DACS: FROM HISTORICAL INERTIA TO FORWARD EVIDENCE FOR BEHAVIORAL AUTHENTICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In behavioral biometric authentication under user drift, forward-only adaptation is not merely an optimization problem; it is an information-selection problem under a restricted forward budget ($\leq$ 20 evaluations): should a drifted verifier be updated using historical training movement or measured forward responses? Testing historical movement reveals that trajectory transfer provides little benefit, yielding only $+0.504\,\mathrm{pp}$ on H-MOG and severe negative transfer on HuMI ($-5.940\,\mathrm{pp}$), showing that historical movement is not a reliable adaptation signal. We introduce DACS, a response-grounded alternative that replaces trajectory inheritance with two complementary sources of forward evidence. Controlled Component Attribution (ComAtt) identifies responsive components from development interventions, while current-user search resolves independent signed edits through a disjoint validation gate. On BehaveFormer authentication cohorts, four-component DACS improves balanced accuracy by $3.037\,\mathrm{pp}$ on H-MOG and $1.646\,\mathrm{pp}$ on HuMI within twenty objective evaluations, with factorial controls confirming a positive attribution–direction interaction. Evaluated through dual lenses, DACS delivers: (a) a lightweight seventeen-evaluation discrete variant retaining 76.5–79.7% of continuous-search gains; and (b) a full-cohort intention-to-treat audit achieving 5.3$\times$/10.3$\times$ larger population-level gains than Random-k while limiting observed severe deterioration to 2.82% on H-MOG and 6.75% on HuMI. These results support an information-anchor shift: exposing plasticity through measured component responses and directing it using present-user evidence.

1 INTRODUCTION

Forward-only behavioral adaptation must decide which information source should govern neural plasticity under distribution drift. Continual learning suggests that historical parameter trajectories supply a prior for online adaptation. Yet under realistic edge constraints—forward-only inference, no backward gradients, and a budget of $B \leq 20$ objective evaluations—historical movement risks directing plasticity into uninformative or harmful regimes. What information can reliably restore verification accuracy under behavioral drift?

Behavioral biometric authentication provides an ideal testbed: mobile verifiers continuously authenticate enrolled users via touch, motion, and keystroke dynamics (Sitóvá et al., 2016; Ray-Dowling et al., 2022; Stragapede et al., 2024), with architectures like BehaveFormer learning temporal representations (Senarath et al., 2023). Enrolled representations degrade within days as device-use habits drift. Adapting on-device is critical, yet edge verifiers lack backpropagation and historical training data, leaving forward responses as the only available signal.

Continual learning frameworks like Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and Synaptic Intelligence (SI) (Zenke et al., 2017) established that historical training information (e.g., Fisher importance or trajectory path integrals) can anchor parameter plasticity to preserve past capabilities. In task-incremental settings with stationary distributions, protecting historically sensitive parameters mitigates catastrophic forgetting. In online behavioral adaptation under drift, however, the challenge shifts from protecting the past to discovering what movement restores the

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.

present user. Relying on historical training inertia implicitly assumes that past optimization trajectories remain aligned with directions needed to counteract future user drift.

We isolate this premise failure through a controlled probe: the Historical Movement Guided Search Strategy (HMGSS), which directly reuses offline parameter movement under our forward-only edge budget. Rather than restoring performance, historical transfer yields no significant gain on H-MOG ($+0.504\,\text{pp}, p = 0.677$) and triggers severe negative transfer on HuMI ($-5.940\,\text{pp}, p < 0.001$). This demonstrates that in forward-only edge adaptation under behavioral drift, historical parameter trajectories fail to provide reliable corrective update directions.

Forward-only adaptation cannot be reduced to choosing an optimizer; it must decide which evidence deserves scarce evaluations: *historical inertia or forward evidence of model behavior?* Historical movement records where optimization has been; controlled forward interventions reveal where the deployed model remains functionally responsive, and current-user evaluations determine which response improves verification now. DACS operationalizes this shift through Controlled Component Attribution (ComAtt) and independent signed search (Figure 1). ComAtt converts development interventions into a frozen ranking of consequential components; current-user evaluations then determine how much to amplify or attenuate each component, and a disjoint validation split authorizes deployment.

