Tempora: Characterising the Time-Contingent Utility of Online Test-Time Adaptation

Sudarshan Sreeram[1] Young D. Kwon[2] Cecilia Mascolo[1]

A machine learning (ML) model deployed on a mobile or edge device often encounters data dissimilar to what it saw during training. For example, a species classifier on a wildlife camera trap sees site-specific flora, fauna, lighting, and weather; an on-device photo classifier sees images shaped by a particular user's camera and habits. These out-of-distribution datapoints can degrade model quality: a ResNet-50 model that achieves 76% accuracy on ImageNet drops to 18% under common corruptions. Retraining a model over time for every deployment is impractical, given the energy, compute, data, labelling, and privacy constraints. It is thus desirable that a model efficiently adjusts itself in-situ without supervision.

Test-time adaptation (TTA) is a compelling remedy that improves model generalisation on-the-fly using only unlabelled inputs, with no access to source data or remote fine-tuning. This flexibility suits a range of real deployments, but there is a cost: adaptation adds latency overhead on the critical path of every prediction. However, conventional evaluations of TTA methods unrealistically assume unbounded overhead and overlook the accuracy-latency trade-off. As ML underpins more latency-sensitive and user-facing use cases, temporal pressure constrains the viability of adaptive inference; a correct prediction arriving after a user has closed an app is futile. Correctness and timeliness are jointly necessary.

Prior attempts at quantifying the trade-off only focus on the net effect of providing slow TTA methods with fewer samples. This evaluation addresses a *discrete* temporal scenario where missing a deadline yields no value, but real deployments span more scenarios. To cover them, we introduce *Tempora*, a framework for evaluating TTA under time constraints. It comprises temporal scenarios (contextualising deployments), evaluation protocols (operationalising measurement), and time-contingent utility metrics (quantifying the accuracy-latency trade-off). We instantiate this framework with three metrics for distinct scenarios: ❶ **discrete** utility for hard deadlines, ❷ **continuous** utility for interactive settings (e.g., a photo classifier that pauses too long to adapt may frustrate users), and ❸ **amortised** utility for budget-limited settings (e.g., a drone performing a real-time crop survey may lose flight time by overspending compute to adapt).

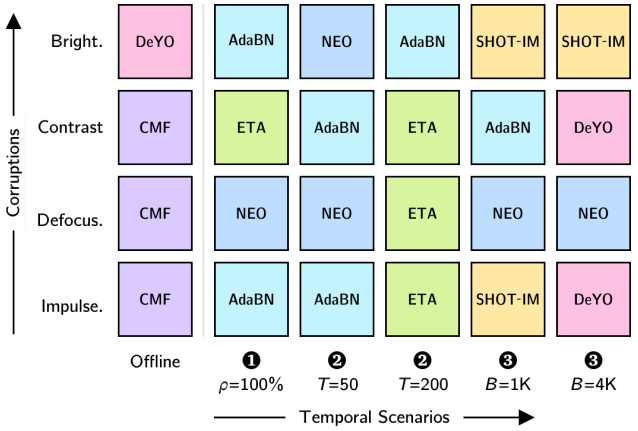

Figure 1. Rank instability persists across evaluations. *Offline*: No time constraints. Cells show the highest utility method across scenarios and corruptions. No method dominates under temporal pressure.

By applying Tempora to nine TTA methods on ImageNet-C (ResNet-50), we find that rank instability is more persistent than previously reported (Fig. 1). **CMF, a state-of-the-art method in the offline setting, no longer ranks first in 211/240 temporal evaluations and underperforms standard inference in 79 cases.** Our utility metrics decompose to reveal three failure modes: availability collapse, accuracy erosion, and harmful overhead. Deployable adaptation requires three properties absent from current methods: compute allocation conditioned on the difficulty of the input shift, time-aware scaling that bounds response delay, and anytime performance so elapsed overhead yields gains over not adapting at all.

Tempora reveals when and why rankings change; it offers practitioners a lens to identify which methods fit their temporal constraints and researchers a target for designing ones that earn their place on a device. **Genuinely deployable methods justify their overhead through accuracy gains; computationally-insolvent ones consume budgets no accuracy benefit can repay.** As TTA methods grow increasingly sophisticated, temporal evaluation becomes essential; without it, the field risks chasing offline accuracy while creating methods that cannot serve real-world systems where adaptation is most needed.

Tempora was accepted at the 43rd International Conference on Machine Learning (ICML). Our pre-print can be found at https://arxiv.org/abs/2602.06136.

[1]Department of Computer Science and Technology, University of Cambridge, UK. [2]Samsung AI Center, Cambridge, UK.
Correspondence to: Sudarshan Sreeram (ss3122@cam.ac.uk).