VISITING THE STABILITY-PLASTICITY DILEMMA IN NON-STATIONARY TEST-TIME ADAPTATION FROM A BAYESIAN PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliably adapting foundation models (FMs) to non-stationary medical environments remains unsolved, as test distributions evolve dynamically through changing acquisition protocols, patient populations, and disease prevalence. Existing test-time adaptation (TTA) methods often degrade below the frozen source model. We trace this failure to *the stability-plasticity dilemma*: plasticity for tracking the evolving target and stability for preserving generic knowledge compete for the same parameter space. From a Bayesian perspective, we decompose the predictive posterior on test data into a class prior that encodes stable class-level knowledge and a class-conditional likelihood that carries per-input evidence, with their balance governed by the uncertainty of online evidence. This yields a **Bayesian Generalization Error Bound (BGEB)** that factorizes target risk into finite-batch uncertainty, prior mismatch, conditional shift, and parameter drift, thereby turning the dilemma into a principled trade-off. Guided by the error bound, we propose an effective algorithm **SynerTTA**, which rectifies the online prior, restricts updates to a low-displacement subspace, and weights adaptation by instance reliability. Across eight medical benchmarks and three modalities, SynerTTA outperforms thirteen baselines, surpassing the second-best method by 7.75% in accuracy and 5.56% in macro-F1, while achieving the best calibration and robustness under unreliable online evidence.

1 INTRODUCTION

Reliably adapting pre-trained foundation models (FMs) for real-world reasoning has emerged as a significant yet challenging topic in medical image computing (Moor et al., 7956; Zhang & Metaxas, 2024; Radford et al., 2021; Zhang et al., 2022; Wang et al., 2022b; Tiu et al., 2022; Huang et al., 2023; Lu et al., 2024; Silva-Rodríguez et al., 2025a). These models are typically trained with the fixed and stationary datasets to learn generic knowledge for deliver sophisticated reasoning on downstream tasks. However, they are deployed in real-world applications experiencing *dynamic evolving distribution shifts* due to varying hardware acquisition protocols, patient's individuality and disease prevalence. In such cases, continuously updating the parameters as new data arrives is a potentially feasible paradigm for FMs to consistently adapt the dynamic evolving distribution shifts (Ovadia et al., 2019; Rabanser et al., 2019).

Test-time adaptation (TTA) provides a promising means of FMs updating, by maintaining the learned generic knowledge across different datasets during inference (Sun et al., 2020; Wang et al., 2021; Liang et al., 2020; Iwasawa & Matsuo, 2021; Wang et al., 2022a; Gong et al., 2022; Yuan et al., 2023; Döbler et al., 2023; Niu et al., 2023). However, existing methods struggle with dynamic evolving distribution shifts under non-stationary clinical deployment, characterized by variable batch sizes, domain discrepancy and mixed data from different clinical centers (Silva-Rodríguez et al., 2025b; Zanella et al., 2025). For example, existing TTA methods suffer from pronounced performance degradations (Figure 1a) on the disease diagnosis task across eight clinical datasets. It shows that CoTTA is particularly fragile, reaching only 27.87% worst-case accuracy, 0.00% minimum Macro-F1, and drops of 8.33/20.82 points under mixed data. It highlights a significant observation that following TTA updating, the reasoning ability of FMs undergoes severe degradation, substantially underperforming the source models in non-stationary clinical deployments.

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.

(a) Stress-test summary

Method	Batch size	Class imbalance	Mixed data	
	Worst ↑ / Range ↓	Mean ↑ / Min ↑	ΔAcc ↑ / ΔF1 ↑	#F1↓ ↓
Tent	50.40 / 17.41	50.36 / 12.96	+2.72 / -1.64	1/3
StatA	50.40 / 12.74	53.84 / 38.89	+6.83 / +1.74	2/3
SAR	50.11 / 18.47	54.17 / 20.37	+5.17 / -0.83	2/3
CoTTA	27.87 / 20.48	19.54 / 0.00	-8.33 / -20.82	3/3
RoTTA	49.83 / 10.23	53.84 / 30.00	+2.00 / +2.13	**0/3**
SynerTTA	**65.39 / 8.46**	**74.32 / 57.07**	**+19.22 / +10.42**	**0/3**

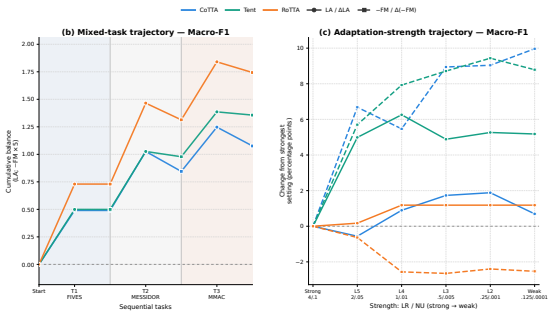


Figure 1: **Empirical evidence of the stability-plasticity dilemma in non-stationary TTA.** (a) Robustness to batch-size variation, class imbalance, and mixed data. (b) Cumulative Macro-F1 LA and -FM trajectories across sequential tasks at a fixed adaptation strength, FM is enlarged 5x for visualization. (c) Relative changes in LA and stability, Δ(-FM), as adaptation strength decreases. CoTTA and Tent vary the learning rate, whereas RoTTA varies NU.

This motivates the fundamental questions: **Q1: Why has TTA become detrimental in the very setting where adaptation is most needed**? It originates from a key finding: the key obstacle to reliable adaptation is the *stability-plasticity dilemma*, in which no single choice achieves a satisfactory trade-off. While memory *stability* prevents catastrophic forgetting of the generic knowledge under evolving data distributions, adaptation *plasticity* enables the model to follow new non-stationary test data in dynamic environments, yet the two compete for the same parameter space in FMs (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; Zenke et al., 2017; Parisi et al., 2019). As illustrated in Figure 1b and c, the trajectories reveal a systematic conflict between stability and plasticity: the FMs aggressively update parameters to adapt new data distribution, which risks eroding parameter space that underpin memory stability and incur knowledge forgetting. Conversely, conservative updates preserve generic knowledge by maintaining memory stability, but impede the model from effectively adapting to newly involved data (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018).

Q2: Is it possible to stably maintain the generic knowledge while addressing the dynamic evolving distribution shifts? Under dynamic evolving distribution shift, TTA methods simultaneously follow the evolving target (plasticity) and resist unreliable online updates (stability), giving rise to the stability-plasticity dilemma. On one hand, non-stationary TTA emphasizes instantaneous response to maintain plasticity, focusing on how adaptively a model can recover representations of non-stationary test streams after a distribution shift; on the other hand, they require stable enough to consolidate generic knowledge from past distributions by resisting non-stationary test data, biased pseudo-labels and continual parameter drifts (Wang et al., 2022a; Niu et al., 2022; Yuan et al., 2023; Niu et al., 2023). Unfortunately, non-stationary test data provides limited evidence through *extremely small batches*, biased evidence through *shifting class proportions*, and heterogeneous evidence through *mixed-center data* acquired with different scanners, protocols, and patient populations (Silva-Rodríguez et al., 2025b; Zanella et al., 2025). Together, these challenges call for a novel perspective that can balance the stability and plasticity in an unified framework.

In this paper, rather than relying on empirical trial and error, we analyze the underlying mechanisms of stability and plasticity in non-stationary TTA from a Baysian perspective. Observing that stability and plasticity are entangled within a single parameter update, we decompose this entanglement into two parts using Bayes' rule: the class prior encoding class-level knowledge we aim to preserve, the class-conditional likelihood carries per-input evidence to drive adaptation to the evolving shift, and the balance between them is governed by the uncertainty of online evidence (Ben-David et al., 2006; 2010; Mansour et al., 2009; Germain et al., 2016; Zhao et al., 2019; Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Garg et al., 2020). This yields a **Bayesian Generalization Error Bound (BGEB)** that makes the roles of both plasticity and stability explicit, and guides how to balance them in non-stationary TTA algorithm design.

Guided by the theoretical analysis, **Synergistic Test-Time Adaptation (SynerTTA)**, a lightweight algorithm is proposed to achieve reliably adaptation of FMs on non-stationary test stream by jointly targeting the dominant terms in the BGEB. SynerTTA estimates the online target prior to mitigate prior mismatch and debias pseudo-labels, applies Gradient-Sensitive Parameter Masking (GSPM) to

restrict unstable updates and reduce finite-batch variance and drift, and uses a reliability-weighted dynamic loss to balance plastic fitting on confident samples with semantic anchoring on uncertain samples. Instead of choosing between aggressive adaptation and conservative regularization, SynerTTA coordinates both sides of the trade-off according to the statistical reliability of the incoming stream. The main contributions of this work are summarized as follows: 1) We identify the failure of TTA in dynamic evolving distribution shift as a stability-plasticity dilemma using LA/FM analysis and adaptation-strength diagnostics. 2) For the first time, we derive a Bayesian target-risk upper bound (BGEB) that decomposes this dilemma into prior mismatch, likelihood shift, finite-batch uncertainty, pseudo-label risk, and parameter drift to balance the trade-off. 3) We propose a lightweight theory-driven algorithm (SynerTTA) that improves both adaptation and stability across non-stationary medical benchmarks.

2 METHOD

In this section, we first formalize non-stationary TTA and identify the stability-plasticity dilemma it encounters under dynamically evolving distribution shifts. From a Bayesian perspective, we then derive BGEB, a generalization bound on the test streams of the pre-trained FMs that explicitly balances stability and plasticity. Finally, guided by BGEB, we design SynerTTA, a theory-driven algorithm that achieves reliable adaptation of FMs on non-stationary test streams.

2.1 PROBLEM FORMULATION

Definition (Non-stationary Test-Time Adaptation). Let $h_0 = h_{\theta_0}$ be a foundation model pre-trained on the large-scale data draw from a source distribution P_s. During deployment, the model is test on an unlabeled data stream $\{\mathcal{B}_t\}_{t=1}^{T}$, where each mini-batch $\mathcal{B}_t = \{x_{t,i}\}_{i=1}^{B_t}$ is sampled from a non-stationary target distribution $P_t(x)$. The non-stationary means time-varying, *i.e.*, $P_t \neq P_{t+1}$, so the distribution shifts are dynamic and evolves over time, potentially accumulating irreversibly. At each step t, the model adapts its parameters $\theta_{t-1} \to \theta_t$ by optimizing the target risk function Eq. (1) only based on the current online evidence B_t from current mini-batch $\mathcal{B}_t$, without access to source data and target label (Wang et al., 2022a; Gong et al., 2022; Yuan et al., 2023; Döbler et al., 2023)

$$\epsilon_t(h) := \mathbb{E}_{(X,Y)\sim P_t}[\ell(h(X), Y)], \qquad \bar{\epsilon}_T := \frac{1}{T}\sum_{t=1}^{T} \epsilon_t(h_t). \tag{1}$$

while retaining performance on previously seen distributions.

Stability-Plasticity Dilemma. For the test data $\{x_{t,i}\}_{i=1}^{B_t}$, the adaptation is optimized to improve the learning accuracy and control $\bar{\epsilon}_T$ using only small, unlabeled, and temporally ordered batches (Eq. (1)). It concerns how rapidly h_t lowers the current-domain risk ϵ_t after a shift, *i.e.*, plasticity. Meanwhile, the model concerns whether this adaptation erodes performance on distributions encountered earlier. However, the model cannot access source data and labels to maintain the performance on source domain (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018). For a past domain $\tau < t$, we quantify the memory forgetting

$$F_{\tau\to t} := [\epsilon_\tau(h_t) - \epsilon_\tau(h_\tau)]_+, \qquad \mathrm{FM}_t := \frac{1}{t-1}\sum_{\tau=1}^{t-1} F_{\tau\to t}. \tag{2}$$

Let $\overline{F}_T = (T-1)^{-1}\sum_{t=2}^{T}\mathrm{FM}_t$. Therefore, non-stationary TTA aims to solve the conflicting bi-criterion problem by jointly minimizing $(\bar{\epsilon}_T, \overline{F}_T)$, where optimizing $\bar{\epsilon}_T$ means changing the parameters to adapt to the evolving data distribution for improving adaptation plasticity, but this change in turn causes the greater forgetting $\overline{F}_T$ to increase for lowering memory stability.

2.2 BAYES-FACTORIZED GENERALIZATION BOUND (BFGB)

To deal with the conflict optimization of $(\bar{\epsilon}_T, \overline{F}_T)$, we now proceed to develop bounds on the target generalization performance of the given pre-trained model h_{θ_0}. We first show how to bound the target error $\bar{\epsilon}_T$ to balance the adaptation plasticity and memory stability. Since the target labels are unknown, we cannot directly compute the target risk in Eq. (1) on the non-stationary test streams.

