AVERT: W1-ANCHORED VERIFICATION FOR RELIABLE TEST-TIME ADAPTATION IN REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) for regression lacks the entropy-based confidence signal that classification TTA relies on, so naive adaptation can shift predictions in arbitrary directions. Analyzing four recent regression and time-series TTA methods, we identify four failure modes—negative transfer on low-shift data, unreliable low-order moment matching, over-adaptation, and *label-shift confusion*, where a near-perfect source model is corrected toward the source *prediction* distribution, moving predictions *away* from the target truth (on Concrete, SSA inflates MAE by $4.1\times$). We propose **AVERT**, a safe TTA framework built on three principles. (1) *W1-anchoring*: because target labels are standardized with source statistics, the 1-Wasserstein distance from a corrected prediction to the source *label* distribution is a label-free verification signal for regression TTA that tracks the true MAE change at $+0.964$ for AVERT's gated corrections over 15 benchmarks, and $+0.72$ on the raw natural-shift scale, with a finite-sample concentration analysis. (2) *Verification-gated adaptation*: a non-parametric quantile-transport bridge (QTB) combined with a gated test-time self-supervised representation adaptation (TT-SRA), both verified and gated by the W1 anchor. (3) *No-harm safeguards*: a safe-return gate detects label-shift confusion and refuses adaptation. Across 15 regression benchmarks and three backbones, AVERT significantly outperforms Original (p=7.8×10^{-5}) and all six adaptation baselines (up to p=4×10^{-22}), achieves the lowest average MAE (0.490 vs. 0.559 for Original) with the best average rank on all three backbones, is the only method that never severely degrades on any of the 45 dataset–backbone combinations, and transfers to a fully external dataset with zero retuning.

1 INTRODUCTION

Deep regression models deployed in the real world routinely face distribution shift between training (source) and test (target) environments. Test-time adaptation (TTA) adapts a pre-trained model using only unlabeled target inputs, without revisiting source data (Wang et al., 2021; 2022). While TTA has matured for classification—entropy minimization (Wang et al., 2021), weight averaging and stochastic restoration (Wang et al., 2022), sample-aware efficiency (Niu et al., 2022), sharpness-aware minimization (Niu et al., 2023), and augmentation marginalization (Zhang et al., 2022)—regression TTA remains underexplored despite its prevalence in energy forecasting, air quality prediction, and materials science (Adachi et al., 2025; Medeiros et al., 2025; Kim et al., 2025b; Im & Kwon, 2026). Classification TTA exploits the softmax output as a natural confidence measure and entropy minimization as a reliable unsupervised signal (Wang et al., 2021); regression lacks both, so distributional alignment must be used instead (Sun & Saenko, 2016), introducing new risks.

Through systematic analysis of four recent regression and time-series TTA methods (SSA (Adachi et al., 2025), PETSA (Medeiros et al., 2025), TAFAS (Kim et al., 2025b), and COSA (Im & Kwon, 2026)), we identify four failure modes. (i) *Negative transfer from unconditional adaptation*: on a low-shift benchmark (Household Power, CNN1D), all four baselines degrade, with COSA inflating MAE $20\times$ (0.681 vs. 0.033; Table B.1). (ii) *Unreliable correction directions from low-order moment matching*: SSA, PETSA, TAFAS, and CORAL (Sun & Saenko, 2016) align at most two statistics, missing multi-modal or heavy-tailed discrepancies. (iii) *Over-adaptation and catastrophic degradation*: aggressive adaptation collapses the prediction distribution (COSA inflates Air-Quality MAE $2.8\times$ on CNN1D). (iv) *Label-shift confusion*, previously unrecognized: when a model is near-perfect

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 the source domain but the target prediction mean shifts, methods that align target predictions to the source *prediction* distribution move them *away* from the target truth—on Concrete, SSA inflates MAE from 0.43 to 1.78 ($4.1\times$), while a QTB-style alignment inflates MAE by 14–36%. The root cause is that distribution alignment targets the source-prediction distribution, which inherits the model's systematic bias; the correct target is the source *label* distribution.

These observations indicate that regression test-time adaptation requires verification before correction. AVERT is designed around three principles: adapt only when the measured shift is sufficient, verify the correction direction against a label-free anchor, and bound the correction magnitude. We propose **AVERT**, a framework built on *W1-anchoring*, *quantile-transport adaptation*, and *no-harm safeguards*.

Our contributions are as follows:

- **We introduce *W1-anchoring*, a novel label-free verification signal for regression TTA.** The 1-Wasserstein distance from a corrected prediction to the source *label* distribution is a label-free proxy for MAE that tracks the true MAE change at $+0.964$ for AVERT's gated corrections over 15 benchmarks, backed by a bounded-error analysis (Appendix A.1). Unlike signals built on a model's own prediction stability, it references an external quantity—the source *label* distribution.
- **We build AVERT on this anchor: verification-gated adaptation with no-harm safeguards.** A Quantile Transport Bridge aligns the full empirical quantile function, and a gated test-time self-supervised representation adaptation fine-tunes the feature extractor by masked-input reconstruction; both are gated and blended under the W1 anchor. A data-driven safe-return gate detects label-shift confusion and refuses adaptation, and a per-sample anomaly clamp bounds tail events; our theory provides design-level motivation, and Section 4.5 tests the mechanisms outside the consistency assumption.
- **We evaluate AVERT on 15 benchmarks and three backbones using Wilcoxon and Friedman tests (n=135).** On natural temporal shifts evaluated on the raw scale, and on a fully external dataset with zero retuning, AVERT significantly outperforms Original (p=7.8×10^{-5}) and all six adaptation baselines ($p\leq3.3\times10^{-8}$), achieves the best average rank on all three backbones, and reduces the risk of large degradations relative to every baseline.

2 Related Work

Research most relevant to this work spans four threads: test-time adaptation for classification, label-free accuracy estimation, test-time and source-free adaptation for regression, and domain adaptation for regression. We review each and identify the limitation that motivates AVERT.

Test-time adaptation for classification. TTA adapts a pre-trained model to shifted test data without source access. TENT (Wang et al., 2021) minimizes prediction entropy by updating BatchNorm parameters; subsequent work adds weight averaging and stochastic restoration (CoTTA (Wang et al., 2022)), sample-aware efficiency (EATA (Niu et al., 2022)), sharpness-aware minimization (SAR (Niu et al., 2023)), augmentation marginalization (MEMO (Zhang et al., 2022)), and recent continual and architectural advances (Ni et al., 2025; Zhang et al., 2025; Chen et al., 2026; Liu et al., 2025; Kim et al., 2025a). All of these rely on entropy- or augmentation-based signals that are undefined or unreliable in regression. **Label-free accuracy estimation.** A complementary line of work estimates a model's accuracy without labels: AETTA (Lee et al., 2024) proposes *prediction disagreement*—the consistency of target predictions across dropout inference passes—as a label-free accuracy estimate for TTA, and shows it outperforms prior estimators across several classification TTA methods. Such signals target classification accuracy and rest on the stability of a model's own outputs; in regression there is no discrete correctness to calibrate against, and a confident but systematically biased model can exhibit low disagreement while remaining far from the target labels. W1-anchoring instead references an external quantity—the source *label* distribution—so its signal measures the distance to the (approximately stationary) target labels rather than the model's internal stability; this is what lets it expose label-shift confusion, where a low-disagreement model is nonetheless corrected away from the truth (Section 3.5). **Test-time and source-free adaptation for regression.** SSA (Adachi et al., 2025) aligns feature distributions within a significant subspace;

TAFAS (Kim et al., 2025b) uses partially observed ground truth with gated calibration; PETSA (Medeiros et al., 2025) employs low-rank adapters with dynamic gating; COSA (Im & Kwon, 2026) is a lightweight output-space adapter. They share two limitations: unconditional adaptation (risking negative transfer on low-shift data) and low-order moment matching that misses multi-modal and heavy-tailed discrepancies. AVERT instead aligns the full empirical quantile function, gated by a domain-shift threshold. **Domain adaptation for regression.** In the source-data-available regime, Deep CORAL (Sun & Saenko, 2016) aligns second-order feature statistics, DARE (Rosenfeld et al., 2022) performs a domain-specific adjustment of ERM that is provably robust to distribution shift, and Scatter Component Analysis (Ghifary et al., 2017) unifies domain adaptation and domain generalization; recent work includes TabFSBench (Cheng et al., 2025), distributionally robust multi-source UDA (Wang et al., 2026), and score-based distribution matching (Gong et al., 2025). These methods require source samples at deployment, which AVERT avoids by relying on compact source summaries computed once at training time.

3 METHOD

AVERT adapts through two mechanisms gated by the measured domain shift, a quantile-transport bridge (QTB, Section 3.3) and a test-time self-supervised representation adaptation (TT-SRA, Section 3.4), protected by two safety layers: a safe-return gate that detects label-shift confusion and refuses adaptation entirely (Section 3.5), and a W1 validation clamp plus a per-sample anomaly clamp that bound the damage of any accepted correction (Section 3.6). Algorithm 1 summarizes the procedure.

3.1 PROBLEM FORMULATION

Let f_θ be a regression model trained on labeled source data $\{(x_i^s, y_i^s)\}$ from P_S. At test time we receive unlabeled target inputs from a shifted distribution $P_T \neq P_S$ and must produce adapted predictions $\hat{y}_j^t$ that minimize $|\hat{y}_j^t - y_j^t|$ without access to target labels or source data during adaptation. We assume access to the pre-trained parameters and to compact source summary statistics computed once before deployment: the source-prediction quantile function, the source *label* distribution y^s, the source-prediction mean and standard deviation, and, for feature-level alignment, the source feature matrix $Z_s \in \mathbb{R}^{n_s \times d}$.