We make three contributions shifting forward-only adaptation from optimizer selection to evidence selection:

1. **Refuting the historical-inertia premise.** We challenge and empirically refute the assumption that historical training trajectories provide reliable directional guidance under forward-only drift. While continual learning (e.g., EWC, SI) leverages historical sensitivity to protect past capabilities, relying on past movement fails to guide new adaptation: a controlled probe (HMGSS) yields no significant gain on H-MOG ($+0.504\,\text{pp}, p = 0.677$) and severe negative transfer on HuMI ($-5.940\,\text{pp}, p < 0.001$), establishing forward responses as the necessary informational anchor.
2. **A response-grounded plasticity mechanism.** We instantiate this principle via DACS, decoupling offline ComAtt (which isolates an editable subspace from forward interventions) from online signed search (which resolves user adjustments without backpropagation), with factorial controls confirming positive attribution–direction interaction.
3. **Empirical validation through two lenses.**
   - *Performance lens (GP/TPE):* Across both cohorts within 20 forward evaluations, DACS consistently outperforms all five alternative derivative-free baselines, supported by attribution–direction interaction intervals strictly excluding zero.
   - *Deployment lens (17-eval discrete + full-cohort ITT audit):* A seventeen-evaluation discrete variant retains 76.5–79.7% of continuous gains without surrogate fitting; full-cohort ITT audit achieves 5.3×/10.3× population gains over Random-k while holding observed severe deterioration to 2.82% on H-MOG and 6.75% on HuMI. Response-grounded editing further transfers to a ResNet-18 vision backbone under noise, suggesting a domain-general principle for forward-only adaptation.

Sections 2–6 review related work, formalize DACS, present empirical findings, and conclude.

2 Related Work

Historical information as a conceptual foundation. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and Synaptic Intelligence (SI) (Zenke et al., 2017) protect learned capabilities by penalizing deviations along historically sensitive parameters during backward training. While EWC and SI use historical curvature as a structural prior to constrain plasticity, extending historical inertia to provide directional update priors fails under open-world drift—a premise failure that undermines historical anchoring. HMGSS isolates this directional limitation under forward-only edge constraints, testing whether historical trajectories supply actionable corrective directions rather than evaluating multi-task regularization itself.

Test-time adaptation and the gradient-free shift. Test-time adaptation incorporates contemporary observations into updates. Gradient-based methods require backpropagation (Wang et al., 2021;

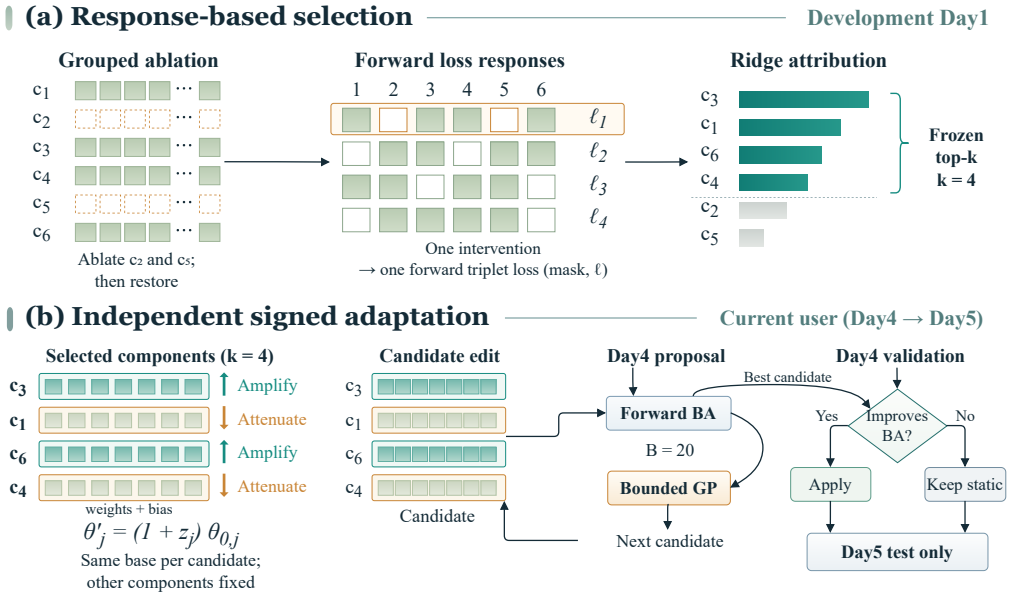


Figure 1: **DACS anchors behavioral adaptation in forward evidence.** (a) Development grouped ablations and forward triplet-loss responses support ridge attribution, yielding a frozen top-k component set. (b) Current-user search independently amplifies or attenuates selected weights and biases relative to the base model. Day4 proposal search uses $B = 20$ evaluations; disjoint Day4 validation selects whether to apply the edit or retain Static. Day5 is test-only. Component identities and masks are illustrative.

Niu et al., 2022; 2023; Wang et al., 2022). Zeroth-order methods adapt models without gradients (Malladi et al., 2023; Feng et al., 2025; Chen et al., 2025), but focus on optimizers rather than the underlying informational foundation, leaving open what evidence should guide adaptation under scarce forward budgets.

Forward evidence as an actionable adaptation anchor. Derivative-free search optimizes objective evaluations (Bergstra & Bengio, 2012; Snoek et al., 2012), while component attribution identifies functionally consequential units from interventions (Shah et al., 2024). DACS links these mechanisms: forward interventions expose consequential components, and current-user responses determine their signed edits. By making this previously implicit choice explicit and testable, DACS shifts forward-only adaptation from optimizer selection to evidence selection: isolating the premise failure of historical inertia and anchoring plasticity in measured model responses.