Moreover, under dynamical evolving distribution shifts, the conventional approach of characterizing a single source-to-target domain discrepancy is also infeasible in the process of non-stationary adaptation (Ben-David et al., 2006; 2010; Mansour et al., 2009; Germain et al., 2016; Zhao et al., 2019) Therefore, we factorize the target risk (Eq. (1)) into class prior and class-conditional likelihood with Bayes' rule, and then estimate the online evidence to bound the generalization to the evolving shifts (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Garg et al., 2020).

Bayes-factorized shift. We write the source and target joint laws as $P_s(X, Y) = P_s(X \mid Y)\pi_s(Y)$ and $P_t(X, Y) = P_t(X \mid Y)\pi_t(Y)$, where π_t is the target class prior. Given an online estimate $\widehat{\pi}_t$, define

$$\Delta_{\text{prior}}(t) := D_{\text{TV}}(\pi_t, \widehat{\pi}_t), \qquad \Delta_{\text{cond}}(t) := \mathbb{E}_{Y\sim\widehat{\pi}_t}[D_{\text{TV}}(P_t(X \mid Y), P_s(X \mid Y))]\,,$$

where $D_{\text{TV}}(P, Q) = \frac{1}{2}\|P - Q\|_1$ (Gibbs & Su, 2002). These quantities are analytical descriptors rather than assumed observables. The first measures residual prior-estimation error; the second captures class-conditional shift that cannot be removed by prior correction. We further denote the source risk reweighted by the estimated target prior as $\epsilon_{s,\widehat{\pi}_t}(h) := \sum_y \widehat{\pi}_t(y)\mathbb{E}_{X\sim P_s(X|y)}[\ell(h(X), y)]$ (Lipton et al., 2018; Garg et al., 2020; Alexandari et al., 2020).

Analytical setup. For the analysis, $\ell \in [0, M]$ is a bounded metric-type loss satisfying the triangle inequality; cross-entropy is used only as its optimization surrogate. Conditional on the stream history $\mathcal{F}_{t-1}$, the pre-update pseudo-labeler g_t and time-varying hypothesis (admissible update class) $\mathcal{H}_t$ are fixed. If a calibration map or parameter mask is selected from the same mini-batch, it is included in the corresponding joint loss class (Bartlett & Mendelson, 2002; Shalev-Shwartz & Ben-David, 2014). We define the target-marginal pseudo-label fitting risk and its empirical counterpart

$$\epsilon_t(h, g_t) := \mathbb{E}_{X\sim P_t^X}\left[\ell(h(X), g_t(X))\right], \widehat{\epsilon}_t(h, g_t) := \frac{1}{B_t}\sum_{i=1}^{B_t}\ell(h(x_{t,i}), g_t(x_{t,i}))\,.$$

Finally, we assume that the source-reweighted risk is locally $L_{s,t}$-Lipschitz along the adaptation trajectory. These conditions are local to each time step and impose no stationarity on the sequence $\{P_t\}_{t=1}^T$. We make use of this condition to state the bound on the target risk of the pre-trained FMs, on the non-stationary test streams.

Theorem 2.1 (Bayes-factorized generalization error bound)**.** *Under the setup above, let $\mathcal{L}_t = \{x \mapsto \ell(h(x), g_t(x)) : h \in \mathcal{H}_t\}$, where $\mathcal{H}_t$ is the time-varying hypothesis (admissible update class). Then, with probability at least $1 - \delta_t$, simultaneously for all $h \in \mathcal{H}_t$,*

$$\begin{aligned}\epsilon_t(h) \le\ & \underbrace{\widehat{\epsilon}_t(h, g_t)}_{\textit{current fitting}} + \underbrace{2\mathfrak{R}_{B_t}(\mathcal{L}_t) + M\sqrt{\frac{\log(1/\delta_t)}{2B_t}}}_{\textit{finite-batch uncertainty}} \\ & + \underbrace{\epsilon_{s,\widehat{\pi}_t}(h_0)}_{\textit{source anchor}} + \underbrace{L_{s,t}\|\vartheta_t - \theta_0\|_2}_{\textit{model drift}} + \underbrace{2M\Delta_{\text{prior}}(t)}_{\textit{prior mismatch}} + \underbrace{2M\Delta_{\text{cond}}(t)}_{\textit{conditional shift}}.\end{aligned} \tag{3}$$

Proof See *Appendix*.

Derivation. *Pseudo-label bridge.* For every (x, y), the triangle inequality gives $\ell(h(x), y) \le \ell(h(x), g_t(x)) + \ell(g_t(x), y)$. Taking the target expectation therefore yields $\epsilon_t(h) \le \epsilon_t(h, g_t) + \eta_t(g_t)$, where $\eta_t(g_t) := \mathbb{E}_{(X,Y)\sim P_t}[\ell(g_t(X), Y)]$ is the pseudo-label error. This is an inequality rather than an additive identity: fitting pseudo-labels and trusting them are two distinct sources of target error (Arazo et al., 2020).

Finite-batch control. The fitting term is defined under the actual target marginal P_t^X. Consequently, the ordinary incoming mini-batch is a valid sample for estimating $\epsilon_t(h, g_t)$; it is neither drawn from nor claimed to estimate a $\widehat{\pi}_t$-reweighted distribution. Conditional on $\mathcal{F}_{t-1}$, standard symmetrization and concentration uniformly bound it by $\widehat{\epsilon}_t(h, g_t) + 2\mathfrak{R}_{B_t}(\mathcal{L}_t) + M\sqrt{\log(1/\delta_t)/(2B_t)}$ (Shalev-Shwartz & Ben-David, 2014). This term makes the role of batch size explicit: even perfect optimization of the empirical objective need not imply reliable adaptation when B_t is small.

Bayes-factorized transfer. The remaining pseudo-label error inherits the classwise factorization of $P_t(X, Y)$. Inserting the estimated prior $\widehat{\pi}_t$ and the source conditional $P_s(X \mid Y)$, and then applying

the variational characterization of total variation, gives (Gibbs & Su, 2002)

$$\eta_t(g_t) \le \epsilon_{s,\widehat{\pi}_t}(g_t) + 2M\Delta_{\mathrm{prior}}(t) + 2M\Delta_{\mathrm{cond}}(t).$$

Local Lipschitz continuity then gives $\epsilon_{s,\widehat{\pi}_t}(g_t) \le \epsilon_{s,\widehat{\pi}_t}(h_0) + L_{s,t}\|\vartheta_t - \theta_0\|_2$. Combining the pseudo-label bridge, finite-batch control, and Bayes-factorized transfer proves Theorem 2.1. The order of this argument is essential: the ordinary batch estimates only target-marginal fitting, whereas prior calibration enters when the unobservable pseudo-label error is connected to the source model. Full measure-theoretic details are deferred to the appendix.

2.3 Stability-Plasticity Implications

Theorem 2.1 is instantaneous, but its terms have different temporal roles. Current fitting and shift correction determine how quickly the model can respond to P_t, whereas finite-batch uncertainty and parameter drift determine whether that response is statistically reliable and compatible with previously acquired knowledge. The drift term therefore provides a direct bridge from current-domain generalization to the forgetting quantity in Eq. (2).

Corollary 2.1.1 (Forgetting under cumulative adaptation)**.** *If $\epsilon_\tau(h_\theta)$ is locally L_τ-Lipschitz in θ,*

$$F_{\tau\to t} \le L_\tau\|\theta_t - \theta_\tau\|_2 \le L_\tau \sum_{k=\tau+1}^{t} \|\theta_k - \theta_{k-1}\|_2. \tag{4}$$

Corollary 2.1.1 shows that stability depends on the entire adaptation path rather than only on the final parameter vector. Each update consumes part of a cumulative stability budget; an unreliable step can increase forgetting at every later time. Conversely, suppressing all movement trivially shortens this path but may leave the current-fitting and shift terms in BGEB large. This is the risk-level form of the stability-plasticity dilemma.

To expose the same tension over the full stream, we specialize to the common online choice $g_t = h_{t-1}$ and aggregate the instantaneous bounds. Define $\bar{\epsilon}_T = T^{-1}\sum_{t=1}^{T}\epsilon_t(h_t)$ and

$$\Gamma_t(\delta, T) := 2\mathfrak{R}_{B_t}(\mathcal{L}_t) + M\sqrt{\frac{\log(T/\delta)}{2B_t}}.$$

Corollary 2.1.2 (Sequential non-stationary risk)**.** *Let $g_t = h_{t-1}$ and $L_{s,t} \le L_s$. Applying Theorem 2.1 with $\delta_t = \delta/T$ and taking a union bound gives, with probability at least $1-\delta$,*

$$\begin{aligned}\bar{\epsilon}_T \le \frac{1}{T}\sum_{t=1}^{T}\left[\widehat{\epsilon}_t(h_t, g_t) + \Gamma_t(\delta, T) + \epsilon_{s,\widehat{\pi}_t}(h_0)\right] \\ + \frac{L_s}{T}\sum_{k=1}^{T-1}(T-k)\|\theta_k - \theta_{k-1}\|_2 + \frac{2M}{T}\sum_{t=1}^{T}\left[\Delta_{\mathrm{prior}}(t) + \Delta_{\mathrm{cond}}(t)\right].\end{aligned} \tag{5}$$

Corollary 2.1.2 makes two asymmetries explicit. First, current fitting, finite-batch uncertainty, and distribution shift are paid at each time step, so plasticity must be continually renewed as the stream evolves. Second, the displacement made at time k is weighted by $T-k$: an early unstable update affects more subsequent models and carries a larger sequential cost than an equally sized late update. Thus, merely achieving low loss on the current batch is insufficient; the update must also be justified by the amount and reliability of the evidence that produced it.

Together, Eqs. (3)–(5) identify three controllable interfaces. Prior correction can reduce the separable Δ_{prior} term, restricting the update subspace can reduce both effective complexity and cumulative displacement, and reliability-aware supervision can decide when target fitting is preferable to source anchoring. SynerTTA instantiates these interfaces jointly: reliable evidence is used plastically, while uncertain evidence triggers conservative updates that preserve stability.

2.4 SynerTTA

In this section, we propose a theory-driven algorithm **SynerTTA** to achieve reliable adaptation of FMs on non-stationary test streams. The proposed SynerTTA operationalizes the controllable terms

in BGEB without storing source data or maintaining a memory bank. Given $\mathcal{B}_t$, it first estimates the evolving target prior and rectifies the logits with Reliability-Weighted Logits Adjustment (RWLA). The rectified probabilities provide pseudo-labels and instance reliability scores. A reliability-aware objective then favors target fitting on confident samples and source-feature anchoring on uncertain samples. Finally, Gradient-Sensitive Parameter Masking (GSPM) restricts the update to a low-displacement normalization-parameter subspace.

2.4.1 Reliability-Weighted Logits Adjustment (RWLA)

Bayesian prior rectification. Let $p_{i,k} = \mathrm{Softmax}(z_i)_k$ be the pre-update posterior for class k and let π_{ref} denote the model's effective training prior. Under the label-shift condition $P_t(X \mid Y) = P_{\mathrm{ref}}(X \mid Y)$, Bayes' rule yields (Saerens et al., 2002; Lipton et al., 2018)

$$P_t(Y = k \mid x_i) = \frac{p_{i,k}\, \pi_t(k)/\pi_{\mathrm{ref}}(k)}{\sum_{j=1}^{K} p_{i,j}\, \pi_t(j)/\pi_{\mathrm{ref}}(j)}. \tag{6}$$

Replacing the unknown target prior by $\widehat{\pi}_t$ gives the tempered correction

$$\widetilde{p}_i = \mathrm{Softmax}(z_i + \tau\,[\log \widehat{\pi}_t - \log \pi_{\mathrm{ref}}])\,, \tag{7}$$

where $\tau \in [0, 1]$ controls correction strength. Eq. (7) is the exact Bayesian correction at $\tau = 1$ under label shift. If π_{ref} is uniform, its log is a class-independent Softmax constant and the rule reduces to the desired $z_i + \tau \log \widehat{\pi}_t$. When class-conditional shift is also present, RWLA should be interpreted as targeting the separable prior-mismatch component in BGEB rather than exactly recovering the full target posterior.

Entropy-weighted online prior estimation. We estimate $\widehat{\pi}_t$ from the streaming predictions using one online EM update (Saerens et al., 2002; Alexandari et al., 2020; Garg et al., 2020). Given $\widehat{\pi}_{t-1}$, the E-step computes

$$q_{i,k} = \frac{p_{i,k}\, \widehat{\pi}_{t-1}(k)/\pi_{\mathrm{ref}}(k)}{\sum_{j=1}^{K} p_{i,j}\, \widehat{\pi}_{t-1}(j)/\pi_{\mathrm{ref}}(j)}, \qquad \overline{\pi}_t(k) = \frac{1}{B_t} \sum_{i=1}^{B_t} q_{i,k}. \tag{8}$$

To prevent an uncertain mini-batch from overwriting the historical estimate, we define its quality as $c_t = \left[1 - (B_t \log K)^{-1} \sum_i H(q_i)\right]_{[0,1]}$ and update

$$\widehat{\pi}_t = (1 - \alpha c_t)\widehat{\pi}_{t-1} + \alpha c_t \overline{\pi}_t, \tag{9}$$

where α is the base update rate. Thus, confident batches rapidly update the prior, whereas ambiguous batches preserve historical evidence (Grandvalet & Bengio, 2004; Wang et al., 2021; Guo et al., 2017; Ovadia et al., 2019).