Label-distribution consistency. Targets are standardized with source statistics, so under covariate-dominated shift the target label distribution stays close to the source: $\mathrm{W}_1(y^t, y^s) \leq \varepsilon$ for small ε (median $0.18\,\sigma_s$ across our benchmarks). As long as this standard condition of label-shift-robust adaptation holds—and it is directly checkable per deployment from the empirical $\mathrm{W}_1(y^t, y^s)$—the W1 anchor yields label-free verification.

3.2 W1-ANCHORING: LABEL-FREE VERIFICATION

The central difficulty of regression TTA is that there is no label-free signal that directly reflects regression error—estimators built on a model's own prediction stability (Lee et al., 2024) measure confidence rather than correctness. We show that the 1-Wasserstein distance to the source *label* distribution provides such a signal.

Concretely, define the verification signal of a candidate prediction $\hat{y}$ as $v(\hat{y}) := \mathrm{W}_1(\hat{y}, y^s)$, its distance to the source *label* distribution. The design rests on one observation: targets are standardized with source statistics, so under covariate-dominated shift the target label distribution stays close to the source label distribution—across our benchmarks the median $\mathrm{W}_1(y^t, y^s)$ is $0.18\,\sigma_s$. Whenever the label shift is bounded, $\mathrm{W}_1(y^t, y^s) \leq \delta$, the triangle inequality of W_1 yields, for *any* candidate prediction $\hat{y}$,

$$|\,\mathrm{W}_1(\hat{y}, y^t) - v(\hat{y})\,| \;\leq\; \delta, \qquad \mathrm{MAE}(\hat{y}, y^t) \;\leq\; v(\hat{y}) + \delta + \kappa(\hat{y}), \tag{1}$$

where $\kappa(\hat{y}) := \mathrm{MAE}(\hat{y}, y^t) - \mathrm{W}_1(\hat{y}, y^t) \geq 0$ is the *sortedness cost*: the excess of the ℓ_1 error over the minimal transport cost, which vanishes when the ranks of $\hat{y}$ agree with those of y^t. QTB is a monotone transformation, so it does not increase κ. Equation (1) states that the label-free signal v tracks the true target error up to $\delta + \kappa$: selecting the candidate with the smallest v selects

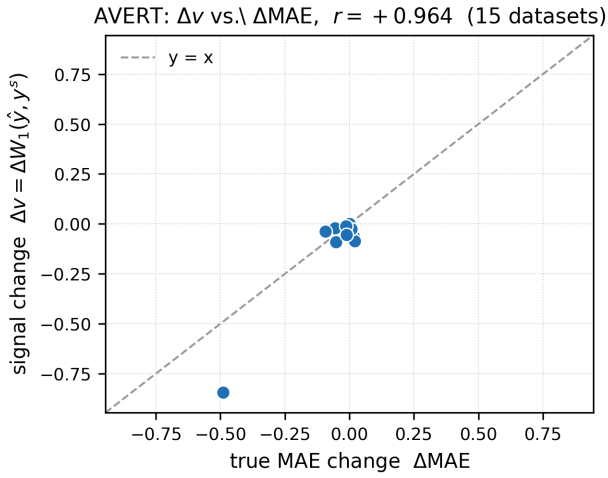

Figure 1: W1-anchoring: for AVERT's gated corrections, the change in the W1 distance from the prediction to the source label distribution correlates with the true MAE change at $+0.964$ across the 15 benchmark datasets. This label-free signal enables verification of any candidate correction.

the one closest to the target labels, up to a $\pm\delta$ slack (Theorem 3, Appendix A.1). A finite-sample concentration analysis strengthens this to the empirical plug-in signal: two candidates whose population signals differ by Δ are ranked correctly with probability $1-\delta'$ once the target batch satisfies $n_t = \Omega\big(\frac{R^2}{\Delta^2}\ln\frac{1}{\delta'}\big)$, with R the support length (Theorem 5, Appendix A.2).

Empirically, a correction's signal change Δv tracks its true MAE change at Pearson $r = +0.964$ for AVERT's gated corrections across the 15 benchmarks (Figure 1)—the co-movement that the verification layers of Section 3.6 exploit (Section 4.4 analyzes when this fails).

3.3 Quantile Transport Bridge (QTB)

Given source predictions $\hat{y}^s = f_\theta(X_s)$ and target predictions $\hat{y}^t = f_\theta(X_t)$, QTB maps each target prediction through the inverse of the target empirical quantile function and the forward of the source one:

$$\hat{y}_j^{\mathrm{QTB}} = Q_s \circ Q_t^{-1}(\hat{y}_j^t), \qquad Q_s(\tau) = \mathrm{Quantile}_\tau(\hat{y}^s), \quad Q_t(\tau) = \mathrm{Quantile}_\tau(\hat{y}^t). \tag{2}$$

Operationally this is a piecewise-linear interpolation of $\hat{y}_j^t$ through the paired quantile levels $(Q_t(\tau_k), Q_s(\tau_k))$. QTB is non-parametric: it captures every empirical quantile—bimodality, heavy tails, asymmetry—that low-order moment matching cannot.

Feature-level QTB (high shift). When $\mathcal{D} > 0.25$, prediction-level transport may be too coarse because the regression head collapses information; AVERT then transports penultimate-layer features per dimension and projects back through the frozen head g_θ. A per-dimension shift score $\delta_k = \max\big(|\bar{Z}_s^{(k)} - \bar{Z}_t^{(k)}|/(\sigma(Z_s^{(k)}) + \sigma(Z_t^{(k)})),\ \frac{1}{2}\,\mathrm{KS}(Z_s^{(k)}, Z_t^{(k)})\big)$ gates the transport (dimensions with $\delta_k \le 0.08$ are left untouched), and an adaptive smoothing $\mathrm{sm}_k = \mathrm{clip}(1-2\delta_k,\ 0.02,\ 0.15)$ lets strongly shifted dimensions transport more while weakly shifted ones stay conservative:

$$Z_{j,k}^{\mathrm{QTB}} = (1-\mathrm{sm}_k)\,\mathrm{interp}\big(Z_{j,k}^t;\ Q_t^{(k)},\ Q_s^{(k)}\big) + \mathrm{sm}_k\, Z_{j,k}^t, \tag{3}$$

with $\hat{y}_j^{\mathrm{QTB}} = 0.85\, g_\theta(Z_j^{\mathrm{QTB}}) + 0.15\,\hat{y}_j^t$ guarding against head-induced drift.

Low-shift branch. When $\mathcal{D} \le 0.25$, AVERT uses prediction-level QTB directly, smoothed toward the original by $\hat{y}_j^{\mathrm{QTB}} \leftarrow (1-\rho)\,\hat{y}_j^{\mathrm{QTB,raw}} + \rho\,\hat{y}_j^t$ with $\rho = 0.3 + (1-\min(2\mathcal{D}, 1))\cdot 0.4$—a continuous shrinkage that is provably non-degrading in the limit $\rho \to 1$.

3.4 TEST-TIME SELF-SUPERVISED REPRESENTATION ADAPTATION (TT-SRA)

A second, more aggressive mechanism, *test-time self-supervised representation adaptation* (TT-SRA), fine-tunes the feature extractor itself on the unlabeled target batch, borrowing the test-time-training principle of Sun et al. (2020; 2025) and the consistency regularization of Wang et al. (2022).

Given the target inputs X_t, TT-SRA masks a random 30% of the feature dimensions of each sample and trains, for a few gradient steps, a lightweight reconstruction head $h : \mathbb{R}^d \to \mathbb{R}^m$ (from the frozen hidden representation back to the input space) jointly with the feature extractor f_θ:

$$\min_{\theta,h} \underbrace{\mathbb{E}_{x\sim X_t}\big[\|h(f_\theta(x\odot m)) - x\odot \bar{m}\|^2\big]}_{\text{masked reconstruction}} + \beta\cdot \underbrace{\mathbb{E}_{x\sim X_t}\big[(f_\theta(x) - f_{\theta^0}(x))^2\big]}_{\text{consistency regularization}}, \tag{4}$$

where $m \in \{0,1\}^d$ masks a random subset of the d feature dimensions ($\bar{m} = 1 - m$) and β limits drift from f_{θ^0}; the reconstruction term adapts the representation to the target inputs while the consistency term prevents catastrophic drift. We use 4 Adam steps, learning rate 10^{-3}, $\beta = 1.0$.

TT-SRA is *gated* by the domain shift and the source-domain model quality: it runs only when $\mathcal{D} > 0.15$ and (the source model is not already highly accurate, $\mathrm{src_mae}/\sigma_s > 0.05$, or the shift is large, $\mathcal{D} \geq 0.3$). This gate prevents unnecessary fine-tuning when the source model is already accurate. On Household Power and SGEMM, where the source model is nearly perfect, fine-tuning degrades predictions. $\hat{y}^{\mathrm{TT}}$ is *blended* with the conservative QTB output ($\alpha = 0.5$), and the W1 validation clamp (Section 3.6) reverts the blend when it moves away from the source-label anchor. TT-SRA contributes most on high-shift datasets: on Energy Efficiency, AVERT reaches 0.591 (MLP) and 0.690 (ResNet1D) versus Original's 1.022 and 1.720 (Table 1; Table B.2).