3 DACS: Adaptation Anchored in Behavioral Response

This section operationalizes the information-anchor principle into an actionable procedure. Controlled development interventions rank atomic components to define an editable subspace; current-user responses determine signed movement within that subspace, followed by validation gating before deployment (Algorithm 1).

3.1 Formalizing the Information-Anchor Shift

Let M_0 with parameters θ_0 be an enrolled verifier. An adaptation policy must choose which evidence governs plasticity. While a historical anchor $a_{\text{hist}} = F(\theta^{(0)}, \ldots, \theta^{(T)})$ inherits priorities from training checkpoints, DACS replaces trajectory inheritance with forward response: $a_{\text{DACS},t} = \big(s(D_a), \mathcal{H}_t\big)$, where $\mathcal{H}_t = \{(z_b, J_t(M_{\theta(z_b,S)}; D_t))\}_{b=1}^{B}$. Specifically, the HMGSS probe instantiates the historical directional prior as $\theta(z_b) = \theta_0 + z_b(\theta_{\text{enroll}} - \theta_{\text{base}})$ with scalar

Algorithm 1 DACS Adaptation Procedure

Phase 1: Development Component Attribution (ComAtt, Offline)
On development data D_a, evaluate grouped forward ablation interventions $r^{(q)} \in \{0,1\}^m$.
Fit ridge surrogate $(\hat{\beta}_0, \hat{\beta})$ via Eq. equation 1; set scores $s_j = -\hat{\beta}_j$.
Select and freeze top-k components $S = \arg\max_{|S|=k} \sum_{j \in S} s_j$.
Phase 2: Present-User Signed Search (Online, Forward-Only)
Initialize candidate pool $\mathcal{H}_t \leftarrow \{(0, J_t(0; D_t))\}$ on proposal data D_t.
for evaluation step $b = 2, \ldots, B$ **do**
    Propose candidate $z_b \in [-\rho, \rho]^k$ via continuous GP-EI or discrete Attribution-8.
    Temporarily apply candidate update $\theta_j(z_{b,j}) = (1 + z_{b,j})\theta_{0,j}$ for $j \in S$.
    Evaluate $J_t(z_b; D_t)$ via forward inference; update $\mathcal{H}_t \leftarrow \mathcal{H}_t \cup \{(z_b, J_t(z_b; D_t))\}$; restore M_0.
end for
Identify best observed candidate $\tilde{z} = \arg\min_{(z,y) \in \mathcal{H}_t} y$.
Phase 3: Validation Gating and Deployment
if $\mathrm{BA}(M_{\theta(\tilde{z},S)}; V_t) - \mathrm{BA}(M_0; V_t) > 0$ on disjoint validation split V_t **then**
    Deploy adapted weights $\hat{\theta} \leftarrow \theta(\tilde{z}, S)$.
else
    Reject candidate and retain Static baseline $\hat{\theta} \leftarrow \theta_0$.
end if

$z_b \in [-\rho, \rho]$, testing whether past optimization directions resolve drift. In contrast, DACS uses development data D_a from separate identities only for the frozen intervention ranking $s(D_a)$, leaving fresh responses on current proposal behavior D_t to determine user-specific updates.

Each seeded run fits one ranking before target evaluation and holds it fixed across target users. At deployment, the adapter receives proposal data D_t and a disjoint validation split V_t, evaluates the model through forward passes, and uses at most B proposal-objective evaluations. Both D_t and V_t pair buffered current genuine windows with a cached on-device development impostor gallery (under a fixed negative cap), evaluated against the frozen enrollment threshold τ_{auth}. Test outcomes enter neither the ranking, search, nor deployment decision.

3.2 Controlled attribution defines the editable subspace

Direct zeroth-order search across all $D > 6 \times 10^4$ model parameters is intractable within $B \leq 20$ evaluations, while gradient importance metrics are unavailable without backpropagation. We select the top-k components ($k = 4$) to fix search dimensionality. Parameters are partitioned into m atomic components $\mathcal{C} = \{C_1, \ldots, C_m\}$ (convolutional channels, linear rows, attention projection rows with bias; $m = 1{,}529$ on H-MOG, $1{,}214$ on HuMI; Supplement Section 2). Grouped forward interventions zero complementary groups under binary mask $r^{(q)} \in \{0,1\}^m$ to evaluate triplet loss $\ell^{(q)}$. A ridge surrogate assigns scores $s_j = -\hat{\beta}_j$:

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum_q \left(\ell^{(q)} - \beta_0 - r^{(q)\top}\beta\right)^2 + \lambda \|\beta\|_2^2, \qquad s_j = -\hat{\beta}_j. \tag{1}$$

A positive s_j means retaining component j predicts lower loss. The top-k scores define the editable subspace S. For $j \in S$, an independent scalar mode z_j scales the component: $\theta_j(z_j) = (1 + z_j)\theta_{0,j}$, where $z \in [-\rho, \rho]^k$ and objective $J_t(z) = 1 - \mathrm{BA}(M_{\theta(z,S)}; D_t)$. For disjoint projections P_j, update displacement satisfies $\|\Delta\theta(z, S)\|_2^2 \leq \rho^2 \sum_{j \in S} \|P_j \theta_0\|_2^2$ (Supplement Section 1). Thus k controls search dimension and ρ bounds relative movement.