2.4.2 Gradient-Sensitive Parameter Masking (GSPM)

Theorem 2.1 and Corollary 2.1.1 motivate controlling both the effective update class and the cumulative parameter path. We group the affine parameters of each adaptable normalization layer as $\theta^{(j)} = (w^{(j)}, b^{(j)})$ and score its instantaneous sensitivity by (Wang et al., 2021; Niu et al., 2022; Tian & Lyu, 2024; Kirkpatrick et al., 2017; Zenke et al., 2017)

$$s_t^{(j)} = \max(\|\nabla_{w^{(j)}} \mathcal{L}_t\|_2, \|\nabla_{b^{(j)}} \mathcal{L}_t\|_2)\,.$$

Let $k = \lceil \rho N \rceil$, where N is the number of adaptable groups and ρ is the update ratio. GSPM activates the k groups with the smallest sensitivity scores, forming a binary mask m_t, and performs

$$\theta_t = \theta_{t-1} - \eta_t\,(m_t \odot \nabla_\theta \mathcal{L}_t)\,. \tag{10}$$

This selection is deliberately conservative: among masks containing k groups, it minimizes the applied gradient norm and hence the instantaneous displacement. Substituting Eq. (10) into Corollary 2.1.1 gives

$$F_{\tau \to t} \le L_\tau \sum_{r=\tau+1}^{t} \eta_r\, \|m_r \odot \nabla_\theta \mathcal{L}_r\|_2\,. \tag{11}$$

For any fixed realized mask, the induced class is a subset of the full adaptation class and therefore has no larger complexity. The path-length guarantee in Eq. (11) holds regardless of how the mask is selected. It establishes drift control, rather than claiming that small-gradient selection is universally optimal for current-domain learning; plasticity is supplied by the reliability-aware objective below.

2.4.3 RELIABILITY-AWARE ADAPTATION OBJECTIVE

RWLA provides the rectified posterior $\widetilde{p}_i$ and pseudo-label $\widehat{y}_i = \arg\max_k \widetilde{p}_{i,k}$. We quantify instance reliability by normalized negative entropy, $r_i = 1 - H(\widetilde{p}_i)/\log K \in [0,1]$. Let $\phi_\theta(x)$ and $\phi_0(x)$ denote features from the adapted and frozen source models. We combine pseudo-label fitting and source-feature anchoring as

$$\mathcal{L}_t = \frac{1}{B_t}\sum_{i=1}^{B_t}\Big[r_i \underbrace{\mathrm{CE}(p_\theta(x_{t,i}), \widehat{y}_i)}_{\mathcal{L}_{\mathrm{fit}}^{(i)}} + \lambda(1-r_i)\underbrace{\|\phi_\theta(x_{t,i}) - \phi_0(x_{t,i})\|_2^2}_{\mathcal{L}_{\mathrm{anchor}}^{(i)}}\Big]. \tag{12}$$

For reliable samples, Eq. (12) emphasizes current-domain fitting and therefore plasticity. For uncertain samples, it suppresses confirmation-prone fitting and increases functional anchoring, thereby limiting drift (Tarvainen & Valpola, 2017; Sohn et al., 2020; Chen et al., 2022; Arazo et al., 2020). The objective does not assume that entropy equals the unobservable pseudo-label error; rather, it uses entropy as an operational reliability surrogate for balancing the two sides of BGEB. After computing Eq. (12), GSPM applies the masked update in Eq. (10). SynerTTA therefore coordinates prior correction, reliability-aware supervision, and drift-controlled optimization in a single online step.

3 EXPERIMENTS AND RESULTS ANALYSIS

To empirically validate our theoretical framework (Theorem 2.1) and evaluate the efficacy of the proposed SynerTTA, we conduct extensive experiments on clinical medical imaging streams. These non-stationary streams inherently exhibit severe long-tailed prior shifts (e.g., rare pathologies) and extreme covariate shifts, representing a safety-critical regime where standard CTTA models typically collapse (Silva-Rodríguez et al., 2025b; Zanella et al., 2025). We compare SynerTTA against the zero-shot baseline and state-of-the-art TTA methods (TENT, EATA, CoTTA, RoTTA, SAR) (Wang et al., 2021; Niu et al., 2022; Wang et al., 2022a; Yuan et al., 2023; Niu et al., 2023). Performance is evaluated using Accuracy (Acc), Macro-F1 (to penalize minority-class failures), and Brier Score (to assess probabilistic calibration) (Brier, 1950; Maier-Hein et al., 2024).

3.1 EXPERIMENTAL CONFIGURATION

Foundation Models and Downstream Benchmarks. To validate the effectiveness of SynerTTA under non-stationary clinical inference, extensive experiments are conducted on three non-stationary medical inference scenarios across heterogeneous medical modalities, where representative medical foundation models and unseen downstream benchmarks are adopted. *Ophthalmology scenario (Ophth.)* FLAIR is adopted as the foundation model, which employs a ResNet-50 vision encoder aligned with expert-driven textual supervision for retinal imaging (Silva-Rodríguez et al., 2025a; He et al., 2016). Performance is evaluated on MESSIDOR, MMAC, and FIVES (Decencière et al., 2014; Qian et al., 2024; Jin et al., 2022). *Histology scenario (Hist.)* CONCH is used as the backbone, featuring a ViT-B/16 visual encoder pre-trained on large-scale histopathology data (Lu et al., 2024; Dosovitskiy et al., 2021). Adaptation performance is assessed on SICAPv2 and SkinCancer (Silva-Rodríguez et al., 2020; Kriegsmann et al., 2022). *Chest X-ray (CXR) scenario* CONVIRT serves as the foundation model, combining a ResNet-50 visual encoder with a BioClinicalBERT text encoder pre-trained on MIMIC-CXR (Zhang et al., 2022; He et al., 2016; Alsentzer et al., 2019; Johnson et al., 2019). Evaluation is performed on NIH-LT and Covid (Wang et al., 2017; Holste et al., 2022; Chowdhury et al., 2020; Rahman et al., 2021).

3.2 MAIN RESULTS: ADAPTING TO NON-STATIONARY CLINICAL STREAMS

Overall Performance Evaluation. Table 1 presents a comprehensive quantitative comparison of SynerTTA against thirteen baseline methods across eight medical imaging datasets. Our proposed SynerTTA achieves the superior overall average performance with an Accuracy of 48.83% and an F1-score of 41.31%, marking a significant improvement over the Source model (36.70%/33.67%) and outperforming the second-best method, StatA, by 7.75% in Accuracy and 5.56% in F1-score. A detailed modality-wise analysis reveals that SynerTTA demonstrates exceptional adaptability in the face of non-stationary clinical shifts. In the Ophthalmology modality, SynerTTA attains the highest Accuracy on MESSIDOR (72.00%) and MMAC (36.81%), maintaining a narrow margin with

Table 1: Quantitative comparisons across three medical imaging modalities. SynerTTA consistently outperforms standard test-time adaptation methods and foundation model–specific approaches.

Method	Ophth.			Hist.			CXR		Overall
	MESSIDOR	MMAC	FIVES	SICAPv2	Skin	NCT-CRC	NIH-LT	Covid	Avg.
	Acc/F1	Acc/F1	Acc/F1	Acc/F1	Acc/F1	Acc/F1	Acc/F1	Acc/F1	Acc/F1
Source	55.14/58.24	35.35/25.15	67.50/66.14	23.49/19.60	22.80/18.81	38.69/35.47	21.75/20.23	28.85/25.75	36.70/33.67
TENTWang et al. (2021)	60.00/47.35	34.88/21.39	30.63/26.66	24.95/19.77	3.26/0.42	4.64/1.02	12.28/9.03	20.09/8.36	23.84/16.75
CoTTAWang et al. (2022a)	33.14/18.64	32.09/10.26	27.50/20.96	22.31/16.26	12.28/1.42	10.47/3.62	3.86/0.89	30.36/11.74	21.50/10.47
RoTTAYuan et al. (2023)	60.57/58.51	35.35/26.01	67.50/66.14	23.54/19.64	22.52/18.72	38.69/35.46	22.00/18.98	35.91/32.24	38.26/34.46
SARNiu et al. (2023)	58.00/47.51	34.88/20.54	31.87/28.77	23.07/19.75	21.33/17.93	41.55/36.56	13.33/10.66	29.18/26.69	31.65/26.05
TPTShu et al. (2022)	57.26/15.90	32.56/9.82	25.62/12.33	30.35/11.64	9.58/1.20	10.32/2.12	4.56/0.49	32.73/15.80	25.37/8.66
VTEDöbler et al. (2025)	58.40/14.75	31.62/11.68	25.00/10.00	30.35/11.64	3.30/0.40	8.51/3.07	5.26/0.53	30.30/11.63	24.09/7.96
BCAZhou et al. (2025)	65.43/49.79	34.88/25.73	35.62/28.94	0.28/11.63	4.23/1.35	7.19/6.55	12.63/6.33	50.67/32.63	26.37/20.37
BatclipMaharana et al. (2025)	57.71/45.85	34.88/20.61	28.72/25.17	37.08/13.56	4.46/3.35	11.43/6.73	11.58/11.49	32.67/28.69	27.32/19.43
StatAZanella et al. (2025)	59.14/56.94	35.81/27.87	71.25/70.12	34.72/12.89	24.01/25.69	**64.90/56.31**	20.00/19.13	12.70/12.03	40.32/35.12
O-TPTSharifdeen et al. (2025)	55.14/58.24	35.35/25.15	67.50/66.14	23.49/19.60	22.81/18.02	38.69/35.47	21.75/20.83	28.82/25.72	36.69/33.65
DOTAHan et al. (2025)	56.28/58.73	36.28/**29.00**	69.37/68.55	21.88/18.96	28.10/25.32	43.05/37.27	23.16/20.93	28.76/26.02	38.36/35.60
TNTImam et al. (2025)	69.42/61.31	33.49/19.05	27.50/14.75	25.04/17.60	7.43/3.79	27.56/26.04	14.43/13.51	34.48/20.14	29.92/22.02
SynerTTA	**72.57/65.47**	**36.81**/27.95	**73.18/71.93**	**37.97/19.98**	**33.38/27.78**	42.90/38.52	**22.81/20.97**	**71.03/57.85**	**48.83/41.31**

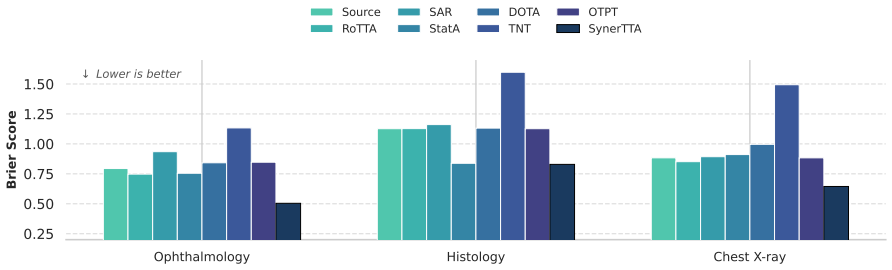


Figure 2: SOTA comparison of calibration performance across three medical modalities. SynerTTA consistently achieves the lowest scores across all domains, demonstrating its robustness in providing well-calibrated clinical inferences under non-stationary distributions.

specialized methods on FIVES. Notably, on the Histology-Skin and CXR-Covid datasets, which are characterized by extreme class imbalance and severe domain divergence, SynerTTA exhibits a decisive advantage. Specifically, on the Skin dataset, while conventional TTA methods like TENT and CoTTA suffer from catastrophic failure (F1-scores of 0.42% and 1.42%, respectively) due to error accumulation in unstable streams, SynerTTA achieves a robust F1-score of 27.78%. Similarly, in the Covid benchmark, SynerTTA yields a prominent 71.03% Accuracy, more than five times higher than that of StatA (12.70%), which, despite its high performance on NCT-CRC (F1 56.31%), shows extreme volatility and fails to generalize across diverse clinical scenarios. Furthermore, compared to foundation model-specific approaches such as TPT, BCA, and Batclip, SynerTTA consistently yields higher average metrics, suggesting that its synergistic mechanism effectively leverages multi-level information more holistically than simple prompt-based or batch-norm-based adaptation. Although SynerTTA is not the top performer in every single metric (e.g., slightly trailing DOTA in MMAC F1 and StatA in NCT-CRC), its ability to provide consistent, high-quality predictions across all eight benchmarks without collapsing on any individual dataset underscores its superior stability and generalization capability, making it highly suitable for reliable deployment in complex, real-world clinical inference environments.