3.5 SAFE-RETURN GATE

Label-shift confusion occurs when three conditions hold simultaneously: (a) the source model is near-perfect ($\mathrm{src_mae}/\sigma_s < 0.25$), (b) the target prediction mean shifts substantially relative to the source ($|\Delta|/\sigma_s > 0.8$), and (c) the W1 distance to the source labels is moderate ($\mathrm{W}_1(\hat{y}^t, y^s) < 2.0$). The three thresholds are ratios of computable statistics rather than absolute constants, and Section 4.5 shows performance varies smoothly under $\pm 30\%$ perturbation of each except at regime crossings (Appendix F). Under these conditions the predictions are already close to the target truth (Concrete: prediction mean -1.7 vs. label mean -1.8), and any alignment toward the source *prediction* distribution—which sits near zero—moves predictions away from the truth. AVERT detects this regime and returns the original prediction unchanged (Algorithm 1, safe-return step).

The gate is safe by construction: when it fires, the output is $\hat{y}^t$ itself, so the model is never worse than no adaptation. When the gate does not fire and adaptation is accepted, a bounded-gain analysis (Appendix A.3) shows that

$$\mathrm{MAE}(\hat{y}^{\mathrm{SR}}, y^t) \;\leq\; \mathrm{MAE}(\hat{y}^t, y^t) - \Delta v + 2\delta + \kappa(\hat{y}^{\mathrm{adapt}}), \tag{5}$$

where $\Delta v := v(\hat{y}^t) - v(\hat{y}^{\mathrm{adapt}}) > 0$ is the measured improvement of the verification signal. The gate thus converts a *measured* signal decrease into a *bounded* error decrease of at least $\Delta v - 2\delta - \kappa$; since QTB is monotone and δ is small on our benchmarks, the improvement is approximately Δv, consistent with the ablation results (Table 3).

3.6 W1 VALIDATION CLAMP AND ANOMALY CLAMP

After adaptation, AVERT validates the corrected prediction against the W1 anchor (bounded-error analysis, Appendix A.1). If the correction moves the prediction *away* from the source label distribution beyond a tolerance, it is partially reverted:

$$\begin{aligned} &\text{if } \mathrm{W}_1(\hat{y}^{\mathrm{mix}}, y^s) > 1.05\cdot \mathrm{W}_1(\hat{y}^t, y^s):\\ &\quad \hat{y}^{\mathrm{final}} \leftarrow \mathrm{frac}\,\hat{y}^{\mathrm{mix}} + (1-\mathrm{frac})\,\hat{y}^t, \end{aligned} \tag{6}$$

with $\mathrm{frac} = \mathrm{clip}\big(1 - \frac{\mathrm{W}_1(\hat{y}^{\mathrm{mix}}, y^s) - \mathrm{W}_1(\hat{y}^t, y^s)}{\mathrm{W}_1(\hat{y}^t, y^s)},\, 0,\, 1\big)$.

Algorithm 1 AVERT Test-Time Adaptation Pipeline

Require: Pre-trained f_θ; source predictions $\hat{y}^s$; source labels y^s; target inputs X_t; source features Z_s; domain shift $\mathcal{D}$

$\hat{y}^t \leftarrow f_\theta(X_t)$
if $\mathcal{D} < 0.015$ **then**
  return $\hat{y}^t$ {no adaptation}
end if
$\Delta \leftarrow \mathrm{mean}(\hat{y}^t) - \mathrm{mean}(\hat{y}^s)$; $\sigma_s \leftarrow \mathrm{std}(\hat{y}^s)$; $\sigma_{\mathrm{res}} \leftarrow \mathrm{std}(\hat{y}^s - y^s)$
if $\mathrm{src_mae}/\sigma_s < 0.25$ **and** $|\Delta|/\sigma_s > 0.8$ **and** $\mathrm{W}_1(\hat{y}^t, y^s) < 2.0$ **then**
  return $\hat{y}^t$ {safe-return: label-shift confusion}
end if
Compute $\hat{y}^{\mathrm{QTB}}$ via QTB (Eq. 2, prediction- or feature-level by $\mathcal{D}$); **if** TT-SRA gated on (Section 3.4): $\hat{y}^{\mathrm{TT}} \leftarrow \text{TT-SRA}(X_t)$
$\hat{y}^{\mathrm{mix}} \leftarrow \hat{y}^{\mathrm{QTB}}$ {or $0.5\,\hat{y}^{\mathrm{TT}} + 0.5\,\hat{y}^{\mathrm{QTB}}$ when TT-SRA runs}
if $\mathrm{W}_1(\hat{y}^{\mathrm{mix}}, y^s) > 1.05 \cdot \mathrm{W}_1(\hat{y}^t, y^s)$ **then**
  {W1 validation clamp}
  $\mathrm{frac} \leftarrow \mathrm{clip}(1 - \Delta\mathrm{W}_1/\mathrm{W}_1(\hat{y}^t, y^s), 0, 1)$
  $\hat{y}^{\mathrm{mix}} \leftarrow \mathrm{frac}\,\hat{y}^{\mathrm{mix}} + (1 - \mathrm{frac})\,\hat{y}^t$
end if
$\hat{y}_j^{\mathrm{final}} \leftarrow (1 - \alpha_j)\,\hat{y}_j^t + \alpha_j\,\hat{y}_j^{\mathrm{mix}}$, with $\alpha_j = \mathrm{clip}\big(2\mathcal{D}^{1.5}(1 - c_j)\,s(n_t),\ 0.02,\ 0.5\big)$
  {c_j: per-sample transport confidence; $s(n_t) = \mathrm{clip}(\log_{10}(n_t{+}1)/3,\ 0.3,\ 1)$ (Appendix A.4)}
Anomaly clamp: if $|\hat{y}_j^{\mathrm{final}} - \hat{y}_j^t| > 3\max(\sigma_{\mathrm{res}}, 0.5\sigma_t)$, set $\hat{y}_j^{\mathrm{final}} \leftarrow 0.85\hat{y}_j^t + 0.15\hat{y}_j^{\mathrm{mix}}$
return $\hat{y}^{\mathrm{final}}$

Finally, a per-sample anomaly clamp bounds single-sample deviations. Let $\sigma_{\mathrm{res}} = \mathrm{std}(\hat{y}^s - y^s)$ be the source residual standard deviation (distinct from the source prediction standard deviation σ_s used above) and $\sigma_t = \mathrm{std}(\hat{y}^t) + \epsilon$ the target prediction standard deviation. Any sample with $|\hat{y}_j^{\mathrm{final}} - \hat{y}_j^t| > 3\max(\sigma_{\mathrm{res}},\ 0.5\sigma_t)$ is replaced by

$$\hat{y}_j^{\mathrm{final}} \leftarrow 0.85\,\hat{y}_j^t + 0.15\,\hat{y}_j^{\mathrm{mix}}, \tag{7}$$

which is at most $0.15\,|\hat{y}_j^{\mathrm{mix}} - \hat{y}_j^t|$ away from the original prediction. This prevents single-sample catastrophic deviations that would otherwise dominate the per-batch MAE.

Per-sample blending is safe by construction: writing $e_j^t = |\hat{y}_j^t - y_j|$ and e_j^{mix} for the routed correction, the blend satisfies $e_j^{\mathrm{final}} \le e_j^t + 0.5\,(e_j^{\mathrm{mix}} - e_j^t)$, so a harmful correction degrades a sample by at most $0.5\,(e_j^{\mathrm{mix}} - e_j^t)$, and the anomaly clamp bounds any single-sample deviation by $0.15\,|\hat{y}_j^{\mathrm{mix}} - \hat{y}_j^t|$ (derivation in Appendix A.4). The W1 validation clamp (6) is likewise monotone: the further a correction moves the prediction from the source label distribution, the smaller the retained fraction, so anchor-aligned corrections are preserved and misaligned ones are suppressed (Appendix A.4).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on 15 regression benchmarks. Because TTA targets deployment under distribution shift, we include only datasets whose source–target split reflects a real deployment boundary rather than a random partition, using three criteria: (i) the data carry an explicit temporal, spatial, or structural axis, and the split follows that axis—chronologically, across sites, across production years, or across operating regimes; (ii) the induced covariate shift is statistically measurable (per-feature Kolmogorov–Smirnov statistics between source and target); and (iii) the source model is competent on its own domain, so that adaptation—not model repair—is the problem. Under these rules we select 15 datasets jointly covering large-shift and moderate-shift regimes ($\mathcal{D}$ from 0.10 to 0.45): Air Quality, Beijing Multi-Site Air Quality, California Housing, Concrete, Energy Efficiency (Tsanas & Xifara, 2012), Gas Turbine NOx, Gas Turbine CO Emissions, Household Power (Hébrail & Bérard, 2012), NWP Temperature (cross-station), PGCB Demand, Seoul Bike Sharing, SGEMM GPU, Solar Plant, and Tetouan Power Consumption. All datasets are public (UCI Machine Learning Repository; the solar-plant logs from a public plant-telemetry release). Features are standardized with source statistics; targets are normalized on the source domain.

Table 1: Main results (MLP backbone): MAE (3-seed mean) on 15 datasets. Lower is better. Best per row in **bold**, second-best underlined (ties share the same rank). Avg Rank (1=best, 8=worst, tied ranks averaged).

