RETHINKING MEAN ACCURACY COMPARISONS IN PEDESTRIAN ATTRIBUTE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pedestrian attribute recognition (PAR) methods are commonly compared using mean accuracy (mA), the macro-average of per-attribute balanced accuracy. Our paper-and-code audit finds heterogeneous label-imbalance interventions paired with a small set of numerical inference thresholds. Because these interventions can change score meaning, native mA comparisons can involve different operating policies. We develop a decision-coordinate framework that derives objective-implied coordinates for identifiable loss and sampling components, and estimates trained predictors' score-space coordinates through validation-set isotonic calibration. The framework complements native mA with objective and realized-policy normalization toward the common balanced-accuracy decision target. We evaluate controlled methods with ResNet-50 and ViT-B/16 backbones and published PAR methods, holding each predictor fixed across evaluation views and using disjoint test data. On RAP, the gap between inverse-prior reweighting and unweighted binary cross-entropy decreases from 8.99 to 0.12 percentage points with ResNet-50 and from 8.16 to 0.28 with ViT-B/16 under realized-policy normalization. Comparisons among published methods using the same backbone also exhibit reversals in mean ordering, while some gaps persist or widen. These results show that the magnitude and direction of method advantages can depend on operating policy, supporting joint reporting of native and normalized mA.

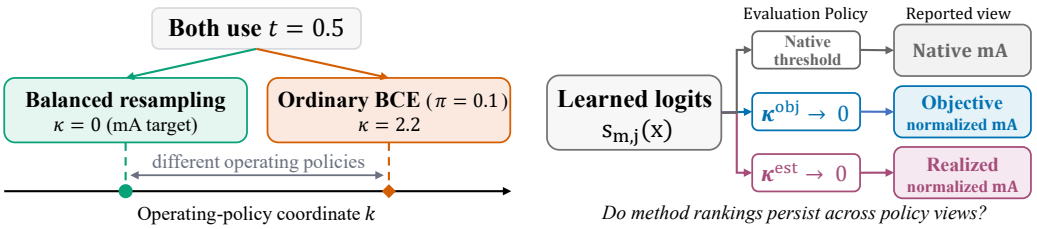
1 INTRODUCTION

Pedestrian attribute recognition (PAR) predicts multiple semantic attributes from a pedestrian image. Since the benchmark introduced by Deng et al. (2014), PAR has progressed from task-specific convolutional and relation-aware architectures (Wang et al., 2022) to transformers and general-purpose human-centric pretraining. Despite these changes, PAR evaluation continues to center on mean accuracy (mA), the macro-average of per-attribute balanced accuracy and a principal criterion for cross-method comparison. This convention persists in recent human-centric benchmarks, pretrained representations, and generalist models (Tang et al., 2023; Zuo et al., 2024; Wang et al., 2025b). Instance-level accuracy, precision, recall, and F1 provide complementary measures.

At the same time, label imbalance is a central concern in PAR training. Representative approaches include class reweighting (Li et al., 2015), focal and asymmetric objectives (Sarafianos et al., 2018; Fan et al., 2024), and balanced classifier retraining through FRDL (Zhou et al., 2025b). These interventions are designed to improve learning under imbalance, but they can also change what numerical scores mean and which decision boundary a fixed threshold implements. The same threshold, typically 0.5, can therefore correspond to different operating policies across training objectives. Native $\mathrm{mA}@0.5$ consequently need not compare methods under a common policy. We ask whether this mismatch materially changes the conclusions drawn from PAR method comparisons.

Our paper-and-code audit establishes the practical scope of this question. We examine a primary cohort of 66 formally published full-text-verifiable closed-set PAR papers from 2019 through our audit cutoff (Section 3.2). Standard balanced mA is verifiable in 38 papers, and 44 explicitly document label-imbalance interventions; numerical final thresholds are rarely stated. Among 28 audited official implementations, 27 use probability boundaries of 0.3, 0.45, or 0.5, and one uses autoregressive decoding. The implementations that threshold scores combine heterogeneous training interventions with a small set of numerical cutoffs whose operating-policy meaning is largely implicit.

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.



(a) Same threshold, different policies **(b)** Native and policy-normalized reporting

Figure 1: **Operating-policy mismatch and policy-aware evaluation.** (a) At their population optima, balanced resampling and ordinary BCE imply $\kappa = 0$ and $\kappa = 2.20$, respectively, for $t = 0.5$ and $\pi = 0.1$; mA targets $\kappa = 0$. (b) We report native mA alongside mA after threshold alignment targeting $\kappa = 0$, using objective-implied boundaries or validation-estimated score-space policies.

To see why this matters, consider an attribute with positive prevalence π. At its population optimum, ordinary binary cross-entropy (BCE) estimates the natural posterior (Reid & Williamson, 2010), so thresholding at 0.5 implements a prior-dependent likelihood-ratio boundary. Inverse-prior weighting makes the same numerical threshold implement the balanced-accuracy boundary instead. We denote the induced log-likelihood-ratio boundary by κ. Prior work relates class priors, loss weighting, and decision thresholds in cost-sensitive and class-imbalanced classification (Elkan, 2001; Menon et al., 2013; Koyejo et al., 2014; Menon et al., 2021). We use the induced boundary as a measurement coordinate for auditing the operating policies underlying PAR comparisons. In this population coordinate, mA has the method-independent target $\kappa^{\star}_{\mathrm{mA}} = 0$, providing a shared decision reference for objectives with different score coordinates. Fig.1.a illustrates this distinction at a shared threshold.

We operationalize this coordinate at two complementary levels. For classification components with identifiable population boundaries, we derive the objective-implied policy, κ^{obj}, from the training specification and inference threshold. Because finite predictors need not realize these boundaries, we also estimate a score-space policy, κ^{est}, through a monotone score-to-posterior map (Zadrozny & Elkan, 2002). These coordinates support two complementary ways to align decision thresholds: objective normalization uses the derived population boundaries, while realized-policy normalization uses the policies estimated from trained predictors. We retain native mA and complement it with both normalized views targeting $\kappa = 0$, as summarized in Fig.1.b.

Across three PAR benchmarks, we evaluate 9 controlled methods with ResNet-50 and ViT-B/16 backbones and 7 published PAR methods across evaluation views. On RAP, the absolute mA gap between inverse-prior reweighting and ordinary BCE falls from over 8 to below 0.5 percentage points under both normalization views on both backbones. Comparisons among published methods within architecture groups also exhibit ordering reversals, while some advantages persist or widen. Together, these findings support joint reporting of native and normalized mA to identify which comparative conclusions persist across evaluation views and which depend on operating-policy choice.

Our contributions are threefold. First, we systematically audit mA reporting, label-imbalance interventions, and decision thresholds in contemporary PAR papers and code. Second, we develop a decision-coordinate framework for mA comparisons: we derive objective-implied coordinates κ^{obj} for identifiable loss and sampling components under specified inference thresholds, and estimate trained predictors' score-space coordinates κ^{est} from validation data. Third, paired evaluations characterize how operating policy affects method gaps and ordering, and assess the sensitivity of normalized comparisons to calibration choices and validation data.

2 Related Work

PAR models and evaluation protocols. PAR research explores visual attention and attribute relationships, as summarized by Wang et al. (2022). Label2Label models attribute dependencies (Li et al., 2022), while PARFormer combines Transformer-based feature learning with auxiliary objectives (Fan et al., 2024). Human-centric pretraining (Yuan et al., 2023; Tang et al., 2023), vision–language fusion (Wang et al., 2025a), unified modeling across datasets and modalities (Xu et al., 2026), and generative data augmentation (Ayuso-Albizu et al., 2026) further broaden the design

space. Evaluation protocols have also been reassessed: Jia et al. (2021) examine pedestrian-identity overlap and introduce identity-disjoint benchmarks. We examine a complementary aspect of PAR evaluation: the operating policies underlying mA comparisons.

Imbalanced and long-tailed learning. Reweighting (Cui et al., 2019), logit adjustment (Menon et al., 2021), and decoupled training (Kang et al., 2020) address unequal class frequencies. In multi-label learning, asymmetric loss addresses positive–negative imbalance (Ridnik et al., 2021), while distribution-balanced loss also accounts for label co-occurrence (Wu et al., 2020). PAR employs frequency-based weighting and focal modulation (Li et al., 2015; 2020). Label co-occurrence also motivates feature-level resampling (Zhou et al., 2025b) and interventions against correlation bias (Zhou et al., 2025a; Zhang & Hu, 2026). Under mA, positive and negative recall receive equal weight within each attribute, regardless of label frequencies. This equal weighting motivates examining the operating policies jointly induced by imbalance interventions and inference rules.

Decision theory and threshold selection. Decision theory characterizes how costs and evaluation criteria determine optimal classification rules (Elkan, 2001; Koyejo et al., 2014). Composite-loss theory relates surrogate objectives to probability estimation (Reid & Williamson, 2010). Menon et al. (2013) establish balanced-accuracy consistency for posterior thresholding and balanced risk minimization. Threshold-selection studies further examine how operating conditions and metric choice govern the conversion of scores into decisions (Hernández-Orallo et al., 2012; Pillai et al., 2013). Building on these foundations, we develop a decision-coordinate framework for PAR and derive objective-implied coordinates for identifiable loss and sampling components under specified inference thresholds, using the mA target as a common reference.

Calibration and operating-policy estimation. Accurate classification need not yield reliable probability estimates, motivating post-hoc score calibration (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). Population-optimal scores may differ from posterior probabilities, as illustrated by analyses of focal loss (Charoenphakdee et al., 2021). MiSLAS addresses class-dependent miscalibration in decoupled long-tailed learning (Zhong et al., 2021). Within our framework, we define score-space policy coordinates and estimate them through validation calibration, enabling native and policy-normalized mA comparisons with each trained predictor held fixed.

3 Operating-Policy Coordinates for mA

3.1 The mA Target in Likelihood-Ratio Coordinates

Let $X \in \mathcal{X}$ denote a pedestrian image and $Y = (Y_1, \dots, Y_C) \in \{0,1\}^C$ its C binary attributes. The population analysis assumes a common distribution for the original training, validation, and test data; Appendix B.5 discusses distribution shifts. For attribute j, define its positive prevalence, class-conditional densities, and log-likelihood ratio as

$$\pi_j = \Pr(Y_j = 1), \qquad f_{y,j}(x) = p(x \mid Y_j = y), \qquad \lambda_j(x) = \log \frac{f_{1,j}(x)}{f_{0,j}(x)}. \tag{1}$$

Writing the natural posterior as $\eta_j(x) = \Pr(Y_j = 1 \mid X = x)$, Bayes' rule gives

$$\operatorname{logit} \eta_j(x) = \lambda_j(x) + \operatorname{logit} \pi_j, \qquad \operatorname{logit}(u) = \log \frac{u}{1-u}. \tag{2}$$

Method m produces a logit score $s_{m,j}(x)$. Using the sigmoid σ and threshold $t_{m,j}$, its prediction is

$$\hat{y}_{m,j}(x) = 1 \quad \Longleftrightarrow \quad \sigma\big(s_{m,j}(x)\big) > t_{m,j} \quad \Longleftrightarrow \quad s_{m,j}(x) > \operatorname{logit}\big(t_{m,j}\big). \tag{3}$$

When method m has a population-optimal likelihood-ratio threshold rule, we write

$$\hat{y}^{\star}_{m,j}(x) = 1 \quad \Longleftrightarrow \quad \lambda_j(x) > \kappa_{m,j}. \tag{4}$$

In this case, $\kappa_{m,j}$ is the population operating-policy coordinate: a likelihood-ratio threshold jointly determined by training and inference, rather than by $t_{m,j}$ alone.

The mA target. For a binary predictor $\hat{y}_j : \mathcal{X} \to \{0,1\}$, the balanced accuracy of attribute j is

$$\begin{aligned} \mathrm{BA}_j(\hat{y}_j) &= \frac{1}{2} \Pr\big(\hat{y}_j(X) = 1 \mid Y_j = 1\big) + \frac{1}{2} \Pr\big(\hat{y}_j(X) = 0 \mid Y_j = 0\big) \\ &= \frac{1}{2} \int \hat{y}_j(x) f_{1,j}(x)\, dx + \frac{1}{2} \int [1 - \hat{y}_j(x)] f_{0,j}(x)\, dx. \end{aligned} \tag{5}$$

Pointwise maximization yields a Bayes-optimal rule satisfying $\hat{y}_j^\star(x) = 1 \iff f_{1,j}(x) > f_{0,j}(x) \iff \lambda_j(x) > 0$, with arbitrary tie-breaking when $\lambda_j(x) = 0$. PAR mA averages these quantities as $\mathrm{mA} = \frac{1}{C}\sum_{j=1}^{C} \mathrm{BA}_j$; hence its method-independent population target is

$$\boxed{\kappa^\star_{\mathrm{mA},j} = 0, \qquad j = 1, \ldots, C.} \tag{6}$$

Thus, $\kappa = 0$ is the balanced-accuracy target for all attributes (Brodersen et al., 2010; Elkan, 2001). We use this boundary to characterize the operating policies induced by training and inference.

A shared probability threshold therefore does not ensure policy comparability. Under the population threshold setting above, a common policy requires the same κ. We next audit how contemporary PAR papers document training interventions and inference thresholds governing this coordinate.