Bayesian Calibration Analysis. To rigorously quantify the reliability of probabilistic predictions across diverse clinical modalities, we perform a Bayesian inference analysis on the Brier scores presented in Figure 2. In the context of medical decision-making, the Brier score serves as a critical measure of calibration, where lower values indicate more trustworthy probability estimates (Brier, 1950; Guo et al., 2017). Given the heterogeneity of the three modalities ($m \in \{\text{Ophth, Hist, CXR}\}$), we model the Brier score of each method j as a random variable and estimate the posterior distribution of the performance gap relative to our proposed method, defined as $\Delta_j = \mathbb{E}[B_j - B_{\text{SynerTTA}}]$. We adopt a Bayesian hierarchical setup with a Gaussian likelihood $D_{m,j} \sim \mathcal{N}(\Delta_j, \sigma_j^2)$ and a weakly-informative Jeffreys prior to ensure the inference is data-driven (Jeffreys, 1007).

Table 2: **Ablation study of SynerTTA across three medical modalities.** Bold values indicate the best performance, demonstrating that the full synergistic framework achieves optimal accuracy and Macro-F1 across all modalities. Variations such as "w/o Weighting" and "w/o adapt" are included to justify the necessity of the weighting mechanism and the adaptive refinement.

Method	Ophth.		Hist.		CXR		Avg.	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Source	52.80	50.15	23.16	19.23	25.30	22.99	36.47	33.55
Source+GSPM	57.09	52.49	30.82	17.07	27.04	24.06	40.99	34.25
Source+GSPM+Naive	59.06	53.31	31.20	21.56	32.40	23.13	43.48	35.61
Source + GSPM+RWLA(w/o Weighting)	59.49	53.95	35.16	22.54	35.03	26.53	45.55	37.14
Source + GSPM+RWLA	60.54	54.58	35.68	23.88	46.31	38.27	47.51	39.11
Source + GSPM+RWLA+Consistency(w/o adapt)	60.72	54.93	36.90	25.74	46.66	39.58	48.09	40.08
SynerTTA	**60.85**	**55.12**	**38.08**	**28.76**	**46.92**	**39.41**	**48.62**	**40.68**

The posterior analysis confirms that **SynerTTA** achieves a superior calibration state compared to all baselines. Specifically, SynerTTA yields the lowest Brier scores across all modalities (Ophth: 0.5037, Hist: 0.8287, CXR: 0.6442). Compared to the Source model, SynerTTA achieves a posterior mean improvement of $\Delta = 0.276$ with a 95% credible interval (CrI) of $[0.063, 0.488]$, and the probability of SynerTTA being better calibrated than the Source is $P(\Delta > 0) = 0.992$. Even when compared to the strongest baseline in calibration, StatA, SynerTTA maintains a clear advantage with a posterior mean difference of 0.175 (95% CrI: $[0.084, 0.266]$, $P > 0.99$). In contrast, TNT exhibits the most severe miscalibration, with a posterior mean Brier score 0.750 higher than SynerTTA (95% CrI: $[0.485, 1.014]$), suggesting that its adaptation process may introduce significant overconfidence or predictive instability. These results provide strong Bayesian evidence that SynerTTA effectively synergizes multi-level information to yield not only more accurate but also more reliable and well-calibrated clinical inferences across non-stationary distributions.

Ablation Studies. We conduct a comprehensive ablation study to verify the effectiveness of each proposed component, namely GSPM, RWLA, and Consistency, as summarized in Table 2. Starting from the baseline Source model (36.47% Avg. Acc and 33.55% Avg. F1), the progressive integration of our modules leads to consistent performance gains across all three modalities. The inclusion of GSPMprovides an initial boost in average accuracy to 40.99%, establishing a solid foundation for cross-modal adaptation. The introduction of RWLAfurther enhances the results, where the importance of our weighting mechanism is clearly demonstrated; specifically, the full RWLAoutperforms the version without weighting (w/o Weighting) by 1.96% in average accuracy and 1.97% in F1-score. This improvement is most pronounced in the CXR modality, where accuracy jumps significantly from 35.03% to 46.31%, underscoring the module's ability to selectively prioritize reliable features under severe distribution shifts. Furthermore, the addition of Consistency, particularly its adaptive mechanism, proves essential for maximizing the model's discriminative power. While the version without the adaptive mechanism (w/o adapt) achieves 48.09% average accuracy, the complete SynerTTAreaches the peak performance of 48.62% Acc and 40.68% F1. These findings highlight the synergistic effect of the three modules, where GSPMprovides initial alignment, RWLAoptimizes information weighting, and Consistencyrefines the final adaptation, collectively ensuring superior robustness and stability in non-stationary clinical inference environments. More experiments and results analysis see *Appendix*.

4 Conclusion

In this work, we revisited test-time adaptation under non-stationary shifts from a Bayesian perspective. We identify the *stability–plasticity dilemma* as the root cause of existing TTA failures and show that Bayes' rule decomposes it into a class prior and a class-conditional likelihood. This leads to a generalization bound that factorizes target risk into finite-batch uncertainty, prior mismatch, conditional shift, and parameter drift, converting the dilemma into a principled trade-off. Guided by this bound, we propose SynerTTA, which rectifies the online prior, restricts updates to a low-displacement parameter subspace, and weights adaptation by instance reliability. Across eight medical benchmarks and three modalities, SynerTTA consistently outperforms state-of-the-art baselines with well-calibrated predictions under dynamic clinical shifts.

AI USE STATEMENT

Generative AI tools were used only for language polishing, manuscript organization, reference formatting, and assisting with the design and visual presentation of figures/charts for experimental results. They were not used for theoretical derivation, mathematical proof development, literature search or summarization, experimental design, data generation, data analysis, or interpretation of empirical results. All experimental data and numerical results were produced by the authors through actual experiments; no experimental data were generated by AI. The authors manually reviewed and revised all AI-assisted text, independently checked all mathematical derivations, verified citations against the original sources, and validated all reported numerical results against the experiment outputs. The authors take full responsibility for the final content, claims, and artifacts in this work.

ETHICS STATEMENT

This work uses previously released medical-imaging benchmarks and does not involve new data collection, intervention, or interaction with human participants. We follow the access conditions and licenses of the respective datasets and use only the information required for the reported research. Although SynerTTA is designed to improve robustness under clinically relevant distribution shifts, the experiments do not constitute clinical validation. Performance may still vary across institutions, acquisition devices, demographic groups, and rare disease categories. The proposed method should therefore be independently validated, monitored for calibration and subgroup performance, and reviewed by qualified clinicians before any real-world clinical deployment.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete theoretical derivations of the Bayesian Generalization Error Bound and online prior rectification in Appendices B and C. Appendix D documents the benchmark datasets and official splits, modality-specific medical vision–language models, preprocessing procedures, source-free online adaptation protocol, non-stationary stream construction, baseline configurations, implementation details, and evaluation metrics. Appendices E and F further report hyperparameter sensitivity analyses and additional stability–plasticity results. All methods are initialized from the same modality-specific checkpoints and evaluated using identical batch partitions and sample orders, with target labels used exclusively for evaluation. To facilitate exact reproduction, we preserve the configuration files, sampled stream indices, and random seeds used to generate all reported tables and figures, and will release them together with the implementation.

REFERENCES

A. Alexandari, A. Kundaje, and A. Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 222–232, 2020.

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: *International Joint Conference on Neural Networks*, 2020.

K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. Regularized learning for domain adaptation under label shifts. In: *International Conference on Learning Representations*, 2019.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482, 2002.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In: *Advances in Neural Information Processing Systems*, vol. 19, pp. 137–144, 2006.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning* **79**, 151–175, 2010.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1), 1–3, 1950.

A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: *European Conference on Computer Vision*, pp. 532–547, 2018.

D. Chen, D. Wang, T. Darrell, and S. Ebrahimi. Contrastive test-time adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, M. B. I. Reaz, and M. T. Islam. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **8**, 132665–132676, 2020.

E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, J.-R. Ordóñez Varela, P. Massin, A. Erginay, B. Charton, and J.-C. Klein. Feedback on a publicly distributed image database: The Messidor database. *Image Analysis & Stereology* **33**(3), 231–234, 2014.

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

M. Döbler, R. A. Marsden, T. Raichle, and B. Yang. A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models. In: *Computer Vision–ECCV 2024 Workshops*, pp. 117–133, 2025.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*, 2021.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In: *International Conference on Learning Representations*, 2021.

S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton. A unified view of label shift estimation. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 3290–3300, 2020.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-bayesian perspective on domain adaptation. In: *Proceedings of the 33rd International Conference on Machine Learning*, pp. 859–868, 2016.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review* **70**(3), 419–435, 2002.

T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In: *Advances in Neural Information Processing Systems*, vol. 35, pp. 27253–27266, 2022.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In: *Advances in Neural Information Processing Systems*, vol. 17, 2004.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Z. Han, J. Yang, G. Wang, J. Li, Q. Xu, M. Z. Shou, and C. Zhang. DOTA: Distributional test-time adaptation of vision-language models. In: *Advances in Neural Information Processing Systems*, vol. 38, 2025.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

G. Holste, S. Wang, Z. Jiang, T. C. Shen, G. Shih, R. M. Summers, Y. Peng, and Z. Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: *Data Augmentation, Labelling, and Imperfections*, pp. 22–32, 2022.

Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine* **29**, 2307–2316, 2023.

R. Imam, A. Hanif, J. Zhang, K. W. Dawoud, Y. Kementchedjhieva, and M. Yaqub. Noise is an efficient learner for zero-shot vision-language models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 5879–5888, 2025.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

Y. Iwasawa and Y. Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In: *Advances in Neural Information Processing Systems*, vol. 34, pp. 2427–2440, 2021.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186**, 453–461 (1946), 1007.

K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, and J. Ye. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data* **9**, 475, 2022.

A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317, 2019.

J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo* . doi:10.5281/zenodo.1214456, 2018.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**(13), 3521–3526, 2017.

K. Kriegsmann, F. Lobers, C. Zgorzelski, J. Kriegsmann, C. Janßen, R. R. Meliß, T. Muley, U. Sack, G. Steinbuss, and M. Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology* **12**, 1022967, 2022.

J. Liang, D. Hu, and J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 6028–6039, 2020.

Z. C. Lipton, Y.-X. Wang, and A. J. Smola. Detecting and correcting for label shift with black box predictors. In: *Proceedings of the 35th International Conference on Machine Learning*, pp. 3122–3130, 2018.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In: *Advances in Neural Information Processing Systems*, vol. 30, 2017.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In: *International Conference on Learning Representations*, 2019.

M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang, and F. Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874, 2024.

S. Maharana, B. Zhang, L. Karlinsky, R. Feris, and Y. Guo. BATCLIP: Bimodal online test-time adaptation for CLIP. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1569–1579, 2025.

L. Maier-Hein, A. Reinke, P. Godau, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods* **21**, 195–212, 2024.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In: *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In: Bower, G. H. (ed.): *Psychology of Learning and Motivation*, vol. 24, pp. 109–165. Academic Press, 1989.

M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023), 7956.

S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan. Efficient test-time model adaptation without forgetting. In: *Proceedings of the 39th International Conference on Machine Learning*, pp. 16888–16905, 2022.

S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan. Towards stable test-time adaptation in dynamic wild world. In: *International Conference on Learning Representations*, 2023.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: *Advances in Neural Information Processing Systems*, vol. 32, 2019.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks* **113**, 54–71, 2019.

B. Qian, B. Sheng, H. Chen, et al. A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms. *JAMA Ophthalmology* **142**(11), 1006–1015, 2024.

S. Rabanser, S. Günnemann, and Z. C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In: *Advances in Neural Information Processing Systems*, vol. 32, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury. Exploring the effect of image enhancement techniques on COVID-19 detection using chest x-ray images. *Computers in Biology and Medicine* **132**, 104319, 2021.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation* **14**(1), 21–41, 2002.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

A. Sharifdeen, M. A. Munir, S. Baliah, S. Khan, and M. H. Khan. O-TPT: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19942–19951, 2025.

M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In: *Advances in Neural Information Processing Systems*, vol. 35, pp. 14274–14289, 2022.

J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine* **195**, 105637, 2020.

J. Silva-Rodríguez, H. Chakor, R. Kobbi, J. Dolz, and I. Ben Ayed. A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. *Medical Image Analysis* **99**, 103357, 2025a.

J. Silva-Rodríguez, F. Shakeri, H. Bahig, J. Dolz, and I. Ben Ayed. Few-shot, now for real: Medical VLMs adaptation without balanced sets or validation. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2025*, pp. 237–247, 2025b.

K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In: *Proceedings of the 37th International Conference on Machine Learning*, pp. 9229–9248, 2020.

