PERSISTENT-STATE MANAGEMENT FOR STREAMING TEST-TIME ADAPTATION IN OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming test-time adaptation (TTA) changes the model while it is being evaluated, so deployment quality depends on the adaptation state carried from earlier samples. We study this persistent state as a first-class object in open-vocabulary semantic segmentation and introduce Profiled State Recovery, a practical framework for managing it. Profiled State Recovery uses a shared monitor–trigger–recovery–resume lifecycle around the native TTA update, while specializing recovery semantics to each adaptation mechanism. A profile is calibrated once on a labeled development stream, frozen, and then deployed label-free without target-stream retuning. Across MLMP, SAR, and TENT, three datasets, two ViT scales, and a ConvNeXt backbone, frozen profiles improve four of five systems, and all four improve on both fresh validation seeds, with mean gains of $+4.204$, $+0.316$, $+0.491$, and $+2.881$ mIoU points for S1, S2, S3, and S5. All three tested cross-dataset transfers are positive without target-stream retuning: $+0.448$ (MLMP), $+0.602$ (SAR), and $+0.739$ (TENT). The same frozen MLMP profile improves both ViT-L/14 and ViT-B/16. Mechanistic analysis shows that recovery granularity is a central design factor: local/progressive, selective late-state, and global recovery regimes emerge for MLMP, SAR, and TENT, respectively. These results establish persistent adaptation-state management as a structured design axis for streaming TTA and provide a compact, transferable implementation framework.

1 INTRODUCTION

Test-time adaptation changes a model while it is being evaluated, but many TTA evaluations treat each sample or short segment as if adaptation state disappears afterward. In a real deployment stream, an update made for x_t changes the state used for $x_{t+1}, x_{t+2}, \ldots$. The model, optimizer, and method history therefore become part of the input to future predictions (Figure 1). This persistent-state view is especially consequential in open-vocabulary semantic segmentation, where long deployment streams allow adaptation memory to accumulate and affect future predictions.

Existing TTA methods provide strong mechanisms for entropy minimization, continual learning, stabilization, and restoration (Wang et al., 2021; 2022; Niu et al., 2023; Press et al., 2023; Lim et al., 2026); the next design object is the state-management contract that governs their long-horizon behavior. This yields a clean decomposition: a native TTA update rule plus an explicit persistent-state management policy. A related study analyzed persistent-state diagnosis and Freeze-to-Confirm recovery in an MLMP-centered setting (Wang, 2026). The present paper asks the broader systems question: can persistent state be profiled into a compact recovery policy for each adaptation mechanism, and can that policy transfer to new streams without target-stream retuning?

We introduce Profiled State Recovery, a shared lifecycle for persistent-state management: monitor, trigger, method-conditioned recovery, and resume around the native TTA update (Figure 2). The lifecycle is common, while the recovery profile is allowed to follow the internal state structure of the adaptation mechanism. Offline profiling uses one labeled calibration stream to select a compact profile specifying trigger persistence, cooldown, recovery granularity, severity mapping, optimizer semantics, and method-history semantics. Once frozen, the profile makes label-free online decisions

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.

on new streams. The resulting profiles are local/progressive for MLMP, selective late-state for SAR, and global for TENT.

The resulting profiles produce broad gains across methods, streams, and backbone scales. The frozen profiles improve four of five primary systems, and all four improve on both fresh validation seeds: S1 MLMP ViT-L/14 on YTVIS by $+4.204$, S2 SAR ViT-L/14 on COCO by $+0.316$, S3 TENT ConvNeXt-B on ACDC by $+0.491$, and S5 MLMP ViT-B/16 on YTVIS by $+2.881$ mIoU points. All three source-to-target profile transfers are positive without target-stream retuning, and the same MLMP profile transfers across the tested ViT-L/14 and ViT-B/16 scales. The accompanying anatomy and headroom analyses explain why these gains are structured: the best recovery actions are often multi-group, the effective granularity follows the adaptation mechanism, and recoverable opportunity varies by system.

Contributions.

- We identify persistent adaptation state as a first-class design object in streaming TTA and show that it materially changes long-horizon behavior across adaptation mechanisms and streams.
- We introduce Profiled State Recovery, a shared state-management lifecycle that profiles recovery semantics once per adaptation mechanism and freezes them for label-free deployment.
- We provide broad empirical evidence across MLMP, SAR, and TENT: four of five primary systems improve, and all four improve on both fresh validation seeds; all three tested cross-dataset transfers are positive without target-stream retuning, and the MLMP profile transfers across ViT scales.
- We reveal structured recoverability: mixed useful/harmful components, multi-group hindsight actions, and method-specific local/progressive, selective, and global recovery regimes define a practical design space for streaming TTA.

2 Stateful Streaming TTA and Related Work

Recent persistent-TTA work studies long-horizon adaptation and collapse monitoring directly, while risk-monitoring work develops label-free alerts for changing predictive systems (Hoang et al., 2024; Schirmer et al., 2025). Profiled State Recovery complements these directions by making the recovery action itself a frozen, method-conditioned deployment object. Let a stream be $x_{1:T}$ and let $\mathcal{S}_t = (\theta_t, o_t, h_t)$ denote the mutable adaptation state after processing x_t.

$$\boldsymbol{\mathcal{S}_t = F(\mathcal{S}_{t-1}, x_t)}, \qquad \boldsymbol{\hat{y}_t = G(\mathcal{S}_t, x_t)}, \tag{1}$$

where θ_t contains adaptable parameters, o_t optimizer state, and h_t method-specific history. Persistent evaluation retains $\mathcal{S}_t$ between samples. Reset-style evaluation instead replaces it by an initial state after each evaluation segment. This is not a cosmetic protocol choice: the two protocols induce different trajectories and can change method rankings.