3.3 Current behavior determines the signed update

The search evaluates zero inside its budget; every subsequent candidate evaluates $J_t(z)$ and restores M_0. Independent signed coordinates permit components to move in opposite directions. Continuous search fits a Gaussian process to accumulated proposal responses and selects bounded candidates by Expected Improvement (standard EI formulation; Supplement Section 1). After B evaluations, the best proposal is $\tilde{z} = \arg\min_{(z,y) \in \mathcal{H}_t} y$, retaining zero on ties. A proposal is deployed only when it strictly improves validation BA ($100[\mathrm{BA}(M_{\theta(\tilde{z},S)}; V_t) - \mathrm{BA}(M_0; V_t)] > 0$); otherwise the verifier remains Static ($\hat{z} = \mathbf{0}$).

3.4 CONTINUOUS ANALYSIS AND DISCRETE DEPLOYMENT

The continuous configuration uses bounded GP at $k = 4, B = 20$. Attribution-8 provides a compact deployment configuration with $k = 8$: it normalizes selected score magnitudes, combines them with four fixed sign patterns and their opposites (8 directions), and halves step size across two rounds ($1 + 8 \times 2 = 17$ evaluations) around the incumbent without surrogate fitting or recurring inference overhead. Algorithm 1 summarizes the adaptation procedure across its three phases.

All controls receive the same budget within their configuration: twenty evaluations for continuous search and seventeen for discrete search. The continuous GP configuration is the response-adaptive analysis reference; Attribution-8 is the deployment realization, and gain retention quantifies how much of the attributed-subspace value survives discretization. An objective evaluation may require several model batches. Offline ComAtt preparation is shared across the target cohort and is accounted for separately from per-user search.

4 EXPERIMENTS

4.1 BEHAVIORAL AUTHENTICATION AND EVALUATION PROTOCOL

We evaluate behavioral adaptation on H-MOG and HuMI using the keystroke branch of BehaveFormer (Senarath et al., 2023) (5 encoder layers with temporal/sensor attention, 50-step windows with 10 H-MOG or 3 HuMI features, 64-dim head). We initialize the verifier from an existing best pretrained BehaveFormer checkpoint; no backbone retraining is performed in adaptation. The checkpoint's original training used Adam (learning rate 10^{-3}, batch size 128, Euclidean triplet margin 1; 5,000 steps on H-MOG and 3,000 on HuMI on an NVIDIA RTX 3090). Base parameters θ_0 provide the invariant reference for updates, while Day1 enrollment data calibrate the embedding center and decision threshold. We report balanced accuracy changes, $\mathrm{BA} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR})$, relative to Static in percentage points (pp).

We pre-specify two analysis populations. The triggered-user analysis in the main adaptation tables includes users with consistent degradation on Day2 and Day3 and a mean decline $\geq 2\,\mathrm{pp}$, yielding 15 H-MOG and 60 HuMI target users. The intent-to-treat analysis in Table 4 includes all 55 H-MOG and 129 HuMI users; users outside the trigger retain Static and contribute zero policy gain.

Day1 development identities supply the frozen ComAtt ranking. Candidates are searched on Day4 proposal data and gated on disjoint Day4 validation data before Day5 testing. Main comparisons fix $k = 4$ (balancing expressiveness with $2^4 = 16$ sign corners), $\rho = 0.5$ (bounding relative displacement), and $B = 20$ (matching edge constraints), including zero (sweeps in Figure 3 and Supplement Tables S1, S3, S5). To rigorously capture stability and heterogeneity, we report two-level statistics: users are first averaged within each seed to compute cohort means, and $\mathrm{SD}_{\mathrm{seed}}$ measures sample variability across the ten optimization seeds. Individual heterogeneity is separately captured by user SDs ($1.842\,\mathrm{pp}$ H-MOG, $1.265\,\mathrm{pp}$ HuMI) and user-level CIs after averaging seeds within user. Inferential paired tests evaluate within-user differences to prevent pseudoreplication. All five DACS-vs-baseline comparisons remain significant under Holm correction (adjusted $p \leq 0.005$ H-MOG, $p = 0.0102$ HuMI; Supplement Section 4 and Table S7).

4.2 MAIN RESULTS: RESPONSE-BASED ADAPTATION

DACS improves BA by $3.037\,\mathrm{pp}$ on H-MOG and $1.646\,\mathrm{pp}$ on HuMI, achieving the highest mean gain among all six evaluated searchers in the same ComAtt subspace (Table 1). Its advantage over the runner-up is $0.826\,\mathrm{pp}$ on H-MOG ($p = 0.0025$) and $0.245\,\mathrm{pp}$ on HuMI ($p = 0.0102$).

