PER-SAMPLE OR NOT: A GRANULARITY CRITERION FOR TEST-TIME CORRECTION IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A linear head on a frozen visual foundation model provides strong AI-generated image (AIGI) detection, but its performance degrades under unseen generators, content domains, and dissemination shifts. We first establish a rank-invariance criterion showing that test-time corrections that only adjust thresholds or apply sample-shared feature translations cannot improve AUROC because they preserve the relative score ordering of real and generated samples. Motivated by the channel-wise non-uniformity and target specificity of forensic feature shifts, we propose Forensic-Shift Conditional Test-Time Affine Adaptation (FS-CTAA). Trained solely on source data with simulated forensic shifts, FS-CTAA predicts sample-specific channel-wise affine parameters from each test feature, the source training-set feature mean, and the current test-subset feature mean, requiring neither target labels nor gradient updates. Across four heterogeneous frozen backbones, FS-CTAA improves macro-average AUROC by 1.67–3.49 percentage points and, averaged over three training seeds, outperforms a sample-shared modulator trained under the same protocol by 1.22–1.58 percentage points, showing an additional ranking benefit from sample-specific rather than merely target-conditioned adaptation. Gains concentrate on challenging shifts where the frozen probe leaves substantial ranking errors, while performance remains stable on already near-saturated cases. FS-CTAA further improves balanced accuracy by 2.3–8.5 percentage points over the frozen probe and exceeds the threshold-only target-label oracle on three backbones. These results establish sample-conditioned feature adaptation as an effective approach to robust AIGI detection under distribution shifts.

1 INTRODUCTION

AI-generated image (AIGI) detection underpins content authentication, platform governance, and digital forensics (Verdoliva, 2020; Corvi et al., 2023; Guarnera et al., 2020). A linear head on a frozen visual foundation model transfers better than a backbone adapted to one detection dataset (Ojha et al., 2023; Zhou et al., 2026), yet still degrades under new generators, content domains, and dissemination conditions that training data cannot enumerate in advance (Guillaro et al., 2025; Li et al., 2025). We therefore study correcting such a fixed detector from unlabeled target images alone. Test-time methods adapt normalization, parameters, statistics, scores, or thresholds (Wang et al., 2020a; Liang et al., 2020; Schneider et al., 2020; Yang et al., 2026; Li et al., 2026). Each test image receives a scalar score and is labelled generated above a threshold, the operating point; AUROC is the fraction of generated-real pairs the resulting score ordering places correctly, so it depends on that ordering alone. A correction can therefore raise accuracy in two ways: by moving the operating point, which leaves the ordering intact, or by reordering images. Label-free threshold rules are of the first kind (Yang et al., 2026). Only reordering changes AUROC and can lift accuracy beyond the best available threshold.

We summarize a distribution by its mean feature: $\bar{\boldsymbol{\mu}}_s$ over the source training images, computed once before deployment, and $\boldsymbol{\mu}_t$ over the unlabeled target images seen at inference. Their absolute difference in one channel c is the displacement $\delta_c = |\bar{\mu}_{s,c} - \mu_{t,c}|$. On SigLIP2 the top 10% of channels carry 48.3–52.8% of it, yet these top-10% sets only partly overlap across target subsets (mean

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.

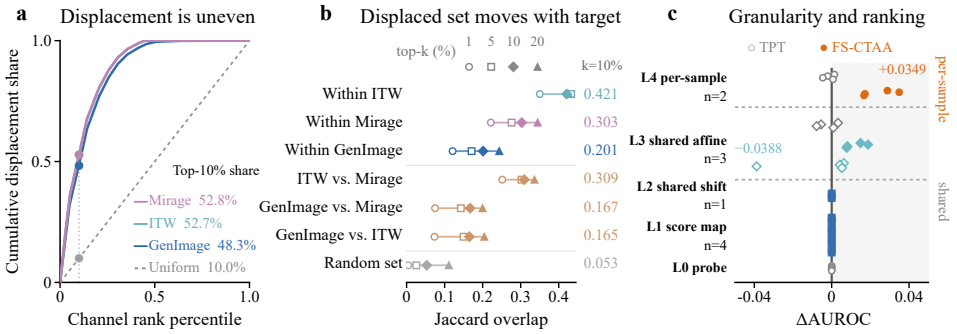

Figure 1: Target shift and correction granularity in a frozen AIGI detector. (a) Cumulative share of the per-channel mean displacement and (b) mean Jaccard overlap of the top-$k\%$ displaced channel sets on SigLIP2; annotations in (b) show $k = 10\%$. (c) ΔAUROC from Table 1 for all four backbones; filled diamonds/circles denote the trained shared modulator/FS-CTAA. Panel details in Section A.7.

Jaccard overlap 0.17–0.42 against approximately 0.05 at random; Figure 1a,b). These observations motivate channel-wise correction conditioned on the target distribution, while leaving open whether a shared correction is sufficient (Section A.7).

A rank-invariance criterion identifies corrections that preserve sample ordering. A correction is rank-invariant when the corrected score is a strictly increasing function of the original score; every pair of test images then keeps its order, so AUROC is unchanged exactly. Checkable from the functional form alone, the criterion induces a hierarchy under a frozen linear head: score calibration (L1), shared feature translation (L2), shared channel-wise correction (L3), and sample-conditioned correction (L4). It certifies L1 and L2 within a target subset, where monotone score calibration and mean-translation alignment preserve ordering. Both L3 and L4 can reorder. Under the same training protocol, sample-conditioned correction yields larger AUROC gains than the trained shared modulator on all four backbones (Figure 1c).

Guided by this criterion, we propose Forensic-Shift Conditional Test-Time Affine Adaptation (FS-CTAA). Source-side episodes, each applying one simulated condition such as compression or re-encoding, train a modulator between the frozen encoder and head. For each image in an unlabeled target subset, the modulator receives three inputs, the image's own feature and the two mean features $\bar{\boldsymbol{\mu}}_s$ and $\boldsymbol{\mu}_t$ above, and predicts a channel-wise scale and offset. The threshold is the mean corrected score; no model parameters are updated at test time.

- **A rank-invariance criterion and granularity hierarchy.** It certifies that monotone score recalibration and shared feature translation cannot change AUROC under a frozen linear head, so their accuracy gains are operating-point effects, confirmed to numerical precision on four backbones.
- **A forensic-shift correction framework.** FS-CTAA predicts a per-image channel-wise correction from the source and target means, learned from source-side episodes and anchored at the rank-invariant mean translation.
- **A portable plug-in.** With backbone and head unchanged and no target-time gradients, it raises AUROC on four backbones by 1.67–3.49 percentage points over the frozen probe and, averaged over three training seeds, by 1.22–1.58 percentage points over a shared modulator trained under the same protocol, with balanced accuracy above the frozen probe's label-oracle calibration on three backbones.

2 Related Work

Generalizable AIGI detectors use forensic traces or transferable visual representations (Wang et al., 2020b; Frank et al., 2020; Wang et al., 2023; Tan et al., 2024b; Chen et al., 2024; Radford et al., 2021). Frozen foundation-model features support strong linear probes, whereas specialized heads and low-rank adaptation (LoRA) fine-tuning can reduce cross-domain transfer (Ojha et al., 2023;

Yan et al., 2025; 2024; Chen et al., 2025; Guo et al., 2025; Zhou et al., 2026), so we keep representation and head fixed. Training-time robustness broadens generator coverage, augments transformations, or removes format and compression shortcuts before deployment (Guillaro et al., 2025; Park & Owens, 2025; Grommelt et al., 2024; Li et al., 2025); our setting is complementary, since correction uses the unlabeled target distribution after deployment.

Test-time adaptation updates model parameters, normalization, or prompts (Wang et al., 2020a; Niu et al., 2022; 2023; Sun et al., 2020; Wang et al., 2022; Liang et al., 2020; Chen et al., 2022; Nguyen-Le et al., 2025; Shu et al., 2022; Li et al., 2026). Gradient-free alternatives adjust feature statistics, scores, or thresholds (Schneider et al., 2020; Yang et al., 2026), or replace the classifier with target-time prototypes (Iwasawa & Matsuo, 2021), which the frozen head rules out. FS-CTAA instead predicts a feature correction without target-time optimization. Its affine form follows conditional modulation (Perez et al., 2018; Huang & Belongie, 2017) and its anchor is the mean-alignment step of statistics-based adaptation (Schneider et al., 2020; Tomar et al., 2024), but it conditions that modulation on the individual feature as well as the two means, anchors it at the rank-invariant mean translation and penalizes departures from it, and draws its supervision from source-side forensic-shift episodes, which teach the module how to read a shift in the target statistics rather than enumerate future targets.

3 METHOD

3.1 RANK INVARIANCE AND CORRECTION GRANULARITY

Let E be a frozen encoder and $f(\boldsymbol{z}) = \boldsymbol{w}^\top \boldsymbol{z} + c$ its frozen linear head, which classifies a sample as generated when $f(\boldsymbol{z}) > \tau$; the threshold τ is the operating point. A target subset is the unlabeled test set from one target distribution; transductive inference sees the whole subset before classifying any image in it. For a source distribution $\mathcal{D}_s$ and a target subset $\mathcal{D}_t = \{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$ whose labels are never observed,

$$\boldsymbol{\mu}_s = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_s}[E(\boldsymbol{x})], \qquad \boldsymbol{\mu}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} E(\boldsymbol{x}_i^t), \tag{1}$$

Let T act on a representation before the frozen head, and let $s_i = f(\boldsymbol{z}_i)$ be the detection score for sample i.

Proposition 1 (Rank-invariance criterion)**.** *Suppose there exists a strictly increasing $g : \mathbb{R} \to \mathbb{R}$ with*

$$f(T(\boldsymbol{z})) = g(f(\boldsymbol{z})) \qquad \textit{for all } \boldsymbol{z}. \tag{2}$$

Then for every pair (i, j), $f(T(\boldsymbol{z}_i)) > f(T(\boldsymbol{z}_j)) \iff f(\boldsymbol{z}_i) > f(\boldsymbol{z}_j)$. Consequently every rank-only functional of the scores is identical before and after T, on any dataset and for any label assignment.

Proof. Strict monotonicity preserves pairwise comparisons and ties. AUROC depends only on the resulting counts of concordant, discordant, and tied pairs, so it is unchanged without assumptions on the data or labels. □

Proposition 1 is the invariance of rank statistics under strictly increasing maps, stated for the score function a correction induces; its use is that the condition is checkable from the functional form of T alone, and it is sufficient but not necessary. Let $\boldsymbol{\phi}$ denote the correction parameters, instantiated in Section 3.2 by the modulator weights $\boldsymbol{\theta}$, and $\mathcal{C}_t$ the target context. Beyond the uncorrected L0 probe, we define four granularities,

$$\begin{aligned} &\text{L1} \quad s_i' = h_t(s_i), && \text{L2} \quad \boldsymbol{z}_i' = \boldsymbol{z}_i + \boldsymbol{b}_t, \\ &\text{L3} \quad \boldsymbol{z}_i' = T_{\boldsymbol{\phi}(\mathcal{C}_t)}(\boldsymbol{z}_i), && \text{L4} \quad \boldsymbol{z}_i' = T_{\boldsymbol{\phi}(\boldsymbol{z}_i, \mathcal{C}_t)}(\boldsymbol{z}_i), \end{aligned} \tag{3}$$

where h_t is strictly increasing. L1 is monotone, while L2 adds the same constant $\boldsymbol{w}^\top \boldsymbol{b}_t$ to every score. Proposition 1 therefore certifies both as rank invariant within a target subset, where h_t and $\boldsymbol{b}_t$ are fixed; the main transductive rankings are evaluated within subsets. The rank-invariance guarantee assumes a fixed context for the compared scores; it does not apply automatically to scores emitted

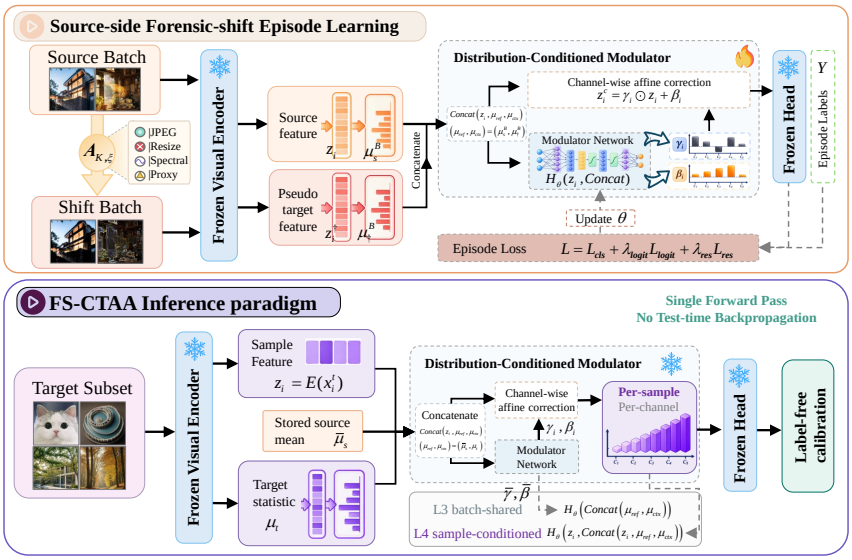

Figure 2: FS-CTAA. Source-side episodes pair each batch with a pseudo-target under one forensic condition; the modulator maps each representation and the source and target statistics to per-sample channel-wise affine parameters, with the modulator weights fixed at test time.

at different times under a changing streaming context. L3 shares one operator within a target subset, but a channel-wise scale $\boldsymbol{\gamma}$ reweights the head to $\boldsymbol{\gamma} \odot \boldsymbol{w}$ and can change score order: with $\boldsymbol{w} = (1, 1)$, $\boldsymbol{z}_1 = (2, 0)$, and $\boldsymbol{z}_2 = (0, 1)$, the shared scaling $\boldsymbol{\gamma} = (0.1, 3)$ maps the scores $(2, 1)$ to $(0.2, 3)$ and reverses the pair. L4 additionally conditions its parameters on each representation. Neither level guarantees an improvement.

Rank invariance ends between L2 and L3, whereas shared parameterization ends between L3 and L4. The hierarchy is defined by parameter conditioning: an L3 operator may preserve ranking by coincidence, and an arbitrarily complex shared operator remains L3 as long as all samples receive the same parameters. We call the rank-invariance test together with this hierarchy the granularity criterion: it certifies L1 and L2 as unable to change AUROC, so any accuracy they gain is an operating-point effect, and admits L3 and L4. Whether sample conditioning is needed is left to Section 4.2. FS-CTAA (Figure 2) realizes L4 as a single feed-forward prediction from a source-trained plug-in module placed between the encoder and its linear head. It requires no target-time gradient updates.