TTA has developed from entropy-based adaptation such as TENT (Wang et al., 2021) to continual and stabilized variants such as CoTTA and SAR (Wang et al., 2022; Niu et al., 2023), together with reset-oriented diagnostics such as RDumb and WHEN/WHERE recovery studies (Press et al., 2023; Lim et al., 2026). This literature establishes the importance of update rules, memory, stabilization, and reset policy. Profiled State Recovery complements it by treating the resulting mutable state as a managed deployment object: the adaptation method remains intact, while a small frozen profile determines when a recovery action is applied and how much state is restored.

We evaluate on open-vocabulary semantic segmentation, using vision-language representations such as CLIP (Radford et al., 2021) and MLMP-style adaptation (Noori et al., 2025); NACLiP supplies one of the segmentation backbones (Hajimiri et al., 2025). Related continual open-vocabulary work, including CARE (Luo et al., 2026), further motivates evaluating long-lived state rather than only isolated samples. Wang's Freeze-to-Confirm study (Wang, 2026) focuses on MLMP-centered recovery

timing and the preservation–protection tradeoff; here we study cross-method state anatomy, method-conditioned recovery profiles, and transfer across streams and backbone scales. This broader framing makes state management comparable across adaptation mechanisms without collapsing their recovery semantics into one rule.

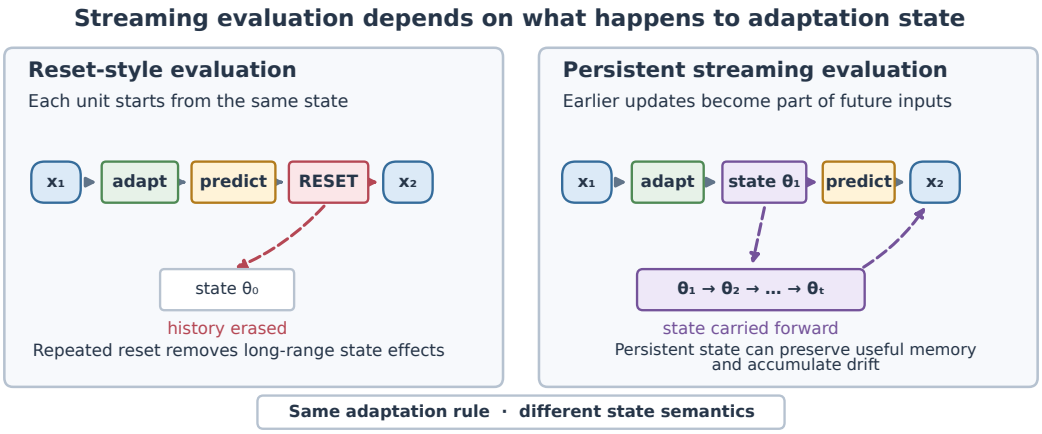


Figure 1: Reset-style and persistent streaming evaluation expose different state dynamics. Resetting erases history between examples; persistent streaming carries θ_t forward, so earlier adaptation affects later predictions.

3 Persistent-State Anatomy

Throughout this analysis, S1–S5 denote the five primary systems specified in Section 5, and H32 denotes a 32-sample causal hindsight replay horizon. We treat state as an object to measure rather than an implementation detail. Our analysis spans persistence, active restoration, and recovery-action structure. Persistent adaptable parameter state is the dominant measured carrier of recovery gain, while useful and harmful components coexist within the same stream: across the 281-event structural-component factorial population, 203 events (72.2%) contain both beneficial and harmful structural components. Persistent state is therefore compositional, making targeted restoration a natural complement to native adaptation.

The recovery action space partitions adaptable state into ordered normalized structural groups G1–G4, from early to late adaptable structure. The partition is semantic: for TENT it covers adaptable normalization state, while MLMP and SAR map their method-defined adaptable state into the same normalized structural groups. Across the fixed 16-action space—`NO_OP` plus all 15 nonempty subsets of G_1–G_4—only 34.9% of events select a singleton optimum; 46.3% select a multi-group optimum, while 18.9% favor `NO_OP`. Restricting recovery to at most two groups remains within 0.1 utility points of the unrestricted empirical action optimum in 92.2% of events, and allowing at most three groups reaches 99.3%, establishing a compact multi-group recovery structure.

Freeze-only prevention and active recovery are distinct interventions. In the separate 10-event S3 active-restoration versus prevention-only comparison, nine events are H32-recoverable; among these nine, active restoration outperforms prevention-only freezing in all nine, with a mean advantage of +1.322 mIoU. Recovery can therefore repair already-encoded state while preserving the native adaptation process.