Dataset	Original	SSA	PETSA	TAFAS	COSA	CORAL	TENT	AVERT
AirQual	0.431	0.389	0.386	0.383	0.945	0.409	1.079	**0.382**
BeijingMS	0.191	0.271	0.245	0.279	0.806	0.240	0.667	**0.191**
CalHousing	0.412	0.402	0.404	**0.402**	0.793	0.409	0.786	0.412
Concrete	**0.433**	1.775	1.784	1.718	1.678	1.838	1.546	**0.433**
EnergyEff	1.022	1.264	1.242	1.261	1.335	1.455	1.243	**0.591**
GasTurbine	1.070	0.922	**0.891**	0.894	1.025	0.924	0.913	0.980
GTEmiss	**0.377**	0.413	0.410	0.443	0.516	0.433	0.549	0.385
Power	**0.035**	0.104	0.092	0.118	0.682	0.102	0.407	0.035
NWP-s24	0.586	0.407	0.416	**0.400**	0.884	0.492	0.852	0.490
NWP-s34	1.267	**0.507**	0.519	0.517	0.852	0.540	0.943	1.247
PGCB	**0.254**	1.931	1.937	1.932	1.964	1.949	1.906	**0.254**
SeoulBike	**0.798**	1.135	1.095	1.150	1.438	1.052	1.009	0.825
SGemm	0.215	0.654	0.664	0.690	0.622	0.850	**0.214**	0.221
SolarP2	0.216	**0.199**	0.202	0.200	0.888	0.211	0.759	0.214
Tetouan	0.700	0.737	0.751	0.736	1.029	0.783	0.771	**0.692**
Avg Rank	3.60	3.93	3.87	4.00	7.00	5.40	5.33	2.87

Baselines. We compare against Original (no adaptation) and six adaptation methods: SSA (Adachi et al., 2025), PETSA (Medeiros et al., 2025), TAFAS (Kim et al., 2025b), COSA (Im & Kwon, 2026), CORAL (Sun & Saenko, 2016), and TENT (Wang et al., 2021). SSA and PETSA are offline batch TTA methods applied directly with their recommended hyperparameters. TAFAS and COSA were designed for online forecasting with streaming ground truth; we adapt them to the offline label-free setting by replacing the ground-truth signal with unsupervised feature-distribution matching (TAFAS) and by training the output-space adapter on unlabeled target predictions (COSA). We distinguish two information regimes: *source-free* methods access only the trained model and the target batch (TENT), whereas *source-statistics* methods access compact source summaries precomputed once before deployment, never the raw source data afterwards (SSA, PETSA, TAFAS, COSA, AVERT, CORAL). CORAL uses the same source statistics as AVERT; its large degradations on label-shift-confusion datasets (Table 1) show the effect of verification over unverified alignment. TENT adapts BatchNorm affine parameters by entropy minimization (applicable to our MLP and CNN1D backbones, which contain BatchNorm layers).

Protocol. We evaluate three backbones: a 2-layer MLP (128 hidden units, BatchNorm), a 3-block 1D-CNN (CNN1D), and a 4-block 1D-ResNet (ResNet1D), each trained for 100 epochs with Adam (Kingma & Ba, 2015) (lr=0.001). We report MAE averaged over 3 random seeds for all methods. Statistical significance is assessed via the Friedman test (Friedman, 1937) and the Wilcoxon signed-rank test (Wilcoxon, 1945), following Demšar (2006). AVERT's hyperparameters are fixed: safe-return thresholds (0.25, 0.8, 2.0), W1 validation tolerance 1.05, and anomaly clamp factor 3. All experiments run on a single NVIDIA GeForce RTX 3080 Laptop GPU (16 GB).

4.2 Main Results

Table 1 reports MAE on the MLP backbone (CNN1D and ResNet1D: Tables B.1–B.2, Figure B.1). AVERT achieves the lowest average MAE pooled over the three backbones (0.490 vs. 0.559 for Original and 0.687 for TENT) and the best average rank on each backbone. The pooled Wilcoxon test (n=135) confirms that AVERT significantly outperforms *every* baseline including Original (Table 2), and the Friedman test over the 45 dataset–backbone 3-seed means is likewise significant.

Two observations are notable. First, on Energy Efficiency ($\mathcal{D} = 0.45$, the highest shift in the benchmark), AVERT improves MAE by 42.2% over Original on MLP (1.022 $\rightarrow$ 0.591), the largest single-dataset gain in the benchmark, where every adaptation baseline degrades. Second, on the label-shift-confusion datasets, the safe-return gate is decisive: on Concrete, SSA and PETSA inflate MAE by more than $4\times$ (0.43 $\rightarrow$ 1.78) and COSA by $3.9\times$, while AVERT detects the confusion and returns predictions unchanged; similarly on SGEMM, all four TTA baselines degrade by $\sim 3\times$ while AVERT stays within 3% of Original. AVERT is the only method whose MAE is never inflated by more than 16.4% on any of the 45 dataset–backbone combinations (it matches or improves Original on 35 of 45), whereas each baseline degrades by more than $1.5\times$ on at least 12 of them. On the MLP backbone, TENT is 70% worse than no adaptation on average (0.910 vs. 0.534).

Table 2: Statistical significance: Wilcoxon signed-rank test (one-sided, AVERT < baseline). Pooled n=135 (15 datasets × 3 backbones × 3 seeds); per-backbone n=45. Pairs are matched per (dataset, backbone, seed); observations sharing a dataset are dependent. ** $p<0.01$, * $p<0.05$, ns = not significant.

Method	Pooled	MLP	CNN1D	ResNet1D
Original	7.8×10^{-5}**	1.1×10^{-2}*	3.5×10^{-3}**	4.4×10^{-2}*
SSA	2.1×10^{-10}**	2.2×10^{-3}**	2.1×10^{-4}**	5.0×10^{-7}**
PETSA	2.2×10^{-9}**	2.6×10^{-3}**	9.6×10^{-5}**	3.8×10^{-5}**
TAFAS	2.6×10^{-9}**	2.0×10^{-3}**	2.8×10^{-4}**	1.6×10^{-5}**
COSA	4.1×10^{-22}**	9.0×10^{-10}**	5.4×10^{-13}**	5.9×10^{-12}**
CORAL	1.6×10^{-9}**	2.6×10^{-5}**	7.0×10^{-4}**	2.5×10^{-4}**
TENT	3.3×10^{-8}**	3.3×10^{-8}**	ns	1.8×10^{-2}*

Table 3: Ablation study (MAE, 3-seed mean averaged over three backbones). "Full" is AVERT re-run within the ablation harness; its overall mean (0.490) coincides with the main-table pooled average. Best per row in **bold**.

	Full	w/o Verify	w/o Gate
Mean MAE (15 ds)	**0.490**	0.549	0.554
Mean Rank (45 pairs)	**1.86**	2.12	2.02
Key cases (mean MAE):			
Concrete (confusion)	**0.507**	**0.507**	0.873
EnergyEff (large shift)	0.682	1.499	**0.677**
AirQual (mixed)	**0.370**	0.384	0.378

Statistical significance. AVERT is significantly better in 20 of 21 comparisons (7 baselines, including Original, × 3 backbones); the sole exception is TENT on CNN1D, where AVERT still attains the better average rank. Table 2 reports the full grid.

TENT's BatchNorm recalibration and architectures. TENT's BatchNorm recalibration is the strongest single mechanism among the baselines (second-best average rank on CNN1D, 3.47, and ResNet1D, 3.27); its largest win is Energy Efficiency on CNN1D (0.579 vs. AVERT's 0.776), where unverified recalibration acts as online feature re-centering matching this near-pure covariate shift. The same unverified update, however, is harmful where the shift changes the label–feature relationship: on PGCB Demand (CNN1D) TENT inflates MAE by 3.4× (0.317 → 1.088) and on Concrete by 2×, while AVERT's gate abstains and matches Original. AVERT remains architecture-agnostic.

4.3 Ablation Study

Table 3 reports a component-necessity analysis (3-seed mean, averaged over three backbones). Removing the W1-anchored verification signal from *every* decision degrades mean MAE from 0.490 to 0.549 and mean rank from 1.86 to 2.12 (p=3.0×10^{-2}), concentrated in the large-shift regime: on Energy Efficiency the gate then fires spuriously and adaptation is refused, so MAE collapses from 0.682 to 1.499 (+120%)—indicating that the W1 signal is the gate's discriminative core. Removing the safe-return gate instead degrades mean MAE to 0.554 and mean rank to 2.02 (p=1.0×10^{-2}), driven by Concrete (0.507 → 0.873, +72%), the label-shift-confusion case.

4.4 Case Study and Reliability Boundary

Sensor drift under deployment. We simulate sensor drift (zero-drift bias and sensitivity loss) on three physical-sensing benchmarks and compare no adaptation, naive unconditional QTB, and AVERT (six scenarios per dataset; Appendix C). On *Concrete* (unsafe regime: near-perfect source model with moderate prediction shift—the label-shift-confusion condition), unconditional QTB is always harmful (+14–36%) and the gate refuses it, so AVERT matches no adaptation in every scenario. On *Beijing PM2.5* (adaptable regime) the gate stays silent and AVERT reduces MAE by 9–29%. On *Air Quality* (mixed regime) the gate is mostly silent but fires conservatively in one scenario (0.759 vs. naive 0.686); this is the only scenario in which unconditional adaptation is marginally better.