3.2 FS-CTAA

Forensic-Shift Episode Learning. Because target labels are unavailable, we construct paired correction supervision from source data. For a batch $\mathcal{B} = \{(\boldsymbol{x}_i^s, y_i)\}_{i=1}^B$ of size B, we draw one forensic condition $\mathcal{A}_{k,\xi}$ from four families, namely compression, resize-reencoding, spectral-response adjustment, and generator-style proxy, and apply it to the whole batch, with transformation details and parameter ranges given in Section A.2,

$$\boldsymbol{z}_i^s = E(\boldsymbol{x}_i^s), \quad \boldsymbol{z}_i^\dagger = E(\mathcal{A}_{k,\xi}(\boldsymbol{x}_i^s)), \quad \boldsymbol{\mu}_s^{\mathcal{B}} = \frac{1}{B}\sum_i \boldsymbol{z}_i^s, \quad \boldsymbol{\mu}_\dagger^{\mathcal{B}} = \frac{1}{B}\sum_i \boldsymbol{z}_i^\dagger. \tag{4}$$

One condition per batch makes $\boldsymbol{\mu}_\dagger^{\mathcal{B}}$ a coherent target-like context. Sample identities and labels remain aligned, so the paired features supervise only the modulator. The encoder and linear head remain frozen.

Distribution-Conditioned Modulator. The modulator maps a representation, a source reference $\boldsymbol{\mu}_{\mathrm{ref}}$, and a target context $\boldsymbol{\mu}_{\mathrm{ctx}}$ to channel-wise affine parameters,

$$\begin{aligned}
&\boldsymbol{h}_i = H_{\boldsymbol{\theta}}(\mathrm{concat}(\boldsymbol{z}_i, \boldsymbol{\mu}_{\mathrm{ref}}, \boldsymbol{\mu}_{\mathrm{ctx}}))\,, \\
&\boldsymbol{\gamma}_i = G_{\boldsymbol{\gamma}}(\boldsymbol{h}_i), \qquad \boldsymbol{\beta}_i = \boldsymbol{\mu}_{\mathrm{ref}} - \boldsymbol{\mu}_{\mathrm{ctx}} + G_{\boldsymbol{\beta}}(\boldsymbol{h}_i), \\
&\boldsymbol{z}_i^c = T_{\boldsymbol{\theta}}(\boldsymbol{z}_i; \boldsymbol{\mu}_{\mathrm{ref}}, \boldsymbol{\mu}_{\mathrm{ctx}}) = \boldsymbol{\gamma}_i \odot \boldsymbol{z}_i + \boldsymbol{\beta}_i.
\end{aligned} \tag{5}$$

Training evaluates equation 5 at $(\boldsymbol{z}_i^\dagger; \boldsymbol{\mu}_s^{\mathcal{B}}, \boldsymbol{\mu}_\dagger^{\mathcal{B}})$; inference uses $(E(\boldsymbol{x}_i^t); \bar{\boldsymbol{\mu}}_s, \boldsymbol{\mu}_t)$, where $\bar{\boldsymbol{\mu}}_s$ is the empirical estimate of $\boldsymbol{\mu}_s$ in equation 1 over the source training set. At initialization, $\boldsymbol{\gamma}_i = \mathbf{1}$ and $G_{\boldsymbol{\beta}}(\boldsymbol{h}_i) = \mathbf{0}$, so the module reduces to the rank-invariant L2 mean translation, and training learns a sample-conditioned departure from this anchor. Here, $\boldsymbol{\theta}$ instantiates $\boldsymbol{\phi}$ in equation 3. At inference, $\boldsymbol{\mu}_{\mathrm{ref}} = \bar{\boldsymbol{\mu}}_s$ and $\boldsymbol{\mu}_{\mathrm{ctx}} = \boldsymbol{\mu}_t$. Two L3 controls follow from the same family. The trained shared modulator $H_{\boldsymbol{\theta}}(\boldsymbol{\mu}_{\mathrm{ref}}, \boldsymbol{\mu}_{\mathrm{ctx}})$ receives only the two statistics and outputs one pair $(\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t)$ per target subset; it is trained exactly as FS-CTAA and is the fair shared counterpart. Shared FS-CTAA averages the predicted $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$ of the per-sample module within a target subset after training and is a post-hoc control.

3.3 Training and Transductive Inference

Training compares the full correction with its L2 anchor, the rank-invariant fallback from which the modulator departs, $\boldsymbol{z}_i^g = \boldsymbol{z}_i^\dagger + \boldsymbol{\mu}_s^{\mathcal{B}} - \boldsymbol{\mu}_\dagger^{\mathcal{B}}$. We interpolate between them using one coefficient per episode,

$$\widehat{\boldsymbol{z}}_i = (1 - \alpha_{\mathcal{B}})\boldsymbol{z}_i^g + \alpha_{\mathcal{B}}\boldsymbol{z}_i^c, \qquad \alpha_{\mathcal{B}} \sim p_\alpha. \tag{6}$$

Inference always uses the full correction. The training objective is

$$\begin{aligned}
&\mathcal{L}_{\mathrm{cls}} = \frac{1}{B}\sum_i \ell_{\mathrm{bin}}(f(\widehat{\boldsymbol{z}}_i), y_i), \qquad \mathcal{L}_{\mathrm{logit}} = \frac{1}{B}\sum_i \big(f(\widehat{\boldsymbol{z}}_i) - \mathrm{sg}(f(\boldsymbol{z}_i^s))\big)^2, \\
&\mathcal{L}_{\mathrm{res}} = \frac{1}{B}\sum_i \|\boldsymbol{z}_i^c - \boldsymbol{z}_i^g\|_2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{logit}}\mathcal{L}_{\mathrm{logit}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}}.
\end{aligned} \tag{7}$$

Here, ℓ_{bin} is the binary classification loss and sg is stop-gradient. The three terms preserve labels, align corrected and clean logits, and regularize departure from the L2 anchor; only $\boldsymbol{\theta}$ receives gradients (details in Section A.2).

Transductive inference. We compute $\boldsymbol{\mu}_t$, apply equation 5, and obtain corrected scores $s_i^c = f(\boldsymbol{z}_i^c)$. We select the operating point without labels by directly setting the threshold to the mean corrected score,

$$\tau_t^\star = \frac{1}{N_t}\sum_{i=1}^{N_t} s_i^c, \qquad \widehat{y}_i = \mathbb{I}[s_i^c > \tau_t^\star]. \tag{8}$$

This rule uses neither labels nor an external class prior; it presumes that both classes are present, and its operating point can vary with target-set composition (Section A.12), affecting balanced accuracy but not AUROC, since threshold subtraction is an L1 translation. The context mean is estimated from $B = 128$ images during training and from the complete target subset at inference.

Deployment modes. Since $\boldsymbol{\mu}_t$ enters the modulator as an input rather than a fitted parameter, deployment requires only a d-dimensional target statistic: images need not be retained, and no labels, gradients, or retraining are involved. In a batch setting, $\boldsymbol{\mu}_t$ is the mean of the available target images; in a stream, the same statistic can be updated online as images arrive. The main experiments use the complete benchmark-defined target subset. Appendix analyses examine finite-sample and boundary-free streaming estimates (Section A.17) and the source-mean fallback when no target statistic is available (Section A.19).

Table 1: Diagnostic suite under frozen encoders and heads. Acc. is the three-collection macro-average balanced accuracy; ΔAUROC is measured against each frozen VFM probe, values below 10^{-7} being numerical precision. Averages are computed at full precision and rounded only for display. Bold marks the best label-free correction per column. A dash marks the probe row, which is that column's own reference.

		SigLIP2		PE-CLIP		MetaCLIP		DINOv2	
Correction	Level	Acc.	ΔAUROC	Acc.	ΔAUROC	Acc.	ΔAUROC	Acc.	ΔAUROC
Reference conditions									
Frozen VFM probe	L0	0.827	—	0.797	—	0.692	—	0.657	—
Prior-matching calibration	L1	0.855	0	0.853	0	0.736	0	0.666	0
Label-oracle calibration	L1	0.858	0	0.865	0	0.741	0	0.681	0
Global mean alignment	L2	0.856	$< 10^{-7}$	0.857	$< 10^{-7}$	0.726	$< 10^{-7}$	0.666	$< 10^{-7}$
Published test-time objectives, frozen-head form unless marked Rel.; the scalar-offset TENT row is a degenerate control									
TENT (scalar logit offset) (Wang et al., 2020a)	L1	0.612	0	0.584	0	0.541	0	0.508	0
TENT, LayerNorm affine updated (Wang et al., 2020a)	Rel.	0.814	−0.0061	0.781	−0.0084	0.676	−0.0127	0.650	−0.0048
Closed-form calibration (Yang et al., 2026)	L1	0.856	0	0.854	0	0.737	0	0.666	0
Test-time normalization (Tomar et al., 2024)	L3	0.860	+0.0033	0.851	−0.0055	0.720	−0.0078	0.666	+0.0009
TPT (frozen-head reduction) (Shu et al., 2022)	L4	0.831	+0.0014	0.804	−0.0021	0.688	−0.0046	0.660	+0.0007
Ours									
Trained shared modulator	L3	0.876	+0.0149	0.862	+0.0188	0.734	+0.0081	0.669	+0.0078
Shared FS-CTAA (post-hoc)	L3	0.868	+0.0063	0.867	+0.0042	0.732	−0.0388	0.670	+0.0053
FS-CTAA	L4	**0.891**	**+0.0288**	**0.882**	**+0.0349**	**0.752**	**+0.0172**	**0.680**	**+0.0167**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate frozen SigLIP2, PE-CLIP, MetaCLIP, and DINOv2 visual foundation model (VFM) detectors on three collections: the eight GenImage generators (Zhu et al., 2023); four in-the-wild (ITW) benchmarks, namely Chameleon (Yan et al., 2025), WildRF (Cavia et al., 2024), SocialRF and CommunityAI (Zhou et al., 2025); and the four two-class Mirage subdomains Animal, Human, Object, and Scene (Xia et al., 2025). Each of these 16 subsets is one target subset: its unlabeled images form the batch from which $\boldsymbol{\mu}_t$ and the label-free threshold are estimated. Probes and correction modules are trained on the GenImage SD-v1.4 training split; the encoder and linear head remain frozen throughout. Balanced accuracy (Acc.) is the mean of the true-positive and true-negative rates at the operating point and so reflects threshold placement, whereas AUROC measures ranking alone. Overall performance is the unweighted mean of the three collection averages, computed from the per-subset values reported in the appendix.

4.2 DIAGNOSTIC SUITE AND MATCHED CONTROLS

Table 1 tests the granularity criterion under frozen encoders and heads. As predicted, all L1 corrections change the operating point but leave AUROC unchanged, while global mean alignment produces only numerical-precision deviations at L2. These results provide a direct implementation check of Proposition 1. Crossing the rank-invariance boundary at L3 makes reordering possible, but does not guarantee improvement: test-time normalization and the frozen-head reduction of test-time prompt tuning (TPT) move AUROC in either direction. In contrast, the trained shared modulator and FS-CTAA improve AUROC on all four backbones, with FS-CTAA achieving the larger gains of $+0.0167$ to $+0.0349$. FS-CTAA also exceeds the label-oracle threshold baseline in balanced accuracy on SigLIP2, PE-CLIP, and MetaCLIP, showing that these gains cannot be explained by operating-point selection alone; on DINOv2 it nearly matches the oracle (0.680 vs. 0.681) while still improving AUROC by $+0.0167$. The label-oracle row is therefore a threshold-only reference rather than a competing method, while scalar-offset TENT is the degenerate frozen-head form of entropy minimization.

The matched L3 controls isolate whether target conditioning alone is sufficient. The trained shared modulator $H_{\boldsymbol{\theta}}(\boldsymbol{\mu}_{\mathrm{ref}}, \boldsymbol{\mu}_{\mathrm{ctx}})$ receives the same source and target statistics and follows the same training protocol, but outputs one affine correction for the entire target subset. It already improves AUROC on every backbone, by $+0.0078$ to $+0.0188$, confirming that target-conditioned shared correction is useful. FS-CTAA nevertheless exceeds it on all four backbones, by $+0.0089$ to $+0.0161$ AUROC, indicating an additional benefit from sample-specific parameters. A parameter-matched post-

Table 2: Detection under distribution shift. Bold and underline mark the best and second-best value per column; italics mark the full-protocol TPT re-evaluation, which optimizes at test time and is therefore not a frozen detector.

	Balanced Accuracy				AUROC			
Method	GenImage	ITW	Mirage	Overall	GenImage	ITW	Mirage	Overall
Published AIGI detectors								
OmniAID (Guo et al., 2025)	0.959	0.680	0.518	0.719	0.9793	0.7728	0.4501	0.7341
DDA (Chen et al., 2025)	0.890	0.852	0.603	0.782	0.9763	0.9255	0.5678	0.8232
NPR (Tan et al., 2024b)	0.838	0.559	0.448	0.615	0.9068	0.6098	0.3089	0.6085
SAFE (Li et al., 2025)	0.836	0.553	0.499	0.629	0.9714	0.5477	0.4235	0.6476
Effort (Yan et al., 2024)	0.807	0.553	0.473	0.611	0.9947	0.6398	0.2464	0.6270
AIDE (Yan et al., 2025)	0.768	0.565	0.366	0.566	0.9665	0.5918	0.3628	0.6404
Per-sample test-time adaptation (TTA) baselines								
IAPL (Li et al., 2026)	0.959	0.581	0.497	0.679	**0.9971**	0.5839	0.2980	0.6263
TPT (full-protocol re-evaluation) (Shu et al., 2022)	*0.924*	*0.801*	*0.758*	*0.828*	*0.9968*	*0.9560*	*0.7550*	*0.9026*
Frozen detectors under our protocol								
SigLIP2 probe (Tschannen et al., 2025)	0.922	0.798	0.763	0.827	0.9891	0.9052	0.8505	0.9149
+ closed-form calibration (Yang et al., 2026)	0.960	0.835	0.774	0.856	0.9891	0.9052	0.8505	0.9149
+ FS-CTAA	0.963	0.888	**0.821**	**0.891**	0.9897	0.9434	**0.8983**	**0.9438**
PE-CLIP probe (Bolya et al., 2026)	0.942	0.861	0.586	0.797	0.9968	**0.9836**	0.6895	0.8900
+ closed-form calibration (Yang et al., 2026)	**0.982**	**0.945**	0.635	0.854	0.9968	0.9836	0.6895	0.8900
+ FS-CTAA	0.976	0.942	0.727	0.882	0.9952	0.9832	0.7963	0.9249
MetaCLIP probe (Chuang et al., 2026)	0.775	0.662	0.640	0.692	0.9124	0.7552	0.7228	0.7968
+ closed-form calibration (Yang et al., 2026)	0.856	0.690	0.666	0.737	0.9124	0.7552	0.7228	0.7968
+ FS-CTAA	0.853	0.720	0.682	0.752	0.9097	0.7915	0.7409	0.8140
DINOv2 probe (Oquab et al., 2023)	0.839	0.636	0.496	0.657	0.9231	0.6849	0.4925	0.7002
+ closed-form calibration (Yang et al., 2026)	0.864	0.638	0.497	0.666	0.9231	0.6849	0.4925	0.7002
+ FS-CTAA	0.869	0.664	0.506	0.680	0.9268	0.7202	0.5037	0.7169

hoc control reaches the same conclusion: Shared FS-CTAA reuses the trained per-sample weights but averages γ_i and β_i within each subset, converting the module to L3 without changing its parameter count. It remains below FS-CTAA on every backbone and reverses the MetaCLIP gain to -0.0388. Thus, the L4 advantage cannot be explained simply by a larger parameter count. Independently trained budget-matched L3/L4 variants and three-seed repeats preserve the same ordering (Section A.6).