The factorial analysis independently supports this interpretation. Here V_{ij} is the H32 future-average foreground utility, with i indicating parameter restoration and j optimizer-state reset: V_{00} is Continue, V_{10} parameter-only restoration, V_{01} optimizer-only reset, and V_{11} both. Let $A_\theta = V_{10} - V_{00}$ denote the adaptable-parameter intervention effect and $G = V_{11} - V_{00}$ the total measured state-management gain. Their H32 ranks correlate at Spearman $\rho = 0.979$, 0.990, and 0.976 for S1, S2, and S3, respectively, linking the measured recovery effect to the adaptable parameter state.

4 PROFILED STATE RECOVERY

4.1 FORMAL STATE-MANAGEMENT OBJECTIVE

At deployment step t, the adaptation mechanism m receives x_t and a persistent state $\mathcal{S}_{t-1}$. A recovery profile π_m augments the native transition with an optional state action:

$$\boldsymbol{\mathcal{S}_t = \mathcal{U}_m(x_t, \mathcal{S}_{t-1}; \pi_m)}, \qquad \boldsymbol{\hat{y}_t = f_m(x_t; \mathcal{S}_t)}, \qquad \boldsymbol{a_t \in \mathcal{A}_m}. \tag{2}$$

The action space is $\mathcal{A}_m = \{\text{NO_OP}\} \cup \{G : G \subseteq \{G_1, G_2, G_3, G_4\}, G \neq \varnothing\}$. Thus the framework can express selective, progressive, and full restoration while retaining the original method's optimizer and history semantics. A profile is selected from a small method-specific family Π_m on the labeled calibration stream D_{42}:

$$\begin{aligned} \boldsymbol{\pi_m^*} &= \boldsymbol{\arg\max_{\pi \in \Pi_m} J(D_{42}; \pi)} \\ \text{subject to } \boldsymbol{V(D_{42}; \pi) = \text{valid};} &\qquad \boldsymbol{\pi_m^*} \text{ is then frozen before deployment.} \end{aligned} \tag{3}$$

Here J is the seed-42 full-stream development mIoU objective. Candidates must satisfy the predefined validity checks before entering profile selection. For SAR, candidates within the predefined 0.10-mIoU effective-tie margin are resolved by the lower-intervention rule, favoring fewer recoveries and then smaller restored cardinality. The optimization is offline; online deployment evaluates the frozen trigger and action rule without target labels.

4.2 COMMON LIFECYCLE AND STATE ABSTRACTION

The lifecycle is shared, while recovery semantics follow the adaptation mechanism's state structure:

Monitor $\rightarrow$ *Trigger* $\rightarrow$ *Recover* $\rightarrow$ *Native Update* $\rightarrow$ *Resume.*

In ordinary deployment, the frozen controller performs a read-only routing forward, computes the online state signal, evaluates the frozen threshold, persistence, and cooldown rules to obtain the profile action, and, when triggered, restores the selected source-anchored adaptable parameter subset before the method's native TTA update. Here, the source anchor is the immutable adaptable-parameter state at stream initialization, before any test-time update. Optimizer state, method history, and random state are preserved. A common state abstraction defines normalized recovery groups across methods and also supports controlled replay analyses. The StateRecoveryAdapter provides full-state capture and restoration for counterfactual replay, including optimizer, method-history, and random-state semantics where required; these replay capabilities are separate from the ordinary deployment intervention.

Normalized structural groups. We represent adaptable state using ordered groups G_1–G_4 from early to late adaptable structure. For the NACLiP ViT systems, the groups approximate normalized module/depth quartiles; for SED ConvNeXt-B, they follow the backbone's semantic stages. The mapping is defined over method-specific adaptable state rather than fixed parameter indices. For TENT, the groups cover adaptable normalization state; for MLMP and SAR, they cover the method-defined adaptable state. This provides normalized structural correspondence across compatible backbone variants without requiring identical parameter membership.

4.3 ONE-TIME PROFILING AND FROZEN DEPLOYMENT

For method m, the RecoveryProfiler uses one seed-42 calibration stream to choose a small profile. The profile includes the primary normalized state-drift statistic, a threshold τ, persistence q, cooldown, a recovery regime, group actions, severity escalation, optimizer semantics, and method-history semantics.

For each adaptable group g, drift is the source-relative ℓ_2 change $d_g = \|\theta_g - \theta_g^0\|_2 / \|\theta_g^0\|_2$; the controller aggregates these as $d = (\sum_g d_g^2)^{1/2}$ and standardizes d using the frozen calibration center and scale before applying τ.

Ground-truth labels are used offline for calibration and evaluation; no target labels are used in online decisions. After freezing the profile, seed-47/48 validation and cross-dataset transfer use it unchanged.

The selected frozen profiles differ across methods. MLMP uses $\tau = 2.0$, $q = 2$, cooldown 4, and progressive G1 $\rightarrow$ G1+G2 recovery. SAR uses $\tau = 3.0$, $q = 3$, cooldown 8, and G4 $\rightarrow$ G1+G4. TENT uses $\tau = 1.0$, $q = 1$, cooldown 4, and FULL adaptable-state recovery. The resulting profiles make the method-specific recovery contract explicit while preserving a common execution interface.