Reliability boundary of the signal. The signal is reliable only within a bounded region, as the proxy analysis (Appendix A.1) predicts. Correlating each method's signal change Δv with its true MAE change across the 15 datasets (MLP, seed 42), AVERT tracks the true error at $+0.964$ (Figure 1), whereas SSA, PETSA, and TAFAS are *negative* (-0.60, -0.61, -0.59) and COSA negative (-0.36): AVERT stays inside the reliability region by design (Appendix D).

4.5 Robustness and Efficiency Analyses

Efficiency. AVERT runs in 0.26 s on average ($17\times$ faster than SSA, $14\times$ than PETSA and TAFAS, $12\times$ than COSA) as closed-form quantile interpolation rather than iterative optimization; AVERT and Original are the only Pareto-optimal methods in the accuracy–speed plane (Appendix E).

Threshold sensitivity and shift response. AVERT gains most on the highest-shift benchmark and substantially on Air Quality ($0.431 \rightarrow 0.382$ on MLP), while leaving low-shift datasets untouched (Household Power: 0.035 vs. 0.035)—showing that the domain-shift gate prevents negative transfer and that the adaptation strength grows with $\mathcal{D}$. Sweeping each safe-return threshold by $\pm 30\%$ changes MAE by at most 11% in six of eight perturbations, with the only large effects where the W1 threshold crosses the dataset's own anchor distance (Appendix F).

Natural label shift. Because the main benchmark standardizes target labels with source statistics, we re-evaluate the signal on three datasets with natural temporal shifts (Air Quality, Beijing PM2.5, Household Power), evaluating entirely on the *raw* target scale. Across 42 adaptation decisions, the scale-normalized signal change tracks the true error change at $+0.723$ with 94% sign agreement, harmful adaptations raising the signal and helpful ones lowering it. The safeguards convert this into safety: AVERT never degrades MAE by more than 0.5%, whereas baselines inflate PM2.5 by 12–37% and Household Power by an order of magnitude (Appendix G).

4.6 Zero-Tuning Evaluation on an External Dataset

To test whether AVERT's configuration depends on its development benchmark, we evaluate the *identical* fixed configuration on Airfoil Self-Noise (UCI), which was not used for hyperparameter selection: MLP, CNN1D, and ResNet1D (3 seeds each) trained on a frequency-defined source split and adapted on the target split ($\mathcal{D} = 0.26$; genuine label and covariate shift); no hyperparameters are retuned. AVERT achieves the best average MAE of all eight methods (1.12 vs. 1.66 for Original and 1.31–1.57 for the baselines), improving over Original in 7 of 9 runs (improvements of 31–55%; the two remaining runs degrade by at most 4.2%); the adaptive baselines' worst-case degradations span 23–98%. Although the label-shift signature is strong ($\mathrm{W}_1(\hat{y}^t, y^s) = 2.0$–$3.8$), the gate correctly does not fire: the source model is accurate ($\mathrm{src_mae}/\sigma_s < 0.25$), and it is the moderate-anchor condition ($\mathrm{W}_1 < 2.0$) that fails. This is genuine shift, not confusion, so AVERT refuses adaptation only in the confusion regime for which it is designed.

5 Conclusion

We presented AVERT, a safe test-time adaptation framework for regression built on three principles: *W1-anchoring*, a label-free verification signal for regression TTA (correlation $+0.964$ with true MAE change, Figure 1); *verification-gated adaptation* by quantile transport; and *no-harm safeguards*—a safe-return gate and an anomaly clamp that on the two label-shift-confusion benchmarks refuse adaptation based on three computable statistics. Across 15 benchmarks and three backbones, AVERT significantly outperforms Original and all six adaptation baselines and transfers to a fully external dataset with zero retuning. The design principle is to verify against the W1 anchor, adapt by quantile transport, and refrain from adaptation when unsafe.

Limitations and future work. (i) W1-anchoring assumes label-distribution consistency: on the raw scale the signal–error correlation drops to $+0.72$ (94% sign agreement); an online label-shift estimator with a streaming anchor is a natural extension. (ii) The gate thresholds are fixed ratios of computable statistics, insensitive to $\pm 30\%$ perturbations except at the two regime crossings (Appendix F); deriving them from source-error quantiles would yield self-calibration. (iii) A conditional residual model could exploit the source residual's per-sample structure. Code and configurations will be released.

Declaration on the Use of AI

Large language models (LLMs) were employed exclusively to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design, data analysis, or any intellectual contribution were generated through LLMs, nor did such assistance influence any reported results.

Ethics Statement

This work uses only publicly available benchmark datasets (UCI Machine Learning Repository and OpenML); no human subjects or personally identifiable data are involved, and no new data were collected. The purpose of this work is to make regression test-time adaptation safer by reducing the risk of deploying silently degraded models, and we foresee no misuse potential beyond that inherent to standard machine-learning practice. There are no dataset-release, conflict-of-interest, sponsorship, discrimination/bias/fairness, privacy/security, or legal-compliance concerns raised by this study, and all experiments comply with the terms of the data sources used.

Reproducibility Statement

We provide everything needed to reproduce our results. The supplementary material contains the full training and evaluation code with the fixed hyperparameters and the per-dataset splits; the theoretical claims (proxy bound, finite-sample concentration, and the no-harm/bounded-gain guarantees) and their proofs are given in Appendix A; all 15 benchmark datasets and the external Airfoil Self-Noise dataset are publicly available (UCI Machine Learning Repository and OpenML), with the preprocessing described in Section 4.1. All experiments were run on a single NVIDIA GeForce RTX 3080 Laptop GPU (16 GB).

References

Kazuki Adachi, Shin'ya Yamaguchi, Atsutoshi Kumagai, and Tomoki Hamagami. Test-time adaptation for regression by subspace alignment. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=SXtl7NRyE5.

Xiao Chen, Jiazhen Huang, Zhiming Liu, Qinting Jiang, Fanding Huang, Jingyan Jiang, and Zhi Wang. Test-time distillation for continual model adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7593–7604, 2026.

Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. TabFSBench: Tabular benchmark for feature shifts in open environments. In *International Conference on Machine Learning (ICML)*, pp. 10025–10089, 2025.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

Muhammad Ghifary, David Balduzzi, W. Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2017.

Ziyu Gong, Jim Lim, and David I. Inouye. Expressive score-based priors for distribution matching with geometry-preserving regularization. In *International Conference on Machine Learning (ICML)*, 2025.

Georges Hébrail and Antoine Bérard. Individual household electric power consumption data set. UCI Machine Learning Repository, 2012.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=L7Z5wBMPrW.

Hyeongyu Kim, GeonHui Han, and Dosik Hwang. Buffer layers for test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025b. doi: 10.1609/aaai.v39i17.33965.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Taeckyung Lee, Sorn Chottananurak, Taesik Gong, and Sung-Ju Lee. AETTA: Label-free accuracy estimation for test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28643–28652, 2024. doi: 10.1109/CVPR52733.2024.02706.

Yingnan Liu, Rui Qiao, Mong-Li Lee, and Wynne Hsu. Test-time adaptation by causal trimming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Heitor R. Medeiros, Hossein Sharifi-Noghabi, Gabriel L. Oliveira, and Saghar Irandoust. Accurate parameter-efficient test-time adaptation for time series forecasting. In *Proceedings of the 2nd Workshop on Test-Time Adaptation at the International Conference on Machine Learning (ICML)*, 2025.

Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15319–15328, 2025. doi: 10.1109/CVPR52734.2025.01427.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, 2022.

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 (ICLR)*, 2023.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, 2016.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at Test Time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 57503–57522, 2025.

Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

Dequan Wang, Evan Shelhamer, Shaoting Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Qineng Wang, Ola Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhenyu Wang, Peter Bühlmann, and Zijian Guo. Distributionally robust learning for multi-source unsupervised domain adaptation. *The Annals of Statistics*, 54(2):570–596, 2026.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. DPCore: Dynamic prompt coreset for continual test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2025.

CONTENTS

A Theoretical Analysis **14**

A.1 Proxy Bound and Ranking Preservation 14

A.2 Finite-Sample Concentration 14

A.3 No-Harm and Bounded Gain of the Safe-Return Gate 15

A.4 Per-Sample Blending and Clamp Properties 15

B Full Per-Backbone Results **16**

C Case Study: Sensor Drift Under Deployment **17**

D Reliability Boundary of the Verification Signal **17**

E Efficiency Analysis **18**

F Threshold Sensitivity: Full Sweep **19**

G W1 Signal under Natural Label Shift: Full Protocol and Diagnostics **19**

This appendix contains the theoretical analysis (Appendix A) and the extended experimental results (Appendices B–G) referenced in the main text.

A Theoretical Analysis

This appendix collects the formal statements underlying the design arguments in the main text. Throughout, $v(\hat{y}) := \mathrm{W}_1(\hat{y}, y^s)$ denotes the verification signal.

A.1 Proxy Bound and Ranking Preservation

Assumption 1 (Bounded label shift)**.** The source and target label distributions satisfy $\mathrm{W}_1(y^t, y^s) \leq \delta$ for a constant $\delta \geq 0$.

Definition 2 (Sortedness cost)**.** For any prediction $\hat{y}$ in target space, let $\kappa(\hat{y}) := \mathrm{MAE}(\hat{y}, y^t) - \mathrm{W}_1(\hat{y}, y^t) \geq 0$ be the *sortedness cost*: the excess of the ℓ_1 error over the minimal transport cost. It vanishes iff $(\hat{y}, y^t)$ admits a monotone coupling, i.e., the ranks of $\hat{y}$ agree with the ranks of y^t.

