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

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

[1]Department of Computer Science and Technology, University of Cambridge, UK [2]Samsung AI Center, Cambridge, UK

Overview

- **Gap**. Existing evaluations crucially ignore real-world temporal pressure.
- **Framework**. Tempora tests TTA methods under diverse constraints.
- **Finding**. Rankings collapse more broadly than previously observed.
- **Impact**. Failure modes and practical guidance for method design.

Correctness and *timeliness* are jointly necessary for realistic evaluation!

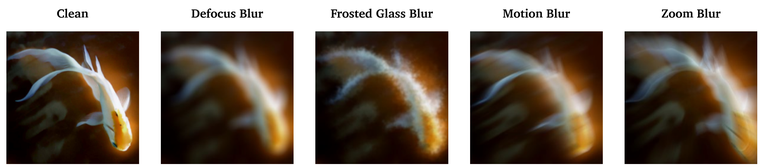

Figure 1. Domain shift induced by various blur corruptions on an ImageNet sample ("goldfish").

Tempora

We define *utility* metrics based on three archetypes: ❶ discrete (environment-led), ❷ continuous (user-led), and ❸ amortised (resource-led).

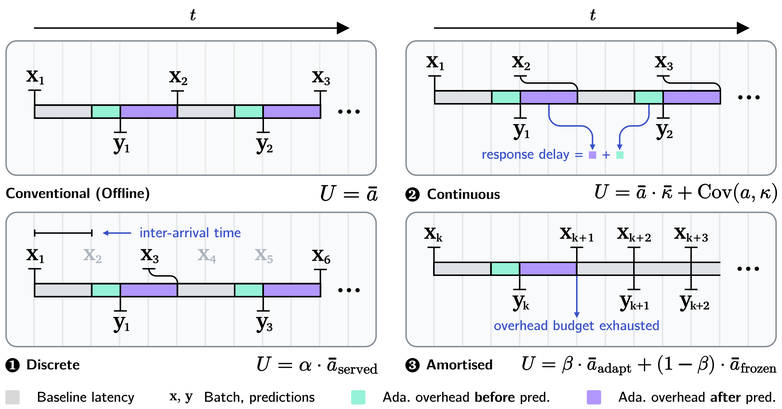

Figure 2. Evaluation protocols for measuring time-contingent utility.

- **Offline.** Stream waits for adaptation to complete to release next batch.
- **Discrete.** Asynchronous arrivals; batches skipped when pipeline is busy.
- **Continuous.** Stream releases $\mathbf{x}_{i+1}$ upon receiving $\mathbf{y}_i$ (greedy user).
- **Amortised.** Follows offline till budget exhaustion, then frozen inference.

Each metric couples accuracy with a scenario-appropriate penalty.

Evaluation Settings

- **Datasets.** ImageNet-C, ImageNet-V2, ImageNet-R, CIFAR-10-C.
- **Models.** ResNet-50, ViT-B/16, ResNet-18.
- **Hardware.** Nvidia RTX 4080 GPU, Raspberry Pi 5 (16 GB).
- **Methods.** 11 Fully TTA methods. *Gradient-free:* AdaBN, LAME, NEO. *Gradient-based:* Tent, ETA, SHOT-IM, SAR, CMF, DeYO, SPA, ZeroSIAM.
- **Scenarios.** 750+ temporal evaluations spanning all protocols.

Evaluation Results

Table 1. Rank instability; cells show the highest-utility method under temporal pressure.

(a) ResNet-50, ImageNet-C; 240 evaluations (160 presented).

	Offline	Discrete (ρ %)					Continuous (T ms)					Amortised (B s)					
Corruption		100	70	50	35	25	50	100	200	400	1k	1	2	4	8	16	32
Gauss. noise	C	A	E	E	E	C	A	E	E	E	E	S	S	D	E	C	C
Impul. noise	C	A	E	E	E	C	A	E	E	E	E	S	S	D	E	E	C
Defoc. blur	C	N	N	E	E	C	N	N	E	E	E	N	N	N	C	C	C
Glass blur	C	A	E	E	E	C	A	E	E	E	E	A	S	D	C	C	C
Zoom blur	C	A	A	E	E	C	A	A	E	E	E	A	S	S	E	E	C
Snow	C	A	A	E	E	D	A	A	E	E	E	S	S	D	E	E	C
Fog	D	A	A	E	E	D	A	A	A	E	E	A	S	S	E	E	C
Brightness	D	A	A	A	E	D	N	A	A	A	A	S	S	S	C	E	C
Contrast	C	E	E	E	E	C	A	E	E	E	E	A	A	D	E	E	C
JPEG comp.	C	A	A	E	E	D	N	A	E	E	E	S	S	D	E	C	C

(b) ViT-B/16, ImageNet-C; 255 evaluations (150 presented).