4.3 Detection under Distribution Shift

Table 2 shows that the benefit of FS-CTAA is concentrated where the frozen detector still has substantial ranking headroom. Across the twelve backbone–collection pairs, all seven pairs with a frozen-probe AUROC below 0.91 improve after correction, with gains ranging from $+0.0112$ to $+0.1068$. In contrast, the five pairs already close to saturation change by only -0.0027 to $+0.0037$. The contrast is particularly clear for PE-CLIP: on Mirage, where the frozen probe reaches only 0.6895 AUROC, FS-CTAA raises it to 0.7963 ($+0.1068$), whereas its already strong GenImage and ITW rankings remain essentially unchanged. Thus, the average gain is not obtained by broadly perturbing already reliable rankings; it is concentrated where distribution shift has degraded the relative ordering of real and generated samples. Mirage is the most severe instance of this regime in our evaluation, and several published detectors obtain AUROC below 0.5 on Mirage, indicating a severe ranking failure rather than merely a misplaced threshold.

Under the matched frozen-detector protocol, this ranking improvement is consistent across backbones. FS-CTAA improves the overall AUROC of SigLIP2, PE-CLIP, MetaCLIP, and DINOv2 by $+0.0288$, $+0.0349$, $+0.0172$, and $+0.0167$, respectively, while also increasing overall balanced accuracy on all four. The AUROC gain also remains positive for every backbone across all three training seeds (Section A.14). Relative to closed-form calibration, which can move the operating point but leaves AUROC unchanged, FS-CTAA further improves balanced accuracy by $+0.014$ to $+0.035$. The improvement is not uniform at every individual subset: Chameleon AUROC decreases slightly on SigLIP2 and PE-CLIP, and PE-CLIP balanced accuracy on CommunityAI decreases from 0.971 to 0.959. FS-CTAA should therefore be interpreted as improving robustness on average, with its largest effect under shifts that most strongly disrupt the frozen ranking, rather than as uniformly dominating every target subset.

For context, the full-protocol TPT re-evaluation reaches 0.828 balanced accuracy and 0.9026 AUROC overall. It is stronger than FS-CTAA when paired with MetaCLIP and DINOv2, but weaker

Table 3: Structural and context ablations on frozen PE-CLIP and SigLIP2. Acc. is the three-collection macro-average balanced accuracy, and ΔAUROC is measured against the frozen probe with paired-bootstrap 95% intervals. Bold marks the full model.

	PE-CLIP		SigLIP2	
Variant	Acc.	ΔAUROC [95% CI]	Acc.	ΔAUROC [95% CI]
FS-CTAA	**0.882**	**+0.0349** [+0.0324, +0.0374]	**0.891**	**+0.0288** [+0.0267, +0.0309]
Correction structure				
w/o channel-wise	0.854	+0.0153 [+0.0121, +0.0185]	0.861	+0.0127 [+0.0102, +0.0152]
w/o β	0.834	+0.0096 [+0.0063, +0.0129]	0.845	+0.0072 [+0.0045, +0.0099]
w/o γ	0.863	+0.0201 [+0.0168, +0.0234]	0.870	+0.0173 [+0.0145, +0.0201]
w/o mean anchor	0.860	+0.0230 [+0.0199, +0.0261]	0.866	+0.0202 [+0.0175, +0.0229]
Target context				
w/o target-mean input	0.866	+0.0192 [+0.0160, +0.0224]	0.874	+0.0168 [+0.0141, +0.0195]
w/o target context	0.853	+0.0068 [+0.0032, +0.0104]	0.850	+0.0051 [+0.0020, +0.0082]
w/o correct subset	0.794	−0.0038 [−0.0077, +0.0001]	0.819	−0.0049 [−0.0086, −0.0012]
Unconditional adapter	0.821	+0.0039 [−0.0002, +0.0080]	0.843	+0.0034 [−0.0008, +0.0076]

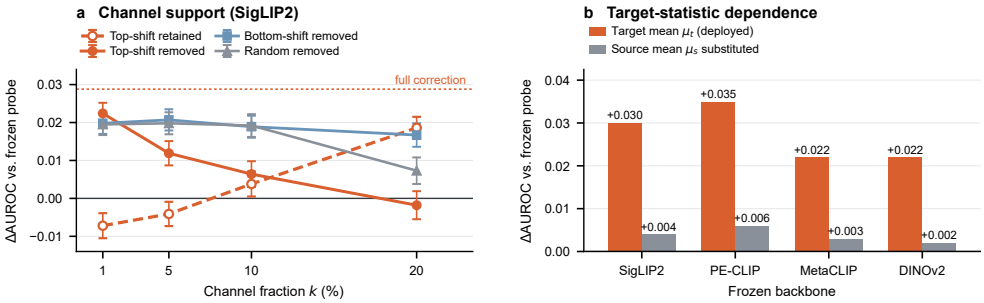

Figure 3: What the ranking gain of FS-CTAA depends on (ΔAUROC against the frozen probe). (a) Channel support on frozen SigLIP2: the correction restricted to the top-k% most displaced channels, or with the top-shift, bottom-shift, or a random k% removed from it with paired-bootstrap 95% intervals at the main seed. (b) Target-statistic dependence on all four backbones: the deployed setting against the same trained modulator run with the source mean substituted for $\boldsymbol{\mu}_t$ at inference. Results are averaged over three training seeds.

with SigLIP2 and PE-CLIP. This comparison is not protocol matched: TPT performs gradient-based prompt optimization for every test image, whereas FS-CTAA keeps both the backbone and classification head frozen and requires only one additional feed-forward pass at inference. We therefore treat TPT as a strong test-time adaptation reference rather than evidence that one protocol uniformly dominates the other.

4.4 Mechanism and Ablation

Table 3 separates the structural and contextual factors behind the correction. Both metrics are reported because balanced accuracy can improve through the operating point alone, whereas AUROC requires a change in sample ordering. The affine structure is important throughout: on PE-CLIP, removing the translation branch reduces the AUROC gain from +0.0349 to +0.0096, while sharing a single scale and bias across channels reduces it to +0.0153. Removing the scale branch or the mean anchor is less damaging, with the same qualitative ordering on SigLIP2. Thus, the gain is not explained by a generic feature adapter; it depends on channel-wise correction with both multiplicative and additive flexibility.

The context ablations identify two complementary sources of information. Removing $\boldsymbol{\mu}_t$ from the modulator input while retaining the mean-translation anchor preserves part of the gain (+0.0192 on PE-CLIP and +0.0168 on SigLIP2), showing that a source-trained per-sample mapping can already repair some ranking errors. Target context alone is also insufficient: the trained shared modulator improves AUROC but remains below the sample-conditioned FS-CTAA, while an unconditional adapter that sees only the image feature yields small gains whose confidence intervals include zero. Conversely, supplying the mean of an incorrect target subset is the only ablation whose point estimates fall below the frozen probe on both metrics. Together, these controls show that the full gain

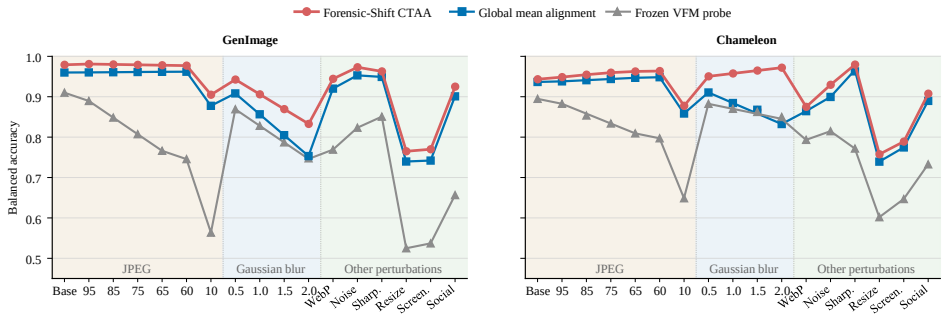

Figure 4: Balanced accuracy across 17 post-processing conditions on matched GenImage and Chameleon manifests (frozen PE-CLIP). All curves use one source-fixed threshold and no target labels, so the differences reflect the feature correction under the same operating point.

relies on both sample-specific evidence and a target statistic corresponding to the current distribution. A separate context-granularity audit further shows graceful degradation when that statistic is pooled over coarser target groups (Section A.13).

4.5 Attributing the Ranking Gain

We next intervene on the deployed correction to identify what produces its ranking gain. Figure 3a tests whether the most displaced channels are sufficient or necessary. Restricting the correction to the top-20% most displaced channels retains $+0.0186$ AUROC, below the full $+0.0288$, whereas removing those channels nearly eliminates the gain (-0.0018); removing a random 20% still retains $+0.0073$. The highly displaced channels therefore carry disproportionate correction signal, but do not explain the gain alone: useful correction remains distributed across the representation. Figure 3b directly tests the role of target context while keeping the trained modulator fixed. Replacing $\boldsymbol{\mu}_t$ with the stored source mean reduces the AUROC gain from $+0.022$–$+0.035$ to only $+0.002$–$+0.006$ across all four backbones. Because the model and affine form are unchanged, this drop isolates the target statistic as a major source of the deployed model's distribution-specific ranking correction.

4.6 Robustness Across Post-Processing Conditions

We further test whether the source-learned correction remains effective under controlled post-processing shifts. Figure 4 evaluates the same images under 17 conditions, including JPEG compression, Gaussian blur, and six dissemination perturbations, using a single threshold fixed on the source validation split and never re-tuned. FS-CTAA achieves the highest balanced accuracy at every plotted condition on both GenImage and Chameleon. Its advantage is particularly visible under dissemination transformations, while severe JPEG compression and Gaussian blur reduce the absolute accuracy of all methods. In contrast, global mean alignment falls to or below the frozen probe under the strongest blur conditions on Chameleon. These results show that the learned correction remains effective across a broad range of post-processing conditions without target-specific threshold tuning. The complete condition list and evaluation protocol are given in Section A.8.

5 Conclusion

We studied test-time correction for AIGI detectors with a frozen encoder and linear head. A rank-invariance criterion identifies corrections that only recalibrate the operating point, while a parameter-sharing axis distinguishes sample-shared from sample-conditioned corrections. Guided by channel-wise, target-dependent shifts, FS-CTAA learns per-sample corrections from source-side forensic-shift episodes without target-time gradients. It improves macro-average AUROC on all four frozen backbones beyond a shared modulator trained under the same protocol. Rank-invariant mean translation accounts for much of the balanced-accuracy gain, while learned non-rank-invariant modulation improves ranking; sample conditioning provides further ranking gains over shared correction.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, LaTeX formatting, and limited drafting and revision. They also supported the presentation and consistency checking of experimental results. All AI-assisted outputs were reviewed and verified by the authors against the underlying sources. The authors determined and validated the methodology, analysis, and conclusions, and take full responsibility for the final content.

ETHICS STATEMENT

This work uses existing research benchmarks and does not involve human-subject experimentation or the collection of new personal data. Automated predictions can be unreliable under distribution shifts and should not be treated as definitive evidence in high-stakes decisions. Appropriate human review and independent validation remain necessary for practical use.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the method and experiments through the paper, appendix, and accompanying materials. These materials include the source code, configurations, evaluation procedures, and implementation details. The anonymous repository is available at: `https://anonymous.4open.science/r/fs-ctaa-CF26`.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35:24597–24610, 2022.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision*, pp. 80–95. Springer, 2024.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Fighting deepfake by exposing the convolutional traces on images. *IEEE access*, 8:165085–165098, 2020.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694. IEEE, 2025.

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. Omniaid: Decoupling semantic and artifacts for universal ai-generated image detection in the wild. *arXiv preprint arXiv:2511.08423*, 2025.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 2405–2414, 2025.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066. IEEE, 2020.

Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. Think twice before adaptation: Improving adaptability of deepfake detection via online test-time adaptation. *arXiv preprint arXiv:2505.18787*, 2025.

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*, pp. 16888–16905. PMLR, 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. *arXiv preprint arXiv:2302.12400*, 2023.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257. IEEE, 2025.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024b.

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Un-mixing test-time normalization statistics: Combatting label temporal correlation. In *International Conference on Learning Representations*, volume 2024, pp. 50967–50988, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Cheng Xia, Manxi Lin, Jiexiang Tan, Xiaoxiong Du, Yang Qiu, Junjun Zheng, Xiangheng Kong, Yuning Jiang, and Bo Zheng. Mirage: Towards ai-generated image detection in the wild. *arXiv preprint arXiv:2508.13223*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

Muli Yang, Gabriel James Goenawan, Henan Wang, Huaiyuan Qin, Chenghao Xu, Yanhua Yang, Fen Fang, Ying Sun, Joo Hwee Lim, and Hongyuan Zhu. Your ai-generated image detector can secretly achieve sota accuracy, if calibrated. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11622–11630, 2026.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, Jinhua Zeng, and Bin Li. Brought a gun to a knife fight: Modern vfm baselines outgun specialized detectors on in-the-wild ai image detection. *arXiv preprint arXiv:2509.12995*, 2025.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Simplicity prevails: The emergence of generalizable aigi detection in visual foundation models. *arXiv preprint arXiv:2602.01738*, 2026.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Additional Analyses and Implementation Details

A.1 Notation

Table 4 collects the symbols used in the main text.

Table 4: Notation.