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

J. Tian and F. Lyu. Parameter-selective continual test-time adaptation. In: *Proceedings of the Asian Conference on Computer Vision*, pp. 1384–1400, 2024.

E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* **6**, 1399–1406, 2022.

D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In: *International Conference on Learning Representations*, 2021.

Q. Wang, O. Fink, L. Van Gool, and D. Dai. Continual test-time domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022a.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017.

Z. Wang, Z. Wu, D. Agarwal, and J. Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b.

L. Yuan, B. Xie, and S. Li. Robust test-time adaptation in dynamic scenarios. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

M. Zanella, C. Fuchs, C. De Vleeschouwer, and I. Ben Ayed. Realistic test-time adaptation of vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25103–25112, 2025.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In: *Proceedings of the 34th International Conference on Machine Learning*, pp. 3987–3995, 2017.

S. Zhang and D. Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis* **91**, 102996, 2024.

Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In: *Proceedings of the 7th Machine Learning for Healthcare Conference*, pp. 2–25, 2022.

H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon. On learning invariant representations for domain adaptation. In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 7523–7532, 2019.

L. Zhou, M. Ye, S. Li, N. Li, X. Zhu, L. Deng, H. Liu, and Z. Lei. Bayesian test-time adaptation for vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29999–30009, 2025.

A Related Work

Test-time adaptation under non-stationary streams. Test-time adaptation updates a pretrained model using unlabeled target samples available only during inference. Early approaches introduced auxiliary self-supervision, entropy minimization, source-hypothesis transfer, or lightweight classifier adjustment to adapt without retraining on labeled target data (Sun et al., 2020; Wang et al., 2021; Liang et al., 2020; Iwasawa & Matsuo, 2021). Continual TTA subsequently extended this setting from a fixed target distribution to temporally evolving streams. CoTTA combines a weight-averaged teacher with stochastic restoration, NOTE addresses temporal correlation, RoTTA maintains robust online statistics and a memory bank, and RMT uses mean-teacher consistency for gradual adaptation (Wang et al., 2022a; Gong et al., 2022; Yuan et al., 2023; Döbler et al., 2023). Other methods improve reliability by filtering redundant or unreliable samples, constraining forgetting, seeking flat minima, or recovering the model after unstable updates (Niu et al., 2022; 2023). These mechanisms substantially improve robustness, but their stability controls are largely introduced as algorithm-specific heuristics. They do not provide a unified account of how current-domain fitting, finite-batch uncertainty, distributional change, and cumulative parameter drift jointly determine target risk. Our work complements this literature by formulating their competing effects as a stability-plasticity dilemma and connecting both sides through a common risk bound.

Medical vision–language models and test-time adaptation. Medical foundation models increasingly acquire transferable visual representations through image–text pretraining. Representative systems include ConVIRT and MedCLIP for radiology, CheXzero for zero-shot chest X-ray diagnosis, PLIP and CONCH for pathology, and FLAIR for retinal imaging (Zhang et al., 2022; Wang et al., 2022b; Tiu et al., 2022; Huang et al., 2023; Lu et al., 2024; Silva-Rodríguez et al., 2025a). Despite their strong transferability, reliable deployment remains difficult when target cohorts are small, imbalanced, or unavailable for validation (Moor et al., 7956; Zhang & Metaxas, 2024; Silva-Rodríguez et al., 2025b). Recent VLM adaptation methods optimize prompts, calibrate prompt directions, update visual and textual branches, or explicitly model predictive distributions at test time (Shu et al., 2022; Sharifdeen et al., 2025; Maharana et al., 2025; Zhou et al., 2025; Han et al., 2025; Imam et al., 2025). Comparative and realistic evaluations further show that adaptation behavior can vary sharply with the stream construction and deployment protocol (Döbler et al., 2025; Zanella et al., 2025). However, these studies primarily focus on improving empirical adaptation or calibration. They do not explicitly characterize why a medical VLM can learn a newly arrived clinical domain yet simultaneously lose previously useful knowledge, particularly under the small-batch, mixed-center streams considered here.

Generalization theory under distribution shift. Classical domain-adaptation theory bounds target risk through source risk, a discrepancy between domains, and an irreducible joint-error term (Ben-David et al., 2006; 2010; Mansour et al., 2009). Related analyses study invariant representations or derive PAC-Bayesian guarantees over distributions of hypotheses (Zhao et al., 2019; Germain et al., 2016). These formulations provide important foundations, but an aggregate source–target discrepancy does not isolate the distinct changes in disease prevalence and class-conditional appearance that occur along a non-stationary clinical stream. Label-shift studies instead use Bayes' rule to correct predictions under changing class priors and develop estimators based on confusion matrices or maximum likelihood (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019; Garg et al., 2020; Alexandari et al., 2020). Their usual assumption of invariant class-conditional distributions, however, does not cover simultaneous acquisition and population shifts, nor does it account for finite-batch uncertainty and cumulative model drift. BGEB draws on both lines of work: unlike PAC-Bayesian bounds over hypothesis posteriors, our Bayesian perspective factorizes the target distribution into its class prior and class-conditional likelihood, then incorporates this decomposition into an online target-risk bound that exposes the competing requirements of plasticity and stability.

B Proof of the Bayesian Generalization Error Bound

We give a conditional, per-step bound for a non-stationary test stream. The argument does not assume a stationary target distribution or a fixed transition law between successive batches.

Setting and notation. Let P_s be the source joint distribution and P_t the target joint distribution at time t, on a common input space $\mathcal{X}$ and a finite label space $\mathcal{Y}$. Write

$$P_t(dx, y) = \pi_t(y)P_t(dx \mid y), \qquad P_s(dx, y) = \pi_s(y)P_s(dx \mid y). \tag{13}$$

We assume that the source class-conditionals $P_s(\cdot \mid y)$ are defined for every class under consideration. Conditional on the history $\mathcal{F}_{t-1}$, the current unlabeled batch $\mathcal{B}_t = (X_{t,1}, \ldots, X_{t,B_t})$ consists of independent draws from the input marginal P_t^X. The distribution P_t and the hypothesis classes below may depend on $\mathcal{F}_{t-1}$, but are fixed before the draws in $\mathcal{B}_t$.

Let h_0 be the frozen source predictor, $h_t \in \mathcal{H}_t$ the adapted predictor, and $g_t \in \mathcal{G}_t$ the predictor producing pseudo-labels. Both h_t and g_t may be selected using the current batch. The loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ obeys

$$\ell(a, c) \le \ell(a, b) + \ell(b, c) \quad \text{for all } a, b, c \in \mathcal{Y}. \tag{14}$$

The 0–1 classification loss is one example. The theorem below does not apply directly to cross-entropy, which does not satisfy equation 14 in this form.

For any class-prior vector $\rho \in \Delta_{|\mathcal{Y}|}$, define the target risk, the source risk reweighted by ρ, and the source prediction disagreement by

$$\varepsilon_t(h) := \mathbb{E}_{(X,Y)\sim P_t}\ell(h(X), Y), \tag{15}$$

$$\varepsilon_{s,\rho}(h) := \sum_{y\in\mathcal{Y}} \rho(y)\mathbb{E}_{X\sim P_s(\cdot|y)}\ell(h(X), y), \tag{16}$$

$$D_{s,\rho}(g, h_0) := \sum_{y\in\mathcal{Y}} \rho(y)\mathbb{E}_{X\sim P_s(\cdot|y)}\ell(g(X), h_0(X)). \tag{17}$$

Let $\hat{\pi}_t \in \Delta_{|\mathcal{Y}|}$ denote an arbitrary, possibly batch-dependent estimate of the current class prior. The observable empirical pseudo-label risk is

$$\hat{\varepsilon}_{\mathcal{B}_t}(h, g) := \frac{1}{B_t}\sum_{i=1}^{B_t} \ell(h(X_{t,i}), g(X_{t,i})). \tag{18}$$

It is evaluated on samples from P_t^X; no reweighting of the unlabeled batch is assumed. Write $D_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$ and define the conditional likelihood discrepancy

$$\Delta_{\mathrm{like},t}(\rho) := \sum_{y\in\mathcal{Y}} \rho(y)D_{\mathrm{TV}}(P_t(\cdot \mid y), P_s(\cdot \mid y)). \tag{19}$$

Finally, let $\mathcal{A}_t = \{x \mapsto \ell(h(x), g(x)) : h \in \mathcal{H}_t,\ g \in \mathcal{G}_t\}$ and let $\mathfrak{R}_{B_t}^{(t)}(\mathcal{A}_t)$ be its expected Rademacher complexity under P_t^X, conditional on $\mathcal{F}_{t-1}$.

Theorem B.1 (Bayesian Generalization Error Bound)**.** *Under the preceding assumptions, for any $\delta \in (0, 1)$, with conditional probability at least $1 - \delta$ over $\mathcal{B}_t$, the following holds simultaneously for every $h_t \in \mathcal{H}_t$, $g_t \in \mathcal{G}_t$, and $\hat{\pi}_t \in \Delta_{|\mathcal{Y}|}$, including choices made after observing the batch:*

$$\varepsilon_t(h_t) \le \hat{\varepsilon}_{\mathcal{B}_t}(h_t, g_t) + \underbrace{2\mathfrak{R}_{B_t}^{(t)}(\mathcal{A}_t) + M\sqrt{\frac{\log(1/\delta)}{2B_t}}}_{\mathcal{V}_{B_t}(\delta)} + \varepsilon_{s,\hat{\pi}_t}(h_0) + D_{s,\hat{\pi}_t}(g_t, h_0)$$
$$+ 2MD_{\mathrm{TV}}(\pi_t, \hat{\pi}_t) + 2M\Delta_{\mathrm{like},t}(\hat{\pi}_t). \tag{20}$$

Proof. By equation 14, pointwise in (X, Y),

$$\ell(h_t(X), Y) \le \ell(h_t(X), g_t(X)) + \ell(g_t(X), Y). \tag{21}$$

Thus $\varepsilon_t(h_t) \le \mathbb{E}_{X\sim P_t^X}\ell(h_t(X), g_t(X)) + \varepsilon_t(g_t)$. This is an inequality, rather than an equality, even for the 0–1 loss.

Condition on $\mathcal{F}_{t-1}$. Symmetrization and bounded differences give the uniform deviation inequality (Bartlett & Mendelson, 2002; Shalev-Shwartz & Ben-David, 2014)

$$\sup_{f\in\mathcal{A}_t}\left\{\mathbb{E}_{P_t^X} f(X) - \frac{1}{B_t}\sum_{i=1}^{B_t} f(X_{t,i})\right\} \le 2\mathfrak{R}_{B_t}^{(t)}(\mathcal{A}_t) + M\sqrt{\frac{\log(1/\delta)}{2B_t}} \tag{22}$$

with conditional probability at least $1-\delta$. Because the event is uniform over $\mathcal{A}_t$, it also covers a student and pseudo-labeler chosen using $\mathcal{B}_t$. Consequently, the first expectation in equation 21 is at most $\hat{\varepsilon}_{\mathcal{B}_t}(h_t, g_t) + \mathcal{V}_{B_t}(\delta)$.

For the remaining pseudo-label error, expand $\varepsilon_t(g_t)$ using equation 13 and insert the estimated prior:

$$\begin{aligned}\varepsilon_t(g_t) &= \sum_y \hat{\pi}_t(y)\mathbb{E}_{P_t(\cdot|y)}\ell(g_t(X), y) + \sum_y \big(\pi_t(y) - \hat{\pi}_t(y)\big)\mathbb{E}_{P_t(\cdot|y)}\ell(g_t(X), y)\\ &\le \sum_y \hat{\pi}_t(y)\mathbb{E}_{P_t(\cdot|y)}\ell(g_t(X), y) + 2MD_{\mathrm{TV}}(\pi_t, \hat{\pi}_t).\end{aligned} \tag{23}$$

For every y, the bounded-loss total-variation inequality gives (Gibbs & Su, 2002)

$$\mathbb{E}_{P_t(\cdot|y)}\ell(g_t(X), y) \le \mathbb{E}_{P_s(\cdot|y)}\ell(g_t(X), y) + 2MD_{\mathrm{TV}}(P_t(\cdot \mid y), P_s(\cdot \mid y))\,. \tag{24}$$

This comparison is valid for the realized g_t and $\hat{\pi}_t$; it makes no independence claim about either quantity. Averaging equation 24 under $\hat{\pi}_t$ and using equation 14 on the source distribution yields

$$\varepsilon_t(g_t) \le \varepsilon_{s,\hat{\pi}_t}(h_0) + D_{s,\hat{\pi}_t}(g_t, h_0) + 2MD_{\mathrm{TV}}(\pi_t, \hat{\pi}_t) + 2M\Delta_{\mathrm{like},t}(\hat{\pi}_t). \tag{25}$$

Combining equation 21, equation 22, and equation 25 proves equation 20. Since equation 23–equation 25 are deterministic for any $\hat{\pi}_t$, estimating the prior from the current batch introduces no additional concentration step into this inequality; its estimation error remains explicit in the prior-mismatch term. □

Finite-batch and parameter-drift forms. The capacity term in equation 20 can be specialized to a familiar finite-batch expression. For the 0–1 loss, if the binary disagreement class $\{x \mapsto \mathbf{1}[h(x) \ne g(x)]\}$ has VC dimension $1 \le d_t \le B_t$, a standard growth-function bound gives (Shalev-Shwartz & Ben-David, 2014)

$$\mathcal{V}_{B_t}(\delta) \le 2\sqrt{\frac{2d_t\log(eB_t/d_t)}{B_t}} + \sqrt{\frac{\log(1/\delta)}{2B_t}}. \tag{26}$$

Here d_t is the capacity of the *student–teacher disagreement class*, not automatically the number of trainable parameters. A reduced capacity can be invoked for a parameter mask only if the restricted class is specified independently of the current batch, or a uniform bound over the possible masks is used.