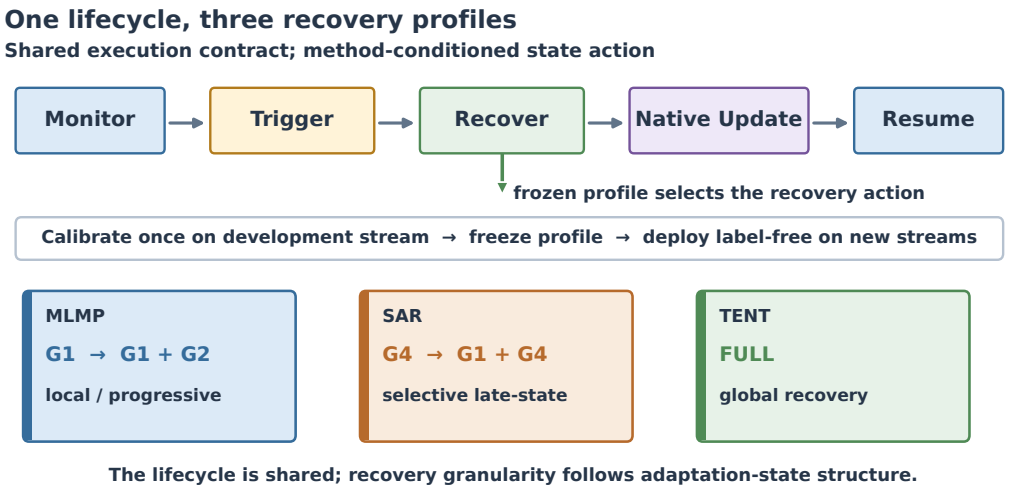


Figure 2: Profiled State Recovery shares a monitor–trigger–recovery–native-update–resume lifecycle while conditioning the recovery action on the adaptation mechanism. When triggered, the frozen profile restores the prescribed source-anchored adaptable parameter subset before the native TTA update. The examples instantiate the frozen MLMP, SAR, and TENT recovery regimes.

4.4 Deployment Algorithm

The online procedure preserves the method's native streaming update and invokes the frozen profile after trigger persistence is confirmed.

Profiled State Recovery (online). For each sample x_t, perform the read-only routing forward and compute normalized state drift. Evaluate the frozen threshold, persistence, and cooldown rules to obtain the profile action. If recovery is triggered, restore the prescribed source-anchored adaptable parameter subset while preserving the native optimizer, method history, and random state. Then execute the method's native TTA update, evaluate the resulting state, and continue the persistent stream. If no recovery is triggered, execute the native update unchanged.

Source-anchored group restoration keeps the intervention compact and reproducible, while the frozen profile determines when and how much adaptable state is recovered. This separation is useful in practice because it keeps the shared controller small while allowing the adaptation mechanism to determine the recovery granularity.

5 Experimental Setup

We evaluate five primary systems: S1 MLMP with NACLiP ViT-L/14 on YTVIS natural; S2 SAR with NACLiP ViT-L/14 on COCO mixed; S3 TENT with SED ConvNeXt-B on ACDC; S4 TENT with NACLiP ViT-L/14 on ACDC; and S5 MLMP with NACLiP ViT-B/16 on YTVIS natural.

YTVIS natural denotes the YouTube-VIS 2019 validation foreground stream with seeded video selection and frames emitted in video/frame order (Yang et al., 2019); COCO mixed denotes the COCO 2017 foreground-80, thing-only validation stream formed from 800 unique validation images in 16 mixed corruption blocks (12,800 stream samples) (Lin et al., 2014). ACDC uses seeded recurrent 50-sample blocks over fog, night, rain, and snow across two cycles (Sakaridis et al., 2021); SED denotes the external ConvNeXt-B open-vocabulary semantic-segmentation model (Xie et al., 2024). The main comparison is Persistent TTA versus frozen Profiled State Recovery on seeds 47 and 48. Each method profile is selected on its designated seed-42 development stream, the profile is frozen before seed-47/48 execution, and the validation streams are processed label-free. This cleanly separates one-time calibration from deployment-time adaptation.

We additionally test source-calibrated profile transfer on MLMP YTVIS $\rightarrow$ COCO, SAR COCO $\rightarrow$ YTVIS, and TENT ACDC $\rightarrow$ COCO. These transfers deploy the source profile unchanged and use no target-stream retuning. The structure analysis uses the same method and model while changing only the dataset and evaluates the fixed 16-action space over H32 replay horizons. Finally, hindsight empirical recovery headroom measures the best action available at deterministic stream positions; it quantifies how much recoverable opportunity remains at those positions.

6 Main Results and Generality

Table 1 gives full-stream foreground mIoU on the two fresh seeds. PSR produces broad gains across adaptation mechanisms: four of five systems improve, and all four improve on both fresh validation seeds. MLMP yields the largest effects, with +4.204 for S1 and +2.881 for S5 under the same frozen profile across ViT-L/14 and ViT-B/16. SAR contributes +0.316, and TENT improves the ConvNeXt system by +0.491. The remaining TENT-ViT setting lies in a recovery-limited regime with only 0.016 mean H32 headroom, reinforcing recoverability itself as a property of the target adaptation dynamics.