All six searchers obtain positive mean gains, confirming that selected components support adaptation under various policies. GP uses the 20 evaluations most effectively. Figure 2 illustrates the intended score-level effect: restoring drifted genuine scores toward acceptance while preserving impostor separation.

Comparators are matched within-protocol reimplementations under the identical BehaveFormer verifier, ComAtt subspace, bounds, incumbent, splits, and budget. Table 1 reports matched benchmark results, as published cross-task numbers from different domains are not commensurable; Supplement Section 1 details baseline implementations and hyperparameters.

Table 1: **Main adaptation results on BehaveFormer.**

Searcher	Cohort mean $\pm$ SD_{seed} $\uparrow$	raw p	User 95% CI
H-MOG ($N = 15$, 10 seeds)			
DACS (GP)	$\mathbf{3.037} \pm 0.418$	—	$[2.017, 4.057]$
CMA-ES	2.211 ± 0.385	0.0025	$[1.255, 3.167]$
SPSA	2.199 ± 0.442	0.0042	$[1.152, 3.246]$
ZO-SGD	1.847 ± 0.460	< 0.001	$[0.758, 2.936]$
Random Search	1.505 ± 0.512	< 0.001	$[0.574, 2.436]$
TPE	1.421 ± 0.496	< 0.001	$[0.474, 2.368]$
HuMI ($N = 60$, 10 seeds)			
DACS (GP)	$\mathbf{1.646} \pm 0.288$	—	$[1.319, 1.973]$
TPE	1.401 ± 0.274	0.0102	$[1.088, 1.714]$
CMA-ES	1.315 ± 0.295	0.0011	$[0.983, 1.647]$
Random Search	1.147 ± 0.310	< 0.001	$[0.806, 1.488]$
SPSA	0.889 ± 0.342	< 0.001	$[0.525, 1.253]$
ZO-SGD	0.659 ± 0.365	< 0.001	$[0.284, 1.034]$

ComAtt, $k = 4$, $B = 20$; gains/intervals in pp. Cohort mean $\pm$ SD_{seed} across 10 seed averages. User CIs and paired tests aggregate seeds within user. All five DACS-vs-baseline comparisons remain significant after Holm correction (adjusted $p \leq 0.005$ H-MOG; 0.0102 HuMI). Comparators are matched within-protocol reimplementations under identical conditions; Supplement Section 1 details implementations and Table S7 gives user SDs and t statistics.

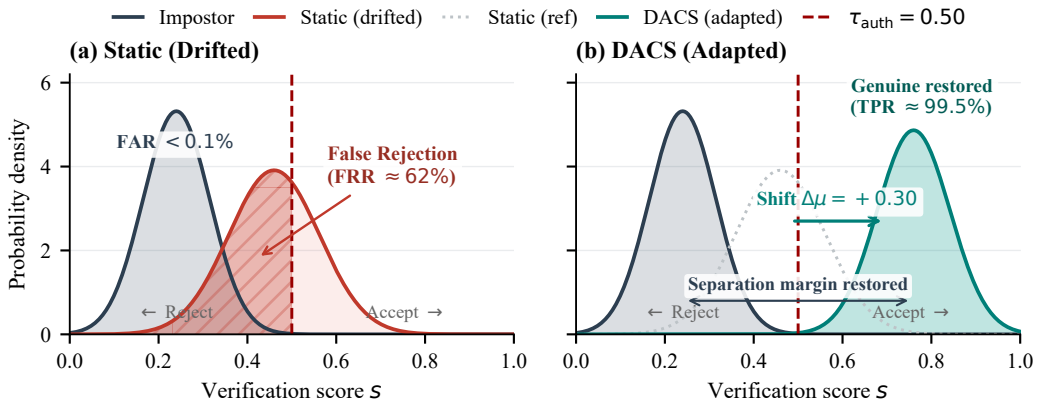


Figure 2: **Conceptual score-shift illustration under behavioral drift and adaptation.** Schematic density profiles of genuine and impostor similarity scores relative to enrollment threshold $\tau_{\text{auth}} = 0.50$. In **(a) Static (Drifted)**, drift shifts genuine scores below threshold, inducing false rejections. In **(b) DACS (Adapted)**, response-grounded adaptation selectively restores genuine scores rightward while preserving impostor rejection margin, eliminating drift without altering the security decision threshold.

4.3 Computational Efficiency

DACS updates existing parameters without architectural changes or additional inference operations (Table 2). Verification throughput changes by less than 1.5%, leaving recurring edge evaluation costs virtually unchanged.

The computational overhead is concentrated in one-time offline preparation and occasional search. ComAtt requires $\sim$ 45 seconds per ranking and is amortized across target users. On the representative-user RTX 3090 benchmark, mean search times span 0.916–1.099 seconds across all six optimizers; GP is within 4.0% of the fastest searcher while delivering the highest BA gain (Supplement Table S8 details full timing conditions and operation counts).