The source disagreement in equation 20 is a functional measure of drift and requires no smoothness assumption on a neural network's hard predictions. To recover the parameter-distance form used to interpret continual updates, assume in addition that $g_t = g_{\theta_t^-}$, $h_0 = g_{\theta_0}$, and that every class-conditional source risk of the pseudo-labeler satisfies

$$\big|\mathbb{E}_{P_s(\cdot|y)}\ell(g_\theta(X), y) - \mathbb{E}_{P_s(\cdot|y)}\ell(g_{\theta_0}(X), y)\big| \le L\|\theta - \theta_0\|_2 \quad \text{for all } y. \tag{27}$$

Then $\varepsilon_{s,\hat{\pi}_t}(g_t) \le \varepsilon_{s,\hat{\pi}_t}(h_0) + L\|\theta_t^- - \theta_0\|_2$, so the proof of Theorem B.1 gives the corresponding BGEB with $D_{s,\hat{\pi}_t}(g_t, h_0)$ replaced by $L\|\theta_t^- - \theta_0\|_2$. If pseudo-labels come from the pre-update model, then $\theta_t^- = \theta_{t-1}$; if they come from a distinct calibrated teacher, its own drift must be controlled. Condition equation 27 is an additional assumption and is not automatic for hard 0–1 predictions.

Relation to stability and plasticity. The empirical pseudo-label term rewards fitting the current batch, whereas the finite-batch term and pseudo-labeler drift penalize unreliable adaptation. Prior mismatch and class-conditional shift describe two different obstacles to current-domain performance. Theorem B.1 bounds the *current target risk*; it is not itself a bound on the experimental

forgetting measure (FM). For any labeled reference distribution Q on an earlier center, equation 14 separately implies

$$\varepsilon_Q(h_t) - \varepsilon_Q(h_0) \le \mathbb{E}_{X\sim Q^X}\,\ell(h_t(X), h_0(X)). \tag{28}$$

This shows how prediction drift can affect past-domain error. It does not directly control Macro-F1-based FM, which is a different, non-decomposable metric.

The reweighted source error, the two total-variation discrepancies, and the source disagreement are theoretical quantities that need not be observable during source-free deployment. Thus BGEB motivates the prior estimator, selective updating, and reliability-aware loss in SynerTTA, but does not claim that the algorithm computes or minimizes every term of the bound exactly. In particular, a practical cross-entropy or semantic-alignment objective is a surrogate for, rather than the loss appearing in, Theorem B.1; identifying the two requires a separate surrogate-to-classification-risk argument.

C Bayesian Justification of Online Prior Rectification

This section formalizes the prior-estimation and logit-rectification component of SynerTTA. Its purpose is to connect the online update to the prior-mismatch term in Theorem B.1; the analysis is conditional on the current target batch and does not require a stationary stream.

From source scores to target posteriors. Let $K = |\mathcal{Y}| \ge 2$, and assume $\pi_s(k) > 0$ for every class k. Write $r_s(k \mid x) = P_s(Y = k \mid X = x)$ for a calibrated source posterior. Whenever the relevant densities are positive, Bayes' rule gives (Saerens et al., 2002; Lipton et al., 2018)

$$r_s(k \mid x) = \frac{\pi_s(k)p_s(x \mid k)}{p_s(x)}, \qquad p_t(k \mid x) = \frac{\pi_t(k)p_t(x \mid k)}{p_t(x)}. \tag{29}$$

Thus, with $c_{t,k}(x) := p_t(x \mid k)/p_s(x \mid k)$, the exact target posterior is

$$p_t(k \mid x) = \frac{r_s(k \mid x)\,[\pi_t(k)/\pi_s(k)]\,c_{t,k}(x)}{\sum_{j=1}^{K} r_s(j \mid x)\,[\pi_t(j)/\pi_s(j)]\,c_{t,j}(x)}. \tag{30}$$

The ratio $p_s(x)/p_t(x)$ cancels across classes in the normalization; equality of the source and target input marginals is unnecessary. Under pure label shift, $p_t(x \mid k) = p_s(x \mid k)$ and hence $c_{t,k}(x) = 1$. If $z_k(x)$ are source logits with $\mathrm{softmax}(z(x))_k = r_s(k \mid x)$, then for any candidate prior $\rho \in \Delta_K$ with positive entries,

$$q_\rho(k \mid x) := \frac{r_s(k \mid x)\rho_k/\pi_s(k)}{\sum_j r_s(j \mid x)\rho_j/\pi_s(j)} = \mathrm{softmax}(z(x) + \log\rho - \log\pi_s)_k\,. \tag{31}$$

Under calibrated source predictions and pure label shift, $q_{\pi_t}(k \mid x) = p_t(k \mid x)$. The source prior must therefore be removed before the target prior is inserted. If π_s is assumed uniform, the $-\log\pi_s$ term is class-independent and cancels in the softmax. Replacing π_s by a uniform prior when it is not uniform, or using a strength coefficient other than one on the log-prior correction, yields a tempered heuristic rather than the exact Bayesian posterior (Saerens et al., 2002; Alexandari et al., 2020; Garg et al., 2020).

Prior *removal* has a different interpretation from posterior *correction*. If logits $z(x)$ already represent the current target posterior, then $z_k(x) - \log\pi_t(k)$ is proportional, up to an x-dependent constant, to $\log p_t(x \mid k)$. Consequently, $\mathrm{softmax}(z - \log\hat{\pi}_t)$ is a prior-neutralized class score when $\hat{\pi}_t \approx \pi_t$. Applied instead to source-posterior logits, this expression is not the target-posterior correction in equation 31. When class-conditional appearance changes, the unknown $c_{t,k}(x)$ in equation 30 prevents exact correction by priors alone; this residual shift is represented separately by $\Delta_{\mathrm{like},t}$ in equation 20.

Online EM and entropy-controlled smoothing. Fix a batch $\mathcal{B}_t = \{X_{t,i}\}_{i=1}^{B_t}$ and strictly positive class scores $a_{t,i,k}$ during the prior update. Under calibrated source probabilities, one may take $a_{t,i,k} = r_s(k \mid X_{t,i})/\pi_s(k)$, which is proportional across k to $p_s(X_{t,i} \mid k)$. Dividing by $\pi_s(k)$ is necessary here: using $r_s(k \mid X_{t,i})$ itself as a likelihood would count the source prior twice. Let $\rho_{t-1} \in \mathrm{int}(\Delta_K)$ denote the preceding estimate. The batch mixture objective is (Saerens et al., 2002;

Alexandari et al., 2020)

$$\mathcal{L}_t(\rho) := \sum_{i=1}^{B_t} \log\left(\sum_{k=1}^{K} \rho_k a_{t,i,k}\right). \tag{32}$$

The additive terms that would turn the proportional scores into normalized densities do not depend on ρ and therefore do not affect its maximizer. If the scores instead come from an adapted, uncalibrated model, the same updates optimize a surrogate mixture objective, without an exact likelihood interpretation. One EM step computes (Saerens et al., 2002; Alexandari et al., 2020)

$$q_{t,i,k} := \frac{\rho_{t-1,k} a_{t,i,k}}{\sum_j \rho_{t-1,j} a_{t,i,j}}, \qquad \tilde{\pi}_{t,k} := \frac{1}{B_t} \sum_{i=1}^{B_t} q_{t,i,k}. \tag{33}$$

The update uses a batch-dependent smoothing coefficient to limit the influence of uncertain small batches. Let $u_{t,i} \in \Delta_K$ be the predictive probability vector used to assess confidence, $H(u) := -\sum_k u_k \log u_k$, and set

$$Q_t := \left[1 - \frac{1}{B_t \log K} \sum_{i=1}^{B_t} H(u_{t,i})\right]_+, \qquad \alpha_t := \alpha Q_t, \qquad \rho_t := (1-\alpha_t)\rho_{t-1} + \alpha_t \tilde{\pi}_t, \tag{34}$$

where $\alpha \in [0,1]$ and $[v]_+ = \max\{v, 0\}$. Because $0 \le H(u) \le \log K$, both Q_t and α_t lie in $[0,1]$. The confidence vectors may be the pre-update model predictions or the responsibilities in equation 33; the argument below only uses $\alpha_t \in [0,1]$. In Theorem B.1, the resulting online estimate is $\hat{\pi}_t = \rho_t$.

Lemma C.1 (Within-batch ascent)**.** *Suppose the scores $a_{t,i,k}$ remain fixed while equation 33–equation 34 are computed. Then*

$$\mathcal{L}_t(\rho_t) \ge \mathcal{L}_t(\rho_{t-1}). \tag{35}$$

Proof. For fixed responsibilities $q_{t,i,k}$ at ρ_{t-1}, Jensen's inequality gives the usual EM lower bound

$$\mathcal{L}_t(\rho) \ge \sum_{i,k} q_{t,i,k} \log\left(\frac{\rho_k a_{t,i,k}}{q_{t,i,k}}\right), \tag{36}$$

with equality at $\rho = \rho_{t-1}$. Maximizing the right-hand side over the simplex gives $\tilde{\pi}_t$ in equation 33, hence $\mathcal{L}_t(\tilde{\pi}_t) \ge \mathcal{L}_t(\rho_{t-1})$. Since $\mathcal{L}_t$ is concave in ρ and $\alpha_t \in [0,1]$, its value at the convex combination equation 34 is at least $\mathcal{L}_t(\rho_{t-1})$. □

This is an ascent result for the *current batch's fixed-score mixture objective*. It does not establish that the estimate equals the unknown π_t, that predictive accuracy increases, or that the online sequence converges when the target distribution and model scores evolve.

Prior-tracking error and the BGEB term. Define the prior estimation error $e_t := D_{\mathrm{TV}}(\rho_t, \pi_t)$, the change in the true prior $v_t := D_{\mathrm{TV}}(\pi_t, \pi_{t-1})$, and the batch-estimation error $\eta_t := D_{\mathrm{TV}}(\tilde{\pi}_t, \pi_t)$. The convex update equation 34 gives the pathwise recursion

$$e_t \le (1-\alpha_t) e_{t-1} + (1-\alpha_t) v_t + \alpha_t \eta_t. \tag{37}$$

Indeed, subtract π_t from both sides of equation 34, insert and subtract π_{t-1} in the first term, and apply the triangle inequality for total variation. The two terms involving v_t and η_t formalize the cost of lagging a changing prior and the cost of trusting an uncertain batch, respectively.

The latter error can be separated into systematic bias and finite-batch fluctuation. Assume the scoring function $a_t(x,k)$ and ρ_{t-1} are fixed conditional on $\mathcal{F}_{t-1}$, and define

$$\mu_t := \mathbb{E}_{X \sim P_t^X} q_t(X; \rho_{t-1}), \qquad q_t(k \mid x; \rho) := \frac{\rho_k a_t(x,k)}{\sum_j \rho_j a_t(x,j)}. \tag{38}$$

Conditional independence within $\mathcal{B}_t$ and a union of Hoeffding bounds across the K coordinates imply that, with conditional probability at least $1-\delta$, (Shalev-Shwartz & Ben-David, 2014)

$$\eta_t \le D_{\mathrm{TV}}(\mu_t, \pi_t) + \frac{K}{2}\sqrt{\frac{\log(2K/\delta)}{2B_t}}. \tag{39}$$

The first term is an estimation bias. It vanishes in the ideal label-shift case when the source posterior is calibrated, the class-conditional distributions are invariant, and $\rho_{t-1} = \pi_t$: then $q_t(k \mid X; \pi_t) = P_t(Y = k \mid X)$ and the law of total expectation gives $\mu_t = \pi_t$. More generally, even if $\rho_{t-1} = \pi_t$, a class-conditional shift can induce bias. Under the exact source-likelihood score model, the contraction of total variation through the source-mixture posterior gives (Gibbs & Su, 2002)

$$D_{\mathrm{TV}}(\mu_t, \pi_t) \le D_{\mathrm{TV}}\left(P_t^X, \sum_k \pi_t(k) P_s(\cdot \mid k)\right) \le \Delta_{\mathrm{like},t}(\pi_t). \tag{40}$$

If $\rho_{t-1} \neq \pi_t$, the bias in equation 39 also contains the effect of using an outdated prior. These inequalities make no claim that entropy is a calibrated measure of label correctness: overconfident incorrect predictions may yield a large Q_t.