Symbol	Meaning
E, $f(\boldsymbol{z}) = \boldsymbol{w}^\top \boldsymbol{z} + c$, τ	Frozen encoder, frozen linear head, decision threshold (operating point)
$\boldsymbol{z}_i = E(\boldsymbol{x}_i)$, $s_i = f(\boldsymbol{z}_i)$	Representation and detection score of image i
$\mathcal{D}_s$, $\mathcal{D}_t$, N_t	Source distribution, target subset (unlabeled test batch), its size
$\boldsymbol{\mu}_s$, $\bar{\boldsymbol{\mu}}_s$	Source mean representation and its empirical estimate over the source set
$\boldsymbol{\mu}_t$	Mean representation of the current target subset (target context)
$\boldsymbol{\mu}_{\mathrm{ref}}$, $\boldsymbol{\mu}_{\mathrm{ctx}}$	Reference and context inputs of the modulator ($\bar{\boldsymbol{\mu}}_s$ and $\boldsymbol{\mu}_t$ at inference)
T, g, h_t, $\boldsymbol{b}_t$	Correction operator, strictly increasing score map, L1 score map, L2 shared translation
$\boldsymbol{\phi}$, $\boldsymbol{\theta}$	Correction parameters in equation 3 and the modulator weights that instantiate them
$\mathcal{C}_t$	Target context available to a correction (here $\boldsymbol{\mu}_{\mathrm{ref}}$ and $\boldsymbol{\mu}_{\mathrm{ctx}}$)
$H_{\boldsymbol{\theta}}$, $G_{\boldsymbol{\gamma}}$, $G_{\boldsymbol{\beta}}$, $\boldsymbol{h}_i$	Modulator trunk, its two output heads, and the hidden representation of image i
$\boldsymbol{\gamma}_i$, $\boldsymbol{\beta}_i$, $\boldsymbol{z}_i^c$	Per-sample channel-wise scale, offset, and corrected representation
$(\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t)$	One shared scale and offset per target subset (trained shared modulator)
$\mathcal{B}$, B	Source episode batch and its size
$\mathcal{A}_{k,\xi}$, $\boldsymbol{z}_i^\dagger$	Forensic condition of family k with parameters ξ, and the shifted representation
$\boldsymbol{\mu}_s^{\mathcal{B}}$, $\boldsymbol{\mu}_\dagger^{\mathcal{B}}$	Clean and shifted batch means of an episode
$\boldsymbol{z}_i^g$, $\alpha_{\mathcal{B}}$, $\widehat{\boldsymbol{z}}_i$	L2-anchored representation, interpolation coefficient, interpolated training input
$\tau_t^\star$, s_i^c, $\widehat{y}_i$	Label-free threshold (mean corrected score), corrected score, and predicted label
ℓ_{bin}, sg, λ_{logit}, λ_{res}	Binary classification loss, stop-gradient, and the two loss weights of equation 7
d, δ_c	Feature dimension, and the displacement of channel c, $\lvert\bar{\mu}_{s,c} - \mu_{t,c}\rvert$
ΔAUROC, Δ_{acc}, Δ_{corr}, Δ_{cal}, Δ_{comp}	Change against the frozen probe; accuracy decomposition terms of equation 9; composition sensitivity of equation 13
L0–L4	Uncorrected probe; score map; shared translation; shared operator; sample-conditioned operator

A.2 Implementation details

All encoders and linear heads remain frozen. The Distribution-Conditioned Modulator uses two hidden layers of width 512 with LayerNorm and predicts the channel-wise affine parameters from the pooled representation and the source and target statistics; its input is the concatenation of the three d-dimensional vectors and its output is the $2d$ affine parameters. The reported results use the sample-conditioned Distribution-Conditioned Modulator. Two L3 controls share its family. Shared FS-CTAA is a post-hoc control that averages the per-sample predictions of the trained module within each target subset. The trained shared modulator (Section A.6) is a separately trained model with the same hidden layers whose input is the concatenation of $\boldsymbol{\mu}_{\mathrm{ref}}$ and $\boldsymbol{\mu}_{\mathrm{ctx}}$ alone and whose output is one affine correction per target subset. All training runs use an episode batch size of 128. We train the modulator for five epochs with AdamW, a learning rate of 10^{-3}, and weight decay 10^{-4}. The objective uses $\lambda_{\mathrm{logit}} = 0.5$ and $\lambda_{\mathrm{res}} = 0.05$. For each source-side episode, the interpolation coefficient in equation 6 is sampled uniformly from $\{0.25, 0.50, 0.75, 1.00\}$. Every reported value uses seed 0 unless the caption states that the entry is averaged over the three training seeds, which are reported in Section A.14. The complete unlabeled target subset is used to estimate $\boldsymbol{\mu}_t$ in the main transductive evaluation, whereas training estimates the context from the 128 images of an episode; Section A.17 reports the effect of smaller inference contexts.

Table 5: Training ablations on frozen PE-CLIP. Acc. is the three-collection macro-average balanced accuracy, and ΔAUROC is measured against the frozen probe with paired-bootstrap 95% intervals. Bold marks the full model.

Variant	Acc.	ΔAUROC [95% CI]
All four families, full objective	**0.882**	**+0.0349** [+0.0324, +0.0374]
Episode supervision		
w/o Generator-style	0.863	+0.0218 [+0.0188, +0.0248]
w/o Spectral-response	0.855	+0.0170 [+0.0138, +0.0202]
w/o Compression	0.851	+0.0145 [+0.0110, +0.0180]
w/o Resize-reencoding	0.866	+0.0232 [+0.0201, +0.0263]
Unstructured shifts only	0.837	+0.0082 [+0.0045, +0.0119]
Training objective		
w/o $\mathcal{L}_{\mathrm{cls}}$	0.802	+0.0020 [−0.0018, +0.0058]
w/o $\mathcal{L}_{\mathrm{logit}}$	0.844	+0.0126 [+0.0091, +0.0161]
w/o $\mathcal{L}_{\mathrm{res}}$	0.852	+0.0189 [+0.0158, +0.0220]

Forensic-shift families. Each episode samples one of four forensic-shift families: compression, resize–re-encoding, spectral-response adjustment, or the generator-style proxy. It then draws one condition within that family. Compression selects single JPEG, double JPEG, WebP, or coarse PNG quantization and draws the JPEG quality factor from $[25, 95]$. Resize–re-encoding draws a scale factor from $[0.35, 0.90]$ and an interpolation kernel from bilinear, bicubic, Lanczos, and nearest neighbour, then applies either JPEG compression or a screenshot-style degradation to the rescaled image. Spectral-response adjustment selects high-frequency attenuation, low-pass filtering, band-stop filtering, or stochastic spectral dropout, with normalized cutoff frequencies in $[0.25, 0.75]$. The generator-style proxy selects smoothing, sharpening, colour perturbation, additive noise, or denoise-then-sharpen. The families are chosen to span the transformations an image undergoes between generation and observation; they are not intended to imitate any particular generator, and no target-domain image or statistic is used to select or tune them.

Frozen probe recipe. We adopt the standard linear-probing recipe used for frozen visual foundation models in AIGI detection without modification, so that the L0 rows reproduce the conventional baseline and every reported change is attributable to the correction rather than to a stronger probe. Concretely, each backbone is fully frozen and only the linear head is updated; the head is fitted on the GenImage SD-v1.4 training split alone, with no target-domain data and no data from the other seven generators; optimization uses AdamW with learning rate 10^{-3}, batch size 128, for two epochs; and images are resized and centre-cropped to each backbone's native input resolution with no additional augmentation, PE-CLIP being evaluated as ViT-L/14 at 336 px. Holding this recipe fixed across all four backbones is what makes the four columns of Table 1 comparable, and Section A.20 reports what happens when the probe is instead given the forensic-shift augmentations or extra capacity.

A.3 Structural Ablation Conventions

In Table 3, "w/o channel-wise" shares one affine scale and one bias across all channels. The "w/o γ" and "w/o β" variants suppress the corresponding prediction branch, and "w/o mean anchor" removes the explicit $\boldsymbol{\mu}_{\mathrm{ref}} - \boldsymbol{\mu}_{\mathrm{ctx}}$ term from $\boldsymbol{\beta}$ while keeping the target mean as a modulator input. "w/o target-mean input" removes $\boldsymbol{\mu}_t$ from the modulator input but retains the $\boldsymbol{\mu}_{\mathrm{ref}} - \boldsymbol{\mu}_{\mathrm{ctx}}$ anchor, so the target enters only through the rank-invariant translation. "w/o target context" sets $\boldsymbol{\mu}_t = \boldsymbol{\mu}_s$ and zeroes the anchor, so no target information reaches the correction. "w/o correct subset" (the wrong-subset variant) supplies the target mean of a different target subset. "Unconditional adapter" is a parameter-matched module that receives only the representation $\boldsymbol{z}_i$ and neither $\boldsymbol{\mu}_s$, $\boldsymbol{\mu}_t$, nor the anchor. In Table 5, each "w/o" row omits one forensic-shift family from episode training, and "Unstructured shifts only" omits all four structured families. Removing the classification loss leaves the module within 0.005 balanced accuracy of the probe and with an AUROC change whose interval includes zero, so label supervision on source episodes is necessary; the logit-consistency and residual terms contribute to ranking as well as to balanced accuracy. The "Unstructured shifts only" row replaces all four image-level families by an isotropic Gaussian perturbation applied directly to the cached feature, $\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.5^2 \boldsymbol{I})$. It is a feature-space stress test, not a fifth forensic-shift family: no image is rendered, encoded, or re-encoded, and the perturbation has no forensic interpretation. Its purpose is to ask whether the episode structure matters or whether any feature-level jitter of comparable magnitude would train an equally useful modulator. It retains $+0.0082$ against $+0.0349$ for the four structured families, so the answer is that the structure matters.

A.4 Plug-in Parameter Scale

Table 6 quantifies the parameter and computational cost of FS-CTAA. SigLIP2 and DINOv2-Giant use 1536-dimensional features and $4{,}200{,}448$ correction parameters. MetaCLIP-H/14 uses 1280-dimensional features and $3{,}544{,}576$ parameters, while PE-CLIP uses 1024-dimensional features and $2{,}888{,}704$ parameters. These modules account for 0.360%–0.911% of the corresponding backbone parameters. The measured batch latency of the module alone, excluding the backbone forward pass and the computation of $\boldsymbol{\mu}_t$, ranges from 4.23 to 7.27 ms. The gradient-based controls in Table 1 and the full-protocol TPT baseline require an optimization loop per batch or per image, which FS-CTAA does not.

Table 6: Parameter scale and computational overhead of FS-CTAA across frozen visual backbones. Counts are for the full modulator, whose input is the $3d$-dimensional concatenation of the image feature and the two mean vectors and whose widths are $3d \to 512 \to 512 \to 2d$ with two LayerNorms of width 512. Ablations that change only the context input leave these counts unchanged, since the modulator reads $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_t$ through the same $3d$ input slots regardless of their values.

Backbone	Backbone params (M)	Module params (M)	Module / backbone	Module FLOPs/sample	Batch latency ($B = 128$)	Latency/ sample
SigLIP2	1163.7	4.20	0.361%	8.39M	7.27 ms	0.0568 ms
PE-CLIP	317.2	2.89	0.911%	5.77M	5.22 ms	0.0408 ms
MetaCLIP-H/14	986.0	3.54	0.360%	7.08M	5.94 ms	0.0464 ms
DINOv2-Giant	1136.5	4.20	0.370%	8.39M	4.23 ms	0.0331 ms

A.5 Frozen-Detector Diagnostic Suite

This appendix provides the protocol definition and the per-collection results of Table 7 for the diagnostic suite in Table 1. All comparisons use a frozen visual encoder and a frozen linear head. A test-time correction may modify only the representation supplied to the head or the resulting score. Accordingly, published objectives are represented by the correction forms admissible under this interface, rather than by reproductions of their original trainable components.

Marking conventions in Table 1. Bold marks the best value among label-free corrections in each column. The label-oracle row uses target labels and is excluded from the ranking, and the TENT scalar-offset row is a degenerate control. "Rel." marks a control that updates LayerNorm parameters

Table 7: Per-collection results for the diagnostic suite in Table 1, in the same row order. Overall is the unweighted mean of the GenImage, ITW, and Mirage balanced-accuracy averages. ΔAUROC is measured against each backbone's frozen probe, with values below 10^{-7} attributed to numerical precision. Bold marks the best label-free Overall and ΔAUROC in each block. The trained shared modulator is reported at the overall level in Table 8. A dash marks the probe row, which is that column's own reference.

Correction	Level	GenImage	ITW	Mirage	Overall	ΔAUROC
SigLIP2, Frozen VFM probe AUROC = 0.9149						
Frozen VFM probe	L0	0.922	0.798	0.763	0.827	—
Prior-matching calibration	L1	0.959	0.831	0.773	0.855	0
Label-oracle calibration	L1	0.961	0.837	0.776	0.858	0
Global mean alignment	L2	0.959	0.834	0.774	0.856	$< 10^{-7}$
TENT (scalar logit offset)	L1	0.808	0.525	0.504	0.612	0
TENT, LayerNorm affine updated	Rel.	0.842	0.803	0.797	0.814	−0.0061
Closed-form calibration	L1	0.960	0.835	0.774	0.856	0
Test-time normalization	L3	0.958	0.847	0.774	0.860	+0.0033
TPT (frozen-head reduction)	L4	0.846	0.827	0.820	0.831	+0.0014
Shared FS-CTAA (post-hoc)	L3	0.961	0.855	0.788	0.868	+0.0063
FS-CTAA	L4	0.963	0.888	0.821	**0.891**	**+0.0288**
PE-CLIP, Frozen VFM probe AUROC = 0.8900						
Frozen VFM probe	L0	0.942	0.861	0.586	0.797	—
Prior-matching calibration	L1	0.983	0.938	0.636	0.853	0
Label-oracle calibration	L1	0.984	0.948	0.663	0.865	0
Global mean alignment	L2	0.981	0.947	0.643	0.857	$< 10^{-7}$
TENT (scalar logit offset)	L1	0.750	0.503	0.500	0.584	0
TENT, LayerNorm affine updated	Rel.	0.812	0.768	0.763	0.781	−0.0084
Closed-form calibration	L1	0.982	0.945	0.635	0.854	0
Test-time normalization	L3	0.980	0.944	0.629	0.851	−0.0055
TPT (frozen-head reduction)	L4	0.820	0.801	0.791	0.804	−0.0021
Shared FS-CTAA (post-hoc)	L3	0.980	0.944	0.677	0.867	+0.0042
FS-CTAA	L4	0.976	0.942	0.727	**0.882**	**+0.0349**
MetaCLIP, Frozen VFM probe AUROC = 0.7968						
Frozen VFM probe	L0	0.775	0.662	0.640	0.692	—
Prior-matching calibration	L1	0.856	0.687	0.665	0.736	0
Label-oracle calibration	L1	0.858	0.697	0.669	0.741	0
Global mean alignment	L2	0.852	0.676	0.649	0.726	$< 10^{-7}$
TENT (scalar logit offset)	L1	0.622	0.501	0.500	0.541	0
TENT, LayerNorm affine updated	Rel.	0.706	0.669	0.653	0.676	−0.0127
Closed-form calibration	L1	0.856	0.690	0.666	0.737	0
Test-time normalization	L3	0.847	0.673	0.641	0.720	−0.0078
TPT (frozen-head reduction)	L4	0.704	0.686	0.674	0.688	−0.0046
Shared FS-CTAA (post-hoc)	L3	0.851	0.695	0.650	0.732	−0.0388
FS-CTAA	L4	0.853	0.720	0.682	**0.752**	**+0.0172**
DINOv2, Frozen VFM probe AUROC = 0.7002						
Frozen VFM probe	L0	0.839	0.636	0.496	0.657	—
Prior-matching calibration	L1	0.863	0.638	0.498	0.666	0
Label-oracle calibration	L1	0.866	0.639	0.538	0.681	0
Global mean alignment	L2	0.864	0.638	0.497	0.666	$< 10^{-7}$
TENT (scalar logit offset)	L1	0.524	0.504	0.498	0.508	0
TENT, LayerNorm affine updated	Rel.	0.672	0.640	0.638	0.650	−0.0048
Closed-form calibration	L1	0.864	0.638	0.497	0.666	0
Test-time normalization	L3	0.861	0.640	0.498	0.666	+0.0009
TPT (frozen-head reduction)	L4	0.678	0.658	0.644	0.660	+0.0007
Shared FS-CTAA (post-hoc)	L3	0.866	0.644	0.500	0.670	+0.0053
FS-CTAA	L4	0.869	0.664	0.506	**0.680**	**+0.0167**

and so relaxes the frozen-encoder protocol. Probe rows are our linear probes following Zhou et al. (2026).