Table 2: **Computational profile of the fixed-size verifier.**

(a) Verifier and preparation

Property	H-MOG	HuMI
Params	195,174	67,809
MACs	1,138,500	283,800
FLOPs	2,728,604	798,279
FPS, Static	208.9	205.9
FPS, Adapted	206.2	204.8
ComAtt (s)	45.77	45.37

(b) Matched search time (s), $B = 20$

Searcher	H-MOG	HuMI
GP	1.020 ± 0.035	0.953 ± 0.027
TPE	1.099 ± 0.011	1.055 ± 0.037
CMA-ES	1.015 ± 0.033	0.945 ± 0.007
Random	0.997 ± 0.043	0.924 ± 0.040
SPSA	0.993 ± 0.042	0.916 ± 0.009
ZO-SGD	0.981 ± 0.025	0.927 ± 0.034

Panel (a): unchanged parameter counts and per-window operations (batch=1); FLOPs count arithmetic. Panel (b): RTX 3090, FP32; mean ± sample SD of 10 timing repeats at fixed seed 42 on a representative user. Random denotes Random Search. Timing excludes offline attribution, gating, and testing; Supplement Table S8 gives full scope.

4.4 Attribution identifies useful plasticity

Beyond the choice of search algorithm, the selection of editable components substantially governs adaptation performance even when the optimizer is fixed. With GP fixed, ComAtt improves over Random-k by $2.466\,\mathrm{pp}$ on H-MOG and $1.486\,\mathrm{pp}$ on HuMI (Table 3a).

Table 3: **Ablation of component selection and directional freedom.**

(a) Component selection under the same search budget ($k = 4, B = 20$, GP)

Selector	H-MOG (ΔBA pp ↑)	HuMI (ΔBA pp ↑)
Fisher	2.635 ± 0.435	1.265 ± 0.302
Gradient magnitude	2.613 ± 0.450	1.239 ± 0.315
Magnitude	0.042 ± 0.245	0.624 ± 0.260
Random-k	0.571 ± 0.312	0.160 ± 0.185
ComAtt	$\mathbf{3.037} \pm 0.418$	$\mathbf{1.646} \pm 0.288$

(b) Directional freedom and attribution × direction interaction

Dataset	Selector	Signed ↑	Positive-only ↑	Difference	Interaction [95% CI]
H-MOG	ComAtt	$\mathbf{3.037} \pm 0.418$	1.444 ± 0.365	+1.593	$+1.404\ [0.129, 2.423]$
	Random	0.273 ± 0.215	0.084 ± 0.142	+0.189	
HuMI	ComAtt	$\mathbf{1.646} \pm 0.288$	0.380 ± 0.210	+1.266	$+1.187\ [0.099, 1.842]$
	Random	0.100 ± 0.135	0.021 ± 0.088	+0.079	

Panel (a): every selector uses GP, $k = 4$, $B = 20$. ComAtt achieves the highest mean gains ($p_{\mathrm{Holm}} = 0.017$ on H-MOG, $p_{\mathrm{Holm}} < 0.001$ on HuMI; Table S12). Panel (b): cell gains are mean ± seed SD. Difference = signed − positive-only; interaction = ComAtt diff − Random diff; 95% CIs use user-cluster bootstrap. Calculations are performed on unrounded values prior to rounding.

ComAtt exceeds Fisher and gradient magnitude on both datasets (paired t-tests; Holm-adjusted $p = 0.017$ on H-MOG and $p < 0.001$ on HuMI, with $m = 2$ sensitivity comparisons per dataset; Supplement Table S12). Mechanistically, behavioral drift decouples past training curvature from contemporary responses; by measuring forward intervention impact, ComAtt isolates currently responsive coordinates without neural-model derivatives.

4.5 Directional freedom depends on attribution

Signed updates yield much larger gains when applied to attributed components than to random components. Crossing the two selectors with signed and positive-only updates produces positive interactions on both datasets, with confidence intervals excluding zero (Table 3b).

Allowing attenuation alongside amplification adds $1.593\,\mathrm{pp}$ and $1.266\,\mathrm{pp}$ under ComAtt, compared with only $0.189\,\mathrm{pp}$ and $0.079\,\mathrm{pp}$ under Random selection. The asymmetry is central: directional