The concentration statement assumes a per-sample scoring function fixed before the current batch, such as a frozen pre-update predictor. If scores for one sample depend on the other samples through current-batch normalization or a within-batch model update, equation 39 requires an additional dependence argument; the deterministic tracking recursion equation 37 still holds.

Substituting $\hat{\pi}_t = \rho_t$ into equation 20, the prior-mismatch contribution is exactly $2Me_t$. Equations equation 37–equation 39 explain the design of the entropy-controlled update: a larger α_t reacts more quickly to a changed prior but gives more weight to the uncertain batch estimate, while a smaller α_t dampens batch noise but can lag behind a shift. This mechanism addresses one component of BGEB; likelihood shift, pseudo-label errors, and parameter drift remain separate considerations.

D Experimental Setup

D.1 Benchmarks and medical vision–language models

We follow the multi-modality benchmark of Silva-Rodríguez et al. (Silva-Rodríguez et al., 2025b) for the dataset definitions, official data splits, and modality-specific pretrained vision–language models (VLMs). Importantly, we reuse the benchmark resources rather than its supervised few-shot protocol: the reference work adapts a classifier from labeled support examples, whereas our setting is source-free and online. During test-time adaptation, neither source samples nor target labels are available, and target labels are accessed only after prediction for evaluation. Our principal benchmark contains the eight tasks summarized in Table 3. Although the original benchmark also includes CheXpert, it is not included in the averages reported in our main experiments (Irvin et al., 2019).

Table 3: Datasets and modality-specific medical VLMs used in our experiments. "Imb." and "Bal." denote naturally imbalanced and balanced label distributions, respectively.

Modality	Dataset	Prediction task	Classes	Balance	Medical VLM / visual encoder	Input size
Ophthalmology	MESSIDORDecencière et al. (2014)	Diabetic-retinopathy grading	5	Imb.	FLAIR / ResNet-50	512^2
	MMACQian et al. (2024)	Myopic-maculopathy staging	5	Imb.	FLAIR / ResNet-50	512^2
	FIVESJin et al. (2022)	Retinal-disease classification	4	Bal.	FLAIR / ResNet-50	512^2
Histology	NCT-CRCKather et al. (2018)	Colorectal-tissue classification	9	Imb.	CONCH / customized ViT-B/16	448^2
	SICAPv2Silva-Rodríguez et al. (2020)	Prostate Gleason grading	4	Imb.	CONCH / customized ViT-B/16	448^2
	SkinCancerKriegsmann et al. (2022)	Skin-tissue classification	16	Imb.	CONCH / customized ViT-B/16	448^2
Chest X-ray	NIH-LTWang et al. (2017); Holste et al. (2022)	Thoracic-disease classification	19	Imb.	CONVIRT / ResNet-50	224^2
	COVID-19Chowdhury et al. (2020); Rahman et al. (2021)	Pneumonia-type classification	4	Bal.	CONVIRT / ResNet-50	224^2

For histology, we use CONCH with its customized large-scale ViT-B/16 visual encoder (Lu et al., 2024; Dosovitskiy et al., 2021). For ophthalmology, we use FLAIR; for chest X-ray, we use CONVIRT pretrained on MIMIC (Silva-Rodríguez et al., 2025a; Zhang et al., 2022; Johnson et al., 2019). FLAIR and CONVIRT employ ResNet-50 visual encoders and fine-tuned BioClinicalBERT text encoders (He et al., 2016; Alsentzer et al., 2019). Each method starts from exactly the same modality-specific checkpoint and uses the same class names, text prototypes, and image preprocessing. Images are resized to the native input resolution in Table 3; grayscale radiographs are replicated across three channels, and fundus images are cropped to the retinal field of view before resizing. CONCH inputs are normalized with the CLIP mean and standard deviation (Radford et al., 2021). Text prototypes are computed once from fixed modality-specific prompt templates and shared by all methods,

except when prompt adaptation is intrinsic to a baseline. The parameter subset updated at test time follows the original design of each method.

D.2 Source-free non-stationary adaptation protocol

Let h_0 denote the frozen source model. For every dataset and method, we initialize the adaptation process from the same h_0 and feed the target split as a temporally ordered stream of unlabeled mini-batches. Unless stated otherwise, the batch size is $B_t = 5$ and one adaptation step is performed for each incoming batch. Samples are processed once and are not revisited through an offline replay pass. The model is not reset between consecutive batches, so the updated state h_t is carried forward throughout the stream; it is reset to h_0 only when starting a new independent dataset, method, or random-seed run. Ground-truth labels do not participate in model selection, pseudo-label generation, prior estimation, or parameter updates. We retain the natural class frequencies of each target dataset and use the same sample order for all compared methods under a given seed.

This protocol differs from the labeled few-shot setting in the benchmark source in three respects. First, adaptation uses only the current unlabeled mini-batch rather than a labeled support set. Second, the online model state is accumulated across batches, which exposes each method to pseudo-label errors and parameter drift. Third, no target validation set is used for hyperparameter selection. We fix hyperparameters before evaluation and apply the same values to all datasets within a modality unless a method's original specification requires a dataset-independent default.

D.3 Implementation Details.

All experiments follow an online test-time adaptation protocol with a batch size of 5 to simulate streaming clinical inference, conducted on a single NVIDIA GeForce RTX 3090 GPU. Optimization is performed using AdamW with a learning rate of 5×10^{-4} (Loshchilov & Hutter, 2019). The textual prompt template was standardized as "a photo of a $\langle CLS \rangle$" (Radford et al., 2021). Regarding the SynerTTA method. For SynerTTA, the update ratio ρ is fixed to 0.5, while the class prior is initialized uniformly and the temperature coefficient τ is set to 1. Experimental performance is evaluated using Accuracy (Acc), F1-score (F1), and Brier Score (BS).

D.4 Construction of non-stationary clinical streams

We evaluate non-stationarity at both the sample and domain levels. In the dataset-level benchmark, the natural target split forms a small-batch stream, and additional stress tests vary the batch size, class proportions, and center composition. Class-imbalance streams alter the label marginal while retaining the corresponding class-conditional samples; mixed streams interleave examples from different clinical domains, so adjacent batches need not follow a single stationary distribution (Lipton et al., 2018; Rabanser et al., 2019). All stream manipulations are performed before adaptation and are shared across methods. Labels are used only to construct controlled stress conditions and to compute metrics, never by the adapting model.

For the stability–plasticity diagnosis, each dataset is treated as one task/domain and a single model state is propagated through the full modality-specific sequence:

$$\begin{aligned} \text{Ophthalmology: FIVES} &\rightarrow \text{MESSIDOR} \rightarrow \text{MMAC}, \\ \text{Histology: NCT-CRC} &\rightarrow \text{SICAPv2} \rightarrow \text{SkinCancer}, \\ \text{Chest X-ray: COVID-19} &\rightarrow \text{NIH-LT}. \end{aligned} \tag{41}$$

Because tasks within a modality can have different label sets, we construct the union label vocabulary for that modality and activate only the task-specific classes during prediction and evaluation. The adaptation object is instantiated once at the beginning of a sequence and is never reset at a task boundary.

Each task provides two disjoint streams with different roles. The *plasticity stream* contains 200 images sampled from the official test split and is used for online adaptation. The *stability set* contains 300 images sampled from the corresponding training split and is used only for read-only evaluation after each task. Sampling is reproducible for a fixed seed and is performed without replacement whenever the source split is sufficiently large; if a split contains fewer samples than required, shuf-

fled passes through the available examples are repeated until the specified size is reached. This fixed-size construction prevents dataset size from confounding the task-level stability–plasticity comparison. No gradient, optimizer-state update, or prior update is permitted while evaluating a stability set.

D.5 Baselines and implementation details

We compare SynerTTA with the frozen source model and the representative TTA baselines reported in the main paper. All methods use the same pretrained checkpoint, class vocabulary, prompt prototypes, batch partition, and sample order. Baseline-specific components, such as parameter restoration, memory banks, sharpness-aware updates, or prompt optimization, follow their official implementations (Wang et al., 2022a; Yuan et al., 2023; Gong et al., 2022; Niu et al., 2023; Foret et al., 2021; Shu et al., 2022). Unless an operation is intrinsic to a method, we do not provide source images, target labels, or future target batches. We perform one online update per batch and preserve each method's optimizer state across the stream. Automatic mixed precision is enabled for efficient inference and adaptation.

For SynerTTA, we use $B_t = 5$ and a weight decay of 0.01. The learning rates are 5×10^{-4} for ophthalmology and chest X-ray and 1×10^{-4} for histology. The prior temperature, consistency coefficient, prior-rectification coefficient, and update ratio are fixed at $(1.0, 0.5, 0.3, 0.6)$ for ophthalmology, $(1.5, 1.0, 0.5, 0.9)$ for histology, and $(1.0, 0.3, 0.3, 0.3)$ for chest X-ray, respectively. These values are held fixed across all tasks of the same modality. For the adaptation-strength analysis, we vary only the designated strength parameter while leaving every other setting unchanged: the learning rate for TENT and CoTTA, and the update-frequency parameter for RoTTA (Wang et al., 2021; 2022a; Yuan et al., 2023).

All experiments are implemented in PyTorch and run on `[a single NVIDIA GeForce RTX 3090 GPU]`. The final version will release the exact configuration files, sampled stream indices, and random seeds used to produce every table and figure.

D.6 Evaluation metrics

We report Accuracy and Macro-F1 for every task. Macro-F1 assigns equal weight to every class and is therefore the primary complement to Accuracy under changing disease prevalence (Maier-Hein et al., 2024). Dataset-level averages are unweighted means over the eight tasks in Table 3. For stress tests, we additionally report the worst performance and performance range across batch sizes, the mean and minimum performance across class-imbalance settings, and the change relative to the frozen source model under mixed-domain streams.

For the stability–plasticity analysis, let $a_{i,j}$ be the score on the stability set of task j after the model has adapted through task i. The score is computed separately with Accuracy and Macro-F1 (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018). At step i, Learning Accuracy (LA) and Forgetting Measure (FM) are

$$\mathrm{LA}_i = \frac{1}{i}\sum_{j=1}^{i} a_{j,j}, \qquad \mathrm{FM}_i = \frac{1}{i-1}\sum_{j=1}^{i-1}\left(\max_{1\le \ell < i} a_{\ell,j} - a_{i,j}\right). \tag{42}$$

Higher LA indicates stronger plasticity on newly encountered domains, whereas lower FM indicates stronger stability on previously encountered domains. FM can be negative when later adaptation improves an earlier task. In trade-off plots, we use $-\mathrm{FM}$ so that larger values consistently indicate better performance on both axes.

E Hyperparameter Robustness: A Plug-and-Play Guarantee

Hyperparameter Sensitivity Analysis. To evaluate the robustness and optimal configuration of SynerTTA, we conduct a systematic sensitivity analysis on four critical hyperparameters: the update ratio ρ in GSPM(GSPM), the prior strength τ in RWLA(RWLA), and the dual parameters (λ, γ) governing the Consistencyadaptive loss. As illustrated in Figure 3, the update ratio ρ, which determines the proportion of parameters updated during test-time alignment, follows a characteristic

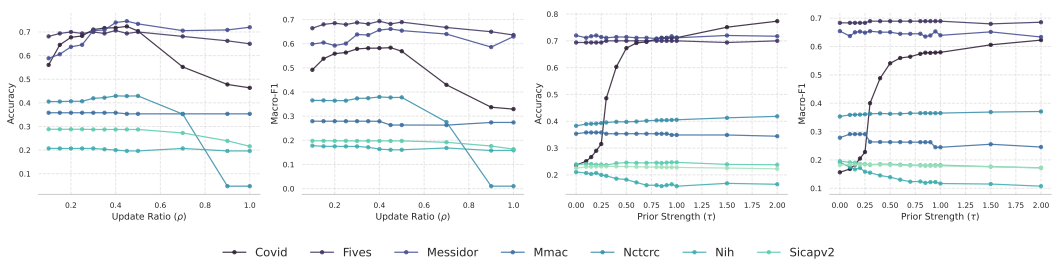


Figure 3: **Sensitivity analysis of hyperparameters ρ and τ across eight clinical datasets.** Left two panels: Performance variation with respect to the update ratio ρ in the GSPM (GSPM) module. A characteristic bell-shaped trend is observed, where moderate updates ($\rho \approx 0.4$) yield optimal adaptation while extreme values lead to knowledge erasure in challenging tasks like NCT-CRC. Right two panels: Impact of the prior strength τ in the RWLA (RWLA) module. The significant performance surge in the Covid dataset underscores the necessity of Bayesian prior reinforcement for correcting predictive biases under severe distribution shifts.