- **Score-level corrections.** The Prior-matching calibration, TENT (scalar logit offset), and Closed-form calibration rows apply score-level or threshold-level changes. They are L1 corrections: they may change threshold-dependent predictions, but cannot change the ordering of samples.
- **Feature-level shared corrections.** Global mean alignment applies the same translation to every sample and is therefore L2. Test-time normalization applies one target-dependent channel-wise transformation to the complete test subset and is classified as L3. The trained shared modulator (Section A.6) learns one affine correction per target subset from the two

statistics and is the trained L3 counterpart of FS-CTAA. Shared FS-CTAA uses the trained per-sample module but averages its predicted parameters within the subset, providing a post-hoc L3 control at an identical parameter count.

- **Sample-conditioned corrections.** FS-CTAA conditions the correction parameters on both the current test subset and the individual representation, and is therefore L4. This sample-conditioned parameterization removes the shared-operator restriction and permits sample ordering to change. The frozen-head reduction of TPT keeps the encoder and head frozen and optimizes a per-sample correction, so it is also L4; it is a named diagnostic control and is reported separately from the full-protocol TPT re-evaluation in the main comparison.
- **Protocol-relaxation control.** TENT with LayerNorm affine parameters updated modifies the encoder and therefore leaves the frozen protocol. It is marked "Rel." and retained as a named diagnostic control.

Functional form of each reduced baseline. Every baseline is reduced to the frozen-head setting so that it acts on the same quantities FS-CTAA acts on. Let s denote the scalar detection score of the frozen head, $\boldsymbol{z}$ the d-dimensional feature, and $\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t$ the source and target channel means and standard deviations.

- **Prior matching.** Assumes a positive-class proportion of 0.5 in the target subset and places the threshold at the median of the target scores, which is the value that realizes that proportion. It moves only the threshold.
- **TENT (scalar logit offset).** Adds one additive offset b to the logit, $s \mapsto s + b$, with b searched over $[-10, 10]$ in steps of 0.05 to minimize the mean binary prediction entropy on the target subset. It is the degenerate frozen-head form of entropy minimization and is included to show that the objective is not what makes TENT ineffective here; the reachable function class is.
- **Closed-form calibration.** Fits a Gaussian kernel density estimate to the one-dimensional target score distribution with Scott's bandwidth rule and subtracts the estimated first moment, giving $s \mapsto s - \hat{\mathbb{E}}[s]$.
- **Global mean alignment.** Applies $\boldsymbol{z} \mapsto \boldsymbol{z} - \boldsymbol{\mu}_t + \boldsymbol{\mu}_s$ to every sample and thresholds the resulting score at 0.
- **Test-time normalization.** Applies the channel-wise standardization $\boldsymbol{z} \mapsto (\boldsymbol{z} - \boldsymbol{\mu}_t) \odot \boldsymbol{\sigma}_s \oslash \boldsymbol{\sigma}_t + \boldsymbol{\mu}_s$, so it does re-scale to source statistics, with the same threshold at 0. One parameter pair is shared by the whole subset, which is why it is L3.
- **Label-oracle calibration.** Chooses the threshold independently for each target subset. The candidate set is the midpoints of adjacent sorted scores, the selected candidate maximizes balanced accuracy using the target labels, and ties are broken toward 0. It is an upper bound on what any threshold rule can achieve at a fixed ordering, not a deployable method.

The frozen-head TPT reduction keeps both the encoder and the linear head untouched and optimizes only a per-sample correction, which is why it is listed as L4 rather than marked "Rel." It is a function-class control derived from the per-sample adaptation interface of TPT, not a reproduction of the original algorithm.

The distinction between the levels is functional rather than architectural. Rank invariance is determined by the score function induced by a correction: if a strictly increasing function g satisfies $f(T(\boldsymbol{z})) = g(f(\boldsymbol{z}))$ for all $\boldsymbol{z}$, then every pairwise sample ordering is preserved. Shared parameterization is a separate property determined by whether the correction parameters vary across samples. A more expressive sample-shared operator remains L3 unless its parameters depend on the individual representation. The per-collection results reproduce the structural separation tested in the main table:

- **L1 controls.** Prior-matching calibration, Label-oracle calibration, TENT (scalar logit offset), and Closed-form calibration leave $\Delta\text{AUROC} = 0$ on every backbone, despite their different balanced-accuracy values.
- **L2 control.** Global mean alignment changes AUROC only at the numerical precision level, consistent with its shared translation of every representation.

Table 8: Trained shared modulator (L3) against FS-CTAA (L4) on four frozen backbones. Acc. is overall balanced accuracy and ΔAUROC is measured against the frozen probe; S0 to S2 are the three training seeds, and seed S0 is the main-table seed. L4−L3 is the seed-averaged difference with a paired-bootstrap 95% interval.

Backbone	L3 Acc. (S0 / S1 / S2)	L3 ΔAUROC (S0 / S1 / S2)	L4−L3 Acc. [95% CI]	L4−L3 ΔAUROC [95% CI]
SigLIP2	0.876 / 0.879 / 0.879	+0.0149 / +0.0182 / +0.0134	+0.0153 [+0.011, +0.019]	+0.0140 [+0.0112, +0.0168]
PE-CLIP	0.862 / 0.863 / 0.859	+0.0188 / +0.0225 / +0.0171	+0.0210 [+0.017, +0.025]	+0.0158 [+0.0131, +0.0187]
MetaCLIP	0.734 / 0.736 / 0.732	+0.0081 / +0.0107 / +0.0092	+0.0190 [+0.014, +0.024]	+0.0122 [+0.0093, +0.0151]
DINOv2	0.669 / 0.671 / 0.667	+0.0078 / +0.0113 / +0.0090	+0.0107 [+0.006, +0.015]	+0.0126 [+0.0094, +0.0158]

Table 9: Capacity-matched controls across the L3/L4 boundary. Collection entries are balanced accuracy and ΔAUROC is measured against the frozen probe. Shrunk L4 narrows the FS-CTAA modulator to the parameter budget of the trained shared modulator, and widened L3 raises that shared modulator to the FS-CTAA budget; each matched width agrees with its target budget to within 0.12%. The native-width references are the L3 and L4 rows of Table 1.

Backbone	Matched variant	GenImage	ITW	Mirage	Overall	ΔAUROC
SigLIP2	Shrunk L4 ($h = 421$)	0.962	0.883	0.816	**0.887**	**+0.0260**
	Widened L3 ($h = 620$)	0.961	0.857	0.789	0.869	+0.0070
PE-CLIP	Shrunk L4 ($h = 426$)	0.975	0.938	0.722	**0.878**	**+0.0310**
	Widened L3 ($h_1 = 581, h_2 = 644$)	0.980	0.944	0.680	0.868	+0.0050
MetaCLIP	Shrunk L4 ($h = 423$)	0.851	0.715	0.678	**0.748**	**+0.0140**
	Widened L3 ($h_1 = 579, h_2 = 655$)	0.850	0.696	0.650	0.732	−0.0430
DINOv2	Shrunk L4 ($h = 421$)	0.867	0.660	0.504	**0.677**	**+0.0140**
	Widened L3 ($h = 620$)	0.866	0.646	0.501	0.671	+0.0060

- **L3 controls.** Test-time normalization, the trained shared modulator, and Shared FS-CTAA can change ranking. Test-time normalization is backbone-dependent in sign. The trained shared modulator is positive on all four backbones and seeds (Table 8) but smaller than FS-CTAA. Shared FS-CTAA changes AUROC by -0.0388 on MetaCLIP despite positive changes on the other three backbones.
- **L4 corrections.** FS-CTAA yields positive AUROC changes on all four backbones, ranging from $+0.0167$ on DINOv2 to $+0.0349$ on PE-CLIP, whereas the frozen-head reduction of TPT changes AUROC by at most 0.005 in either direction. The matched comparison with the trained shared modulator isolates the role of sample-conditioned parameterization under identical training; the comparison with Shared FS-CTAA holds the trained weights fixed while the restriction to one shared operator is imposed after training.

These results connect the functional hierarchy to the empirical diagnostic. Rank-invariant corrections can improve the operating point without changing sample ordering. Shared feature transformations may reorder samples, but do not guarantee a consistent direction. FS-CTAA combines target conditioning with sample-specific channel-wise modulation and is the only correction in this suite that improves ranking across all four frozen backbones by more than its trained shared counterpart.

A.6 Trained Shared Modulator (L3 Control)

The trained shared modulator is the L3 counterpart of FS-CTAA. It receives the same source reference and target context, is trained with the same forensic-shift episodes and objective, but omits the per-sample feature $\boldsymbol{z}_i$ and outputs one affine pair $(\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t)$ for the entire target subset. Because removing $\boldsymbol{z}_i$ narrows the input layer, the native L3 and L4 modules match architecture depth and hidden width but not parameter count.

We therefore use three complementary controls. The main-table Shared FS-CTAA control has exactly the same trained parameters as L4 but averages its predicted affine parameters within each subset. Table 8 tests the independently trained L3 control across three seeds, and Table 9 explicitly matches the trained L3 and L4 parameter budgets in both directions.

Table 9 matches the two levels on parameter count in both directions, with trained models rather than the post-hoc control. Narrowing the FS-CTAA modulator to the shared budget leaves it ahead of the

shared control on every backbone in both metrics: Overall accuracy is 0.887, 0.878, 0.748, and 0.677 against the L3 values 0.876, 0.862, 0.734, and 0.669, and ΔAUROC is $+0.0260$, $+0.0310$, $+0.0140$, and $+0.0140$ against $+0.0149$, $+0.0188$, $+0.0081$, and $+0.0078$. The narrowing costs at most 0.004 Overall accuracy and 0.004 AUROC relative to full-width FS-CTAA.

Across the three training seeds, the trained shared modulator improves AUROC on every backbone, but FS-CTAA remains ahead in both metrics. The seed-averaged L4$-$L3 AUROC advantage ranges from $+0.0122$ to $+0.0158$, and the paired-bootstrap intervals remain above zero on all four backbones.

Explicit parameter matching gives the same conclusion. The shrunk L4 model remains ahead of the widened L3 model on every backbone in both overall balanced accuracy and ΔAUROC, despite using no larger parameter budget. Together with the equal-parameter post-hoc control in Table 1, these results show that the L4 advantage is not an artifact of a particular seed or of additional modulator capacity.

A.7 Target-Specific Channel Sufficiency and Contribution

Panel details. In Figure 1(a), channels are ranked by $\delta_j = |\bar{\mu}_{s,j} - \mu_{t,j}|$, and the curve shows the cumulative share of $\sum_j \delta_j$; the top 10% carry 48.3–52.8%. Panel (b) shows mean Jaccard overlap of top-$k\%$ channel sets within and across collections. Marker shapes identify $k \in \{1, 5, 10, 20\}$; annotations report $k = 10\%$, and horizontal segments span the displayed means, not confidence intervals. The matched random-set mean at $k = 10\%$ is 0.053 over $1{,}000$ pairs. This label-free diagnostic includes 17 subsets, including Mirage-Anime; AUROC uses the 16 two-class subsets.

Panel (c) plots the main-seed ΔAUROC from Table 1 for each correction on all four backbones; n counts methods, and vertical offsets separate points. At L3, grey open, cyan filled, and cyan open diamonds denote test-time normalization, the trained shared modulator, and Shared FS-CTAA, respectively. At L4, open and filled circles denote the frozen-head TPT reduction and FS-CTAA. Dashed lines mark the L2/L3 rank-invariance and L3/L4 parameter-sharing boundaries. All four L1 controls are included; the unclassified "Rel." control is omitted, and L2 deviations below 10^{-7} are displayed at zero. Table 8 separately reports the three-seed L4–L3 comparison.

Channel interventions. Table 10 reports balanced accuracy and ΔAUROC for the interventions in Section 4.5. The sufficiency test adds only the retained channels' correction contributions to the frozen logit; the contribution test subtracts the removed channels' contributions from the fully corrected logit. Both use the linear head: excluded channels revert to their frozen features rather than being zeroed. In Figure 3(a), the dotted line marks the full correction ($+0.0288$); panel (b) is described in Section A.19.

Restricting correction to the top 1% or 5% lowers AUROC below the frozen probe. Retaining the top 20% reaches 0.861 balanced accuracy and $+0.0186$ AUROC, below the full correction. For $k \geq 5\%$, removing top-shift channels is more damaging than removing bottom-shift or random channels in both metrics. Removing the top 20% nearly eliminates the ranking gain (-0.0018), whereas random removal retains $+0.0073$. At $k = 1\%$, this ordering reverses: top-shift removal retains the largest gain ($+0.0224$). Thus, highly displaced channels contribute disproportionately, while useful correction extends beyond the most displaced few.

A.8 Robustness Across Post-Processing Conditions

This section specifies the controlled post-processing evaluation used in Section 4.6. The six dissemination perturbations are WebP conversion, noise, sharpening, resizing, screenshot capture, and social recompression. Together with the JPEG-quality and Gaussian-blur sweeps, they form eight condition groups and 16 perturbed conditions in addition to the unperturbed base. Both GenImage and Chameleon manifests use the full image counts reported in Table 11, with the same images evaluated under every condition.

All methods use the same threshold τ_{src}, fixed once on the source validation split and never re-tuned on the target conditions. The frozen VFM probe uses the uncorrected feature, Global mean alignment applies the L2 correction before the frozen head, and FS-CTAA estimates the target statistic for each test subset and applies its per-sample correction. Because this sweep uses the two matched

Table 10: Sufficiency and contribution of target-specific top-shift channels on frozen SigLIP2. Acc. is overall balanced accuracy, and ΔAUROC is measured against the frozen probe with paired-bootstrap 95% intervals. The first row is the full correction without any channel intervention; random removals report the mean and standard deviation of Acc. over draws. A dash in the k column marks the row with no channel intervention.