3.2 A Protocol Audit of Contemporary PAR

Scope and coding. We audit formally published studies with verifiable full text from 2019 through the cutoff that report PAR results on the mainstream benchmarks PETA, PA100K, or the RAP family (Deng et al., 2014; Li et al., 2016; Liu et al., 2017; Li et al., 2019). PAR may be the primary task or a downstream evaluation. Five eligible pre-2019 studies are retained only as qualitative historical anchors. For each paper in the cohort, we record verifiable evidence of standard balanced mA, explicit label-imbalance interventions, and final decision rules. Official implementations are analyzed separately, so their availability does not affect the paper-level denominator. The audit scope, coding criteria, and paper- and code-level records are provided in Appendix A.

Audit findings. Table 1 shows that 26 papers document both standard balanced mA and an imbalance intervention, while final numerical thresholds are rarely stated. Across heterogeneous training objectives, the audited code uses three numerical probability boundaries and an autoregressive decoder. This motivates a coordinate analysis of scalar score-threshold policies. We derive κ^{obj} for identifiable classification components and define κ^{est} for fixed predictors. Autoregressive decoding is outside this coordinate analysis.

Table 1: Protocol-audit counts. Paper- and code-level evidence use separate denominators.

Paper-level evidence ($N = 66$)	**Count**
Verified standard mA	38
Explicit label-imbalance intervention	44
Verified mA and imbalance intervention	26
Numerical final threshold (0.5)	2
Fixed threshold, value undisclosed	1
Autoregressive decoding stated	2
No final decision-rule statement	61
Code-level evidence ($N = 28$)**: rule/count**	
0.5/19 0.45/7 0.3/1	Autoregressive decoding/1

3.3 Objective-Implied and Realized Policies

We analyze two broad families of classification interventions: fixed class weighting or resampling (Li et al., 2015; Zhou et al., 2025b), and prediction-dependent focal or asymmetric objectives (Sarafianos et al., 2018; Li et al., 2020; Fan et al., 2024). We define κ^{obj} only for an identifiable pointwise classification component with a unique, monotone population-optimal response.

Fixed reweighting and resampling. For attribute j, let $\eta = \eta_j(x)$ and let $a_j, b_j > 0$ denote the positive- and negative-class weights. The conditional weighted-BCE risk and induced boundary are

$$\begin{aligned} \mathcal{R}_j(p;\eta) &= -a_j \eta \log p - b_j(1-\eta)\log(1-p), \\ \kappa^{\mathrm{obj}}_{m,j}(t_{m,j}) &= \operatorname{logit} t_{m,j} + \log\frac{b_j}{a_j} - \operatorname{logit}\pi_j. \end{aligned} \tag{7}$$

Class-dependent resampling that preserves the class-conditional distributions and sets the positive prior to r_j yields, under ordinary BCE, $\kappa^{\mathrm{obj}}_{m,j}(t_j) = \operatorname{logit} t_j - \operatorname{logit} r_j$.

Prediction-dependent objectives. Suppressing the attribute index, let $\ell_+(z)$ and $\ell_-(z)$ denote the positive- and negative-class losses of method m, where z is a scalar logit. At threshold t, define

$$g_\pm(t) = \left.\frac{\partial \ell_\pm(z)}{\partial z}\right|_{\sigma(z)=t}, \qquad \eta^\dagger_m(t) = \frac{g_-(t)}{g_-(t) - g_+(t)}, \qquad \kappa^{\mathrm{obj}}_m(t) = \operatorname{logit}\eta^\dagger_m(t) - \operatorname{logit}\pi. \tag{8}$$

Table 2: Objective-implied coordinates at $t = 0.5$, with $\pi = \pi_j$. Rows summarize reference and audited classification components, not necessarily full methods.

Component	**Parameters**	$\boldsymbol{\kappa^{\mathrm{obj}}(0.5)}$
Ordinary BCE	$a = b = 1$	$-\operatorname{logit}\pi$
Inverse-prior RW_{inv}	$a = 1/\pi,\ b = 1/(1-\pi)$	0
RW_{exp} (Li et al., 2015)	$a = e^{1-\pi},\ b = e^{\pi}$	$2\pi - 1 - \operatorname{logit}\pi$
RW_{sqrt} (Tan et al., 2020)	$a = \pi^{-1/2},\ b = (1-\pi)^{-1/2}$	$-\frac{1}{2}\operatorname{logit}\pi$
FRDL (Zhou et al., 2025b)	$r = 0.5$	0
DIAA Focal (Sarafianos et al., 2018)	$a = b,\ \gamma_{\pm} = 0.5$	$-\operatorname{logit}\pi$
MLFL (Li et al., 2020)	$a = e^{1-\pi},\ b = e^{\pi},\ \gamma_{\pm} = 1$	$2\pi - 1 - \operatorname{logit}\pi$
PARFormer ASL (Fan et al., 2024)	$a = b = 1,\ \gamma_{+} = 0,\ \gamma_{-} = 1$	$\log\frac{1+\log 2}{2} - \operatorname{logit}\pi$

For the conditional risk $\mathcal{R}_m(z;\eta) = \eta\ell_+(z) + (1-\eta)\ell_-(z)$, an interior population optimum at $z = \operatorname{logit} t$ satisfies $\eta g_+(t) + (1-\eta)g_-(t) = 0$, yielding the expression for $\eta_m^{\dagger}(t)$. Under $g_+(t) < 0 < g_-(t)$ and the unique monotone-response condition, $\eta_m^{\dagger}(t)$ identifies the posterior boundary for thresholds attained by the population-optimal response.

Equation 7 applies at any specified threshold, whereas Eq.8 applies where the stated derivative-sign and monotonicity conditions hold. Tab.2 lists the $t = 0.5$ cases derived in Appendix B. When a native threshold is undisclosed, the corresponding entry is conditional rather than a native-policy assignment. At $t = 0.5$, Multi-Label Focal Loss (MLFL) and RW_{exp} share the same κ^{obj} despite different loss forms. This is decision-boundary equivalence, not score or full-method equivalence.

Finite-model realization. A finite network need not attain its objective-implied boundary. For the fixed score $S_{m,j} = s_{m,j}(X)$, define $c_{m,j}(s) = \Pr(Y_j = 1 \mid S_{m,j} = s)$. Under the same reference distribution, Eq.2 identifies $\operatorname{logit} c_{m,j}(s) - \operatorname{logit}\pi_j$ as the score-space log-likelihood ratio. We model $c_{m,j}$ with a nondecreasing calibration map $\widehat{c}_{m,j}$ fitted to validation-set logits and labels. At the native logit threshold $\tau_{m,j} = \operatorname{logit} t_{m,j}$, the estimated policy coordinate is

$$\boxed{\kappa^{\mathrm{est}}_{m,j} = \operatorname{logit}\widehat{c}_{m,j}(\tau_{m,j}) - \operatorname{logit}\pi_j}. \tag{9}$$

3.4 Native and Policy-Normalized mA

All three evaluations use the same frozen checkpoint for a given method, dataset, and seed. Write $\mathrm{mA}_m(\boldsymbol{\tau})$ for mA with attribute decisions $s_{m,j}(x) > \tau_j$. By default, empirical coordinates and normalized thresholds use the training-set estimate $\widehat{\pi}_j^{\mathrm{tr}}$ in place of π_j.

Native evaluation. We retain the native inference threshold specified for each evaluated method. With $\tau^{\mathrm{native}}_{m,j} = \operatorname{logit} t^{\mathrm{native}}_{m,j}$, the native result is

$$\mathrm{mA}^{\mathrm{native}}_m = \mathrm{mA}_m\big(\boldsymbol{\tau}^{\mathrm{native}}_m\big). \tag{10}$$

Objective normalization. For each identifiable classification component, we choose a probability threshold $t^{\mathrm{obj}}_{m,j}$ whose objective-implied coordinate matches the mA target:

$$\kappa^{\mathrm{obj}}_{m,j}\big(t^{\mathrm{obj}}_{m,j}\big) = 0, \qquad \tau^{\mathrm{obj}}_{m,j} = \operatorname{logit} t^{\mathrm{obj}}_{m,j}. \tag{11}$$

The corresponding result is

$$\mathrm{mA}^{\mathrm{obj\text{-}norm}}_m = \mathrm{mA}_m\big(\boldsymbol{\tau}^{\mathrm{obj}}_m\big). \tag{12}$$

The objective-normalized threshold depends only on the documented objective, its fixed hyperparameters, and the class prior.

Realized-policy normalization. Within a fixed predictor's score space, the balanced-accuracy Bayes rule predicts positive when $c_{m,j}(s) > \pi_j$. For nondecreasing $c_{m,j}$, this rule can be implemented by a score threshold with appropriate tie handling. Est. applies the same validation-based threshold-construction procedure to each fixed predictor, selecting attribute-wise thresholds to target the common mA reference $\kappa = 0$:

$$\tau^{\mathrm{est}}_{m,j} = \widehat{c}^{-1}_{m,j}(\pi_j), \qquad \mathrm{mA}^{\mathrm{est\text{-}norm}}_m = \mathrm{mA}_m\big(\boldsymbol{\tau}^{\mathrm{est}}_m\big), \tag{13}$$

where $\widehat{c}_{m,j}^{-1}$ is the left generalized inverse. Maps and thresholds are frozen before evaluation on disjoint test data. Appendix C details interpolation, tie handling, and support conventions.

The three views answer complementary questions. We use objective normalization and realized-policy normalization to test whether cross-method conclusions persist when thresholds are set using objective-implied boundaries or validation-estimated score-space policies, respectively. Together with κ^{obj} and κ^{est}, these views make both performance and its policy dependence explicit.

4 Empirical Evaluation

We empirically examine how native operating policies vary across methods, how validation-estimated score-space coordinates relate to objective-implied coordinates, and whether method comparisons persist after policy normalization.

4.1 Experimental Setup

Datasets. We evaluate on three widely used PAR benchmarks: PETA (Deng et al., 2014), RAP (Li et al., 2016), and PA100K (Liu et al., 2017). Dataset statistics and splits are provided in Appendix D.

Methods and native evaluation. We evaluate two complementary groups of methods. (1) *Controlled label-imbalance interventions.* Under a shared training framework based on the strong PAR baseline (Jia et al., 2021), we compare ERM with RW_{inv}, RW_{exp} (Li et al., 2015), FRDL (Zhou et al., 2025b), MLFL (Li et al., 2020), Asymmetric Loss (ASL) (Ridnik et al., 2021; Fan et al., 2024), Class-Balanced Loss (CB) (Cui et al., 2019), training-time Logit Adjustment (LA) (Menon et al., 2021), and Distribution-Balanced Loss (DB) (Wu et al., 2020). (2) *Published PAR methods.* We use the official implementations of Label2Label (L2L) (Li et al., 2022), HAP (Yuan et al., 2023), PARFormer (Fan et al., 2024), PromptPAR (Wang et al., 2025a), PIL (Zhou et al., 2025a), FRDL+GOAT (Zhou et al., 2025b), and ReSAGE-PAR (Ayuso-Albizu et al., 2026). ReSAGE-PAR is evaluated as data augmentation for the baseline of Jia et al. (2021).

Except for PromptPAR, we train each method with three independent seeds and evaluate the checkpoint saved after convergence for each run. Validation data are used for policy estimation. PromptPAR is evaluated using one officially released checkpoint per dataset. Native evaluation uses a probability threshold of 0.5 for controlled methods and the verified method-specific inference rules for published methods. Each evaluated checkpoint and its test logits remain fixed across evaluation views. Further details are provided in Appendix D.

Policy estimation. For each dataset, we estimate π_j from the training set as $\widehat{\pi}_j^{\mathrm{tr}}$. For each evaluated checkpoint and attribute, we fit the nondecreasing calibration map $\widehat{c}_{m,j}$ by isotonic regression on validation-set logits and labels. We evaluate this map at the native logit threshold to estimate $\kappa_{m,j}^{\mathrm{est}}$ and invert it at $\widehat{\pi}_j^{\mathrm{tr}}$ for realized-policy normalization. Native $\kappa_{m,j}^{\mathrm{obj}}$ and objective normalization refer to identifiable classification components; estimated policies describe the complete predictor. Calibration maps and thresholds are frozen before evaluation on the test set.

4.2 Native and Policy-Normalized Comparisons

Table 3 reports native mA, objective-normalized mA (Obj.), and realized-policy-normalized mA (Est.), where Est. uses validation-estimated score-space policies. Controlled methods are presented separately for ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021), followed by published PAR methods grouped by architecture family. All three views evaluate the same checkpoint for each predictor. We examine how method gaps and relative ordering change across views.

Controlled comparisons across backbones. The controlled results show that the magnitude of an imbalance intervention's advantage over ERM can depend strongly on operating policy. On RAP, for example, RW_{inv} has a native advantage of more than 8 percentage points (pp) over ERM on both backbones. Under both Obj. and Est., the corresponding absolute gaps fall below 0.5 pp. The *Max. gap* summaries further show that the largest within-group mA difference decreases from Native to Est. on all three datasets for both backbones. LA's strong native performance is consistent with its objective-level alignment with the mA target ($\kappa_{\mathrm{LA}}^{\mathrm{obj}} = 0$).

Table 3: Native and policy-normalized mA (%). Obj. and Est. denote objective normalization and realized-policy normalization using validation-estimated score-space policies, respectively.