Table 1: Full-stream foreground mIoU (%) for Persistent TTA and frozen Profiled State Recovery. Δ *denotes PSR minus Persistent.*

System	Seed 47			Seed 48			Mean Δ
	Persistent	PSR	Δ	Persistent	PSR	Δ	
S1 MLMP ViT-L/14 / YTVIS	45.738	50.275	4.537	48.126	51.997	3.871	**4.204**
S2 SAR ViT-L/14 / COCO	31.806	31.851	0.045	31.324	31.912	0.588	**0.316**
S3 TENT ConvNeXt-B / ACDC	29.694	30.239	0.545	29.627	30.064	0.437	**0.491**
S4 TENT ViT-L/14 / ACDC	20.293	20.139	-0.154	20.301	20.133	-0.167	-0.161
S5 MLMP ViT-B/16 / YTVIS	46.256	49.153	2.898	47.268	50.132	2.864	**2.881**

Table 2 shows positive mean transfer deltas of +0.448, +0.602, and +0.739 for MLMP, SAR, and TENT, respectively. All three source-calibrated profiles transfer to the tested target streams without target-stream retuning. The second panel shows the same frozen MLMP profile improving both ViT-L/14 and ViT-B/16 by +4.204 and +2.881. Together, these results make Profiled State Recovery a practical profile-once, deploy-many-streams framework across the tested mechanisms.

Table 2: Transfer evidence from frozen profiles. Panel (a) reports source-to-target deployment without target-stream retuning. Panel (b) reports the MLMP cross-backbone comparison under the same frozen profile.

(a) Cross-dataset profile transfer

Method	Calibration $\rightarrow$ target	Δ47	Δ48	Mean Δ
MLMP	YTVIS $\rightarrow$ COCO	0.659	0.237	**0.448**
SAR	COCO $\rightarrow$ YTVIS	0.363	0.841	**0.602**
TENT	ACDC $\rightarrow$ COCO	0.377	1.101	**0.739**

(b) Cross-backbone profile transfer

Method	Comparison	Source-system mean Δ	Target-system mean Δ
MLMP	ViT-L/14 $\rightarrow$ ViT-B/16	**4.204**	**2.881**

The transfer result is also a statement about the division of labor between profiling and deployment. Calibration identifies the recovery semantics that match an adaptation mechanism; deployment then applies that compact profile to a new stream using only the online state signal and its frozen persistence/cooldown logic. No target labels enter this decision path, and no target-specific action search is required. This makes the framework suitable for settings in which the deployment stream is large, ordered, and unavailable for supervised policy selection.

The same profile representation supports two forms of transfer. Across datasets, the profile is unchanged while the stream statistics and intervention opportunities vary. Across the MLMP backbone scale, the normalized state-group mapping preserves the same G1/G2 semantics across the compatible ViT-L/14 and ViT-B/16 backbones. The resulting evidence suggests a useful practical strategy: profile the adaptation mechanism once, then reuse its recovery semantics through a common state abstraction.

7 Recoverability and State Design

Figure 3 maps the state-management design space. S1 and S5 combine substantial headroom with selective/local recovery regimes. S3 combines substantial headroom with positive FULL utility, revealing a global-recovery regime.

S2 is intermediate and mixed, while S4 occupies a recovery-limited regime. These distinct geometries show why recovery semantics are a central part of a streaming TTA system.