Intervention	k	Acc.	ΔAUROC [95% CI]
None (full FS-CTAA)	—	0.891	$+0.0288$ [$+0.0267$, $+0.0309$]
Sufficiency: correction restricted to the top-k% most displaced channels			
Retain top-shift	1%	0.807	-0.0072 [-0.0105, -0.0039]
	5%	0.809	-0.0041 [-0.0073, -0.0009]
	10%	0.822	$+0.0038$ [$+0.0005$, $+0.0071$]
	20%	0.861	$+0.0186$ [$+0.0157$, $+0.0215$]
Contribution: selected channels removed from the correction			
Remove top-shift	1%	0.868	$+0.0224$ [$+0.0196$, $+0.0252$]
	5%	0.852	$+0.0119$ [$+0.0087$, $+0.0151$]
	10%	0.846	$+0.0064$ [$+0.0030$, $+0.0098$]
	20%	0.834	-0.0018 [-0.0055, $+0.0019$]
Remove bottom-shift	1%	0.863	$+0.0198$ [$+0.0170$, $+0.0226$]
	5%	0.864	$+0.0207$ [$+0.0179$, $+0.0235$]
	10%	0.861	$+0.0189$ [$+0.0160$, $+0.0218$]
	20%	0.859	$+0.0167$ [$+0.0136$, $+0.0198$]
Remove random	1%	0.863 ± 0.0009	$+0.0195$ [$+0.0167$, $+0.0223$]
	5%	0.863 ± 0.0017	$+0.0198$ [$+0.0169$, $+0.0227$]
	10%	0.862 ± 0.0025	$+0.0192$ [$+0.0162$, $+0.0222$]
	20%	0.841 ± 0.0035	$+0.0073$ [$+0.0038$, $+0.0108$]

manifests rather than the 16 target subsets of the main protocol, its unperturbed base should not be compared numerically with the three-collection averages in Table 1.

As shown in Figure 4, FS-CTAA attains the highest balanced accuracy at every plotted condition on both manifests. The advantage is most visible under the dissemination transformations, whereas severe JPEG compression and Gaussian blur reduce the absolute accuracy of all methods. Global mean alignment also becomes less reliable under the strongest blur conditions on Chameleon, where it falls to or below the frozen probe.

A.9 Evaluation-Set Composition

Table 11 reports the class counts of every evaluated subset. The GenImage collection contributes eight generator subsets with 46,000 real and 46,000 generated images in total; the SD-v1.4 subset is the in-distribution reference, and BigGAN is the smallest subset at 2,000 images per class. Mirage forms its own collection and aggregates four binary subdomains. Mirage-Anime contains 6,000 generated images and no real images, so it is excluded from binary accuracy and AUROC aggregation and is not among the 16 target subsets.

A.10 Detailed Detection Results

Tables 12 and 13 report detailed balanced accuracy on the eight GenImage generators and the four ITW benchmarks, and Tables 14 to 16 the detailed AUROC on those subsets and on the four binary Mirage subdomains. In the AUROC tables, bold and underline mark the best and second-best value in each column, and the italic TPT row is included in the ranking.

Marking conventions in Table 2. Bold and underline mark the best and second-best value in each column across all rows; a probe and its calibrated variant share the same AUROC and are marked once, on the probe row. Italics mark the full-protocol re-evaluation of test-time prompt tuning (TPT), which performs test-time optimization and does not use a frozen detector; it is included in the ranking. IAPL denotes image-adaptive prompt learning. The protocol shared by every row is the 16 target subsets with the real and fake counts of Table 11, the same balanced-accuracy and AUROC definitions, and the same three-collection macro-average.

Chameleon illustrates the separation between ranking and operating-point effects within one subset. Its AUROC changes only from 0.9263 to 0.9251 for SigLIP2 and from 0.9921 to 0.9881 for PE-CLIP, while balanced accuracy rises from 0.803 to 0.880 and from 0.881 to 0.942, respectively. The

Table 11: Class counts for every evaluated subset. The Mirage row is the sum of its four binary subdomains, which is how it enters the three-collection macro-average. Mirage-Anime is listed for completeness and excluded from all binary metrics.

Subset	Real	Generated	Total
GenImage (Zhu et al., 2023)			
SD-v1.4	6,000	6,000	12,000
SD-v1.5	8,000	8,000	16,000
ADM	6,000	6,000	12,000
Midjourney	6,000	6,000	12,000
BigGAN	2,000	2,000	4,000
GLIDE	6,000	6,000	12,000
VQDM	6,000	6,000	12,000
Wukong	6,000	6,000	12,000
GenImage total	46,000	46,000	92,000
In-the-wild			
Chameleon (Yan et al., 2025)	14,863	11,170	26,033
WildRF (Cavia et al., 2024)	1,054	1,187	2,241
SocialRF (Zhou et al., 2025)	3,000	3,000	6,000
CommunityAI (Zhou et al., 2025)	6,000	6,000	12,000
Mirage (Xia et al., 2025)			
Mirage-Human	6,000	6,000	12,000
Mirage-Animal	3,500	3,500	7,000
Mirage-Object	6,000	6,000	12,000
Mirage-Scene	6,000	6,000	12,000
Mirage total (binary)	21,500	21,500	43,000
Mirage-Anime (excluded)	0	6,000	6,000

Table 12: Detailed balanced accuracy on the eight GenImage generators.

Method	ADM	BigGAN	Midjourney	VQDM	GLIDE	SD-v1.4	SD-v1.5	Wukong	Acc.
Competitor methods									
CNNSpot (Wang et al., 2020b)	0.507	0.530	0.610	0.501	0.522	0.998	0.996	0.985	0.706
FreqNet (Tan et al., 2024a)	0.608	0.939	0.849	0.626	0.858	0.950	0.949	0.937	0.840
Gram-Net (Liu et al., 2020)	0.576	0.591	0.665	0.528	0.747	0.989	0.988	0.960	0.756
NPR (Tan et al., 2024b)	0.619	0.849	0.850	0.567	0.844	0.996	0.994	0.984	0.838
UnivFD (Ojha et al., 2023)	0.571	0.839	0.795	0.644	0.842	0.959	0.958	0.926	0.817
SAFE (Li et al., 2025)	0.600	0.880	0.829	0.708	0.836	0.964	0.967	0.903	0.836
LaDeDa (Cavia et al., 2024)	0.512	0.583	0.595	0.513	0.585	0.876	0.871	0.820	0.670
Effort (Yan et al., 2024)	0.704	0.749	0.756	0.824	0.807	0.874	0.873	0.867	0.807
DDA (Chen et al., 2025)	0.880	0.741	0.960	0.719	0.862	0.986	0.985	0.986	0.890
AIDE (Yan et al., 2025)	0.602	0.648	0.813	0.694	0.514	0.959	0.957	0.954	0.768
OmniAID (Guo et al., 2025)	0.914	0.987	0.857	0.973	0.987	0.989	0.988	0.981	0.959
Our frozen-detector evaluations									
SigLIP2 probe (Tschannen et al., 2025)	0.800	0.897	0.846	0.933	0.902	0.999	0.998	0.997	0.922
+ FS-CTAA	0.901	0.976	0.932	0.952	0.948	0.999	0.998	0.996	0.963
PE-CLIP probe (Bolya et al., 2026)	0.711	0.997	0.900	0.959	0.972	0.9998	0.999	0.999	0.942
+ FS-CTAA	0.900	0.995	0.952	0.978	0.987	0.9998	0.999	0.999	0.976
MetaCLIP probe (Chuang et al., 2026)	0.549	0.602	0.839	0.653	0.652	0.984	0.981	0.942	0.775
+ FS-CTAA	0.688	0.714	0.898	0.777	0.818	0.987	0.985	0.957	0.853
DINOv2 probe (Oquab et al., 2023)	0.651	0.780	0.811	0.758	0.852	0.967	0.962	0.929	0.839
+ FS-CTAA	0.709	0.833	0.856	0.801	0.876	0.972	0.968	0.937	0.869

correction therefore improves the operating point on this subset without a corresponding ranking gain.

Table 17 decomposes the Mirage accuracy column of Table 2 into its four subdomains. No competitor exceeds 0.57 on any subdomain and none is above chance on more than two of them. Object is the hardest subdomain for three of the four frozen probes and Animal the easiest for three. The correction raises every subdomain on SigLIP2, PE-CLIP and MetaCLIP, the largest single gain being $+0.166$ on PE-CLIP Human; on DINOv2 the probe sits near chance on Human, Object and Scene and stays there.

Table 13: Detailed balanced accuracy on the four ITW benchmarks.

Method	Chameleon	WildRF	SocialRF	CommunityAI	Acc.
Competitor methods					
CNNSpot (Wang et al., 2020b)	0.554	0.625	0.565	0.541	0.571
FreqNet (Tan et al., 2024a)	0.538	0.645	0.549	0.553	0.571
Gram-Net (Liu et al., 2020)	0.518	0.576	0.527	0.523	0.536
NPR (Tan et al., 2024b)	0.523	0.612	0.565	0.537	0.559
UnivFD (Ojha et al., 2023)	0.602	0.665	0.600	0.636	0.626
SAFE (Li et al., 2025)	0.520	0.613	0.548	0.532	0.553
LaDeDa (Cavia et al., 2024)	0.504	0.554	0.524	0.506	0.522
Effort (Yan et al., 2024)	0.588	0.567	0.523	0.533	0.553
DDA (Chen et al., 2025)	0.824	0.904	0.832	0.847	0.852
AIDE (Yan et al., 2025)	0.574	0.584	0.560	0.542	0.565
OmniAID (Guo et al., 2025)	0.749	0.609	0.581	0.782	0.680
Our frozen-detector evaluations					
SigLIP2 probe (Tschannen et al., 2025)	0.803	0.790	0.719	0.880	0.798
+ FS-CTAA	0.880	0.922	0.846	0.903	0.888
PE-CLIP probe (Bolya et al., 2026)	0.881	0.843	0.751	0.971	0.861
+ FS-CTAA	0.942	0.963	0.905	0.959	0.942
MetaCLIP probe (Chuang et al., 2026)	0.644	0.724	0.637	0.643	0.662
+ FS-CTAA	0.728	0.763	0.669	0.720	0.720
DINOv2 probe (Oquab et al., 2023)	0.605	0.704	0.650	0.584	0.636
+ FS-CTAA	0.640	0.738	0.665	0.613	0.664

Table 14: Detailed AUROC on the eight GenImage generators. The final column is the arithmetic mean across the eight subsets.

Method	ADM	BigGAN	Midjourney	VQDM	GLIDE	SD-v1.4	SD-v1.5	Wukong	AUROC
Competitor methods									
NPR (Tan et al., 2024b)	0.6838	0.7963	0.9760	0.8307	0.9700	**1.0000**	0.9988	0.9986	0.9068
SAFE (Li et al., 2025)	0.8429	0.9539	0.9961	0.9931	0.9859	**1.0000**	0.9997	0.9997	0.9714
Effort (Yan et al., 2024)	0.9711	0.9999	0.9895	0.9997	0.9973	**1.0000**	**1.0000**	**1.0000**	0.9947
DDA (Chen et al., 2025)	0.9743	0.9390	0.9953	0.9268	0.9770	0.9993	0.9991	0.9994	0.9763
AIDE (Yan et al., 2025)	0.9910	0.9249	0.8368	0.9827	0.9992	0.9996	0.9989	0.9988	0.9665
OmniAID (Guo et al., 2025)	0.9150	0.9888	0.9434	0.9907	0.9992	0.9998	0.9998	0.9979	0.9793
IAPL (Li et al., 2026)	0.9910	0.9915	**0.9974**	**0.9998**	0.9973	**1.0000**	**1.0000**	0.9999	**0.9971**
TPT (full-protocol re-evaluation) (Shu et al., 2022)	***0.9942***	*0.9971*	*0.9965*	*0.9970*	*0.9968*	*0.9999*	*0.9998*	*0.9931*	*0.9968*
Our frozen-detector evaluations									
SigLIP2 probe (Tschannen et al., 2025)	0.9613	0.9850	0.9826	0.9913	0.9928	**1.0000**	**1.0000**	0.9999	0.9891
+ FS-CTAA	0.9643	0.9955	0.9820	0.9884	0.9873	**1.0000**	**1.0000**	**1.0000**	0.9897
PE-CLIP probe (Bolya et al., 2026)	0.9830	**1.0000**	0.9927	0.9994	**0.9998**	**1.0000**	**1.0000**	**1.0000**	0.9968
+ FS-CTAA	0.9702	0.9999	0.9937	0.9987	0.9992	**1.0000**	**1.0000**	**1.0000**	0.9952
MetaCLIP probe (Chuang et al., 2026)	0.7842	0.8118	0.9593	0.8690	0.8864	0.9987	0.9982	0.9915	0.9124
+ FS-CTAA	0.7615	0.8080	0.9582	0.8582	0.9027	0.9987	0.9983	0.9919	0.9097
DINOv2 probe (Oquab et al., 2023)	0.7798	0.9001	0.9277	0.8712	0.9452	0.9934	0.9908	0.9768	0.9231
+ FS-CTAA	0.7850	0.9065	0.9298	0.8828	0.9438	0.9949	0.9926	0.9793	0.9268

A.11 Accuracy Decomposition

The decomposition of the balanced-accuracy gain referenced in Section 4.4 is defined as

$$\Delta_{\mathrm{corr}} = C_{\mathrm{fixed}} - S_{\mathrm{fixed}}, \qquad \Delta_{\mathrm{cal}} = C_{\mathrm{deploy}} - C_{\mathrm{fixed}},$$
$$\Delta_{\mathrm{acc}} = C_{\mathrm{deploy}} - S_{\mathrm{fixed}}, \tag{9}$$

where S_{fixed} is the uncorrected probe, C_{fixed} is the corrected score before final calibration, and C_{deploy} is the corrected and calibrated score. The ordering follows the computation used to obtain the reported FS-CTAA predictions and separates feature correction from the subsequent operating-point adjustment. The terms satisfy

$$\Delta_{\mathrm{acc}} = \Delta_{\mathrm{corr}} + \Delta_{\mathrm{cal}}. \tag{10}$$

Table 18 separates the change in balanced accuracy into the change produced by the feature correction at the fixed threshold and the change produced by the subsequent label-free threshold. The correction contributions are $+0.056$, $+0.085$, $+0.049$, and $+0.021$ for SigLIP2, PE-CLIP, MetaCLIP, and DINOv2, while final calibration contributes $+0.008$, $+0.000$, $+0.011$, and $+0.002$, respectively; each pair sums to the Δ_{acc} of the same row. Two points follow. First, Δ_{corr} is not a ranking effect. The correction contains the rank-invariant anchor translation $\boldsymbol{\mu}_{\mathrm{ref}} - \boldsymbol{\mu}_{\mathrm{ctx}}$, and Table 1 shows that Global mean alignment, which is exactly this translation, already raises balanced accuracy to

Table 15: Detailed AUROC on the four ITW benchmarks. The final column is the arithmetic mean across the four subsets.