Theorem 3 (Verification proxy bound)**.** *Under Assumption 1, for* any *target-space prediction $\hat{y}$ with verification signal $v(\hat{y}) := \mathrm{W}_1(\hat{y}, y^s)$:*

$$\begin{aligned} |\,\mathrm{W}_1(\hat{y}, y^t) - v(\hat{y})\,| \;&\leq\; \delta, \\ \mathrm{MAE}(\hat{y}, y^t) \;&\leq\; v(\hat{y}) + \delta + \kappa(\hat{y}). \end{aligned} \tag{A.8}$$

Proof. Both statements follow from the triangle inequality of W_1: $\mathrm{W}_1(\hat{y}, y^t) \leq \mathrm{W}_1(\hat{y}, y^s) + \mathrm{W}_1(y^s, y^t) = v(\hat{y}) + \delta$ and $\mathrm{W}_1(\hat{y}, y^t) \geq \mathrm{W}_1(\hat{y}, y^s) - \mathrm{W}_1(y^s, y^t) = v(\hat{y}) - \delta$; the second bound in (A.8) then follows from Definition 2. □ □

Corollary 4 (Ranking preservation)**.** *Let $\mathcal{C}$ be a finite set of candidate predictions. If $\delta < \frac{1}{2}\min_{\hat{y}\neq\hat{y}'} |v(\hat{y}) - v(\hat{y}')|$, then $\arg\min_{\hat{y}\in\mathcal{C}} v(\hat{y}) = \arg\min_{\hat{y}\in\mathcal{C}} \mathrm{W}_1(\hat{y}, y^t)$; in general, minimizing v minimizes $\mathrm{W}_1(\hat{y}, y^t)$ up to a $\pm\delta$ slack. Hence, when the label shift δ is small relative to the spread of the verification signal,* selecting the candidate with the smallest verification signal is equivalent to selecting the one closest to the target labels in W_1.

Theorem 3 and Corollary 4 formalize why the label-free verification signal is reliable: it approximates the true target error up to the label-shift magnitude δ and the sortedness cost κ, both of which are small on well-calibrated regression models (QTB is a monotone transformation, so it does not increase κ). Figure 1 confirms this empirically.

A.2 Finite-Sample Concentration

Theorem 5 (Verification ranking concentration)**.** *Let $y^s_1, \ldots, y^s_{n_s} \sim P_s$ and $\hat{y}^t_1, \ldots, \hat{y}^t_{n_t} \sim P_{\hat{y}}$ be i.i.d. samples, with both distributions supported on an interval of length R (after standardization). Define the* empirical *verification signal $\hat{v}(\hat{y}) = \mathrm{W}_1(\hat{y}, \hat{y}^s)$ and the population signal $v(\hat{y}) = \mathrm{W}_1(P_{\hat{y}}, P_s)$, where $\hat{y}$ and $\hat{y}^s$ denote empirical distributions. For any two candidate predictions P_a, P_b whose population signals differ by $\Delta = |v(P_a) - v(P_b)| > 0$, the empirical ordering agrees with the population ordering with probability at least $1 - \delta$, where*

$$\delta \;=\; 4\exp\!\left(-\frac{\Delta^2}{8R^2}\,\min(n_s, n_t)\right), \tag{A.9}$$

i.e., $n_t = \Omega\big(\frac{R^2}{\Delta^2}\ln\frac{1}{\delta}\big)$ target predictions suffice for a $1-\delta$ correct ranking.

Proof. For a 1D distribution P supported on an interval of length R, $\mathrm{W}_1(\hat{P}_n, P) = \int |\hat{F}_n - F| \leq R \cdot \sup_x |\hat{F}_n - F|$, and the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) gives $P(\sup_x |\hat{F}_n - F| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$, hence $P(\mathrm{W}_1(\hat{P}_n, P) \geq t) \leq 2e^{-2nt^2/R^2}$. By the triangle inequality, $|\hat{v}(P_a) - v(P_a)| \leq \mathrm{W}_1(\hat{P}_a, P_a) + \mathrm{W}_1(\hat{P}_s, P_s)$; ordering is preserved whenever both estimation errors stay below $\Delta/2$. A union bound over the two candidates and both sides with $t = \Delta/4$ gives (A.9). □ □

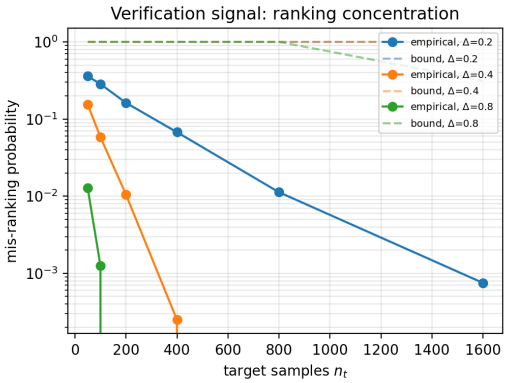

Figure A.2: Monte Carlo verification of Theorem 5: the empirical mis-ranking probability of the verification signal (solid) decays exponentially in the number of target samples n_t for signal separations $\Delta \in \{0.2, 0.4, 0.8\}$, far below the conservative bound of Eq. (A.9) (dashed). A batch of $n_t \approx 400$ suffices for 10^{-3} mis-ranking probability at $\Delta = 0.4$.

Theorem 5 provides the finite-sample analysis that backs the empirical correlation reported in the main text: the verification signal can be trusted to *rank* corrections as long as the target batch is large relative to $R^2/\Delta^2 \cdot \ln(1/\delta)$, which is exactly the regime in which AVERT operates (the domain-shift gate keeps adaptation away from degenerate small-Δ configurations). Figure A.2 verifies the bound numerically.

A.3 No-Harm and Bounded Gain of the Safe-Return Gate

Theorem 6 (No-harm and bounded gain of safe return)**.** *Let $\hat{y}^{\mathrm{SR}}$ be the output of the safe-return gate: $\hat{y}^{\mathrm{SR}} = \hat{y}^t$ if the gate fires (the safe-return condition of Algorithm 1 holds), and $\hat{y}^{\mathrm{SR}} = \hat{y}^{\mathrm{adapt}}$ otherwise. Under Assumption 1:*

1. ***(No harm.)*** *If the gate fires, $\mathrm{MAE}(\hat{y}^{\mathrm{SR}}, y^t) = \mathrm{MAE}(\hat{y}^t, y^t)$, i.e., the model is never worse than no adaptation.*
2. ***(Bounded gain.)*** *If adaptation is accepted, letting $\Delta v := v(\hat{y}^t) - v(\hat{y}^{\mathrm{adapt}}) > 0$,*

$$\mathrm{MAE}(\hat{y}^{\mathrm{SR}}, y^t) \;\le\; \mathrm{MAE}(\hat{y}^t, y^t) - \Delta v + 2\delta + \kappa(\hat{y}^{\mathrm{adapt}}). \tag{A.10}$$

Proof. (1) is immediate from the definition. For (2), from Theorem 3, $\mathrm{MAE}(\hat{y}^{\mathrm{adapt}}, y^t) \le v(\hat{y}^{\mathrm{adapt}}) + \delta + \kappa(\hat{y}^{\mathrm{adapt}})$ and $\mathrm{MAE}(\hat{y}^t, y^t) \ge \mathrm{W}_1(\hat{y}^t, y^t) \ge v(\hat{y}^t) - \delta$; subtracting gives the claim. □ □

Remark 7 (Interpretation). The safe-return gate converts a *measured* decrease Δv of the label-free verification signal into a *bounded* decrease of the true error: the accepted adaptation improves MAE by at least $\Delta v - 2\delta - \kappa$. Since QTB is a monotone transformation, $\kappa(\hat{y}^{\mathrm{adapt}}) \approx \kappa(\hat{y}^t)$ in practice, and on our benchmarks δ is small (Figure 1); hence the improvement is approximately Δv, consistent with the ablation results in Table 3.

A.4 Per-Sample Blending and Clamp Properties

In AVERT the per-sample weight is $\alpha_j = \mathrm{clip}\big(2\mathcal{D}^{1.5}(1 - c_j)\, s(n_t),\, 0.02,\, 0.5\big)$, where $c_j = \big[1 + (d_j/d_{\mathrm{med}})^{\gamma}\big]^{-1}$ is the transport confidence of sample j (transport distance d_j, batch median d_{med}, $\gamma = \mathrm{clip}(1 + 2\mathcal{D}, 1, 3)$) and $s(n_t) = \mathrm{clip}(\log_{10}(n_t{+}1)/3,\, 0.3,\, 1)$; hence $\alpha_j \in [0.02, 0.5]$ grows with the measured shift $\mathcal{D}$ and shrinks with the sample's transport distance.

Proposition 8 (Asymmetric Loss–Gain Bound)**.** *Let $\hat{y}^t$ have error $e^t = |\hat{y}^t - y|$, and $\hat{y}^{\mathrm{mix}}$ the routed correction with error e^{mix}. Under per-sample blending $\hat{y}_j^{\mathrm{final}} = (1 - \alpha_j)\hat{y}_j^t + \alpha_j \hat{y}_j^{\mathrm{mix}}$ with*

Table B.1: Main results (CNN1D backbone): MAE (3-seed mean) on 15 datasets. Lower is better. Best per row in **bold**, second-best underlined (ties share the same rank). Avg Rank (1=best, 8=worst, tied ranks averaged).