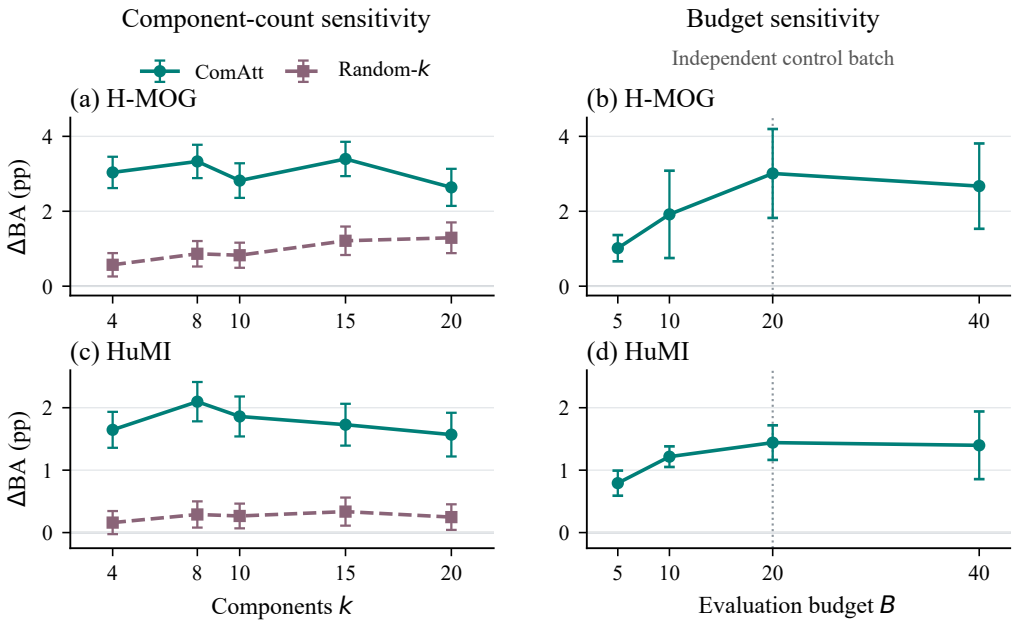


Figure 3: **Component-count and search-budget sensitivity.** (a,c) ComAtt vs. matched Random-k at $B = 20$ (Supplement Table S1). (b,d) Independent search with $k = 4$ in the separate control batch (Supplement Table S3). Markers: mean BA gain over Static; error bars: ± 1 across-seed SD. Points are separate runs; connecting lines are visual guides. Dotted line marks $B = 20$.

freedom cannot recover omitted components, spending its budget assigning signs to uninfluential weights under Random selection. In contrast, ComAtt isolates consequential components, allowing signed search to resolve bidirectional movement. Attribution and direction thus act as complements: the former supplies an informative coordinate system, while the latter resolves movement within it (positive interaction intervals in Table 3b).

Independent modes matter: replacing a shared scalar with component-wise modes improves mean gain by 2.128 pp on H-MOG and 0.241 pp on HuMI. Sensitivity sweeps (Figure 3) confirm ComAtt maintains its advantage across all tested $k \in \{4, 8, 10, 15, 20\}$, with peak performance at $B = 20$ and $\rho = 0.50$ (Supplement Tables S1, S3, S5).

4.6 Attribution remains useful under discrete search

The attribution advantage survives search simplification. With eight components and seventeen evaluations, Attribution-8 retains 76.5% and 79.7% of continuous GP gains over Static (Supplement Table S4). Attribution-8 exceeds Random-8 by 1.689 ± 0.185 pp on H-MOG and 1.296 ± 0.154 pp on HuMI, outperforming coordinate descent on both datasets and Small CMA-ES on H-MOG (2.548 vs. 1.574 pp; within 0.034 pp on HuMI: 1.672 vs. 1.706 pp, Supplement Table S4) without surrogate fitting or covariance tracking. Relative to each regime's random control, Attribution-8 retains 68.5% and 71.7% of the continuous attribution advantage (relative to the matched $k = 8$ continuous GP reference), confirming that the primary benefit is captured directly by the identified subspace.

4.7 Population benefit and deterioration risk

The gains on drift-triggered users translate into positive full-cohort returns: 0.828 pp on H-MOG and 0.765 pp on HuMI ($5.3\times$ and $10.3\times$ over Random-k; paired bootstrap intervals strictly exclude zero, Table 4), while limiting severe-deterioration ($\Delta\text{BA} < -0.50$ pp) rates to 2.82% and 6.75% across all users (non-triggered users retain Static with zero gain).

Table 4: **Population gain and severe-deterioration risk.**

		Users	ΔBA (pp) ↑		Risk (%) ↓
Dataset	Policy	Triggered / all	Triggered	All users	Severe
H-MOG	Random-k	15/55	+0.571	+0.156	0.00
	DACS	15/55	+3.037	+0.828	2.82
HuMI	Random-k	60/129	+0.160	+0.074	3.88
	DACS	60/129	+1.646	+0.765	6.75

Intent-to-treat accounting assigns zero gain to untriggered users (Static). Population net gains of DACS over Random-k are $+0.672$ pp (95% CI $[0.218, 1.185]$) on H-MOG and $+0.691$ pp ($[0.405, 1.012]$) on HuMI, using paired user-level bootstrap over the full cohorts. Severe-deterioration rates ($\Delta\mathrm{BA} < -0.50$ pp) are reported descriptively across ten seeds. Detailed cohort breakdowns are in Supplement Section 3 and Table S6.