Method	Chameleon	WildRF	SocialRF	CommunityAI	AUROC
Competitor methods					
NPR (Tan et al., 2024b)	0.4975	0.7602	0.6430	0.5384	0.6098
SAFE (Li et al., 2025)	0.5120	0.6155	0.5652	0.4981	0.5477
Effort (Yan et al., 2024)	0.6078	0.6966	0.6051	0.6499	0.6398
DDA (Chen et al., 2025)	0.9194	0.9480	0.8904	0.9443	0.9255
AIDE (Yan et al., 2025)	0.5417	0.6830	0.5906	0.5518	0.5918
OmniAID (Guo et al., 2025)	0.8659	0.6982	0.6364	0.8905	0.7728
IAPL (Li et al., 2026)	0.5366	0.7040	0.5784	0.5165	0.5839
TPT (full-protocol re-evaluation) (Shu et al., 2022)	*0.9584*	*0.9692*	*0.9276*	*0.9688*	*0.9560*
Our frozen-detector evaluations					
SigLIP2 probe (Tschannen et al., 2025)	0.9263	0.9276	0.8106	0.9563	0.9052
+ FS-CTAA	0.9251	0.9733	0.9104	0.9647	0.9434
PE-CLIP probe (Bolya et al., 2026)	**0.9921**	0.9899	0.9565	**0.9958**	**0.9836**
+ FS-CTAA	0.9881	**0.9923**	**0.9602**	0.9922	0.9832
MetaCLIP probe (Chuang et al., 2026)	0.7694	0.8047	0.6847	0.7621	0.7552
+ FS-CTAA	0.7942	0.8443	0.7320	0.7955	0.7915
DINOv2 probe (Oquab et al., 2023)	0.6389	0.7792	0.7093	0.6123	0.6849
+ FS-CTAA	0.6894	0.8115	0.7272	0.6528	0.7202

Table 16: Detailed AUROC on the four binary Mirage subsets. The final column is their arithmetic mean.

Method	Animal	Human	Object	Scene	AUROC
Competitor methods					
NPR (Tan et al., 2024b)	0.3045	0.3190	0.3490	0.2631	0.3089
SAFE (Li et al., 2025)	0.4391	0.3844	0.4163	0.4545	0.4235
Effort (Yan et al., 2024)	0.1762	0.2922	0.2352	0.2820	0.2464
DDA (Chen et al., 2025)	0.6300	0.7081	0.5140	0.4189	0.5678
AIDE (Yan et al., 2025)	0.3964	0.3835	0.3453	0.3260	0.3628
OmniAID (Guo et al., 2025)	0.5036	0.5503	0.4009	0.3455	0.4501
IAPL (Li et al., 2026)	0.2501	0.2933	0.3023	0.3465	0.2980
Our frozen-detector evaluations					
SigLIP2 probe (Tschannen et al., 2025)	0.8932	0.8794	0.8016	0.8276	0.8505
+ FS-CTAA	**0.9382**	**0.9079**	**0.8664**	**0.8806**	**0.8983**
PE-CLIP probe (Bolya et al., 2026)	0.7384	0.7762	0.5824	0.6612	0.6895
+ FS-CTAA	0.8090	0.8540	0.7153	0.8071	0.7963
MetaCLIP probe (Chuang et al., 2026)	0.8110	0.7174	0.6679	0.6952	0.7228
+ FS-CTAA	0.8225	0.7396	0.6825	0.7191	0.7409
DINOv2 probe (Oquab et al., 2023)	0.6753	0.4380	0.4403	0.4162	0.4925
+ FS-CTAA	0.6788	0.4523	0.4472	0.4363	0.5037

0.856, 0.857, 0.726, and 0.666, that is, by $+0.029$, $+0.060$, $+0.034$, and $+0.009$ without changing AUROC. A large share of Δ_{corr} is therefore an operating-point effect that is obtained in feature space rather than in score space. Second, a larger accuracy gain does not indicate a larger change in sample ordering. The positive AUROC gains in Table 1 are the direct evidence that FS-CTAA alters ranking, and the balanced-accuracy gains over closed-form calibration in Table 2 ($+0.014$ to $+0.035$) and above the threshold-only label oracle on three backbones are the part of the accuracy that could not have been obtained by moving the threshold.

This distinction is also visible in Table 1. Prior-matching and label-oracle calibration improve balanced accuracy on SigLIP2 while leaving AUROC unchanged. Accuracy alone therefore cannot distinguish a change in operating point from a change in sample ordering, which motivates the rank-based criterion used in this work and the ΔAUROC columns in every ablation table.

A.12 Sensitivity to Target Class Composition

FS-CTAA estimates the target context and operating point from unlabeled test images. We therefore examine whether changing the class composition of the images used for test-time estimation affects the resulting predictions. To separate this effect from the direct influence of class composition on the evaluation metric, we keep the query set fixed and vary only the unlabeled adaptation set.

For each two-class target subset, we first construct a fixed query set $\mathcal{Q}$ containing equal numbers of real and generated images. The remaining images form an adaptation pool $\mathcal{A}$. From this pool, we

Table 17: Detailed balanced accuracy on the four binary Mirage subsets.

Method	Animal	Human	Object	Scene	Acc.
Competitor methods					
NPR (Tan et al., 2024b)	0.4339	0.4574	0.4489	0.4496	0.448
Effort (Yan et al., 2024)	0.5441	0.4603	0.3955	0.4901	0.473
AIDE (Yan et al., 2025)	0.3866	0.3648	0.3249	0.3864	0.366
OmniAID (Guo et al., 2025)	0.5321	0.5673	0.4942	0.4786	0.518
DIRE (Wang et al., 2023)	0.5090	0.5014	0.4926	0.4969	0.500
DRCT (Chen et al., 2024)	0.5360	0.4816	0.4938	0.4880	0.500
Our frozen-detector evaluations					
SigLIP2 probe (Tschannen et al., 2025)	0.7841	0.7938	0.7275	0.7467	0.763
+ FS-CTAA	**0.8689**	**0.8307**	**0.7870**	**0.7992**	**0.821**
PE-CLIP probe (Bolya et al., 2026)	0.6141	0.6110	0.5346	0.5846	0.586
+ FS-CTAA	0.7379	0.7773	0.6578	0.7353	0.727
MetaCLIP probe (Chuang et al., 2026)	0.6964	0.6238	0.5997	0.6381	0.640
+ FS-CTAA	0.7451	0.6778	0.6379	0.6653	0.682
DINOv2 probe (Oquab et al., 2023)	0.6416	0.4392	0.4618	0.4405	0.496
+ FS-CTAA	0.6394	0.4544	0.4707	0.4593	0.506

Table 18: Collection-macro accuracy decomposition of FS-CTAA.

	Overall balanced accuracy			AUROC		Accuracy decomposition	
Backbone	Probe	+ FS-CTAA	Δ_{acc}	Probe	ΔAUROC	Δ_{corr}	Δ_{cal}
SigLIP2	0.827	0.891	$+0.064$	0.9149	$+0.0288$	$+0.056$	$+0.008$
PE-CLIP	0.797	0.882	$+0.085$	0.8900	$+0.0349$	$+0.085$	$+0.000$
MetaCLIP	0.692	0.752	$+0.060$	0.7968	$+0.0172$	$+0.049$	$+0.011$
DINOv2	0.657	0.680	$+0.023$	0.7002	$+0.0167$	$+0.021$	$+0.002$

construct adaptation subsets with generated-image proportions

$$\Pi = \{0.20, 0.35, 0.50, 0.65, 0.80\}. \tag{11}$$

The total number of adaptation images is kept fixed across conditions, and sampling is repeated with $R = 5$ independent random seeds. Within each repetition, the same sampled adaptation subsets are used by all methods. We conduct this audit using PE-CLIP on three representative target subsets, ADM, Chameleon, and Mirage-Human, covering generator, dissemination, and content-domain shifts, respectively.

The $\pi = 0.50$ condition serves as the matched-composition reference. The remaining four ratios define the mismatched conditions. For a method $\mathcal{M}$, the adaptation subset is used only to compute quantities available at test time, including the target statistic and any label-free operating point. Adaptation-set labels are never provided to the method. The visual encoder, linear head, source reference, correction parameters, and checkpoint remain fixed throughout the audit. Query labels are used only to compute balanced accuracy after prediction.

Let $B_{\mathcal{M}}(\pi, r)$ denote the balanced accuracy of method $\mathcal{M}$ on the fixed query set under adaptation ratio π and repetition r. For each mismatched ratio, we report the signed matched-minus-mismatched difference

$$d_{\mathcal{M}}(\pi) = \frac{1}{R} \sum_{r=1}^{R} \left[B_{\mathcal{M}}(0.50, r) - B_{\mathcal{M}}(\pi, r) \right], \qquad \pi \in \Pi \setminus \{0.50\}, \tag{12}$$

together with its standard deviation across the five repetitions.

For consistency with the displayed ratio-specific means, we summarize the composition sensitivity as the mean absolute value of the four aggregated signed differences:

$$\Delta_{\mathrm{comp}}(\mathcal{M}) = \frac{1}{4} \sum_{\pi \in \Pi \setminus \{0.50\}} \left| d_{\mathcal{M}}(\pi) \right|. \tag{13}$$

Thus, Δ_{comp} summarizes the average magnitude of the matched-minus-mismatched differences of Table 19 across the four tested composition ratios.

The observed composition effects are small across the tested range. The largest reported Δ_{comp} is 0.0021 for FS-CTAA, corresponding to a 0.21 percentage-point average difference in balanced

Table 19: Sensitivity to the class composition of the unlabeled adaptation set. The query set is fixed and class balanced. Each entry reports the signed difference between the matched condition ($\pi = 0.50$) and the corresponding mismatched condition, averaged over five sampling repetitions.

Method	$\pi = 0.20$	$\pi = 0.35$	$\pi = 0.65$	$\pi = 0.80$	Δ_{comp}
Frozen VFM probe	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000	0.0000
Global mean alignment	$-$0.0013 $\pm$0.0011	$-$0.0009 $\pm$0.0007	$+$0.0001 $\pm$0.0010	$+$0.0013 $\pm$0.0014	0.0009
FS-CTAA	$-$0.0023 $\pm$0.0012	$-$0.0016 $\pm$0.0011	$+$0.0018 $\pm$0.0007	$+$0.0028 $\pm$0.0015	0.0021

Table 20: Effect of target-context granularity, before label-free calibration. Overall is the collection macro-average of balanced accuracy; ΔAcc. is relative to the per-subset context and ΔAUROC is relative to the frozen probe, so the two columns use different references. Pooling replaces the per-subset mean $\boldsymbol{\mu}_t$ by one mean shared across a collection or across all 16 subsets, and changes nothing else. The Acc. and ΔAUROC entries are both averages over the three training seeds of Table 21, which is why the per-subset column reads $+0.0295$ on SigLIP2 rather than the main-seed $+0.0288$ of Table 1. Being seed-averaged and uncalibrated, the per-subset Overall differs from the calibrated main-seed FS-CTAA row of Table 2 by up to 0.011. Both Overall orderings are monotone in context scope on every backbone. A dash marks the row that serves as that column's own reference.

Backbone	Context scope	GenImage	ITW	Mirage	Overall	ΔAcc.	ΔAUROC
SigLIP2	Frozen VFM probe	0.922	0.798	0.763	0.827	$-$0.0560	—
	Per-subset	0.961	0.879	0.807	**0.883**	—	**+0.0295**
	Collection-pooled	0.950	0.837	0.809	0.865	$-$0.0175	+0.019
	Global-pooled	0.941	0.837	0.794	0.857	$-$0.0255	+0.015
PE-CLIP	Frozen VFM probe	0.942	0.861	0.586	0.797	$-$0.0850	—
	Per-subset	0.974	0.943	0.729	**0.882**	—	**+0.0353**
	Collection-pooled	0.956	0.920	0.731	0.869	$-$0.0128	+0.022
	Global-pooled	0.937	0.880	0.665	0.827	$-$0.0545	+0.012
MetaCLIP	Frozen VFM probe	0.775	0.662	0.640	0.692	$-$0.0490	—
	Per-subset	0.844	0.710	0.667	**0.741**	—	**+0.0215**
	Collection-pooled	0.814	0.702	0.667	0.728	$-$0.0127	+0.013
	Global-pooled	0.761	0.700	0.660	0.707	$-$0.0336	+0.009
DINOv2	Frozen VFM probe	0.839	0.636	0.496	0.657	$-$0.0210	—
	Per-subset	0.866	0.664	0.504	**0.678**	—	**+0.0220**
	Collection-pooled	0.858	0.661	0.507	0.675	$-$0.0029	+0.014
	Global-pooled	0.860	0.646	0.507	0.671	$-$0.0072	+0.011

accuracy between the matched and mismatched conditions. Within the tested range of class proportions, the composition of the unlabeled adaptation set has only a small effect on balanced accuracy. The result is therefore evidence of local robustness, rather than a claim of invariance to arbitrary or extreme class imbalance.

A.13 Target-Context Granularity

The main protocol estimates one target context for each target subset. To determine how strongly the correction depends on this partition, we progressively coarsen the context from per-subset statistics to collection-level and global statistics. Table 20 evaluates the Distribution-Conditioned Modulator before the final score calibration, thereby isolating the effect of context granularity on the feature correction.

Coarsening the target context reduces both metrics monotonically on all four backbones: per-subset context performs best, followed by collection-pooled and then global-pooled context. Nevertheless, the pooled variants retain positive AUROC gains over the frozen probe. Thus, accurate target partitioning improves the correction but is not a binary requirement: performance degrades gradually as

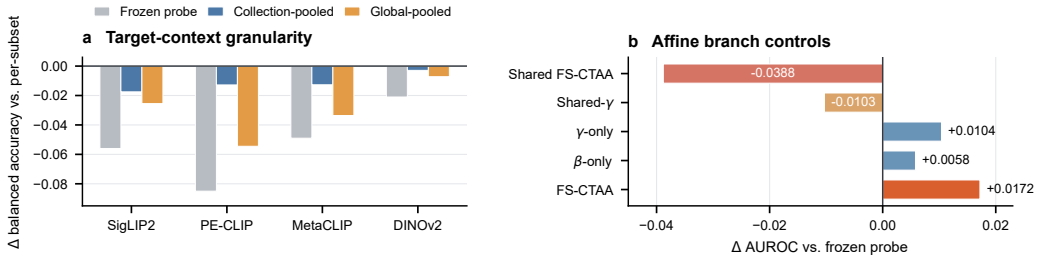

Figure 5: Target-context granularity and affine controls, discussed in Section 4.4. Panel (a) is relative to the per-subset context; panel (b) reports AUROC changes relative to the frozen probe.