Dataset	Original	SSA	PETSA	TAFAS	COSA	CORAL	TENT	AVERT
AirQual	0.366	0.391	0.359	0.356	1.012	0.372	0.361	**0.350**
BeijingMS	0.186	0.264	0.259	0.283	0.805	0.238	0.192	**0.186**
CalHousing	**0.402**	0.404	0.413	0.404	0.797	0.403	0.402	0.402
Concrete	**0.448**	1.766	1.802	1.762	2.065	1.859	0.897	**0.448**
EnergyEff	1.746	1.114	1.122	1.090	1.381	1.537	**0.579**	0.776
GasTurbine	1.110	0.889	0.901	**0.886**	1.017	0.911	1.007	1.018
GTEmiss	0.411	0.437	0.477	0.410	0.516	0.435	0.411	**0.403**
Power	0.033	0.134	0.096	0.146	0.681	0.097	0.040	**0.033**
NWP-s24	0.659	0.516	0.544	0.557	0.833	**0.498**	0.558	0.613
NWP-s34	1.224	0.585	0.652	0.614	0.930	**0.498**	0.753	0.863
PGCB	**0.317**	1.905	1.928	1.893	1.953	1.940	1.088	**0.317**
SeoulBike	**0.769**	1.119	1.192	1.143	1.384	1.002	1.032	0.808
SGemm	**0.223**	0.504	0.594	0.805	0.601	0.794	0.247	0.239
SolarP2	0.246	0.272	0.237	0.234	0.921	**0.231**	0.238	0.235
Tetouan	0.704	0.847	0.847	0.822	0.937	0.844	**0.697**	0.698
Avg Rank	3.87	4.80	5.07	4.20	7.53	4.47	3.47	2.60

Table B.2: Main results (ResNet1D backbone): MAE (3-seed mean) on 15 datasets. Lower is better. Best per row in **bold**, second-best underlined (ties share the same rank). Avg Rank (1=best, 8=worst, tied ranks averaged).

Dataset	Original	SSA	PETSA	TAFAS	COSA	CORAL	TENT	AVERT
AirQual	0.379	0.363	0.372	**0.362**	0.993	0.378	0.371	0.369
BeijingMS	0.198	0.294	0.301	0.343	0.807	0.249	0.200	**0.198**
CalHousing	0.402	**0.402**	0.405	0.407	0.800	**0.402**	0.402	0.402
Concrete	**0.644**	1.697	1.798	1.769	1.877	1.870	1.061	**0.644**
EnergyEff	1.720	1.139	1.347	1.227	1.344	1.537	**0.587**	0.690
GasTurbine	1.145	0.920	**0.870**	1.015	0.997	0.880	1.091	1.128
GTEmiss	0.387	0.434	0.426	0.447	0.521	0.418	**0.378**	0.379
Power	**0.034**	0.106	0.119	0.176	0.684	0.101	0.035	0.035
NWP-s24	0.573	0.583	**0.476**	0.507	0.904	0.530	0.513	0.567
NWP-s34	0.784	0.704	0.607	0.572	0.848	**0.559**	0.695	0.709
PGCB	**0.411**	1.853	1.927	1.725	1.942	1.927	1.068	**0.411**
SeoulBike	**0.703**	1.326	1.262	1.236	1.417	0.957	1.236	0.818
SGemm	**0.237**	0.659	0.672	0.966	0.600	0.645	0.264	0.251
SolarP2	**0.201**	0.218	0.218	0.217	0.900	0.210	0.202	0.202
Tetouan	**0.500**	1.259	1.339	1.320	0.906	0.664	0.671	0.509
Avg Rank	3.60	4.77	5.33	5.00	7.07	4.10	3.27	2.87

$\alpha_j \in [0, \bar{\alpha}]$, $\bar{\alpha} = 0.5$,

$$\begin{aligned} e_j^{\text{final}} &\le (1-\alpha_j) e_j^t + \alpha_j e_j^{\text{mix}} \\ &\le e_j^t + \bar{\alpha}(e_j^{\text{mix}} - e_j^t). \end{aligned} \tag{A.11}$$

Thus, when the correction is harmful on a sample ($e_j^{\text{mix}} > e_j^t$), the error degrades by at most $0.5 \cdot (e_j^{\text{mix}} - e_j^t)$, while when it is beneficial ($e_j^{\text{mix}} < e_j^t$) the error improves by at least $\alpha_j(e_j^t - e_j^{\text{mix}})$. The anomaly-clamp bound is then explicit: a sample flagged by the clamp is replaced by the $0.85/0.15$ blend $\hat{y}_j^{\text{final}} = 0.85\,\hat{y}_j^t + 0.15\,\hat{y}_j^{\text{mix}}$, so $|\hat{y}_j^{\text{final}} - \hat{y}_j^t| = 0.15\,|\hat{y}_j^{\text{mix}} - \hat{y}_j^t|$ and $|\hat{y}_j^{\text{final}} - \hat{y}_j^{\text{mix}}| = 0.85\,|\hat{y}_j^{\text{mix}} - \hat{y}_j^t|$, and the worst-case per-sample error is strictly bounded.

Proposition 9 (W1 Validation Monotonicity)**.** *The W1 validation clamp (Eq. 6) reverts a correction monotonically in the amount by which it moves the prediction away from the source label distribution: the larger $\mathrm{W}_1(\hat{y}^{\text{mix}}, y^s) - \mathrm{W}_1(\hat{y}^t, y^s)$, the smaller the retained fraction* frac*. Consequently, corrections that align with the W1 anchor are preserved, and corrections that misalign are suppressed, with no manual per-dataset tuning.*

B Full Per-Backbone Results

The CNN1D and ResNet1D per-dataset MAE tables and the per-dataset bar chart corresponding to Table 1.

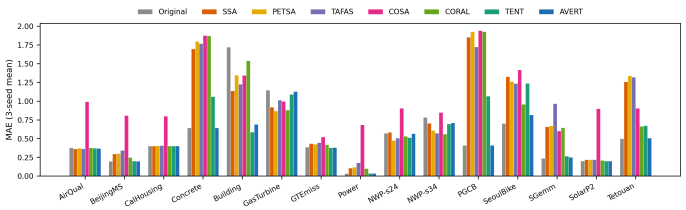

Figure B.1: MAE of all eight methods across the fifteen datasets (ResNet1D backbone). AVERT (blue) matches or improves Original on 10 of 15 datasets and never degrades MAE by more than 16.4%, while avoiding the large degradations seen in COSA.

C Case Study: Sensor Drift Under Deployment

To demonstrate the deployment value of the verification signal, we simulate sensor drift for three physical-sensing benchmarks (Air Quality, Beijing PM2.5, Concrete): a model is trained on normal sensor readings and then deployed on target inputs whose sensor readings drift (zero-drift bias and sensitivity loss), while the underlying quantities being measured stay the same. Because the drift protocol uses its own source–target split (single seed), the no-drift baselines here differ from the main-table values (e.g., Concrete 0.527 vs. 0.433 in Table 1). Table C.1 reports MAE under six drift scenarios for three strategies: no adaptation, naive quantile transport (unconditional QTB), and AVERT.

The three datasets realize three qualitatively different regimes. *Concrete (unsafe regime):* here the source model is near-perfect and the target predictions shift moderately — precisely the label-shift-confusion condition — so unconditional QTB is always harmful, degrading MAE by 14–36% across scenarios. The safe-return gate detects the condition from the verification signal and refuses adaptation; AVERT matches the no-adaptation baseline and beats naive QTB in every scenario. *Beijing PM2.5 (adaptable regime):* drift induces genuine shape/location shift that QTB corrects, and the gate correctly stays silent; AVERT reduces MAE by 9–29% over no adaptation. *Air Quality (mixed regime):* the gate is mostly silent and QTB improves MAE, but under a 0.5-std bias drift the gate fires conservatively and is marginally worse than naive (0.759 vs. 0.686) ; this is the only scenario in which unconditional adaptation is marginally better; the same unconditional adaptation is harmful on Concrete. This asymmetry is a consequence of the safe-return gate: it avoids large degradations at the cost of occasional small conservatism.

D Reliability Boundary of the Verification Signal

The verification signal is reliable only within a bounded region of adaptation, exactly as the proxy analysis Appendix A.1 predicts: $|\, W_1(\hat{y}, y^t) - v(\hat{y}) \,| \leq \delta$ ties the signal to the true error only when the label shift δ is small relative to the signal spread. To make this boundary explicit, we record for every dataset and every method the change in the verification signal, $\Delta v = v(\hat{y}^{\text{adapt}}) - v(\hat{y}^t)$, together with the true MAE change, ΔMAE, and correlate the two across datasets (MLP backbone, seed 42).

For *AVERT*, the signal change tracks the true error change at $+0.964$ across the 15 benchmark datasets — consistent with the $+0.964$ correlation in Figure 1 — because its gate-constrained adaptation keeps Δv small and directed (Figure 1). For *SSA*, *PETSA*, and *TAFAS*, the correlation is *negative* (-0.60, -0.61, -0.59): these methods aggressively pull the prediction distribution toward the source prediction distribution, shrinking $v(\hat{y})$ even when the true target labels lie far from the source labels, so a decreasing signal is associated with *increasing* error. COSA, whose output-space adapter is more conservative, shows a negative correlation (-0.36). The contrast between AVERT and the unconditional methods indicates that signal reliability is a property of the *adaptation magnitude*, not of the method.

