level 1 | **Avg.** |
|---|---|---|---|---|---|---|---|---|
| Source | – | Acc. ↑ | 56.9 | 64.1 | 69.4 | 73.6 | 76.6 | 68.1 |
| | | ECE ↓ | 4.3 | 5.9 | 6.7 | 7.7 | 8.3 | 6.6 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 57.1 | 64.3 | 69.6 | 73.8 | 76.7 | 68.3 |
| | | ECE ↓ | 4.8 | **3.3** | **2.7** | **2.2** | **2.2** | **3.0** |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 57.2 | 64.3 | 69.6 | 73.7 | 76.7 | 68.3 |
| | | ECE ↓ | **3.5** | 4.2 | 5.0 | 5.9 | 6.4 | 5.0 |
| REM (ICML 2025) | ✗ | Acc. ↑ | 58.3 | 65.4 | 70.7 | 74.8 | **77.6** | 69.4 |
| | | ECE ↓ | 6.4 | 6.8 | 7.1 | 7.7 | 8.1 | 7.2 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 57.7 | 64.9 | 69.3 | 73.2 | 76.3 | 68.3 |
| | | ECE ↓ | 7.5 | 10.0 | 9.8 | 9.4 | 9.6 | 9.3 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 55.1 | 62.5 | 68.2 | 72.8 | 76.2 | 67.0 |
| | | ECE ↓ | 6.5 | 4.8 | 3.8 | 2.9 | 2.2 | 4.1 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 57.0 | 64.1 | 69.5 | 73.6 | 76.6 | 68.1 |
| | | ECE ↓ | 36.9 | 31.0 | 26.6 | 23.0 | 20.6 | 27.6 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | **58.9** ± 0.07 | **65.8** ± 0.08 | **70.8** ± 0.05 | **74.9** ± 0.05 | 77.5 ± 0.05 | **69.6** ± 0.04 |
| | | ECE ↓ | 6.2 ± 0.10 | 6.7 ± 0.09 | 7.0 ± 0.10 | 7.0 ± 0.09 | 7.2 ± 0.04 | 6.8 ± 0.07 |

Table 12: **LHA results on ImageNet-3DCC.** Accuracy (Acc., %) and expected calibration error (ECE, %) over 10 repeated corruption cycles (R1–R10) with ViT-Base at severity 5. BP-free denotes backpropagation-free adaptation. Bold indicates the best results; Source is shown for reference only.

| Method | BP-free | Metric | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | – | Acc. ↑ | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 | 56.9 |
| | | ECE ↓ | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 | 7.2 |
| CoTTA (CVPR 2022) | ✗ | Acc. ↑ | 57.0 | 57.0 | 57.1 | 56.9 | 56.8 | 56.8 | 56.9 | 57.1 | 57.1 | 57.0 | 57.0 |
| | | ECE ↓ | 8.1 | 17.3 | 23.8 | 27.4 | 29.5 | 31.1 | 32.1 | 33.2 | 33.7 | 34.3 | 27.0 |
| ViDA (ICLR 2024) | ✗ | Acc. ↑ | 57.1 | 57.9 | 58.3 | 58.6 | 58.8 | 58.9 | 59.0 | 59.0 | 59.1 | 59.1 | 58.6 |
| | | ECE ↓ | **6.7** | **6.0** | **6.8** | 8.9 | 11.2 | 13.1 | 14.7 | 16.1 | 17.2 | 18.1 | 11.9 |
| REM (ICML 2025) | ✗ | Acc. ↑ | 60.0 | 58.1 | 58.7 | 58.1 | 56.8 | 50.5 | 36.7 | 0.3 | 0.1 | 0.1 | 37.9 |
| | | ECE ↓ | 11.1 | 18.1 | 19.7 | 21.7 | 24.4 | 33.5 | 51.3 | 99.5 | 99.9 | 99.9 | 47.9 |
| DPCore (ICML 2025) | ✗ | Acc. ↑ | 59.6 | 59.4 | 58.5 | 58.4 | 58.1 | 57.7 | 57.5 | 57.2 | 56.8 | 56.6 | 58.0 |
| | | ECE ↓ | 9.2 | 8.3 | 8.5 | 8.8 | 9.0 | 9.0 | 9.3 | 9.5 | 9.6 | 9.6 | 9.1 |
| PAID (NeurIPS 2025) | ✗ | Acc. ↑ | 57.3 | 55.1 | 52.8 | 51.0 | 49.4 | 48.4 | 47.1 | 46.0 | 45.1 | 44.2 | 49.6 |
| | | ECE ↓ | 7.6 | 8.1 | 8.3 | 8.4 | 8.6 | 8.8 | 9.1 | 9.4 | 9.7 | 9.8 | 8.8 |
| DOTA (NeurIPS 2025) | ✓ | Acc. ↑ | 58.3 | 58.5 | 58.5 | 58.5 | 58.5 | 58.5 | 58.5 | 58.4 | 58.4 | 58.4 | 58.4 |
| | | ECE ↓ | 37.6 | 38.7 | 38.8 | 38.8 | 38.8 | 38.9 | 38.9 | 38.9 | 38.9 | 38.9 | 38.7 |
| **GAIN** (Ours) | ✓ | Acc. ↑ | **60.5** | **60.8** | **60.8** | **60.8** | **60.8** | **60.8** | **60.8** | **60.8** | **60.8** | **60.7** | **60.7** |
| | | ECE ↓ | 6.9 | 8.0 | 8.2 | **8.3** | **8.3** | **8.4** | **8.4** | **8.4** | **8.4** | **8.4** | **8.2** |

Table 13: **TTA results under domain shifts.** Top-1 accuracy on ImageNet-R, ImageNet-V2, and ImageNet-Sketch. BP-free denotes backpropagation-free adaptation. Bold indicates the best results.

| Method | BP-free | ImageNet-R | ImageNet-V2 | ImageNet-Sketch | **Avg.** |
|---|---|---|---|---|---|
| Source | – | 59.5 | 75.4 | 44.9 | 59.9 |
| Tent (ICLR 2021) | ✗ | 63.9 | 75.2 | 49.1 | 62.7 |
| CoTTA (CVPR 2022) | ✗ | 63.5 | 75.4 | 50.0 | 63.0 |
| SAR (ICLR 2023) | ✗ | 63.3 | 75.1 | 48.7 | 62.4 |
| FOA (ICML 2024) | ✓ | 63.8 | 75.4 | 49.9 | 63.0 |
| REM (ICML 2025) | ✗ | **64.3** | 75.2 | 49.7 | 63.1 |
| **GAIN** (Ours) | ✓ | 63.0 | **75.8** | **51.5** | **63.4** |