Method	PETA			PA100K			RAP		
	Native	Obj.	Est.	Native	Obj.	Est.	Native	Obj.	Est.
Controlled methods—ResNet-50									
ERM	$82.10_{\pm 0.42}$	$87.78_{\pm 0.18}$	$87.93_{\pm 0.09}$	$77.80_{\pm 0.16}$	$83.96_{\pm 0.71}$	$85.97_{\pm 0.42}$	$75.76_{\pm 0.15}$	$85.23_{\pm 0.19}$	$85.39_{\pm 0.04}$
RW_{exp}	$83.23_{\pm 0.17}$	$87.43_{\pm 0.20}$	$87.98_{\pm 0.16}$	$79.12_{\pm 0.29}$	$83.94_{\pm 0.28}$	$\mathbf{86.19_{\pm 0.23}}$	$78.62_{\pm 0.15}$	$85.35_{\pm 0.23}$	$85.58_{\pm 0.27}$
RW_{inv}	$85.97_{\pm 0.12}$	$85.97_{\pm 0.12}$	$87.33_{\pm 0.71}$	$83.31_{\pm 0.61}$	$83.31_{\pm 0.61}$	$85.26_{\pm 0.18}$	$84.75_{\pm 0.25}$	$84.75_{\pm 0.25}$	$85.52_{\pm 0.61}$
FRDL	$85.83_{\pm 0.06}$	$85.83_{\pm 0.06}$	$87.53_{\pm 0.03}$	$83.09_{\pm 0.05}$	$83.09_{\pm 0.05}$	$86.05_{\pm 0.08}$	$83.89_{\pm 0.04}$	$83.89_{\pm 0.04}$	$85.20_{\pm 0.03}$
MLFL	$83.47_{\pm 0.00}$	$87.18_{\pm 0.17}$	$87.99_{\pm 0.10}$	$78.75_{\pm 0.70}$	$83.93_{\pm 0.28}$	$85.59_{\pm 0.62}$	$78.99_{\pm 0.22}$	$85.10_{\pm 0.14}$	$85.60_{\pm 0.34}$
ASL	$84.42_{\pm 0.13}$	$\mathbf{87.84_{\pm 0.09}}$	$87.91_{\pm 0.08}$	$79.83_{\pm 0.34}$	$\mathbf{84.81_{\pm 0.43}}$	$85.90_{\pm 0.51}$	$79.61_{\pm 0.06}$	$85.31_{\pm 0.16}$	$85.29_{\pm 0.11}$
CB	$86.57_{\pm 0.05}$	$87.26_{\pm 0.10}$	$87.81_{\pm 0.18}$	$80.06_{\pm 0.39}$	$84.23_{\pm 0.23}$	$86.17_{\pm 0.54}$	$82.69_{\pm 0.15}$	$85.02_{\pm 0.16}$	$85.23_{\pm 0.30}$
LA	$\mathbf{87.75_{\pm 0.04}}$	$87.75_{\pm 0.04}$	$\mathbf{88.02_{\pm 0.12}}$	$\mathbf{84.11_{\pm 0.39}}$	$84.11_{\pm 0.39}$	$85.89_{\pm 0.30}$	$\mathbf{85.60_{\pm 0.18}}$	$\mathbf{85.60_{\pm 0.18}}$	$\mathbf{85.68_{\pm 0.17}}$
DB	$84.87_{\pm 0.08}$	n/a	$87.49_{\pm 0.24}$	$80.39_{\pm 0.14}$	n/a	$85.33_{\pm 0.11}$	$79.49_{\pm 0.13}$	n/a	$84.81_{\pm 0.18}$
Max. gap (pp)		5.65 → 0.69			6.31 → 0.93			9.85 → 0.87	
Controlled methods—ViT-B/16									
ERM	$83.17_{\pm 0.18}$	$88.22_{\pm 0.11}$	$88.66_{\pm 0.22}$	$78.45_{\pm 0.19}$	$84.33_{\pm 0.31}$	$86.00_{\pm 0.29}$	$77.90_{\pm 0.20}$	$85.75_{\pm 0.27}$	$86.08_{\pm 0.32}$
RW_{inv}	$87.42_{\pm 0.37}$	$87.42_{\pm 0.37}$	$88.29_{\pm 0.49}$	$\mathbf{84.94_{\pm 0.47}}$	$84.94_{\pm 0.47}$	$87.24_{\pm 0.24}$	$86.06_{\pm 0.19}$	$86.06_{\pm 0.19}$	$86.37_{\pm 0.16}$
RW_{exp}	$84.72_{\pm 0.07}$	$88.09_{\pm 0.08}$	$88.69_{\pm 0.09}$	$80.89_{\pm 0.67}$	$\mathbf{85.95_{\pm 1.14}}$	$\mathbf{87.28_{\pm 0.30}}$	$80.98_{\pm 0.41}$	$\mathbf{86.53_{\pm 0.11}}$	$\mathbf{86.81_{\pm 0.13}}$
FRDL	$86.61_{\pm 0.04}$	$86.61_{\pm 0.04}$	$87.98_{\pm 0.39}$	$82.89_{\pm 0.13}$	$82.89_{\pm 0.13}$	$85.89_{\pm 0.14}$	$84.76_{\pm 0.16}$	$84.76_{\pm 0.16}$	$86.20_{\pm 0.06}$
MLFL	$84.43_{\pm 0.06}$	$87.60_{\pm 0.11}$	$88.98_{\pm 0.08}$	$80.11_{\pm 0.28}$	$85.46_{\pm 0.73}$	$86.70_{\pm 0.54}$	$81.11_{\pm 0.73}$	$86.36_{\pm 0.16}$	$86.58_{\pm 0.14}$
ASL	$85.76_{\pm 0.06}$	$88.42_{\pm 0.03}$	$88.64_{\pm 0.08}$	$78.69_{\pm 0.28}$	$83.97_{\pm 0.41}$	$84.60_{\pm 0.26}$	$81.14_{\pm 0.13}$	$86.00_{\pm 0.15}$	$85.97_{\pm 0.10}$
CB	$87.72_{\pm 0.10}$	$88.22_{\pm 0.08}$	$88.92_{\pm 0.29}$	$80.60_{\pm 0.12}$	$84.67_{\pm 0.92}$	$86.48_{\pm 0.18}$	$84.63_{\pm 0.09}$	$86.27_{\pm 0.39}$	$86.46_{\pm 0.20}$
LA	$\mathbf{88.74_{\pm 0.08}}$	$\mathbf{88.74_{\pm 0.08}}$	$\mathbf{89.03_{\pm 0.38}}$	$84.63_{\pm 1.10}$	$84.63_{\pm 1.10}$	$85.82_{\pm 0.91}$	$\mathbf{86.31_{\pm 0.25}}$	$86.31_{\pm 0.25}$	$86.45_{\pm 0.13}$
DB	$85.64_{\pm 0.00}$	n/a	$87.70_{\pm 0.02}$	$81.04_{\pm 0.24}$	n/a	$86.03_{\pm 0.14}$	$80.55_{\pm 0.09}$	n/a	$85.89_{\pm 0.07}$
Max. gap (pp)		5.56 → 1.33			6.49 → 2.67			8.41 → 0.92	
Published methods—CNN-based architectures									
L2L†	$86.06_{\pm 0.05}$	$87.66_{\pm 0.11}$	$87.62_{\pm 0.08}$	$82.02_{\pm 0.17}$	$84.59_{\pm 0.24}$	$85.59_{\pm 0.17}$	$82.03_{\pm 0.20}$	$85.88_{\pm 0.19}$	$\mathbf{86.34_{\pm 0.16}}$
PIL†	$85.92_{\pm 1.09}$	$87.22_{\pm 0.66}$	$87.08_{\pm 0.61}$	$82.57_{\pm 0.06}$	$85.37_{\pm 0.32}$	$86.02_{\pm 0.09}$	$81.66_{\pm 0.04}$	$83.60_{\pm 0.10}$	$83.65_{\pm 0.06}$
FRDL+GOAT†	$\mathbf{86.87_{\pm 0.09}}$	$86.87_{\pm 0.09}$	$88.10_{\pm 0.15}$	$\mathbf{85.16_{\pm 0.12}}$	$85.16_{\pm 0.12}$	$87.28_{\pm 0.10}$	$\mathbf{84.15_{\pm 0.15}}$	$84.15_{\pm 0.15}$	$86.01_{\pm 0.14}$
ReSAGE-PAR†	$84.89_{\pm 0.15}$	$\mathbf{88.27_{\pm 0.14}}$	$\mathbf{88.45_{\pm 0.25}}$	$82.11_{\pm 0.20}$	$\mathbf{87.12_{\pm 0.40}}$	$\mathbf{87.97_{\pm 0.22}}$	$80.46_{\pm 0.22}$	$\mathbf{85.97_{\pm 0.40}}$	$86.33_{\pm 0.24}$
Published methods—Transformer-based architectures									
HAP†	$87.37_{\pm 0.10}$	$88.74_{\pm 0.15}$	$\mathbf{90.26_{\pm 0.40}}$	$85.74_{\pm 0.08}$	$89.62_{\pm 0.47}$	$\mathbf{91.08_{\pm 0.16}}$	$82.28_{\pm 0.18}$	$87.38_{\pm 0.39}$	$87.82_{\pm 0.18}$
PARFormer-B†	$\mathbf{87.99_{\pm 0.15}}$	$88.82_{\pm 0.26}$	$89.24_{\pm 0.23}$	$81.85_{\pm 0.34}$	$85.06_{\pm 0.75}$	$86.83_{\pm 0.57}$	$82.73_{\pm 0.33}$	$86.01_{\pm 0.80}$	$86.49_{\pm 0.53}$
PromptPAR†	87.79	88.99	88.02	**87.23**	**89.89**	90.10	**85.42**	**89.06**	**88.83**
ReSAGE-PAR†	$85.78_{\pm 0.06}$	$\mathbf{89.07_{\pm 0.07}}$	$89.42_{\pm 0.05}$	$79.92_{\pm 0.35}$	$84.31_{\pm 1.09}$	$88.06_{\pm 0.19}$	$81.10_{\pm 0.21}$	$86.89_{\pm 0.10}$	$87.00_{\pm 0.09}$

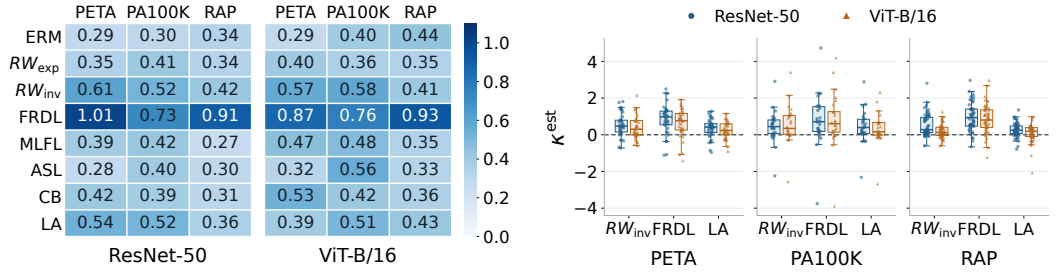
"n/a" indicates that no κ^{obj} is identifiable. † denotes component-level threshold derivation; its scope is detailed in Appendix D.2.2. All views evaluate the same complete fixed predictor. PromptPAR is evaluated using a single officially released checkpoint. Summary rows report the maximum within-group mA gap (Native → Est.), computed from unrounded means. Bold indicates the highest reported mA.

Comparisons of published methods. Even within architecture groups, reported advantages can depend on operating policy. Among CNN-based methods, FRDL+GOAT leads ReSAGE-PAR under Native on all three datasets, but the ordering reverses under both component-derived Obj. and Est. With predictors held fixed across views, this reversal shows that the native lead is conditional on the decision conditions used for evaluation. Other advantages persist or grow under Est.: L2L's advantage over PIL increases to about 2.7 pp on RAP, while HAP retains its lead over ReSAGE-PAR on PETA and RAP within the Transformer-based group. For HAP and ReSAGE-PAR on PETA, Obj. and Est. yield opposite orderings. Joint reporting therefore clarifies how method ordering and mA gaps depend on the evaluation policy and whether reported advantages persist under normalization.

4.3 Objective-Implied and Estimated Coordinates

Correspondence at native thresholds. To investigate the differences between Obj. and Est., we first compare the two coordinates at native thresholds for the controlled methods in Fig.2.a. ERM illustrates the distinction between correspondence to an objective-implied coordinate and alignment with the mA target: it has smaller median coordinate discrepancies than FRDL in all six dataset–backbone settings, while its median $|\kappa^{\text{est}}|$ is larger. On RAP, the latter is 2.68 with ResNet-50 and 2.53 with ViT-B/16. Closer agreement with objective-implied decision conditions therefore does not imply closer alignment with the class-balanced target.

For RW_{inv}, FRDL, and LA, whose objective coordinates are all zero, departures from the objective coordinate are also departures from the common mA target. Despite this shared reference, FRDL has larger median absolute departures than LA in every setting. Fig.2.b further shows attribute means on both sides of zero, with most being positive. Under the fitted calibration maps, positive coordinates correspond to native decisions requiring stronger positive evidence than the mA target. The differing magnitudes and directions show why alignment needs to be examined at the attribute level: a shared objective target can coexist with different estimated decision conditions across methods



(a) Discrepancies between κ^{obj} and κ^{est}

(b) Estimated coordinates for $\kappa^{\text{obj}} = 0$

Figure 2: Objective-implied and estimated coordinates at native thresholds. (a) Median $|\kappa^{\text{est}} - \kappa^{\text{obj}}|$ over attributes and three seeds. (b) Estimated coordinates for RW_{inv}, FRDL, and LA, which share $\kappa^{\text{obj}} = 0$. Dots show per-attribute means over three seeds; boxes show medians and interquartile ranges (IQRs), with 1.5-IQR whiskers. Dashed lines mark the mA target $\kappa = 0$.

and attributes. Predictor-specific policy estimation provides an empirical check on these conditions when interpreting native mA comparisons, even under a shared numerical threshold.