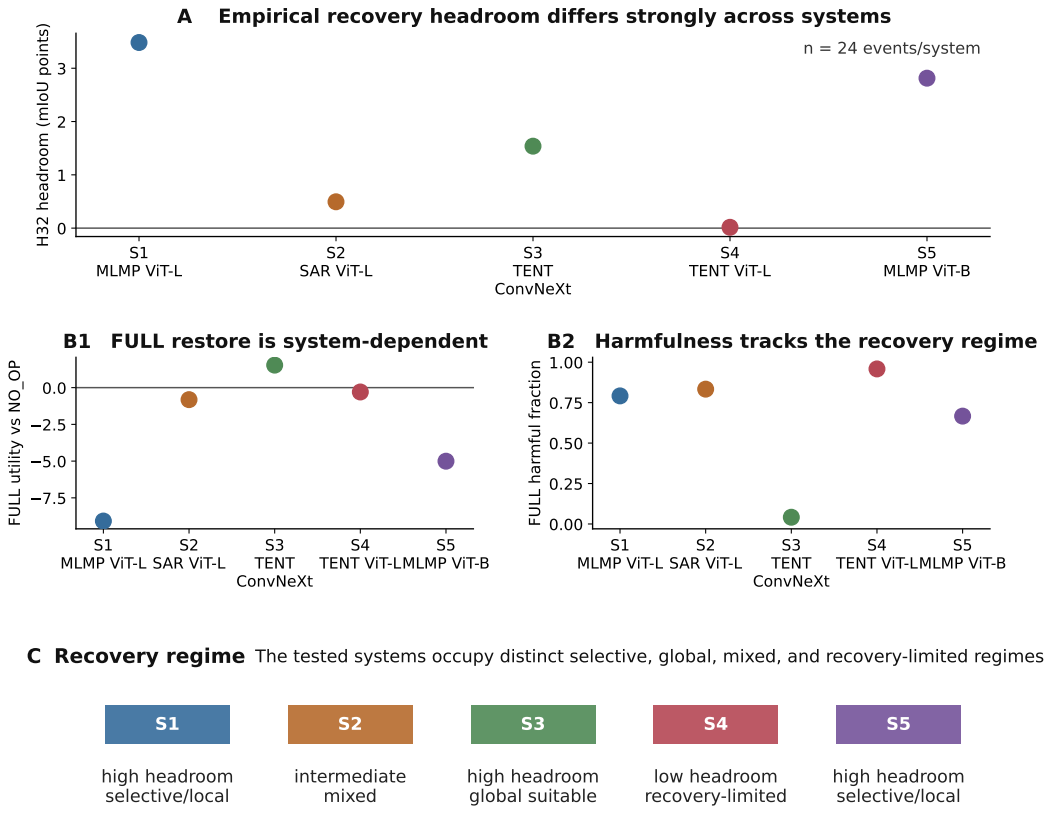


Figure 3: Hindsight H32 recoverability differs across the five systems. FULL restoration is suitable for S3 but harmful for S1, S2, S4, and S5; S4 has almost no substantive recovery headroom (mean 0.016 mIoU; no evaluated event exceeds 0.1 mIoU).

The frozen profile defines the method-level recovery contract, while the recoverability realized under that contract remains system-dependent: S3 and S4 share the same TENT profile but occupy global-recovery and recovery-limited regimes, respectively.

The controlled comparisons in Table 3 show that matching recovery granularity to adaptation-state structure is decisive for MLMP and SAR: MLMP benefits from progressive local recovery, while SAR benefits from selective late-state recovery G4 → G1+G4. The selected TENT profile combines broad restoration with a high-recall trigger. The state-drift versus matched-count periodic differences are +0.020 (S1), +0.074 (S2), -0.065 (S3), and +0.008 (S4). These small matched-count differences show that precise state-drift timing contributes little beyond the recovery structure and intervention frequency already fixed by the profile.

Table 3: Compact design evidence from frozen development/replay protocols. MLMP and SAR rows compare recovery granularity; TENT rows distinguish the selected profile from timing and action diagnostics. The MLMP (S1), SAR (S2), and TENT (S3) granularity/profile rows report differences versus Persistent, whereas the S1–S4 WHEN rows report state-drift minus matched-count periodic. The complementary method-conditioning diagnostic is reported in the appendix.

Family	Variant	Difference (mIoU points)	Diagnostic
MLMP	MLMP-A (G1)	**2.285**	granularity
MLMP	MLMP-B (G1 → G1+G2)	**4.482**	granularity
MLMP	MLMP-C (through G1+G2+G3)	**3.992**	granularity
SAR	SAR-A (G4)	-0.068	granularity
SAR	SAR-B (G4 → G1+G4)	**0.385**	granularity
SAR	SAR-C (frozen)	**0.363**	granularity
TENT	TENT-A (selected profile)	**0.611**	selected profile
TENT	TENT-B (FULL, higher threshold)	0.206	timing
TENT	TENT-C (G3+G4)	0.192	action
S1	drift − periodic	0.020	WHEN
S2	drift − periodic	0.074	WHEN
S3	drift − periodic	-0.065	WHEN
S4	drift − periodic	0.008	WHEN

Headroom analysis reveals substantial recoverable opportunity beyond the current profiles. S1, S3, and S5 exhibit mean H32 hindsight recovery headroom of 3.483, 1.538, and 2.814 foreground macro mIoU points, respectively, alongside the strong gains of the frozen profiles. This remaining opportunity makes persistent-state management a rich design axis for future streaming TTA, with the current framework providing a compact and reproducible practical instantiation.

This taxonomy also explains why state management is a systems contribution rather than a single-action heuristic. In selective/local regimes, a successful intervention must preserve useful adaptation while removing a targeted component of accumulated drift. In the global regime, a broad restore can recover the dominant state carrier. In the intermediate regime, the profile must trade these effects against the stream's intervention budget. The profile format exposes these choices explicitly, allowing future adaptation methods to report not only their update objective but also the state-management semantics under which they are deployed.

Design implication. A streaming TTA method should report two coupled objects: its native update rule and the state-management contract that governs long-horizon deployment. The first determines how state is created; the second determines how state is observed, checkpointed, and selectively recovered. This separation makes recovery behavior measurable across methods without requiring their internal optimizers or histories to be identical.

The frozen profiles illustrate the resulting workflow. A developer calibrates a small profile once, verifies the StateRecoveryAdapter mapping, and freezes the configuration. A deployment then carries the profile unchanged across the tested stream and backbone settings. The framework consequently turns a long-horizon implementation detail into a reusable interface that can be measured, compared, and extended by future TTA methods.

8 Conclusion

Persistent adaptation state is a structured, heterogeneous, and controllable object in streaming TTA. This paper makes state management explicit as a second design object alongside the native update rule and shows that method-specialized recovery profiles yield broad gains across mechanisms,

streams, and MLMP backbone scales. Profiled State Recovery turns long-horizon adaptation state into an explicit, reusable deployment contract. The substantial empirical headroom revealed by our anatomy motivates stronger state-management controllers as a productive next step. More broadly, exposing the state-management contract alongside the native update rule provides a reusable basis for designing and comparing future streaming TTA systems.