4.8 Failure case analysis

Candidate failures associate with sparse adaptation evidence and directional mismatch rather than an uninformative editable subspace. In pre-gating diagnostics, severely deteriorating users have 45.1% fewer Day4 windows on H-MOG (205.0 vs. 373.3) and 46.4% fewer on HuMI (44.3 vs. 82.6), alongside large angular deviations from the hindsight reference ($121.68^\circ \pm 59.55^\circ$ vs. $31.91^\circ \pm 48.80^\circ$ on H-MOG; $78.95^\circ \pm 39.55^\circ$ vs. $50.82^\circ \pm 43.85^\circ$ on HuMI; Supplement Tables S9–S10).

4.9 Additional-backbone evidence under noise

DACS extends to ResNet-18 (He et al., 2016) on severity-5 CIFAR-10-C (Hendrycks & Dietterich, 2019) under a few-shot supervised cross-entropy objective (Supplement Section 7 details mobile CNN exploration), gaining 1.256 ± 0.577 pp under impulse noise, 1.075 ± 0.567 pp under Gaussian noise, and 1.015 ± 0.464 pp under shot noise across five checkpoint folds (averaging five seeds within fold). Gains exceed one percentage point across all three noise types, confirming that response-grounded editing generalizes beyond Transformers (Supplement Table S11 covers all fifteen corruptions and separates fold from seed variability).

5 Discussion

Our findings establish that updates under distribution drift require forward evidence: HMGSS demonstrates the empirical failure of historical trajectories (-5.940 pp on HuMI). DACS resolves this by decoupling reusable component attribution from user-specific signed search, with positive factorial interactions confirming that informative coordinates and directional freedom mutually reinforce.

Limitations and future directions. Fixed drift triggering remains sensitive to sparse evidence volumes, motivating adaptive gating thresholds; ComAtt also relies on representative development interventions whose stability across enrollment splits warrants broader study.

6 Conclusion

Forward-only behavioral adaptation requires establishing which evidence should anchor plasticity. The empirical failure of HMGSS (-5.940 pp) confirms that historical trajectories do not provide reliable corrective directions under drift. DACS operationalizes response-grounded adaptation: with four components and twenty evaluations, it improves balanced accuracy by 3.037 pp on H-MOG and 1.646 pp on HuMI, retaining 76.5–79.7% in discrete deployment while limiting observed severe deterioration to under 7%. Treating the information anchor as a primary design decision transforms forward-only adaptation from blind parameter search into response-grounded control of model plasticity.

LLM USE DISCLOSURE

Language models assisted with manuscript organization, language editing, and conceptual overview illustration. The authors take responsibility for the technical content, data, and conclusions.

REFERENCES

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(10):281–305, 2012.

Jiao Chen, Jiayi He, Fangfang Chen, Zuohong Lv, and Jianhua Tang. Forward-only continual learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4048–4057, Dublin, Ireland, 2025. ACM. doi: 10.1145/3746027.3755302. URL https://doi.org/10.1145/3746027.3755302.

Tao Feng, Wei Li, Didi Zhu, Hangjie Yuan, Wendi Zheng, Dan Zhang, and Jie Tang. ZeroFlow: Overcoming catastrophic forgetting is easier than you think. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16808–16824. PMLR, 2025. URL https://proceedings.mlr.press/v267/feng25j.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075, 2023. doi: 10.52202/075280-2308. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a627810151be4d13f907ac898ff7e948-Abstract-Conference.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

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.

Aratrika Ray-Dowling, Daqing Hou, Stephanie Schuckers, and Abbie Barbir. Evaluating multimodal mobile behavioral biometrics using public datasets. *Computers & Security*, 121:102868, 2022. ISSN 0167-4048. doi: 10.1016/j.cose.2022.102868.

Dilshan Senarath, Sanuja Tharinda, Maduka Vishvajith, Sanka Rasnayaka, Sandareka Wickramanayake, and Dulani Meedeniya. BehaveFormer: A framework with spatio-temporal dual attention transformers for IMU-enhanced keystroke dynamics. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9, 2023. doi: 10.1109/IJCB57857.2023.10448997.

Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44244–44292. PMLR, 2024. URL https://proceedings.mlr.press/v235/shah24a.html.

Zdeňka Sitová, Jaroslav Šeděnka, Qing Yang, Ge Peng, Gang Zhou, Paolo Gasti, and Kiran S. Balagani. HMOG: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, 11(5):877–892, 2016. doi: 10.1109/TIFS.2015.2506542.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2951–2959, 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Giuseppe Stragapede, Paula Delgado-Santos, Ruben Tolosana, Ruben Vera-Rodriguez, Richard Guest, and Aythami Morales. TypeFormer: Transformers for mobile keystroke biometrics. *Neural Computing and Applications*, 36(29):18531–18545, 2024.

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://arxiv.org/abs/2006.10726.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.