Alignment at objective-normalized thresholds. We next evaluate the estimated coordinate at the Obj. threshold, $\kappa_{m,j}^{\text{est,obj}} = \operatorname{logit} \widehat{c}_{m,j}\left(\tau_{m,j}^{\text{obj}}\right) - \operatorname{logit} \pi_j$, using the same validation calibration maps and training priors. We focus on ERM, RW_{exp}, MLFL, ASL, and CB, whose thresholds change under Obj.; the other three identifiable controlled methods retain their native thresholds.

Objective normalization lowers the median absolute estimated coordinate for all five methods shown across the three datasets and both backbones (Fig.3). These reductions support the analytical threshold adjustment as a way to approach the class-balanced reference for the fixed predictors. However, the median $|\kappa^{\text{est,obj}}|$ remains between 0.33 and 0.69, and the median $\kappa^{\text{est,obj}}$ is positive in every setting. A positive $\kappa_{m,j}^{\text{est,obj}}$ means that the Obj. threshold corresponds to a calibrated positive-class probability above π_j, the mA reference level. These residual offsets show that objective-level alignment with the mA target need not translate into estimated alignment for a finite predictor. Obj. and Est. can therefore specify different operating points for the same fixed predictor despite their shared reference target.

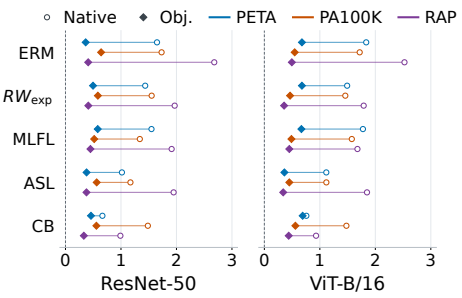


Figure 3: Estimated distances to the mA target. Open circles and filled diamonds show median $|\kappa^{\text{est}}|$ at Native thresholds and $|\kappa^{\text{est,obj}}|$ at Obj. thresholds, respectively.

4.4 Robustness of Policy Estimation and Comparisons

Est. fits validation-set isotonic maps and inverts them at training-set priors. Under our common-reference assumption, training and validation prevalences estimate the same population priors but can differ in finite samples. The PAV–ROC convex-hull equivalence (Fawcett & Niculescu-Mizil, 2007) implies that, with calibration-sample empirical prevalences and appropriate tie handling, thresholding the fitted maps maximizes validation mA over attribute-wise score thresholds. Appendix C details this relationship. We next examine the sensitivity of policy estimation.

Sensitivity to the calibration model. The paired analysis in Fig.4 compares isotonic and monotone sigmoid calibration. Overall Est. mA changes modestly, with sigmoid yielding slightly higher mA in most evaluated method–dataset–backbone combinations. The attribute-level results are more heterogeneous: changes occur in both directions and can partially cancel under mA's macro-average. An increase in overall mA therefore does not imply that every attribute improves. Method comparisons depend on how the estimator changes each method's mA. Small aggregate changes can consequently alter rankings, as shown in Tab.4: LA is replaced as the highest-ranked method by RW_{exp} on RAP with ResNet-50 and by MLFL on PETA with ViT-B/16. The corresponding isotonic Est. gaps are only 0.10 and 0.05 pp, respectively. These results support relative stability in Est. mA, while rank-

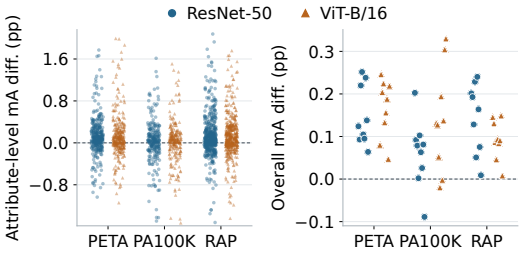


Figure 4: Calibration-model sensitivity. Differences (sigmoid minus isotonic, pp) in (left) per-attribute mA and (right) overall mA.

Table 4: Est. rankings: Isotonic → Sigmoid. Unchanged ranks are shown once.

Method	ResNet-50			ViT-B/16		
	PETA	PA100K	RAP	PETA	PA100K	RAP
ERM	4	4 → 3	5	5 → 6	6 → 5	7
RW_{exp}	3	1	3 → 1	4	1	1
RW_{inv}	9	9 → 8	4	7	2	5
FRDL	7	3 → 4	8 → 7	8	7	6
MLFL	2	7	2 → 3	2 → 1	3	2
ASL	5 → 6	5 → 6	6	6 → 5	9	8
CB	6 → 5	2	7 → 8	3 → 2	4	3 → 4
LA	1	6 → 5	1 → 2	1 → 3	8	4 → 3
DB	8	8 → 9	9	9	5 → 6	9

ings based on small normalized gaps can depend on the calibration rule. Appendices E.1 and E.4 provide coordinate comparisons and raw score–label diagnostics, respectively.

Sensitivity to validation data. We refit both calibrators on 30 shared subsets at each of 25%, 50%, and 75% of the validation-set data. Larger subsets reduce the variability of method gaps, with sigmoid yielding smaller standard deviations than isotonic in every examined comparison and sampling fraction (Fig.5). Across supported refits, the absolute RW_{inv}–ERM gaps on RAP remain below 0.63 pp on both backbones, compared with native gaps above 8 pp. This contraction persists even though the near-zero ResNet-50 gap can change sign. On PETA, HAP retains its Est. lead over ReSAGE-PAR in every supported refit under both calibrators, preserving the opposite ordering to Obj. These support the examined gap contraction and ordering reversal under validation subsampling. Direct validation mA optimization preserves three comparison patterns, with gaps within 0.013 pp of isotonic Est. (Appendix E.2). Reference-prior ablations also preserve the examined RAP gap contraction and HAP's Est. lead over ReSAGE-PAR on PETA, while some small method gaps change sign (Appendix E.3).

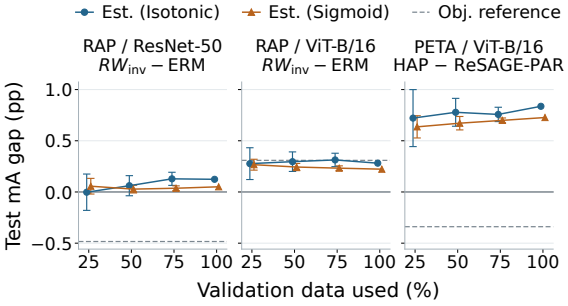


Figure 5: Test mA gaps under validation subsampling. At 25–75%, markers and error bars show means ± SD over valid refits from 30 subsamples per fraction.

5 Policy-Aware Reporting and Conclusion

Reporting protocol. Alongside standard experimental details, report each method's imbalance interventions, prior source, and native inference rule. Report Native and applicable Obj. and Est. mA for the same fixed checkpoint, specifying the component used to derive Obj. thresholds. For Est., we recommend isotonic calibration as the reference: fit a nondecreasing map to each attribute's validation logits and labels, without imposing a sigmoid shape, and invert it at the training prior. Disclose the validation split, threshold and support conventions, and alternative calibration rules. Freeze maps and thresholds before evaluation on disjoint test data; provide per-attribute thresholds and coordinates alongside aggregate results. Interpret small normalized gaps cautiously: rankings can depend on calibration rules, validation samples, and reference priors even with fixed scores.

Conclusion. Our audit, derivations, and paired evaluations on PAR benchmarks show that native mA advantages can depend strongly on operating policy. Normalization can contract gaps or reverse orderings, while some advantages persist or widen. The framework distinguishes conditions implied by training from estimated policies in each predictor's score space under a common mA reference. Joint reporting retains native results while making the decision conditions underlying PAR comparisons explicit. Future work could quantify uncertainty in attribute-level policy estimates, especially under limited positive or negative validation support. Cross-domain evaluations could test the derived transfer conditions and determine how much additional calibration data are needed to update policies when class-conditional score distributions change.

AI USE STATEMENT

We used LLM-based tools, including ChatGPT and Codex, to improve the clarity and fluency of the manuscript and to assist with formatting and layout refinement. All LLM-assisted revisions were reviewed by the authors, who take full responsibility for the manuscript's content and accuracy.

ETHICS STATEMENT

This work examines PAR evaluation using existing benchmark datasets and predictions from trained models. Pedestrian imagery and potentially sensitive attribute annotations raise privacy and fairness concerns; biases in data collection, annotation, and pretrained models may also affect comparisons. We report aggregate analyses to clarify the decision conditions underlying mA comparisons. Class-balanced evaluation alone does not establish demographic fairness or deployment suitability. Practical applications require separate assessments of privacy, performance across relevant population groups, and potential misuse, particularly in surveillance settings. Reuse of benchmark data and model artifacts should respect their access conditions and applicable usage restrictions.

REPRODUCIBILITY STATEMENT

We evaluate published PAR methods using their publicly available official implementations and build the controlled experiments on the public baseline of Jia et al. (2021). Section 3 specifies the decision coordinates, assumptions, and normalization rules. Appendix A documents the audit scope, coding criteria, and supporting evidence. Calibration, threshold inversion, tie handling, and support conventions are detailed in Appendix C. Appendix D documents the code sources, implementation adaptations, dataset partitions, method settings, and validation-subsampling procedure. Each evaluated predictor is held fixed across evaluation views, and calibration maps and thresholds are frozen before evaluation on disjoint test data, supporting paired comparisons under the specified decision conditions.

REFERENCES

Pablo Ayuso-Albizu, Pablo Carballeira, Juan C SanMiguel, and Paula Moral. Resage-par: Representational similarity assessment for generative expansion in pedestrian attribute recognition. *arXiv preprint arXiv:2606.06020*, 2026.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5198–5207, 2021. doi: 10.1109/CVPR46437.2021.00516.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9260–9269. IEEE, 2019.

Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 789–792, 2014. doi: 10.1145/2647868.2654966.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001.

Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. PARFormer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):411–423, 2024. doi: 10.1109/TCSVT.2023.3285411.

Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68 (1):97–106, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012. URL https://jmlr.org/papers/v13/hernandez-orallo12a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://www.jstor.org/stable/4615733.

Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. URL https://arxiv.org/abs/2107.03576.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2744–2752. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/98053046e0dce5c7d946c67b96a85e18-Abstract.html.

Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition*, pp. 111–115, 2015. doi: 10.1109/ACPR.2015.7486476.

Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. URL https://arxiv.org/abs/1603.07054.

Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 28 (4):1575–1590, 2019. doi: 10.1109/TIP.2018.2878349.

Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *European conference on computer vision*, pp. 562–579. Springer, 2022.

Ye Li, Fangyan Shi, Shaoqi Hou, Jipeng Li, Chao Li, and Guangqiang Yin. Feature pyramid attention model and multi-label focal loss for pedestrian attribute recognition. *IEEE Access*, 8: 164570–164579, 2020. doi: 10.1109/ACCESS.2020.3010435.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. HydraPlus-Net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 350–359, 2017. doi: 10.1109/ICCV.2017.46.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 603–611. PMLR, 2013. URL https://proceedings.mlr.press/v28/menon13a.html.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430.

Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065, 2013. doi: 10.1016/j.patcog.2013.01.012.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(83):2387–2422, 2010. URL https://jmlr.org/papers/v11/reid10a.html.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021. doi: 10.1109/ICCV48922.2021.00015.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision*, pp. 708–725, 2018. doi: 10.1007/978-3-030-01252-6_42.

Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12055–12062, 2020. doi: 10.1609/aaai.v34i07.6883.

Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21970–21982. IEEE, 2023.

Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022. doi: 10.1016/j.patcog.2021.108220.

Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. Pedestrian attribute recognition via CLIP-based prompt vision-language fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):148–161, 2025a. doi: 10.1109/TCSVT.2024.3454366.

Yizhou Wang, Yixuan Wu, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, Wanli Ouyang, et al. Hulk: A universal knowledge translator for human-centric tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5672–5689, 2025b.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European conference on computer vision*, pp. 162–178. Springer, 2020.

Minghe Xu, Rouying Wu, Jiarui Xu, Minhao Sun, Zikang Yan, Xiao Wang, ChiaWei Chu, and Yu Li. Unipar: A unified framework for pedestrian attribute recognition. *IET Computer Vision*, 20(1):e70081, 2026.

Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. Hap: Structure-aware masked image modeling for human-centric perception. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50597–50616, 2023.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, New York, NY, USA, 2002. Association for Computing Machinery. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

Xiaokang Zhang and Hai-Miao Hu. ASRL: Correlation-robust pedestrian attribute recognition via fixed orthogonal classifier. *Neural Networks*, 196:108408, 2026. doi: 10.1016/j.neunet.2025.108408.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16484–16493, 2021. doi: 10.1109/CVPR46437.2021.01622.

Yibo Zhou, Hai-Miao Hu, Jinzuo Yu, Haotian Wu, Shiliang Pu, and Hanzi Wang. A solution to co-occurrence bias in pedestrian attribute recognition: Theory, algorithms, and improvements. *International Journal of Computer Vision*, 133(7):4712–4726, 2025a.

Yibo Zhou, Bo Li, Hai-Miao Hu, Xiaokang Zhang, Dongping Zhang, and Hanzi Wang. Heterogeneous feature re-sampling for balanced pedestrian attribute recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2706–2722, 2025b. doi: 10.1109/TPAMI.2025.3526930.

Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. Plip: Language-image pre-training for person representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45666–45702, 2024.

A PROTOCOL AUDIT: SCOPE, CODING, AND EVIDENCE

We detail the audit underlying Section 3.2. Paper-level reporting and official-code behavior are treated as separate evidence layers. The former describes what the audited publications disclose; the latter records the active training components and final decision rules in the audited official implementations.

A.1 AUDIT SCOPE AND ELIGIBILITY

The primary cohort covers formally published studies from 2019 through August 31, 2026. Eligibility requires verifiable full text and closed-set multi-attribute PAR results on PETA, PA100K, or the RAP family, including RAPv1 and RAPv2. PAR may be the primary task or a downstream task of a general human-centric model. For pooled datasets, performance on at least one required source domain must be reported separately. Different versions of the same work contribute one paper-level entry. Tab.5 summarizes the screening results. Only the 66 primary-cohort papers enter the paper-level proportions; the five pre-2019 studies serve as historical anchors.

Table 5: Screening summary. Unresolved records lack sufficient full-text or eligibility evidence.

Category	Count
Primary cohort	66
Historical anchors	5
Unresolved records	135
Other exclusions or duplicate versions	61
Total candidate records	267

Table 6: Cross-counts for the 66 audited papers. M: standard balanced mA verified; I: label-imbalance intervention explicitly documented (1: yes; 0: not verified from the audited publication).

	$I=1$	$I=0$	Total
Standard mA verified ($M=1$)	26	12	38
Standard mA not verified ($M=0$)	18	10	28
Total	44	22	66

A.2 PAPER-LEVEL RECORDS AND FINDINGS

Table 6 summarizes the overlap between verified standard balanced mA and documented label-imbalance interventions across the 66 papers. Standard balanced mA is verified in 38 papers, 44 document an imbalance intervention, and 26 satisfy both criteria. Verification requires a balanced definition or explicit adoption of a traceable standard evaluation protocol. An $M = 0$ code can reflect insufficient evidence, without implying that a different metric was used.

Table 7 lists the individual papers and their metric, intervention, and final decision-rule codes. Only P21 and P25 state numerical final thresholds, both specifying 0.5. P28 states a fixed threshold without reporting its value, while P08 and P63 describe autoregressive decoding. No final decision-rule statement was located in the remaining 61 papers.

Table 7: Audit of the 66 papers in the primary cohort. "PAR" abbreviates pedestrian attribute recognition in the displayed titles. M: standard balanced mA verified; I: imbalance intervention documented (1: verified; 0: not reported or insufficient evidence). T: final decision rule; NS: no final decision-rule statement; F: fixed threshold without its value; D: autoregressive decoding.

ID	Audited paper	M	I	T
P01	Attribute Aware Pooling for PAR	1	0	NS
P02	Improving PAR With Weakly-Supervised Multi-Scale Attribute-Specific Localization	1	1	NS

Continued on the next page

Table 7 (continued)

ID	Audited paper	M	I	T
P03	PARNet: A Joint Loss Function and Dynamic Weights Network for Pedestrian Semantic Attributes Recognition of Smart Surveillance Image	1	1	NS
P04	PAR by Joint Visual-Semantic Reasoning and Knowledge Distillation	0	0	NS
P05	PAR with Graph Convolutional Network in Surveillance Scenarios	1	0	NS
P06	PAR in Surveillance Scenarios Using Multi-Task Lightweight Convolutional Neural Network	1	1	NS
P07	Recurrent Attention Model for PAR	1	1	NS
P08	Visual-Semantic Graph Reasoning for PAR	1	0	D
P09	An attention-based deep learning model for multiple PAR	1	1	NS
P10	Deep Template Matching for PAR with the Auxiliary Supervision of Attribute-wise Keypoints	1	1	NS
P11	Feature Pyramid Attention Model and Multi-Label Focal Loss for PAR	1	1	NS
P12	Multi-Task Learning Via Co-Attentive Sharing For PAR	1	0	NS
P13	Relation-Aware PAR with Graph Convolutional Networks	1	1	NS
P14	LM Filter-Based Deep Convolutional Neural Network for PAR	0	0	NS
P15	Cascaded Split-and-Aggregate Learning with Feature Recombination for PAR	1	1	NS
P16	Multi-Branch Gabor Wavelet Layers for PAR	0	1	NS
P17	PAR using two-branch trainable Gabor wavelets network	0	1	NS
P18	Skeleton-Based Attention Mask for PAR Network	0	0	NS
P19	Spatial and Semantic Consistency Regularizations for PAR	1	1	NS
P20	A Framework for PAR Using Deep Learning	0	1	NS
P21	Learning Disentangled Attribute Representations for Robust PAR	0	1	0.5
P22	PAR in Video Surveillance Scenarios Based on View-attribute Attention Localization	1	1	NS
P23	Reinforced PAR with group optimization reward	1	1	NS
P24	A Solution to Co-occurrence Bias in PAR: Theory, Algorithms, and Improvements	1	1	NS
P25	A Transfer Learning Approach to Recognize Pedestrian Attributes	0	0	0.5
P26	PARFormer: Transformer-Based Multi-Task Network for PAR	0	1	NS
P27	SSPNet: Scale and spatial priors guided generalizable and interpretable PAR	1	1	NS
P28	UPAR: Unified PAR and Person Retrieval	1	1	F
P29	ALFormer: Attribute localization transformer in PAR	1	0	NS
P30	Attribute correlation mask fusion network for PAR	1	1	NS
P31	C2T-Net: Channel-Aware Cross-Fused Transformer-Style Networks for PAR	1	0	NS
P32	CrossPAR: Enhancing PAR with Vision-Language Fusion and Human-Centric Pre-training	1	0	NS
P33	MCGCN: Multi-Correlation Graph Convolutional Network for PAR	1	1	NS
P34	MRG-T: Mask-Relation-Guided Transformer for Remote Vision-Based PAR in Aerial Imagery	1	1	NS
P35	Heterogeneous Feature Re-Sampling for Balanced PAR	1	1	NS
P36	PAR via CLIP-Based Prompt Vision-Language Fusion	0	1	NS
P37	Selective and Orthogonal Feature Activation for PAR	0	1	NS
P38	A Data-Centric Approach to PAR: Synthetic Augmentation via Prompt-Driven Diffusion Models	0	1	NS
P39	AdaGPAR: Generalizable PAR via Test-time Adaptation	1	1	NS
P40	CLEAR: Cross-Transformers with Pre-trained Language Model is All you need for PAR and Retrieval	0	0	NS
P41	Enhanced Visual-Semantic Interaction with Tailored Prompts for PAR	0	0	NS

Continued on the next page

Table 7 (continued)

ID	Audited paper	M	I	T
P42	High-order diversity feature learning for PAR	1	1	NS
P43	Improved PAR Algorithm Based on Image Style Transfer	0	1	NS
P44	Lightweight Attribute Localizing Models for PAR	1	1	NS
P45	PAR: A New Benchmark Dataset and a Large Language Model Augmented Framework	0	1	NS
P46	Rethinking attention mechanism for enhanced PAR	0	1	NS
P47	SNN-PAR: Energy Efficient PAR via Spiking Neural Networks	0	1	NS
P48	ViTA-PAR: Visual And Textual Attribute Alignment With Attribute Prompting For PAR	0	1	NS
P49	Adversarial Semantic and Label Perturbation Attack for PAR	0	1	NS
P50	Cross-Domain PAR: Evaluation Criteria, a New Baseline and Remote Sensor-Based Application	1	1	NS
P51	Joint Implicit and Explicit Language Learning for PAR	0	0	NS
P52	UniPAR: A Unified Framework for PAR	0	1	NS
P53	HAP: Structure-Aware Masked Image Modeling for Human-Centric Perception	1	0	NS
P54	HumanBench: Towards General Human-Centric Perception With Projector Assisted Pretraining	0	0	NS
P55	UniHCP: A Unified Model for Human-Centric Perceptions	0	1	NS
P56	Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks	1	0	NS
P57	PLIP: Language-Image Pre-training for Person Representation Learning	0	0	NS
P58	Hulk: A Universal Knowledge Translator for Human-Centric Tasks	0	0	NS
P59	CLIP-Guided Adaptable Self-Supervised Learning for Human-Centric Visual Tasks	1	0	NS
P60	Part-guided Network for PAR	1	0	NS
P61	Exploring attribute localization and correlation for PAR	1	0	NS
P62	Label2Label: A Language Modeling Framework for Multi-Attribute Learning	1	1	NS
P63	SequencePAR: Understanding Pedestrian Attributes via a Sequence Generation Paradigm	0	1	D
P64	UPAR Challenge: PAR and Attribute-based Person Retrieval – Dataset, Design, and Results	1	1	NS
P65	Visual Attention Consistency Under Image Transforms for Multi-Label Image Classification	0	1	NS
P66	ASRL: Correlation-Robust PAR via Fixed Orthogonal Classifier	1	1	NS

A.3 Official-Code Audit and Interpretation

An official implementation enters the code audit when its provenance, active training components, and final decision rule can be traced in the inspected source. Each record describes the selected training and inference setting within that implementation. Source versions and file locations are retained in the audit records.

Table 8 presents the training components and final decision rules of the 28 audited official implementations. Of these, 21 are associated with primary-cohort papers, three with historical anchors, and four with other audit records. Nineteen use a 0.5 boundary or its equivalent, seven use 0.45, one uses 0.3, and one uses autoregressive decoding.

The paper and code counts describe different evidence: what publications state and what the selected implementations execute. Their denominators are therefore kept separate. Repeated numerical cutoffs do not establish equivalent operating policies across training objectives. The fixed-predictor evaluations in the main text examine how operating-policy choice affects method comparisons.

Table 8: Training components and decision rules in 28 audited official implementations. Numerical boundaries are expressed on the sigmoid-probability scale; asterisks denote zero-logit comparisons or probability rounding. Decoding denotes an autoregressive rule without a scalar probability cutoff.

Method	Training component	Decision rule
ALM	Prior-weighted BCE (positive term)	0.5
GCN-PAR	Unweighted multilabel soft-margin loss	0.5^*
MAN-PAR	Class-weighted BCE	0.5^*
Rethinking baseline	$RW_{\exp}$-weighted BCE	0.5
VTB	$RW_{\exp}$-weighted BCE	0.45
PARFormer	ASL and auxiliary objectives	0.5
SSPNet	$RW_{\exp}$-weighted BCE	0.5
C2T-Net	$RW_{\exp}$-weighted BCE	0.5
PromptPAR	$RW_{\exp}$-weighted BCE	0.45
FRDL	BCE; balanced feature resampling and head retraining	0.5
LLM-PAR	$RW_{\exp}$-weighted BCE	0.45
SNN-PAR	$RW_{\exp}$-weighted BCE	0.45
AttackPAR	$RW_{\exp}$-weighted BCE	0.45
UniPAR	$RW_{\exp}$-weighted BCE	0.45
Label2Label	$RW_{\exp}$-weighted BCE	0.5
MambaPAR	$RW_{\exp}$-weighted BCE	0.45
DeepMAR	$RW_{\exp}$-weighted BCE	0.5^*
WPAL	Weighted BCE	0.5
HAP	$RW_{\exp}$-weighted BCE (PAR fine-tuning)	0.5
PATH	$RW_{\exp}$-weighted BCE (PAR fine-tuning)	0.5
UniHCP	$RW_{\exp}$-weighted BCE (PAR component)	0.5
SOLIDER	$RW_{\exp}$-weighted BCE (PAR fine-tuning)	0.5
Hulk (direct evaluation)	Prior-weighted masked BCE; dataset weighting	0.3
CrossPAR	$RW_{\exp}$-weighted BCE	0.5
PIL	$RW_{\exp}$-weighted BCE; posterior-invariant feature learning	0.5
ReSAGE-PAR (Rethinking)	Prior-weighted BCE; synthetic-data augmentation	0.5
VeSPA	Prior-weighted BCE	0.5
SequencePAR	Token NLL; autoregressive attribute decoding	Decoding

B Derivations of Decision Coordinates

B.1 Reference Distribution and the mA Target

We use the notation of Section 3.1. For each attribute j, assume $0 < \pi_j < 1$ and that the class-conditional distributions admit densities with respect to a common dominating measure. Likelihood ratios may take extended values where one density vanishes. Probability thresholds lie in $(0, 1)$ unless a limiting rule is stated. The reference distribution describes the original data, before any training-time reweighting or resampling. Training, validation, and test observations are assumed to represent this same distribution for the population interpretation.

For a binary decision rule h_j, the expression in Section 3.1 can be written as

$$\mathrm{BA}_j(h_j) = \frac{1}{2} + \frac{1}{2} \int h_j(x)\big[f_{1,j}(x) - f_{0,j}(x)\big]\, dx. \tag{14}$$

The integrand is maximized by predicting positive wherever $f_{1,j} > f_{0,j}$ and negative wherever $f_{1,j} < f_{0,j}$. Decisions on $f_{1,j} = f_{0,j}$ do not affect the population value. In particular, for any such optimal rule $h_j^\star$,

$$\mathrm{BA}_j(h_j^\star) - \mathrm{BA}_j(h_j) = \frac{1}{2} \int |f_{1,j}(x) - f_{0,j}(x)| \mathbf{1}\{h_j(x) \neq h_j^\star(x)\}\, dx. \tag{15}$$

Thus the target is $\lambda_j(x) > 0$, or equivalently $\eta_j(x) > \pi_j$ (Elkan, 2001; Menon et al., 2013). Since mA is the equal-weight average over attributes, optimizing each attribute gives the mA-optimal unconstrained decision vector. This argument requires no independence assumption between attributes. The corresponding boundary is $\kappa^\star_{\mathrm{mA},j} = 0$.