AI USE STATEMENT

ChatGPT/Codex assisted with hypothesis refinement, experiment planning, implementation and debugging support, artifact auditing, result organization and interpretation, manuscript organization, and language editing/drafting. The author reviewed all assisted outputs, independently verified code and reported data where applicable, and retains full responsibility for the scientific content and final submission.

REPRODUCIBILITY STATEMENT

The frozen profiles, seed policy, full-stream result tables, evaluation protocol, and machine-readable results are documented in the supplementary material. An anonymous package contains the common recovery interfaces, immutable profile/config files, figure and table-generation scripts, and lightweight machine-readable result data.

REFERENCES

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, February 2025. doi: 10.1109/WACV61041.2025.00495. URL https://openaccess.thecvf.com/content/WACV2025/html/Hajimiri_Pay_Attention_to_Your_Neighbours_Training-Free_Open-Vocabulary_Semantic_Segmentation_WACV_2025_paper.html.

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent test-time adaptation in recurring testing scenarios. In *Advances in Neural Information Processing Systems*, volume 37, pp. 123402–123442, 2024. doi: 10.52202/079017-3923. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

Taejun Lim, Joong-Won Hwang, and Kibok Lee. When and where to reset matters for long-term test-time adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 74988–75015, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79bdd6fe3f012befcc459ad13de65d13-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Fen Luo, Sen Li, and Zhanqiang Huo. CARE: conflict-aware regulation and evolution for continual test-time open-vocabulary semantic segmentation. *Journal of King Saud University Computer and Information Sciences*, 38:799, 2026. doi: 10.1007/s44443-026-01252-x. URL https://link.springer.com/article/10.1007/s44443-026-01252-x.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Mehrdad Noori, David Osowiechi, Gustavo A. Vargas Hakim, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Test-time adaptation of

vision-language models for open-vocabulary semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74741–74778, 2025. doi: 10.52202/085713-2510. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6c5b82193c5d8e6aa5806239676ddc97-Abstract-Conference.html.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39915–39935, 2023. doi: 10.52202/075280-1735. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7d640f377893fc5f22b5610e175ef7c3-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10765–10775, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Sakaridis_ACDC_The_Adverse_Conditions_Dataset_With_Correspondences_for_Semantic_Driving_ICCV_2021_paper.html.

Mona Schirmer, Metod Jazbec, Christian Andersson Naesseth, and Eric Nalisnick. Monitoring risks in test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 80915–80948, 2025. doi: 10.52202/085713-2705. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/746960ad49ddb47248970a0e1404230c-Abstract-Conference.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022. doi: 10.1109/CVPR52688.2022.00706. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Wenxi Wang. Stateful streaming open-vocabulary segmentation test-time adaptation: Persistent-state diagnosis and freeze-to-confirm recovery. Research Square, 2026. URL https://doi.org/10.21203/rs.3.rs-10866033/v1.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. SED: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3426–3436, 2024. doi: 10.1109/CVPR52733.2024.00329. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xie_SED_A_Simple_Encoder-Decoder_for_Open-Vocabulary_Semantic_Segmentation_CVPR_2024_paper.html.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5188–5197, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yang_Video_Instance_Segmentation_ICCV_2019_paper.html.

A Detailed Protocol and Frozen Profiles

The framework defines a common normalized state-group abstraction with method-specific state semantics. The StateRecoveryAdapter supplies the full-state capture and restoration used in controlled replay analyses. The frozen profiles are calibrated on seed 42 and deployed without target-stream retuning:

Method	Calibration	τ	q	Cooldown (stream samples)	Recovery rule
MLMP	YTVIS	2.0	2	4	G1 $\rightarrow$ G1+G2
SAR	COCO	3.0	3	8	G4 $\rightarrow$ G1+G4
TENT	ACDC	1.0	1	4	FULL

Exact frozen configurations are provided in the anonymous supplementary package.

Profile selection uses a fixed enumerated family of three candidates per method, not a Cartesian full grid. For MLMP, the candidates use $\tau = 1.5/2.0/2.5$, $q = 2/2/3$, cooldown $4/4/8$, and actions $G1$, $G1 \rightarrow G1 + G2$, and $G1 \rightarrow G1 + G2 \rightarrow G1 + G2 + G3$. For SAR, they use $\tau = 2.0/2.5/3.0$, $q = 2/2/3$, cooldown $8/8/8$, and $G4$, $G4 \rightarrow G1 + G4$, and the same action map with the conservative threshold/patience. For TENT, they use $\tau = 1.0/1.5/1.5$, $q = 1/1/1$, cooldown $4/4/4$, and FULL, FULL, and $G3 + G4$. Optimizer and method-history semantics are preserved for every candidate. Selection maximizes full-stream development mIoU; within SAR's predefined 0.10-mIoU effective-tie margin, it favors fewer recoveries and then smaller restored cardinality.

Frozen validity checks require the profile schema fields, patience at least 1, nonnegative cooldown, and valid action names and severity mappings. Candidate selection additionally requires a COMPLETE development summary with a finite full-stream metric; frozen validation verifies expected sample count, protocol identity, profile identity, and no online ground-truth use.