This boundary explains the behavior of the verification signal: the verification signal does not certify *any* correction; it certifies corrections that stay inside its reliability region. AVERT is designed to operate inside that region (domain-shift gate, adaptive blending, safe-return), which is why it is the only method whose signal—error correlation is strongly positive. Unconditional methods leave

Table C.1: Case study: sensor drift under deployment (MLP backbone). A model is trained on normal sensor readings and deployed under injected drift (zero-drift bias b and sensitivity loss σ). "Naive QTB" applies quantile transport unconditionally; "AVERT" additionally runs the safe-return gate. On Concrete, unconditional adaptation is always harmful (naive degrades MAE by 14–36%) and the gate refuses it; on Beijing PM2.5, adaptation is beneficial and the gate stays silent; on Air Quality the gate is mostly silent but occasionally conservative (bias 0.5 case), the only case where naive is marginally better. Best per row in **bold**.

Drift scenario	Original	Naive QTB	AVERT	Safe-return
Concrete (unsafe regime)				
baseline(no drift)	**0.527**	0.602	**0.527**	yes
sensor bias 0.5std	**0.448**	0.540	**0.448**	yes
sensor bias 1.0std	**0.464**	0.631	**0.464**	yes
sensor noise 1.0std	**0.595**	0.689	**0.595**	yes
sensor noise 2.0std	**0.807**	0.917	**0.807**	yes
combined drift	**0.521**	0.621	**0.521**	yes
Beijing PM2.5 (adaptable regime)				
baseline(no drift)	**0.555**	0.555	0.555	—
sensor bias 0.5std	1.108	**1.012**	**1.012**	—
sensor bias 1.0std	2.697	**1.925**	**1.925**	—
sensor noise 1.0std	1.855	**1.562**	**1.562**	—
sensor noise 2.0std	3.885	**2.768**	**2.768**	—
combined drift	1.827	**1.506**	**1.506**	—
Air Quality (mixed regime)				
baseline(no drift)	0.452	**0.442**	**0.442**	—
sensor bias 0.5std	0.759	**0.686**	0.759	yes
sensor bias 1.0std	0.752	**0.631**	**0.631**	—
sensor noise 1.0std	0.668	**0.650**	**0.650**	—
sensor noise 2.0std	1.060	**0.975**	**0.975**	—
combined drift	0.633	**0.623**	**0.623**	—

the region and fail when the signal stops tracking the error. This analysis provides a quantitative, theory-backed explanation of the failure modes in Section 3.

E Efficiency Analysis

Table E.1 reports the average computation time. AVERT runs in 0.26s on average, 17× faster than SSA (4.46s), 14× faster than PETSA (3.51s), and 14× faster than TAFAS (3.70s), and 12× faster than COSA (2.96s). AVERT uses a single closed-form quantile interpolation with a per-dimension loop, avoiding iterative gradient-based optimization. Figure E.1 shows AVERT and Original are the only Pareto-optimal methods: AVERT achieves the best cross-backbone average rank (2.78) while being about 12× faster than the fastest competing adaptation method (COSA); all other adaptation methods are dominated.

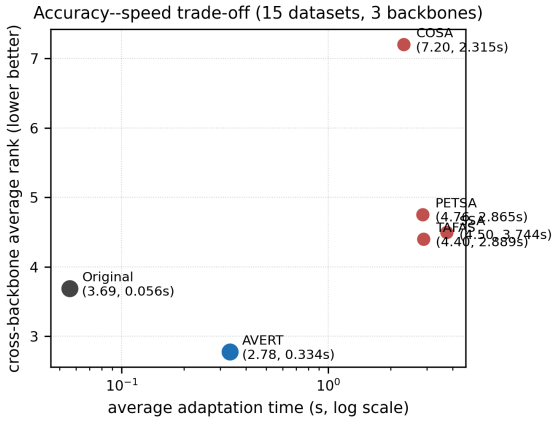

Figure E.1: Accuracy-speed trade-off: average rank (over the 15 benchmark datasets and three backbones) versus average adaptation time. Lower-left is better. AVERT and Original are the only Pareto-optimal methods.

Table E.1: Efficiency comparison (averaged across datasets). Bold indicates the best per column.

Method	Avg Time (s)	Avg Mem (MB)
Original	**0.07**	**0.01**
SSA	4.46	0.5
PETSA	3.51	9.1
TAFAS	3.70	18.2
COSA	2.96	0.7
AVERT	**0.26**	7.9

F Threshold Sensitivity: Full Sweep

AVERT provides the largest gain on Energy Efficiency ($\mathcal{D} = 0.45$, the highest shift in the benchmark), a substantial gain on Air Quality, and no degradation on low-shift datasets, confirming that the domain-shift gate prevents negative transfer. All gate thresholds are fixed globally (Section 3) rather than tuned per dataset: on low-shift datasets the gate keeps AVERT's predictions identical to Original (Household Power: 0.035 vs. 0.035 on MLP, 0.033 vs. 0.033 on CNN1D, 0.035 vs. 0.034 on ResNet1D), and the adaptation strength α grows continuously with the measured shift $\mathcal{D}$, so behavior near the gate boundary degrades gracefully rather than abruptly. Throughout, the domain shift $\mathcal{D}$ is computed as the mean Jensen–Shannon distance (50 equal-width bins) between the source and target histograms, averaged over the first $\min(d, 10)$ feature dimensions—a non-parametric covariate-shift score in $[0, 1]$ (the per-dimension KS statistic of Section 3.3 is a separate measure used for dataset selection and for the feature-transport gate δ_k). We additionally sweep each safe-return threshold one at a time by $\pm 30\%$ ($\mathrm{src_mae}/\sigma_s$: $0.175/0.325$; $|\Delta|/\sigma_s$: $0.56/1.04$; W1: $1.4/2.6$; clamp tolerance: $0.735/1.365$) on three datasets chosen to exercise each mechanism (Concrete: label-shift confusion; Energy Efficiency: high-shift adaptation; SGEMM: near-perfect source; MLP, seed 42). Every perturbation except the two W1-threshold crossings below changes MAE by at most 11% and none causes major degradation. The only large effects occur when the W1 threshold crosses the dataset's own anchor distance: shrinking it to 1.4 disables the safe-return protection on Concrete ($0.390 \rightarrow 0.623$), and loosening it to 2.6 lets the gate fire spuriously on Energy Efficiency ($0.711 \rightarrow 1.201$). Both effects occur at a regime boundary: the thresholds lie on a performance plateau, and the sensitive direction corresponds to the discriminative core identified in the ablation study.

G W1 Signal under Natural Label Shift: Full Protocol and Diagnostics

The main benchmark standardizes target labels with source statistics, and the W1 anchor's near-perfect correlation could be an artifact of this protocol. We therefore re-evaluate the signal on three datasets with natural temporal shifts—Air Quality, Beijing PM2.5, and Household Power—training on the source period and evaluating entirely on the *raw* target scale, without standardizing target labels (natural label drift 0.05–$0.18\,\sigma_s$).

Across all methods, seeds, and datasets (42 adaptation decisions), the scale-normalized signal change $\Delta v/\sigma_s$ tracks the true error change $\Delta\mathrm{MAE}/\sigma_s$ at correlation $+0.723$, and the *sign* of Δv agrees with the sign of $\Delta\mathrm{MAE}$ on 94% of materially changed decisions: harmful adaptations move the signal up by $+0.37\,\sigma_s$ on average, helpful ones move it down by $-0.14\,\sigma_s$. The signal therefore remains decisively discriminative between adaptation that helps and adaptation that harms even when the evaluation protocol makes no use of source-statistic normalization of target labels. The safeguards convert this discrimination into safety: AVERT never degrades MAE by more than 0.5% anywhere in the experiment (worst case across baselines: $+0.69\,\sigma_s$; SSA/PETSA/TAFAS/COSA inflate Beijing PM2.5 by 12–37% and Household Power by more than an order of magnitude), and it improves Air Quality by 11–15%. The W1 anchor thus remains a conservative but still informative signal under raw-scale natural-shift evaluation, with the safe-return gate and anomaly clamp reducing the residual risk. The sign agreement is significant (sign test: 34 of 36 materially changed decisions, $p = 1.9 \times 10^{-8}$), and harmful adaptations exhibit systematically larger signal increases than helpful ones (Mann–Whitney $U = 128$, $p = 1.7 \times 10^{-5}$; mean $\Delta v = +0.37\,\sigma_s$ harmful vs. $-0.14\,\sigma_s$ helpful). A trigger analysis of the two gate layers explains AVERT's zero-

degradation profile here: the safe-return gate never fires on these datasets — their diagnostics ($\mathrm{src_mae}/\sigma_s = 0.02$–$0.43$, $|\Delta|/\sigma_s = 0.006$–$0.28$, $\mathrm{W}_1(\hat{y}^t, y^s) = 0.09$–$0.29$) place them outside the label-shift-confusion regime, i.e., the gate correctly remains silent instead of refusing useful adaptation — and TT-SRA activates exactly on the datasets satisfying its design rule ($\mathcal{D} > 0.15$ with a non-perfect source model: Air Quality), whereas on Household Power the near-perfect source model ($\mathrm{src_mae}/\sigma_s \approx 0.02$) keeps it off.