B.2 Fixed Class Weights and Class-Conditional Resampling

Suppress m, j temporarily. At a fixed input x, write $\eta = \eta_j(x)$ and let z denote a candidate logit, with $p = \sigma(z)$. For positive class weights $a, b > 0$, the conditional weighted-BCE risk is

$$\mathcal{R}(p; \eta) = -a\eta \log p - b(1-\eta)\log(1-p).$$

For $0 < \eta < 1$, strict convexity in p gives the unique minimizer

$$p^*(\eta) = \frac{a\eta}{a\eta + b(1-\eta)}, \qquad \operatorname{logit} p^*(\eta) = \operatorname{logit}\eta + \log\frac{a}{b}. \tag{16}$$

Combining this identity with Eq.2 yields

$$\begin{aligned} p^*(\eta_j(x)) > t \iff \lambda_j(x) &> \operatorname{logit} t + \log\frac{b}{a} - \operatorname{logit}\pi_j, \\ t^{\mathrm{obj}}_{m,j} = \frac{a_j\pi_j}{a_j\pi_j + b_j(1-\pi_j)}, &\qquad \tau^{\mathrm{obj}}_{m,j} = \operatorname{logit}\pi_j + \log\frac{a_j}{b_j}. \end{aligned} \tag{17}$$

The term $\log(a_j/b_j)$ gives the correction to the ordinary-BCE logit threshold $\operatorname{logit}\pi_j$ required to preserve the mA boundary under class weighting.

The special cases used in the analysis follow directly: ordinary BCE has $\tau^{\mathrm{obj}} = \operatorname{logit}\pi$; RW_{inv} has $\tau^{\mathrm{obj}} = 0$; RW_{exp} has $\tau^{\mathrm{obj}} = \operatorname{logit}\pi + 1 - 2\pi$; and RW_{sqrt} has $\tau^{\mathrm{obj}} = \frac{1}{2}\operatorname{logit}\pi$. For the binary class-balanced weighting component (Cui et al., 2019),

$$a_j = \frac{1-\beta}{1-\beta^{n_{+,j}}}, \qquad b_j = \frac{1-\beta}{1-\beta^{n_{-,j}}}, \qquad 0 \le \beta < 1, \tag{18}$$

where both training class counts are positive. Any common positive normalization of a_j, b_j cancels from Eq.17.

For resampling, let r_j be the positive prevalence in the resampled training distribution. If sampling changes only the class prior and preserves $f_{1,j}, f_{0,j}$, its posterior satisfies

$$\operatorname{logit}\eta^{(r)}_j(x) = \lambda_j(x) + \operatorname{logit} r_j.$$

Ordinary BCE on this distribution therefore gives

$$\kappa^{\mathrm{obj}}_{m,j}(t) = \operatorname{logit} t - \operatorname{logit} r_j, \qquad t^{\mathrm{obj}}_{m,j} = r_j. \tag{19}$$

Balanced resampling sets $r_j = 0.5$, yielding $\kappa^{\mathrm{obj}}_{m,j}(0.5) = 0$ for the FRDL classification component in Tab.2 (Zhou et al., 2025b). For classifier retraining on fixed features, the same derivation applies to the class-conditional distributions in that feature space. More generally, if fixed class weights are also applied after prior-only resampling, the two effects combine as

$$\kappa^{\mathrm{obj}}_{m,j}(t) = \operatorname{logit} t + \log\frac{b_j}{a_j} - \operatorname{logit} r_j.$$

B.3 Training-Time Logit Adjustment

Consider BCE applied during training to $\sigma(z + d_j)$, where d_j is a fixed offset, while inference thresholds the unadjusted score z. The population optimum satisfies $z^*(x) + d_j = \operatorname{logit}\eta_j(x)$. Consequently,

$$\begin{aligned} \kappa^{\mathrm{obj}}_{m,j}(t) &= \operatorname{logit} t + d_j - \operatorname{logit}\pi_j, \\ \tau^{\mathrm{obj}}_{m,j} &= \operatorname{logit}\pi_j - d_j. \end{aligned} \tag{20}$$

The binary training-time adjustment $d_j = \operatorname{logit}\pi_j$ used for LA makes the native $t = 0.5$ boundary equal to the mA target (Menon et al., 2021). Thresholding $z + d_j$ at inference instead gives $\kappa^{\mathrm{obj}}_{m,j}(t) = \operatorname{logit} t - \operatorname{logit}\pi_j$.

B.4 PREDICTION-DEPENDENT LOSSES

For differentiable class losses ℓ_+, ℓ_-, let $g_\pm(t) = \partial \ell_\pm(z)/\partial z$ at $z = \operatorname{logit} t$. Assume that the pointwise conditional risk has a unique monotone population-optimal response, that this response attains t at an interior optimum, and that $g_+(t) < 0 < g_-(t)$. The stationarity condition at that crossing is

$$\eta^\dagger(t) g_+(t) + [1 - \eta^\dagger(t)] g_-(t) = 0.$$

Hence

$$\eta^\dagger(t) = \frac{g_-(t)}{g_-(t) - g_+(t)}, \qquad \kappa^{\mathrm{obj}}(t) = \log \frac{g_-(t)}{-g_+(t)} - \operatorname{logit} \pi. \tag{21}$$

Under these conditions, $\eta^\dagger(t)$ identifies the posterior boundary implemented by threshold t. Objective normalization follows by solving $\eta^\dagger(t^{\mathrm{obj}}) = \pi$.

Focal and asymmetric components. For the unshifted focal family (Lin et al., 2017; Charoenphakdee et al., 2021),

$$\ell_+(z) = -a(1-p)^{\gamma_+} \log p, \qquad \ell_-(z) = -bp^{\gamma_-} \log(1-p), \qquad p = \sigma(z),$$

where $a, b > 0$ and $\gamma_+, \gamma_- \geq 0$, differentiation gives

$$\begin{aligned} g_+(t) &= a(1-t)^{\gamma_+} \big[\gamma_+ t \log t - (1-t)\big], \\ g_-(t) &= bt^{\gamma_-} \big[t - \gamma_-(1-t)\log(1-t)\big]. \end{aligned} \tag{22}$$

Both derivative signs hold for $0 < t < 1$. To verify uniqueness, set $h_\gamma(p) = -(1-p)^\gamma \log p$. Its second derivative is

$$h''_\gamma(p) = (1-p)^{\gamma-2} \left[-\gamma(\gamma-1)\log p + \frac{2\gamma(1-p)}{p} + \frac{(1-p)^2}{p^2}\right] > 0.$$

For $\gamma \geq 1$, each term in brackets is nonnegative, with a strictly positive final term. For $0 \leq \gamma < 1$, the inequality $-\log p \leq (1-p)/p$ also gives a strictly positive lower bound. The conditional risk is therefore strictly convex in p. Writing the conditional risk as $\mathcal{R}(p;\eta)$, we have $\partial^2_{p\eta}\mathcal{R} < 0$. Implicit differentiation at its interior minimizer gives

$$\frac{dp^*(\eta)}{d\eta} = -\frac{\partial^2_{p\eta}\mathcal{R}(p^*(\eta);\eta)}{\partial^2_{pp}\mathcal{R}(p^*(\eta);\eta)} > 0.$$

At $t = 0.5$, Eq.21 reduces to

$$\kappa^{\mathrm{obj}}(0.5) = \log \frac{b}{a} + (\gamma_+ - \gamma_-)\log 2 + \log \frac{1 + \gamma_- \log 2}{1 + \gamma_+ \log 2} - \operatorname{logit} \pi. \tag{23}$$

When $\gamma_+ = \gamma_-$, the modulation terms cancel at $t = 0.5$, leaving $\log(b/a) - \operatorname{logit} \pi$. For the symmetric DIAA component, $a = b$ gives $-\operatorname{logit} \pi$; for MLFL, $a = e^{1-\pi}$ and $b = e^\pi$ give $2\pi - 1 - \operatorname{logit} \pi$ (Sarafianos et al., 2018; Li et al., 2020). For PARFormer's ASL classification component, $a = b = 1$, $\gamma_+ = 0$, and $\gamma_- = 1$ give $\log[(1 + \log 2)/2] - \operatorname{logit} \pi$ at $t = 0.5$ (Fan et al., 2024), as recorded in Tab.2.

Objective-normalized thresholds are determined by the full crossing map through $\eta^\dagger(t^{\mathrm{obj}}) = \pi$.

These formulas apply to identifiable pointwise classification components. Results based on a specified component of a composite objective are labeled as component-level normalization. Obj. is reported as unavailable when no scalar coordinate is identified.

B.5 SCORE-SPACE COORDINATES FOR FIXED PREDICTORS

For a fixed $S_{m,j} = s_{m,j}(X)$, write $f^S_{y,m,j}(s) = p(S_{m,j} = s \mid Y_j = y)$, with a density or mass function as appropriate. Bayes' rule gives

$$\operatorname{logit} c_{m,j}(s) - \operatorname{logit} \pi_j = \log \frac{f^S_{1,m,j}(s)}{f^S_{0,m,j}(s)}. \tag{24}$$

The mA-optimal decision among rules that use this score is $c_{m,j}(s) > \pi_j$, with arbitrary population ties. When $c_{m,j}$ is nondecreasing, this rule can be implemented by a score threshold with a specified tie convention. Isotonic calibration estimates the score posterior within the class of nondecreasing functions.

Since $S_{m,j}$ is a function of X, $c_{m,j}(S_{m,j}) = \mathbb{E}[\mathbb{E}[Y_j \mid X] \mid S_{m,j}] = \mathbb{E}[\eta_j(X) \mid S_{m,j}]$. If the score exactly realizes an objective-implied response that is invertible in η_j, then $\eta_j(X)$ is determined by $S_{m,j}$, giving $c_{m,j}(S_{m,j}) = \eta_j(X)$ almost surely. The population score-space coordinate therefore agrees with κ^{obj} at supported thresholds. Score-space policy estimation examines this correspondence for the fixed predictor. Differences between κ^{obj} and κ^{est} may reflect departures from the objective-implied score response as well as calibration error.

Distribution shifts. Fix a predictor and attribute and suppress m, j. For $d \in \{\mathrm{val}, \mathrm{test}\}$, write $c^d(s) = \Pr_d(Y = 1 \mid S = s)$, $\pi^d = \Pr_d(Y = 1)$, and let $f_y^{S,d}(s)$ denote the class-conditional score density or mass function. Assume $0 < \pi^d < 1$. Under prior-only shift, $f_y^{S,\mathrm{test}} = f_y^{S,\mathrm{val}}$ for $y \in \{0, 1\}$, while the class prevalences may differ. Applying Eq. 24 to each distribution gives

$$\operatorname{logit} c^{\mathrm{test}}(s) - \operatorname{logit} \pi^{\mathrm{test}} = \log \frac{f_1^{S,\mathrm{test}}(s)}{f_0^{S,\mathrm{test}}(s)} = \operatorname{logit} c^{\mathrm{val}}(s) - \operatorname{logit} \pi^{\mathrm{val}}. \tag{25}$$

The mA-optimal score-based decision is therefore unchanged. If c^{val} is nondecreasing, a score-threshold rule implementing $c^{\mathrm{val}}(s) > \pi^{\mathrm{val}}$ remains optimal under the test distribution, with appropriate boundary-tie handling.

For a general reference $q \in (0, 1)$, the population decision instead satisfies

$$c^{\mathrm{val}}(s) > q \quad \Longleftrightarrow \quad \log \frac{f_1^{S,\mathrm{test}}(s)}{f_0^{S,\mathrm{test}}(s)} > \operatorname{logit} q - \operatorname{logit} \pi^{\mathrm{val}}. \tag{26}$$

Thus, the reference prior determines the likelihood-ratio boundary even when the fitted predictor is unchanged. The population counterpart of the training-prior rule sets $q = \pi^{\mathrm{tr}}$, where π^{tr} is the original-training prevalence. It targets the zero boundary when $\pi^{\mathrm{tr}} = \pi^{\mathrm{val}}$, as in the common-reference formulation. The empirical Est. rule uses the fitted map and $\widehat{\pi}^{\mathrm{tr}}$ to estimate this operating point.

Under class-conditional shift, the score likelihood ratio can change and its zero crossing can move, even if the class prevalence remains unchanged. Prior adjustment alone therefore does not generally preserve the test mA-optimal decision rule.

C Policy Estimation and Threshold Construction

C.1 Fixed Predictors, Data Splits, and Prior Estimates

For each selected predictor m and attribute j, calibration uses validation pairs $(s_{m,j}(x_i), y_{i,j})$ only. We compute attribute prevalences on the original training and validation sets as

$$\widehat{\pi}_j^{\mathrm{tr}} = \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathcal{D}_{\mathrm{tr}}} y_{i,j}, \qquad \widehat{\pi}_j^{\mathrm{val}} = \frac{1}{n_{\mathrm{val}}} \sum_{i \in \mathcal{D}_{\mathrm{val}}} y_{i,j}. \tag{27}$$

The training estimate $\widehat{\pi}_j^{\mathrm{tr}}$, computed before any training-time resampling, is used in the policy coordinates and normalized thresholds. The validation estimate $\widehat{\pi}_j^{\mathrm{val}}$ is used in the validation-optimality analysis below. Under the common-reference assumption, both estimate the same population prevalence π_j, while their finite-sample values can differ.