	Offline	Discrete (ρ %)					Continuous (T ms)				Amortised (B s)					
Corruption		100	70	50	35	25	200	400	1k	2k	2.5	5	10	20	40	80
Gauss. noise	P	N	N	N	E	C	N	N	N	E	P	E	E	P	P	P
Impul. noise	P	N	N	E	E	C	N	N	N	E	P	P	P	P	P	P
Defoc. blur	C	N	N	E	E	C	N	N	E	E	Z	E	E	E	E	C
Glass blur	P	N	E	E	E	C	N	E	E	E	S	E	E	E	E	P
Zoom blur	P	N	N	E	E	E	N	N	E	E	S	S	E	E	E	P
Snow	P	N	N	E	E	C	N	N	N	E	N	E	P	P	P	P
Fog	P	N	N	N	E	C	N	N	N	N	N	N	N	N	P	P
Brightness	P	N	N	N	E	C	N	N	N	N	S	E	E	P	P	P
Contrast	C	N	N	T	T	C	N	N	T	T	N	E	E	E	C	C
JPEG comp.	P	N	N	N	E	C	N	N	N	E	S	S	S	P	P	P

ETA (103/67) NEO (9/81) AdaBN (55/-) SPA (-/56) CMF (39/17)
SHOT-IM (26/14) DeYO (23/0) Tent (0/4) ZeroSIAM (-/1)

- Gradient-based methods 2.2–5.6× slower than gradient-free methods.

No single method dominates across corruptions and temporal scenarios.

Method	Discrete ($\rho = 100\%$) α	$\bar{a}_{\mathrm{s}}$	U_{d}	Continuous ($T = 50$ ms) $\bar{a}$	$\bar{\kappa}$	U_{c}	Amortised ($B = 1$ s) β	$\bar{a}_{\mathrm{a}}$	$\bar{a}_{\mathrm{f}}$	U_{a}
Standard	100.0	18.2	18.2	18.16	100.0	18.16	100.0	18.16	-	18.16
NEO	100.0	22.1	22.1	22.14	100.0	22.14	100.0	22.14	-	22.14
LAME	98.8	17.5	17.3	17.40	95.7	16.96	100.0	17.40	-	17.40
AdaBN	97.2	31.7	**30.8**	31.72	89.6	**28.42**	100.0	31.72	-	31.72
Tent	41.2	40.4	16.6	42.88	15.1	6.46	2.3	32.11	0.10	0.84
ETA	41.0	45.6	18.7	48.35	15.0	7.22	2.3	33.37	0.10	0.87
SHOT-IM	33.2	40.6	13.5	42.43	11.2	4.73	1.7	32.23	32.22	**32.22**
DeYO	31.8	45.0	14.3	48.76	10.2	4.92	6.7	8.66	0.10	0.67
CMF	25.1	**46.0**	11.5	**49.13**	7.9	3.84	1.7	21.91	0.10	0.48
SAR	20.6	37.6	7.8	44.14	6.2	2.73	0.9	33.54	0.10	0.40

Table 2. Utility decomposition under temporal pressure (RN-50/IN-C).

- **Rank instability.** Offline winner CMF loses in 211/240 evaluations.
- **Computational insolvency**. No accuracy gain can compensate for penalty.
- **Accuracy-latency trade-offs.** Offline rankings recover as pressure eases.

Tempora reveals *when* and *why* rankings change.

Deployable Adaptation

Decompositions show failure modes from **overhead not justified by gains**. Deployable adaptation requires:

- **Corruption-conditioned allocation**
  Allocate compute proportional to difficulty of correcting shift.
- **Time-aware scaling**
  Bound response delay to finish adapting before gains are erased.
- **Anytime performance**
  Always yield gains over standard inference despite when adaptation halts.

Tempora guides practitioners in *method selection* and directs researchers towards deployable *method design*.

Acknowledgements

Our work was supported by Nokia Bell Labs through a donation and by EPSRC through grant EP/Z53447X/1.

Engineering and Physical Sciences Research Council

arXiv GitHub