At each eligible trigger, let $\delta = z_{\mathrm{drift}} - \tau$. The controller maps $\delta \geq 0$ to moderate, $\delta \geq 1$ to strong, and $\delta \geq 2$ to extreme. Trigger persistence is checked before action selection. Severity is recomputed from the current z-score at each eligible trigger rather than incremented merely because triggers repeat; after recovery, streak and recovery-order state are updated and cooldown applies. The same controller logic is used across methods with method-specific action maps.

For reproducibility, the deployed drift standardization is

$$d_g = \frac{\sqrt{\sum_{p \in g} \|p - p^0\|_2^2}}{\max\left(\sqrt{\sum_{p \in g} \|p^0\|_2^2}, 10^{-30}\right)},$$

$$d = \left(\sum_g d_g^2\right)^{1/2},$$

$$z = \frac{d - c_m}{\max(s_m, 10^{-12})}.$$

The constants c_m and s_m are frozen per-method calibration center and scale constants and are never recomputed online:

Method	c_m	s_m
MLMP	0.0018995999125763774	0.0009750520002795382
SAR	0.0029641095316037536	0.0009331262039020657
TENT	0.0011075037546282457	0.00037155140721102274

B Mechanism Evidence

Persistent state changes long-horizon behavior and can change method ranking, but the direction is system-dependent. In the active-restoration comparison, nine of ten H32 events are recoverable; among those nine, active restoration outperforms prevention-only freezing in all nine, with a mean restore-minus-freeze ($G - P$) advantage of +1.322 mIoU. Adaptable parameter state is the dominant measured carrier of recovery gain, and mixed beneficial/harmful components occur in 203 of 281 evaluated events. The multi-group analysis found best cardinality proportions of 18.9% `NO_OP`, 34.9% one group, 29.9% two groups, 12.5% three groups, and 3.9% FULL: 46.3% of all events select a multi-group optimum. The matched-position method comparison was SAR-anchored; exact best-action agreement was 8.3%, best-cardinality agreement 16.7%, mean Spearman correlation

0.3784, beneficial-set Jaccard 0.1707, and harmful-set Jaccard 0.5874. Together, these comparisons motivate treating recovery structure as a method- and stream-dependent property.

C Complete Headroom Interpretation

The H32 analysis uses 24 deterministic stream-wide positions per system and the same 16-action space. Mean H32 hindsight headroom, in foreground macro mIoU points, is 3.4828 for S1, 0.4943 for S2, 1.5377 for S3, 0.0156 for S4, and 2.8136 for S5. The corresponding positive-opportunity rates are 75.0%, 75.0%, 100.0%, 58.3%, and 75.0%. FULL mean utilities are -9.0781, -0.8178, +1.5313, -0.2963, and -5.0062. The full harmful fractions are 79.2%, 83.3%, 4.2%, 95.8%, and 66.7%. We use these hindsight quantities to characterize recoverable opportunity; online deployment itself follows the frozen profile.

D Cross-Dataset Structure Audits

With method and model fixed within each pair, the aggregate Spearman action-rank correlations are 0.5643 for MLMP YTVIS versus COCO, 0.4964 for SAR COCO versus YTVIS, and -0.6571 for TENT ACDC versus COCO. The mean within-method value is 0.1345 in this analysis. Local action rankings shift across datasets: MLMP remains `NO_OP`-dominant in the aggregate regimes, SAR changes from selective-structural to local-progressive, and TENT has opposite aggregate rankings. The frozen profiles remain positively transferable, suggesting that they act as robust recovery priors rather than exact templates for the local hindsight optimum.

E Ablation Details

The method-conditioning diagnostic jointly changes timing and recovery action, so we use it as a descriptive cross-profile comparison rather than isolated causal evidence. The granularity comparison supports recovery granularity using controlled comparisons: MLMP fixed G1 (+2.2853), G1 $\rightarrow$ G1+G2 (+4.4822), and G1 $\rightarrow$ G1+G2+G3 (+3.9916); SAR fixed G4 (-0.0676), G4 $\rightarrow$ G1+G4 (+0.3854), and the frozen selected profile (+0.3626); TENT selected-profile FULL (+0.6109), higher-threshold FULL (+0.2059), and G3+G4 (+0.1918). The TENT selected-profile comparison combines broad restoration with its high-recall timing rule, whereas the MLMP and SAR comparisons isolate progressive/selective action structure under their controlled timing. The matched-count timing diagnostic was run for S1–S4. The state-drift minus matched-periodic differences are +0.0200, +0.0737, -0.0647, and +0.0083 for S1–S4. The matched-count periodic schedule is an offline diagnostic control whose intervention count and action sequence are inherited from the state-drift run; it is not a deployable alternative.

F S4 Diagnosis

S4 has one trigger per seed, as does S3, but S3 local FULL utility is +3.2407/+3.1401 while S4 local FULL utility is -0.4123/-0.3941. Stream-wide S4 headroom is 0.0156, with no event above 0.1, while S3 is 1.5377. The evidence identifies low intrinsic recovery headroom as the primary explanation for S4, with FULL over-restoration as a secondary factor; trigger timing is not the dominant cause.

G Reproducibility and Statistical Discipline

Primary claims use paired seed 47/48 full-stream values. With two seeds, we report direction consistency and mean paired deltas rather than formal significance tests. Machine-readable result tables, sample-order hashes, and reproducibility instructions are included in the anonymous supplementary package.