We use the checkpoints described in Section 4.1. Each predictor and its logits remain fixed across evaluation views. Calibration maps are fitted to validation logits and labels; Est. thresholds are obtained by inverting these maps at the training priors. The maps and thresholds are frozen before evaluation on the test set.

C.2 Isotonic Fitting and Coordinate Evaluation

Fix a predictor and attribute and suppress their indices. Let $u_1 < \cdots < u_K$ be the distinct validation logits, let w_k be the number of observations at u_k, and let $\bar{y}_k$ be their positive fraction. Equal logits are grouped before fitting, so their ordering cannot depend on how tied observations are listed. The weighted isotonic problem is

$$(v_1, \ldots, v_K) = \operatorname*{argmin}_{0 \le v_1 \le \cdots \le v_K \le 1} \sum_{k=1}^{K} w_k (v_k - \bar{y}_k)^2. \tag{28}$$

The pool-adjacent-violators algorithm repeatedly merges adjacent blocks whose fitted levels violate monotonicity and assigns each merged block its weighted positive fraction (Zadrozny & Elkan, 2002; Fawcett & Niculescu-Mizil, 2007). The weights w_k count observations with the same score, preserving the empirical class frequencies of the validation sample.

To specify the map between observed logits, we use linear interpolation between the fitted values at adjacent distinct scores:

$$\widehat{c}(s) = v_k + \frac{s - u_k}{u_{k+1} - u_k}(v_{k+1} - v_k), \qquad u_k \le s \le u_{k+1}. \tag{29}$$

Equal adjacent levels form flat intervals. When $K = 1$, the map is defined only at the single observed score.

For finite numerical log-odds, coordinate evaluation clips the fitted probability to $[\epsilon, 1 - \epsilon]$, where

$$\epsilon = \frac{0.5}{n_{\mathrm{val}} + 1}, \qquad \widehat{c}_\epsilon(s) = \min\{1 - \epsilon, \max\{\epsilon, \widehat{c}(s)\}\}. \tag{30}$$

Thus the empirical coordinates at supported Native and Obj. thresholds are

$$\begin{aligned} \kappa^{\mathrm{est}}_{m,j} &= \operatorname{logit} \widehat{c}_{\epsilon,m,j}(\tau^{\mathrm{native}}_{m,j}) - \operatorname{logit} \widehat{\pi}^{\mathrm{tr}}_j, \\ \kappa^{\mathrm{est,obj}}_{m,j} &= \operatorname{logit} \widehat{c}_{\epsilon,m,j}(\tau^{\mathrm{obj}}_{m,j}) - \operatorname{logit} \widehat{\pi}^{\mathrm{tr}}_j. \end{aligned} \tag{31}$$

Threshold inversion uses the original fitted levels v_k; probability clipping is applied only when computing log-odds coordinates.

C.3 Objective and Estimated Thresholds

Objective normalization. Appendix B gives the analytical thresholds for fixed weighting, resampling, and logit adjustment. All empirical formulas substitute $\widehat{\pi}^{\mathrm{tr}}_j$ for π_j and use the classification component and score convention specified for that method. For prediction-dependent components, the target equation is

$$\eta^{\dagger}_{m,j}(t^{\mathrm{obj}}_{m,j}) = \widehat{\pi}^{\mathrm{tr}}_j, \qquad \tau^{\mathrm{obj}}_{m,j} = \operatorname{logit} t^{\mathrm{obj}}_{m,j}.$$

For MLFL and ASL, the crossing maps are computed from the focal gradients in Eq.22, using the loss parameters in Tab.2. The target equations are solved using 100 bisection iterations, accepting a root when the absolute posterior-target residual is at most 10^{-10}.

Realized-policy normalization. For $q = \widehat{\pi}^{\mathrm{tr}}_j$, the left generalized inverse of the interpolated isotonic map is

$$\tau^{\mathrm{est}}_{m,j} = \inf\{s \in [u_1, u_K] : \widehat{c}_{m,j}(s) \ge q\}. \tag{32}$$

It is supported only when $q \in [v_1, v_K]$. Let $k = \min\{l : v_l \ge q\}$. If $v_k = q$, the threshold is u_k, including $k = 1$. Otherwise $k > 1$ and $v_{k-1} < q < v_k$, and

$$\tau^{\mathrm{est}}_{m,j} = u_{k-1} + \frac{q - v_{k-1}}{v_k - v_{k-1}}(u_k - u_{k-1}). \tag{33}$$

If a flat interval has fitted value q, the inverse selects its left endpoint. Predictions follow the main-text convention $\hat{y}_{m,j}(x) = \mathbf{1}\{s_{m,j}(x) > \tau^{\mathrm{est}}_{m,j}\}$. Observations at the selected endpoint are negative, while those with larger scores on the same plateau are positive.

Support criteria. The training prior is required to lie in $(0, 1)$, and calibration requires both validation classes. Coordinate queries are restricted to $[u_1, u_K]$; isotonic inversion additionally requires $q \in [v_1, v_K]$. Unsupported cases are recorded explicitly. Once fixed, a supported threshold is applied to all test scores. Test mA uses the complete prescribed attribute set, with both test classes present for each attribute. Paired analyses retain only refits with supported thresholds for every required attribute and predictor.

C.4 Connection to Validation-Optimal Threshold Selection

We derive the score cuts that maximize validation mA and relate them to the fitted isotonic map. Set $q = \widehat{\pi}_j^{\mathrm{val}} \in (0, 1)$ and retain the original, unsmoothed binary validation labels. For the sorted distinct scores used above, define

$$W_k = \sum_{l=1}^{k} w_l, \qquad A_k = \sum_{l=1}^{k} w_l \bar{y}_l, \qquad W_0 = A_0 = 0,$$

so $W_K = n_{\mathrm{val}}$ and $A_K = n_{\mathrm{val}} q$. A threshold cut after group k predicts negative on groups $1, \ldots, k$ and positive on the remaining groups. Its empirical per-attribute balanced accuracy is

$$\begin{aligned} \mathrm{BA}_j^{\mathrm{val}}(k) &= \frac{1}{2}\left[\frac{n_{\mathrm{val}} q - A_k}{n_{\mathrm{val}} q} + \frac{W_k - A_k}{n_{\mathrm{val}}(1-q)}\right] \\ &= \frac{1}{2} + \frac{qW_k - A_k}{2 n_{\mathrm{val}} q(1-q)}. \end{aligned} \tag{34}$$

The all-positive and all-negative rules correspond to $k = 0$ and $k = K$ and are included in this optimization.

Weighted PAV constructs the slopes of the greatest convex minorant of the cumulative-sum points (W_k, A_k). Its segment slopes are the fitted block probabilities v_k; this is the cumulative-sum form of the PAV–ROC convex-hull connection (Fawcett & Niculescu-Mizil, 2007). Subtracting the line qW from that convex minorant gives a convex piecewise-linear function with successive slopes $v_k - q$. Its minimum occurs where these slopes change from negative to positive, or along a segment whose slope is zero. A minimizing segment endpoint is an original cumulative-sum point, where the minorant equals A_k; at every other original point the minorant is no higher than A_k. That endpoint therefore also minimizes $A_k - qW_k$ over all score cuts and maximizes Eq. 34.

It follows that thresholding the PAV probabilities at q, assigning the entire target-level plateau to one class, attains a validation-optimal score-threshold decision. Choosing all blocks with probability exactly q as positive, or choosing all of them as negative, gives an optimal boundary. Equal raw scores are never split. Since mA assigns equal weight to attributes, independent optimal cuts also maximize validation mA over attribute-wise score thresholds.

The empirical optimum above uses $\widehat{\pi}_j^{\mathrm{val}}$ and places the cut at a PAV block boundary. The reported Est. rule uses $\widehat{\pi}_j^{\mathrm{tr}}$ and the left inverse in Eq.32. For a target strictly between adjacent fitted levels, linear interpolation locates the threshold between the corresponding validation logits, preserving their cut at that target. For a target equal to a plateau level, strict comparison with the left endpoint can place the cut inside a PAV block.

C.5 Monotone Sigmoid Calibration

To examine sensitivity to the calibration model, we fit $\widehat{c}_{m,j}^{\mathrm{sig}}(s) = \sigma(a_{m,j} s + b_{m,j})$ with a positive slope. The parameters minimize cross-entropy on class-dependent smoothed validation targets (Niculescu-Mizil & Caruana, 2005):

$$\widetilde{y}_+ = \frac{n_+ + 1}{n_+ + 2}, \qquad \widetilde{y}_- = \frac{1}{n_- + 2},$$

where n_+ and n_- are the positive and negative counts in the calibration sample. Validation logits are standardized using their sample mean and standard deviation; the fitted coefficients are then converted back to the original logit scale. The standardized slope is constrained to be at least 10^{-8};

fits at this lower bound are treated as degenerate for inversion. For a supported fit, the estimated threshold is

$$\tau_{m,j}^{\mathrm{est,sig}} = \frac{\operatorname{logit} \widehat{\pi}_j^{\mathrm{tr}} - b_{m,j}}{a_{m,j}}. \tag{35}$$

The same validation-score support restriction and coordinate clipping apply as for isotonic calibration. The sigmoid and isotonic evaluations use the same predictors, validation data, and training priors. Their comparison in Section 4.4 measures sensitivity to the calibration rule under these fixed conditions.

D Experimental Details

D.1 Benchmarks and Data Partitions

We evaluate PETA (Deng et al., 2014), PA100K (Liu et al., 2017), and RAPv1 (Li et al., 2016) using their standard attribute subsets. RAP in the experimental sections refers to RAPv1. PETA and PA100K use the supplied training, validation, and test partitions. Tab.9 summarizes the attribute and image counts.

Table 9: Dataset partitions used in the experiments.

Dataset	Attributes	Training	Validation	Test
PETA	35	9,500	1,900	7,600
PA100K	26	80,000	10,000	10,000
RAP	51	29,572	3,696	8,317

Validation partition for RAP. The released RAP protocol provides five training/test partitions, each with 33,268 training images and 8,317 test images, but no separate validation set (Li et al., 2016). We use the first partition and divide its original training pool into 29,572 training images and 3,696 validation images, approximately an $8{:}1$ split, while retaining the original test set. The three subsets are disjoint and shared across retrained methods and backbones. Training priors are computed from the 29,572-image training subset, and the held-out validation subset is used for policy estimation. All 51 attributes have both classes in the validation subset. The fixed image indices and attribute order are provided in `RAP_dataset_all_train_val_test.pkl` in the supplementary material; experiments use entry zero of its partition lists.

D.2 Method Implementations

D.2.1 Controlled Methods

The controlled experiments use the PAR training framework of Jia et al. (2021), with ImageNet-pretrained ResNet-50 and ViT-B/16 backbones and a linear attribute classifier. Inputs are resized to 256×192. Training applies 10-pixel padding, random cropping back to this resolution, and random horizontal flipping. Evaluation uses deterministic resizing; ImageNet normalization is applied in both training and evaluation.

The single-stage methods are trained for 30 epochs using Adam with a batch size of 64, an initial learning rate of 10^{-4} for both the backbone and classifier, and weight decay 5×10^{-4}. The learning-rate schedule follows the baseline implementation (Jia et al., 2021).

FRDL classifier retraining. The controlled FRDL experiment starts from the converged ERM predictor for the corresponding dataset and backbone. The backbone is frozen, and separate linear attribute heads are trained using balanced positive/negative feature queues (Zhou et al., 2025b). The queue size is 1,024, and feature augmentation is disabled in this controlled setting. Head retraining uses 30 epochs with the same batch size, optimizer settings, and learning-rate schedule as above.

Table 10: Published-method settings and the scope of Obj. threshold derivation. The final column lists method-specific mechanisms outside the scalar derivation.

Method	Backbone	Native t	Component used for Obj.	Outside the scalar derivation
L2L	ResNet-50	0.5	RW_{exp} BCE on label-head outputs	Query-head supervision, label masking, and attribute-dependency modeling
PIL	ResNet-50	0.5	RW_{exp} BCE	Posterior-invariant feature learning
FRDL+GOAT	ResNet-50	0.5	Balanced classifier BCE, $r_j = 0.5$	Feature-extractor training and GOAT feature augmentation
ReSAGE-PAR	ResNet-50 / ViT-B/16	0.5	RW_{exp} BCE	Synthetic-image generation, image–text scoring, and pseudo-labeling
HAP	ViT-B/16	0.5	RW_{exp} BCE for PAR fine-tuning	Structure-aware human-centric pre-training
PARFormer-B	Swin-B	0.5	ASL	Multi-attribute center loss, multi-view contrastive loss, and batch random masking
PromptPAR	CLIP ViT-L	0.45	RW_{exp} BCE	CLIP pretraining and prompt-based vision–language fusion

Loss settings. The RW_{inv} and RW_{exp} weights follow Appendix B.2. CB uses binary effective-number weights with $\beta = 0.9999$, normalized so that the positive and negative weights sum to two for each attribute. LA adds $\operatorname{logit} \widehat{\pi}_j^{\text{tr}}$ during training and evaluates the unadjusted logit. DB uses the non-focal, loss-only implementation with rebalancing parameters $(\alpha, \beta, \mu) = (0.1, 10, 0.3)$, negative scale 5, and bias scale 0.05 (Wu et al., 2020), with a uniform image sampler.