Table 21: Sensitivity to three training seeds. Values are gains relative to the corresponding frozen probe. Std. is the sample standard deviation ($n-1$) across the three training seeds. Mean and Std. are computed from the displayed per-seed values and reproduce exactly from them in every row; the ΔAcc. seed-0 column is the Overall difference of Table 2 at its printed precision.

Backbone	Seed 0	Seed 1	Seed 2	Mean	Std.
ΔAcc.					
SigLIP2	+0.064	+0.067	+0.069	+0.0667	0.0025
PE-CLIP	+0.085	+0.087	+0.084	+0.0853	0.0015
MetaCLIP	+0.060	+0.066	+0.058	+0.0613	0.0042
DINOv2	+0.023	+0.026	+0.019	+0.0227	0.0035
ΔAUROC					
SigLIP2	+0.0288	+0.0337	+0.0261	+0.0295	0.0039
PE-CLIP	+0.0349	+0.0391	+0.0318	+0.0353	0.0037
MetaCLIP	+0.0172	+0.0258	+0.0216	+0.0215	0.0043
DINOv2	+0.0167	+0.0274	+0.0219	+0.0220	0.0054

the supplied context becomes less specific. This complements the main-table ablations, which test whether target context is present and correctly matched, by separately varying its aggregation scope.

A.14 Seed Sensitivity

The main tables report seed 0, which was fixed before the additional runs. We evaluate whether the performance of FS-CTAA remains stable across three independently trained Distribution-Conditioned Modulators. We report gains in Acc. and AUROC relative to the corresponding Frozen VFM probe.

The balanced-accuracy gains are positive across all four backbones and all three training seeds. Their standard deviations range from 0.0015 to 0.0042, indicating modest variation relative to the mean gains. Seed 0 is the lowest of the three seeds for the MetaCLIP and DINOv2 AUROC gains, so the main tables do not report the most favourable seed. The AUROC results are likewise positive for every backbone and seed. Thus, the direction of both the balanced-accuracy and AUROC improvements is reproducible across the three independently trained modulators; the main tables do not rely on a favourable seed.

A.15 Bootstrap Confidence Intervals

We quantify uncertainty in the ΔAUROC values reported in Tables 1 and 3 with paired bootstrap resampling. The resampling unit is the individual image, drawn with replacement within each target subset, so every resample preserves the subset structure and the real/generated class counts of Table 11 exactly; subsets are never resampled against each other, and the aggregation of Section A.16 is then applied unchanged. The two compared methods receive the identical sampled index set in every resample, which is what makes the interval an interval on the difference rather than on two independent quantities, and ties in the score are handled identically for both because both are eval-

Table 22: Paired-bootstrap 95% confidence intervals for the main-table ΔAUROC gains, computed from 10,000 resamples.

Backbone	ΔAUROC	95% CI
SigLIP2	+0.0288	[+0.0267, +0.0309]
PE-CLIP	+0.0349	[+0.0324, +0.0374]
MetaCLIP	+0.0172	[+0.0141, +0.0203]
DINOv2	+0.0167	[+0.0139, +0.0195]

Table 23: Collection-wise and 16-subset equal-weight AUROC changes relative to the frozen probe.

Backbone	GenImage	ITW	Mirage	Collection macro	16-subset equal
SigLIP2	+0.0006	+0.0382	+0.0478	+0.0288	+0.0218
PE-CLIP	−0.0016	−0.0004	+0.1068	+0.0349	+0.0258
MetaCLIP	−0.0027	+0.0363	+0.0181	+0.0172	+0.0122
DINOv2	+0.0037	+0.0353	+0.0112	+0.0167	+0.0135

uated on the same images. We draw 10,000 resamples and report the percentile 95% intervals of Table 22.

Two conventions are worth stating. The main-table intervals are computed at the single fixed main seed, so they measure test-sample uncertainty at a given trained modulator and not uncertainty over training runs. For the L4−L3 differences of Table 8, the paired difference is computed separately for each of the three fixed seeds and then averaged *within* every resample, so the interval is an interval on the seed-averaged difference. Training-seed variability is reported directly as standard deviations in Table 21.

All four intervals lie above zero under the paired resampling protocol. These intervals reflect the variability of the test sample at the fixed training seed 0 and do not include training-seed variability; the latter is reported separately in Table 21, where the seed standard deviations of ΔAUROC (0.0037 to 0.0054) are larger than the bootstrap half-widths. The same paired protocol yields the intervals for the L4−L3 differences in Table 8 and for the ablations in Tables 3, 5, and 10.

A.16 Robustness to Evaluation Aggregation

The main metric gives equal weight to the three collection averages so that generator, dissemination, and semantic shifts contribute equally. As a robustness check, Table 23 also gives equal weight to all 16 binary target subsets, comprising eight GenImage generators, four ITW benchmarks, and four Mirage subdomains. The gain remains positive on every backbone under both aggregation rules. All averages in this paper are computed at full numerical precision and rounded only for display, so recomputing an average from the rounded per-subset entries shown can differ from the reported value in the last displayed digit.

A.17 Target-Context Size and Streaming Adaptation

The main protocol estimates the target context $\boldsymbol{\mu}_t$ from the complete unlabeled target subset. Because $\boldsymbol{\mu}_t$ is an input to the modulator rather than a fitted parameter, it can be estimated from a finite batch or updated incrementally. This section measures the resulting sensitivity to context size and update rule. Table 24 reports a separate sensitivity experiment on frozen SigLIP2 over four target subsets chosen to span the three collections: ADM from GenImage, Chameleon from the in-the-wild benchmarks, and Mirage-Human and Mirage-Anime from Mirage. Each subset is subsampled to at most 128 images and split into 20 random partitions at every context size.

Three details make this table self-contained; they also explain why its absolute values are on a different scale from the main tables. First, the reported figure is plain classification accuracy without the label-free calibration of equation 8, so it isolates the quality of the estimated context from the threshold rule. Second, the aggregate is an equal-weight average over the subset-by-partition records rather than a pooled accuracy over images, so each subset contributes equally regardless of its size. Third, Mirage-Anime contains generated images only (Table 11), so balanced accuracy and AUROC are undefined on it; including it is what forces the plain-accuracy metric here, and it is also why this

Table 24: Target-context sensitivity on frozen SigLIP2 over four target subsets (ADM, Chameleon, Mirage-Human, Mirage-Anime), each subsampled to at most 128 images with 20 random partitions per setting. Aggregate accuracy is plain classification accuracy without label-free calibration, averaged with equal weight over the subset-by-partition records rather than pooled over images. Batch size is the number of unlabeled target images used to estimate $\boldsymbol{\mu}_t$.

Target-statistic setting	Aggregate accuracy
Batch size 1	0.7793
Batch size 4	0.7961
Batch size 8	0.8109
Batch size 16	0.8206
Batch size 32	0.8265
Batch size 64	0.8307
Full diagnostic sample ($n = 128$)	**0.8320**
Cumulative streaming mean	0.8262
Exponential moving average (decay 0.9)	0.8257
Source-mean fallback	0.7988

Table 25: Unknown-boundary mixed stream against fixed contexts, in balanced accuracy. Images from all 16 target subsets are shuffled into one stream and $\boldsymbol{\mu}_t$ is tracked by an exponential moving average, so the context is never matched to the subset of the current image and no boundary is supplied. Overall is the collection macro-average.

Backbone	Target context	GenImage	ITW	Mirage	Overall
SigLIP2	Per-subset	0.963	0.888	0.821	**0.891**
	Unknown-boundary EMA	0.952	0.866	0.813	0.877
	Global-pooled	0.943	0.846	0.803	0.864
PE-CLIP	Per-subset	0.976	0.942	0.727	**0.882**
	Unknown-boundary EMA	0.953	0.908	0.699	0.853
	Global-pooled	0.939	0.879	0.662	0.827

subset is excluded from the 16 target subsets of the main evaluation. The full-sample row of this table should therefore be read as the ceiling of this experiment, not as a restatement of the main-table result.

Tables 25 and 26 remove the subset boundary altogether. Images from all 16 target subsets are interleaved in one stream and $\boldsymbol{\mu}_t$ is tracked by an exponential moving average, with no subset identity available to the estimator. Relative to the per-subset context, the mixed stream retains $+0.0202$ of $+0.0288$ AUROC on SigLIP2 and $+0.0227$ of $+0.0349$ on PE-CLIP, while remaining above the global-pooled context in balanced accuracy. Thus, unknown boundaries reduce but do not eliminate the benefit of target conditioning in this streaming audit.

A.18 Affine Branch Controls

Shared FS-CTAA replaces both sample-specific affine branches with their target-subset averages. Shared-$\boldsymbol{\gamma}$ averages only the scale branch while retaining sample-specific translations, and Shared-$\boldsymbol{\beta}$ averages only the translation branch while retaining sample-specific scales. The $\boldsymbol{\gamma}$-only and $\boldsymbol{\beta}$-only controls retain one sample-specific branch and disable the other. Sharing the scale branch alone is harmful on MetaCLIP, while sharing both affine branches causes a larger degradation; retaining a sample-specific scale with a shared translation remains beneficial. Both single-branch controls improve on the frozen probe, but the complete sample-specific transformation performs best. The degradation is therefore tied to sharing the scale, which reweights the head, rather than to sharing the translation, which is rank-invariant on its own.

A.19 Source-Mean Substitution at Inference

The ablations of Table 3 remove target information before training, so each variant is free to find a different solution. That answers what an alternative objective can reach, not what the deployed model relies on. This section isolates the second question. We take the trained FS-CTAA modulator

Table 26: ΔAUROC of the unknown-boundary mixed stream by collection, measured against the frozen probe. Overall is the unweighted mean of the three collection entries.

Backbone	Target context	GenImage	ITW	Mirage	Overall
SigLIP2	Per-subset	+0.0006	+0.0382	+0.0478	**+0.0288**
	Unknown-boundary EMA	+0.0002	+0.0268	+0.0335	+0.0202
PE-CLIP	Per-subset	−0.0016	−0.0004	+0.1068	**+0.0349**
	Unknown-boundary EMA	−0.0018	−0.0011	+0.0710	+0.0227

Table 27: MetaCLIP AUROC under shared and sample-specific affine controls. Δ is measured against the frozen MetaCLIP probe.

Variant	Overall AUROC	Δ vs. frozen
Frozen MetaCLIP probe	0.7968	0.0000
Shared FS-CTAA (post-hoc; both branches shared)	0.7580	−0.0388
Shared-γ (shared scale, sample-specific translation)	0.7865	−0.0103
Shared-β (sample-specific scale, shared translation)	0.8078	+0.0110
γ-only (sample-specific scale, no translation)	0.8072	+0.0104
β-only (sample-specific translation, no scale)	0.8026	+0.0058
FS-CTAA	**0.8140**	**+0.0172**

without modification and, at inference only, replace the target subset mean $\boldsymbol{\mu}_t$ by the stored source mean $\boldsymbol{\mu}_s$ wherever it enters the modulator input and the anchor. Every weight, the concatenation layout, the per-sample conditioning on $\boldsymbol{z}_i$, the label-free threshold rule, and the transductive per-subset protocol are unchanged; only the numerical value of one of the three input blocks differs. The comparison of Table 28 is therefore a controlled test of target-distribution conditioning rather than of the affine form.

Table 28: Replacing the target mean by the source mean at inference only. "Target mean" is the deployed FS-CTAA setting; "Source mean" reuses the same trained modulator with $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\mu}_s$. ΔAUROC is measured against the frozen probe and averaged over the three training seeds of Table 21.

	ΔAUROC	
Backbone	Target mean	Source mean
SigLIP2	**+0.030**	+0.004
PE-CLIP	**+0.035**	+0.006
MetaCLIP	**+0.022**	+0.003
DINOv2	**+0.022**	+0.002

Replacing the target mean with the source reference removes most of the ranking gain: ΔAUROC falls to +0.002–+0.006 across the four backbones, compared with +0.022–+0.035 using the deployed target statistic. At the main seed, the same substitution also reduces balanced accuracy by 0.017–0.049. Because the trained modulator and affine form are unchanged, this intervention isolates the target statistic as a major source of the deployed model's distribution-specific correction.

This experiment answers a different question from the "w/o target-mean input" ablation in Table 3: that variant is retrained and can compensate for the missing input, whereas the substitution here probes what the already trained deployed model relies on.

A.20 Augmented and Higher-Capacity Source Heads

The main granularity criterion and its diagnostic suite assume a frozen linear head. As a supplementary experiment outside that theoretical setting, we ask whether two stronger source-side alternatives can substitute for target-conditioned correction: training the linear head with the same forensic-shift families used by FS-CTAA, or replacing the linear head with a higher-capacity frozen classifier.

All source heads in Table 29 are trained within the same study. "Aug. linear" applies the four forensic-shift families during source-side head training. "Matched MLP" replaces the linear probe

with a two-layer $d \to h \to 2$ MLP with one GELU-activated hidden layer, whose parameter count is matched to the FS-CTAA modulator of Table 6. We use $h = 2729, 2813, 2763$, and 2729 for SigLIP2, PE-CLIP, MetaCLIP, and DINOv2, respectively, matching the corresponding modulator parameter counts within 0.02%.

For "MLP + FS-CTAA", the linear-head FS-CTAA weights are *not* reused. After the matched MLP is trained, both the backbone and the MLP are frozen, and a fresh FS-CTAA modulator is trained for that detector using the same forensic-shift episode construction, modulator architecture, training objective, and optimization protocol as in the main experiments. The classification and logit-alignment terms of equation 7 are evaluated through the frozen MLP head rather than the frozen linear head. At test time, the resulting modulator uses the same unlabeled target-context interface and label-free threshold rule as the main method.

Table 29: Source-side augmentation and head-capacity controls. Acc. is the three-collection macro-average balanced accuracy. Probe, Aug. linear, and Matched MLP use source-trained heads without target-conditioned correction; "MLP + FS-CTAA" uses a separately trained FS-CTAA modulator for the corresponding frozen MLP head. Its final column reports the threshold-free AUROC change relative to that Matched MLP.

	Acc.				
Backbone	Probe	Aug. linear	Matched MLP	MLP + FS-CTAA	ΔAUROC vs. MLP
SigLIP2	0.827	0.818	0.824	**0.879**	$\mathbf{+0.026}$
PE-CLIP	0.797	0.789	0.797	**0.874**	$\mathbf{+0.030}$
MetaCLIP	0.692	0.681	0.689	**0.744**	$\mathbf{+0.015}$
DINOv2	0.657	0.648	0.654	**0.673**	$\mathbf{+0.013}$

Under this protocol, neither source-side augmentation nor additional source-head capacity consistently improves upon the frozen linear probe. After the higher-capacity MLP is fixed, a separately trained FS-CTAA still adds $+0.013$–$+0.030$ AUROC over that MLP. Thus, the benefit of target-conditioned correction persists after increasing source-head capacity. Because the MLP head is nonlinear, this experiment is an auxiliary empirical check rather than an extension of the linear-head rank-invariance result in Proposition 1.