"bell-shaped" performance curve across most modalities. For instance, in the Covid and Messidor datasets, Accuracy peaks at $\rho = 0.45$ (reaching 0.7236 and 0.7457, respectively) before declining. Notably, in the challenging NCT-CRC benchmark, increasing ρ beyond 0.5 causes a catastrophic performance collapse (Acc dropping from 0.4294 to 0.0474), which suggests that excessive updates in highly non-stationary streams can lead to the erasure of foundational knowledge. Conversely, the prior strength τ exhibits a monotonic or stabilizing trend. Its impact is most decisive in the Covid dataset, where Accuracy surges from 0.2364 at $\tau = 0$ to 0.7733 at $\tau = 2.0$, demonstrating that reinforcing the Bayesian prior through RWLA is essential for correcting predictive biases under extreme distribution shifts. For more stable datasets like FIVES and MMAC, the model maintains a high performance plateau across $\tau \in [0.3, 2.0]$, underscoring the inherent stability of the reliability-weighting mechanism.

Joint Sensitivity Analysis. Furthermore, we investigate the interaction between the consistency weight λ and the proportion weight γ via a grid search on the Messidor dataset, as visualized in the heatmaps in Figure 4. The results reveal a remarkably stable "flat optima" landscape; the Accuracy fluctuates within a negligible range of 0.6% (0.723 to 0.729), and Macro-F1 remains consistently around 0.650. This insensitivity indicates that SynerTTAis not overly dependent on precise weighting of the adaptive losses to achieve competitive results. The optimal joint configuration is identified at $\lambda = 0.3$ and $\gamma = 0.1$ (Acc: 0.729, F1: 0.651), though other combinations such as $(\lambda = 1.0, \gamma = 0.7)$ yield identical peak performance. Based on these comprehensive observations, we recommend $\rho = 0.4$, $\tau = 1.0$, $\lambda = 0.3$, and $\gamma = 0.7$ as a robust default parameter set. This configuration ensures an optimal trade-off between flexible adaptation and the preservation of reliable clinical priors across diverse medical imaging modalities.

F Additional Experimental Results

This section provides additional evidence for the stability–plasticity dilemma and evaluates SynerTTA under complementary non-stationary stress conditions. Unless stated otherwise, all methods follow the source-free online protocol described in Appendix D, use the same incoming batches, and are evaluated without resetting the model within a stream.

F.1 Additional stability–plasticity trajectories

Figure 5 extends the task-wise diagnosis in the main paper from ophthalmology to histology and chest X-ray. To display plasticity and stability in one trajectory, we start from $C_0 = 0$ and alternately add LA and the sign-reversed FM after each task:

$$C_{2i-1} = C_{2i-2} + \mathrm{LA}_i, \qquad C_{2i} = C_{2i-1} - 5\,\mathrm{FM}_i. \tag{43}$$

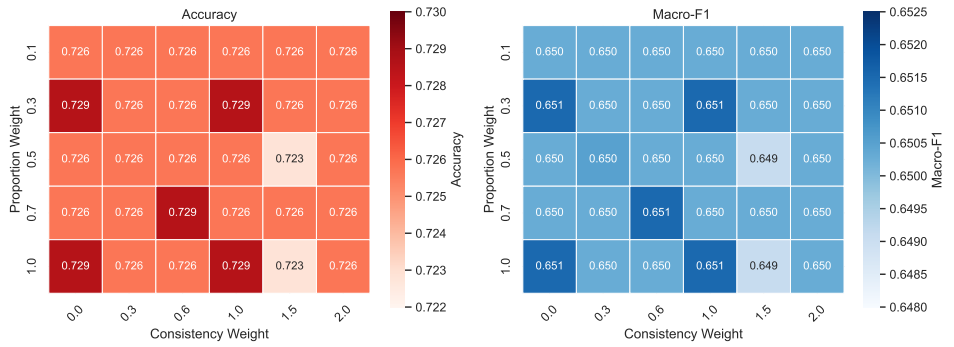


Figure 4: **Joint sensitivity analysis of consistency weight λ and proportion weight γ on the Messidor dataset.** The heatmaps illustrate the Accuracy (left) and Macro-F1 (right) scores under various combinations of the adaptive loss parameters in the Consistency module. The remarkably consistent performance across the grid indicates a "flat optima" landscape, suggesting that SynerTTA is highly robust to the weighting of its adaptive refinement mechanisms and does not require extensive per-dataset tuning.

The factor of five is used only to make forgetting visually discernible on the same scale as LA. Thus, these cumulative curves are a diagnostic visualization rather than a new evaluation metric; the underlying LA and FM retain the definitions in Eq. equation 42.

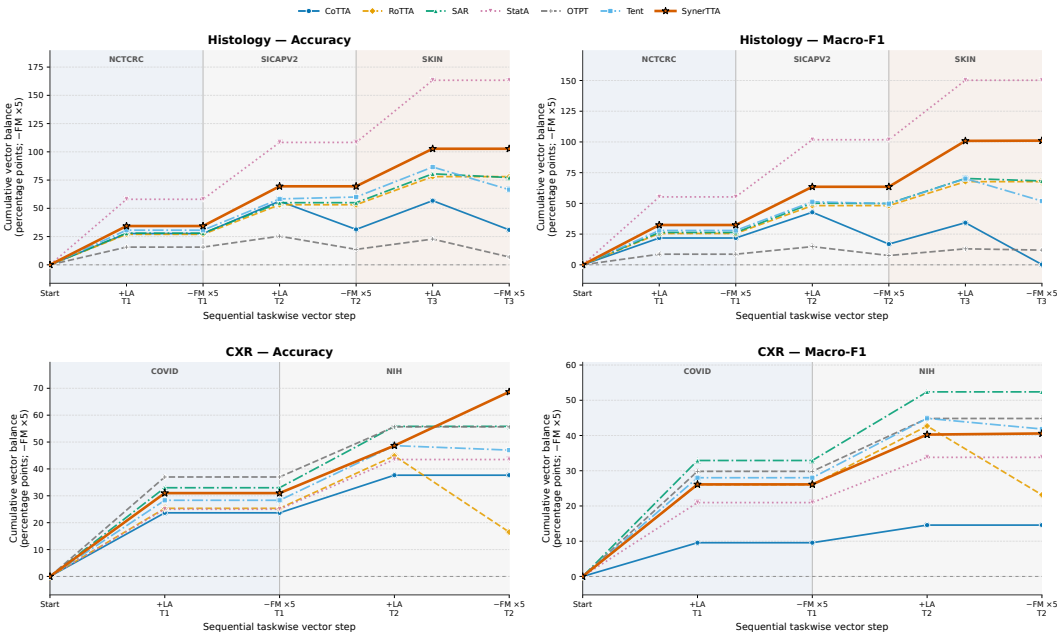


Figure 5: **Task-wise stability–plasticity trajectories on histology and chest X-ray.** Each task contributes a positive LA step followed by a sign-reversed forgetting step, $-\mathrm{FM} \times 5$. Results are shown for both Accuracy and Macro-F1. Upward LA segments indicate adaptation to the newly arrived domain, whereas downward $-\mathrm{FM}$ segments expose the loss of previously acquired performance.

The trajectories reveal that the relative behavior of existing methods changes substantially across modalities and metrics. For example, CoTTA accumulates pronounced forgetting in histology Macro-F1, whereas methods that are stable in one modality need not provide the strongest learning response in another. SynerTTA maintains consistently positive learning contributions while limiting subsequent reductions, yielding a strong final balance across all four panels. On chest X-ray, it obtains the largest final Accuracy balance; for Macro-F1, SAR is stronger, while SynerTTA

remains competitive without the severe degradation observed for several alternatives. These results support the central diagnosis: satisfactory non-stationary adaptation cannot be characterized by current-domain learning or retention alone.

F.2 Joint Accuracy–Macro-F1 comparison

Accuracy can be dominated by prevalent classes, while Macro-F1 gives equal weight to each class (Maier-Hein et al., 2024). We therefore plot both metrics jointly in Figure 6. A desirable method should move toward the upper-right region rather than improving Accuracy by sacrificing minority-class performance, or vice versa.

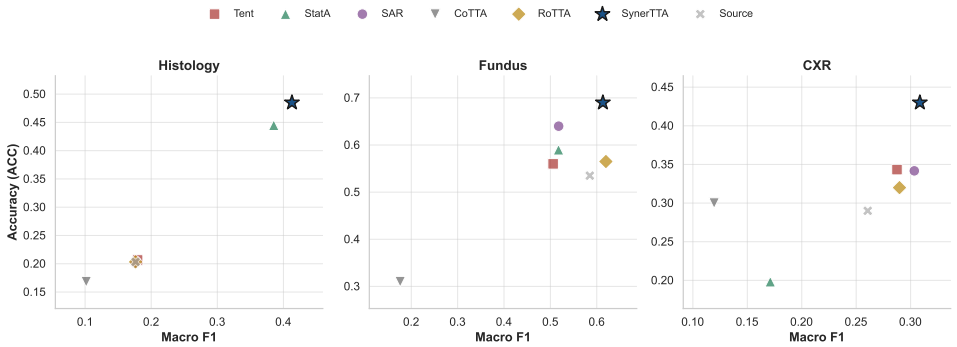


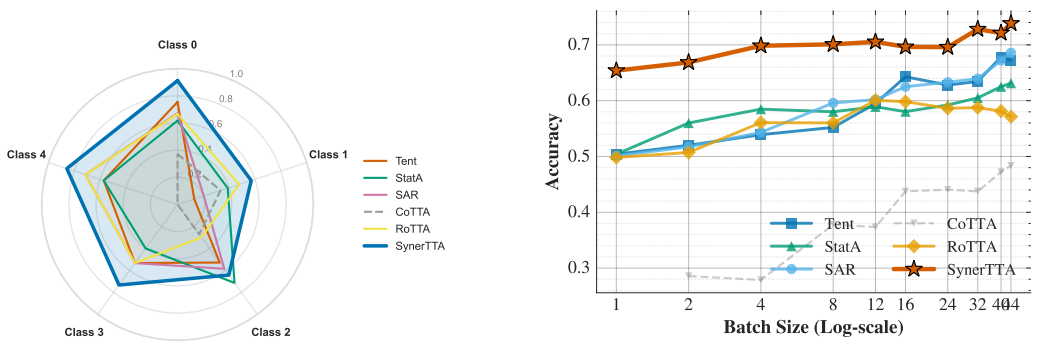
Figure 6: **Joint Accuracy–Macro-F1 comparison across modalities.** Each marker denotes one method under the same modality-specific stress-test protocol; points closer to the upper-right corner provide a better joint operating point. The frozen source model is included as a reference.

SynerTTA lies on or near the upper-right frontier in histology, fundus imaging, and chest X-ray, and improves both metrics relative to the corresponding frozen source model. In contrast, several existing methods exhibit a marked discrepancy between Accuracy and Macro-F1, indicating that an apparently favorable aggregate accuracy can conceal weak class-balanced performance. The result complements the LA/FM analysis by showing that SynerTTA's stability–plasticity balance does not arise from favoring only the majority classes.

F.3 Robustness to class imbalance and small batches

Figure 7 examines two deployment factors that make online evidence unreliable. Figure 7a reports class-wise recall on the five-class MESSIDOR imbalance stress test. Classes 0–4 denote increasing disease severity: no DR, mild, moderate, severe, and proliferative DR (Decencière et al., 2014). SynerTTA achieves the most uniform profile, with a mean recall of 74.32% and a worst-class recall of 57.07%. The strongest baseline reaches 54.17% in mean recall and 38.89% in worst-class recall, demonstrating that SynerTTA's gain is not driven solely by common disease grades.

Figure 7b varies the MESSIDOR test-time batch size from 1 to 44. SynerTTA maintains the highest Accuracy throughout the sweep, with a worst-case Accuracy of 65.39% and a range of only 8.46 percentage points. By comparison, the best baseline worst-case Accuracy is 50.40%, and the smallest baseline range is 10.23 points. The advantage is especially pronounced for extremely small batches, where noisy empirical statistics and pseudo-label errors most strongly affect conventional TTA.



(a) Class-wise recall under label imbalance.

(b) Accuracy across test-time batch sizes.

Figure 7: **Robustness under unreliable online evidence.** (a) SynerTTA preserves a substantially stronger worst-class recall under class imbalance. (b) SynerTTA remains accurate and comparatively insensitive as the amount of evidence in each incoming batch changes.

Together, these experiments show that SynerTTA's gains persist across modalities, evaluation metrics, class-frequency shifts, and batch sizes. They also agree with the BGEB interpretation: prior rectification addresses biased class evidence, while restricted and reliability-aware updates reduce the variance and drift induced by small online batches.