MLFL uses $\gamma_+ = \gamma_- = 1$ and weights $a_j = \exp(1 - \widehat{\pi}_j^{\text{tr}})$ and $b_j = \exp(\widehat{\pi}_j^{\text{tr}})$, with differentiation through the focal factors (Li et al., 2020). Its Obj. threshold satisfies $\widehat{\pi}_j^{\text{tr}} g_+(t) + (1 - \widehat{\pi}_j^{\text{tr}}) g_-(t) = 0$, using the class gradients in Eq.22 and the root-finding procedure in Appendix C.3.

D.2.2 Published PAR Methods

The published-method experiments are based on the authors' official implementations and released method-specific settings, with data loading adapted to the shared partitions and attribute order. Further implementation details are available in the public repositories listed in Tab.8. Tab.10 identifies the classification component used to derive Obj. thresholds and the method-specific mechanisms outside that derivation. These mechanisms affect the learned scores, but their effects are not separately modeled by the scalar coordinate. All three evaluation views use the same complete trained predictor; Obj. applies the component-derived thresholds to its fixed scores. PromptPAR is evaluated using one officially released checkpoint per dataset.

D.3 Validation Subsampling

For each validation fraction in Fig.5, 30 image subsets are sampled uniformly without replacement. The same subsets are used across methods, seeds, and calibration rules within each dataset. Only calibration is refitted; predictors, training priors, and test data remain fixed. Each method gap is computed from test mA averaged across three seeds. A subsampling replicate is retained only when all thresholds required for the comparison satisfy the support criteria in Appendix C.3. Isotonic retains all 30 replicates in every setting. For sigmoid at 25%, the supported counts are 28 on RAP with ResNet-50 and 29 on PETA with ViT-B/16; all other settings retain 30 replicates.

Table 11: Native coordinate distances, shown as Isotonic / Sigmoid. Medians are computed over attributes and three seeds, with absolute distances formed before aggregation. For FRDL and LA, $\kappa^{\mathrm{obj}} = 0$, so their distances to the mA target also equal their objective-coordinate discrepancies.

Dataset	median $\lvert\kappa^{\mathrm{est}} - \kappa^{\mathrm{obj}}\rvert$	median $\lvert\kappa^{\mathrm{est}}\rvert$		
	ERM	ERM	FRDL	LA
ResNet-50				
PETA	0.288 / 0.182	1.649 / 1.711	1.011 / 0.985	0.535 / 0.341
PA100K	0.302 / 0.322	1.732 / 1.533	0.726 / 0.598	0.522 / 0.245
RAP	0.336 / 0.259	2.681 / 2.532	0.906 / 0.739	0.364 / 0.209
ViT-B/16				
PETA	0.288 / 0.186	1.836 / 1.947	0.866 / 0.811	0.395 / 0.325
PA100K	0.401 / 0.338	1.718 / 1.414	0.764 / 0.707	0.507 / 0.344
RAP	0.442 / 0.363	2.525 / 2.455	0.931 / 0.816	0.428 / 0.247

Table 12: Coordinate estimates for ASL on PA100K. Medians are computed over attributes and three seeds.

Backbone	Calibrator	median $\lvert\kappa^{\mathrm{est}}\rvert$	median $\lvert\kappa^{\mathrm{est,obj}}\rvert$	median $\kappa^{\mathrm{est,obj}}$
ResNet-50	Isotonic	1.170	0.562	$+0.225$
	Sigmoid	1.081	0.356	-0.108
ViT-B/16	Isotonic	1.117	0.450	$+0.237$
	Sigmoid	1.047	0.402	-0.119

E Supplementary Robustness Analyses

E.1 Calibration Sensitivity of Decision Coordinates

We compare isotonic and monotone sigmoid estimates at the same Native and Obj. thresholds, using the same predictors, validation data, and training priors. Tab.11 reports the Native coordinate distances underlying the comparisons in Section 4.3. Across all six dataset–backbone settings, both calibrators give ERM a smaller median discrepancy from its objective-implied coordinate than FRDL but a larger median distance to the mA target. FRDL also remains farther from the target than LA despite their shared zero objective coordinate. These relations persist even when the estimated distances change appreciably: for LA on PA100K with ResNet-50, the median distance changes from 0.522 to 0.245. The qualitative comparisons therefore persist across the two calibrators, while the numerical distances remain sensitive to the calibration model.

For ERM, RW_{exp}, MLFL, ASL, and CB, Obj. reduces the median absolute estimated coordinate in all 30 method–dataset–backbone settings under both calibrators. The calibrators agree on the sign of the median $\kappa^{\mathrm{est,obj}}$ in 28 of these settings. The exceptions are ASL on PA100K, detailed in Tab. 12. There, the absolute distance decreases from Native to Obj. under both calibrators, while the signed Obj. medians are positive under isotonic and negative under sigmoid calibration. These results support a consistent reduction in estimated distance under the two calibrators, while the directional interpretation of the remaining offset depends on the fitted map.

E.2 Threshold Selection and Comparison Uncertainty

We include direct validation mA optimization (Val-opt.) as a threshold-selection reference for the same fixed predictors. For each attribute, Val-opt. maximizes per-attribute mA over all realizable validation score cuts, including constant decisions. Among tied optima, the lower median cut is selected; internal thresholds lie midway between adjacent distinct logits. Point estimates use the full validation and test sets and are averaged across three seeds.

Table 13: Signed test mA gaps and 95% bootstrap intervals (pp). Est. and Val-opt. intervals include validation refitting and paired test-image resampling; Native and Obj. thresholds remain fixed. RAP gaps are $RW_{\text{inv}} - \text{ERM}$; PETA gaps are HAP minus ReSAGE-PAR. Each entry shows the full-data gap above its interval.

Setting	Native	Obj.	Est. (Isotonic)	Est. (Sigmoid)	Val-opt.
RAP ResNet-50	$+8.99$ $[+8.69, +9.32]$	-0.48 $[-0.70, -0.28]$	$+0.12$ $[-0.22, +0.36]$	$+0.05$ $[-0.17, +0.23]$	$+0.12$ $[-0.19, +0.35]$
RAP ViT-B/16	$+8.16$ $[+7.86, +8.50]$	$+0.31$ $[+0.15, +0.48]$	$+0.28$ $[+0.02, +0.52]$	$+0.22$ $[+0.08, +0.43]$	$+0.29$ $[+0.02, +0.52]$
PETA ViT-B/16	$+1.59$ $[+1.30, +1.89]$	-0.34 $[-0.61, -0.09]$	$+0.84$ $[+0.44, +1.06]$	$+0.73$ $[+0.45, +0.93]$	$+0.84$ $[+0.43, +1.03]$

We perform 1,000 bootstrap repetitions to capture both threshold-estimation and test-sample variation. In each repetition, validation and test-image rows are independently resampled with replacement at their original sizes. Both calibrators are refitted to construct Est. thresholds, and Val-opt. thresholds are reselected. The bootstrap conditions on the fixed trained predictors; their weights remain unchanged throughout resampling. Training priors and Native and Obj. thresholds remain fixed. Within each comparison, methods, seeds, and evaluation views share the corresponding resampling indices. Tab.13 reports 95% percentile intervals. Est. refits require supported thresholds for every attribute and all six seed predictors in each comparison. Val-opt. requires both validation classes and includes constant decisions. On RAP with ResNet-50, each Est. variant retains 998 of 1,000 comparison replicates; all other intervals use 1,000.

With full-validation threshold selection, Val-opt. and isotonic-based Est. differ by at most 0.035 pp in mean test mA and 0.013 pp in method gaps across these three comparisons. Both recover the RAP gap contraction and the PETA ordering reversal relative to Obj., consistent with the PAV connection in Appendix C.4.

On RAP, the gap contraction persists under joint resampling, while evidence for the residual ordering differs between backbones. The intervals for both Est. variants and Val-opt. include zero with ResNet-50 and remain positive with ViT-B/16, although the isotonic and Val-opt. lower bounds are close to zero. On PETA, the Obj. interval is negative, whereas both Est. variants and Val-opt. have positive intervals. The paired Obj.–Est. change in the gap under isotonic calibration is $+1.18$ pp, with a joint 95% interval of $[0.81,\ 1.39]$ pp. These results support the examined gap contraction and policy-dependent reversal when validation fitting and test sampling vary jointly, conditional on the fixed predictors.

E.3 Sensitivity to Reference Priors

Table 14 summarizes empirical prevalence differences between the training, validation, and test splits. These differences are largest on PA100K. Appendix B.5 gives the population conditions for transferring the mA decision rule across distributions. Here, we isolate the effect of the reference-prior input by replacing $\widehat{\pi}_j^{\text{tr}}$ with $\widehat{\pi}_j^{\text{val}}$ in isotonic inversion, keeping predictors, fitted maps, and evaluation data fixed. The controlled comparison uses the eight methods in the coordinate analysis of Section 4.3, with both backbones and all three datasets. The analysis also includes HAP and ReSAGE-PAR on PETA with ViT-B/16. Test prevalences are used only for descriptive diagnostics.

The resulting test mA changes depend on the dataset and predictor (Tab.15). On PA100K, the median absolute change in logit thresholds, computed over attributes and three seeds, ranges from 0.125 to 0.467 across the 16 controlled method–backbone settings. With ResNet-50, the mean FRDL-minus-RW_{exp} mA gap changes from -0.142 to $+0.034$ pp. A common change in reference priors can therefore affect methods unequally and alter close comparisons. The RW_{inv}–ERM gaps on RAP remain below 0.30 pp in absolute value on both backbones under either prior source, while the HAP–ReSAGE-PAR mean gap on PETA remains approximately 0.84 pp. Across the 48 controlled and two published-method settings, validation-prior inversion differs from Val-opt. by less than 0.007 pp in mean test mA, consistent with the empirical-prior condition in Appendix C.4.

Table 14: Absolute differences in empirical positive-class prevalence (pp), summarized over the evaluated attributes.

	Validation versus training		Test versus training	
Dataset	Mean	Maximum	Mean	Maximum
PETA	0.72	3.28	0.34	1.46
PA100K	3.01	11.26	4.01	14.53
RAP	0.05	0.61	0.32	1.09

Table 15: Change in test mA when validation priors replace training priors in the same isotonic maps (pp). Entries are validation-prior Est. minus training-prior Est., averaged across three seeds.

	ResNet-50			ViT-B/16		
Method	PETA	PA100K	RAP	PETA	PA100K	RAP
ERM	+0.117	-0.301	+0.003	+0.007	-0.242	-0.008
RW_{exp}	-0.026	-0.251	+0.021	+0.093	-0.233	0.000
RW_{inv}	+0.164	-0.039	+0.006	+0.047	-0.195	+0.005
FRDL	+0.073	-0.075	0.000	+0.073	-0.255	+0.012
MLFL	-0.021	-0.325	+0.004	+0.022	-0.433	-0.006
ASL	+0.015	-0.555	+0.004	-0.030	-0.009	+0.002
CB	+0.093	-0.246	+0.015	+0.046	+0.030	+0.012
LA	+0.031	-0.226	-0.002	+0.016	-0.166	+0.004

We additionally perturb the reference-prior input using $q_j = \sigma(\operatorname{logit} \widehat{\pi}_j^{\text{tr}} + \delta_\pi)$, with a common offset $\delta_\pi \in \{0, \pm 0.25, \pm 0.5, \pm 1, \pm 1.25\}$. The tested range covers the empirical train–validation log-odds differences observed across the benchmarks. At each offset, we invert the same isotonic maps at q_j to obtain thresholds and recompute test mA gaps, holding predictors and data partitions fixed. In Fig.6, the RW_{inv}–ERM gaps on RAP remain below 0.46 pp in absolute value at all tested offsets, although their mean ordering can change. On PA100K, the opposite mean orderings for the two backbones persist across the tested offsets. The PETA HAP–ReSAGE-PAR mean gap remains positive, ranging from 0.72 to 1.14 pp. For the RAP comparisons, the contraction relative to Native therefore persists across the tested reference-prior choices, while the direction of the small remaining gaps is less stable.

E.4 Monotonicity Diagnostics

We examine the monotonicity assumption underlying Est. through raw score–label relations for the same 150 fixed predictors as in Appendix E.3. For each predictor–attribute curve, validation-score quantiles define ordered bins, and test images provide bin-wise positive fractions and pointwise 95% Wilson intervals (Wilson, 1927). Ten bins form the primary analysis; five and twenty bins assess sensitivity to resolution. We examine the full score range and the Est. neighborhood, defined as the bin containing the frozen Est. threshold and its immediate neighbors. One-sided Fisher exact tests evaluate decreases between ordered bins. Tested pairs require at least 100 test images in each bin, and attributes require at least 20 observations of each class. Bonferroni correction accounts for bin pairs within each curve and region; Holm correction (Holm, 1979) covers the resulting tests across curves, regions, and binning resolutions at $\alpha = 0.05$.

With ten bins, numerical decreases among eligible bin pairs occur in 3,910 of 5,586 curves over the full score range and 88 (1.6%) in the Est. neighborhood. Among curves with full-range decreases, the largest drop in positive fraction per curve has a median of 0.24 pp and a 95th percentile of 0.84 pp. No curve meets the corrected criterion for a decrease with 5, 10, or 20 bins. These results are consistent with a nondecreasing approximation near the mA decision reference at these resolutions.

As shown in Fig.7, the ten-bin positive fractions are nondecreasing for the displayed attributes, with a single crossing of the training-prior reference within the Est. neighborhood. This ordering holds across all three seeds, and local ordering persists with five and twenty bins. Across the displayed

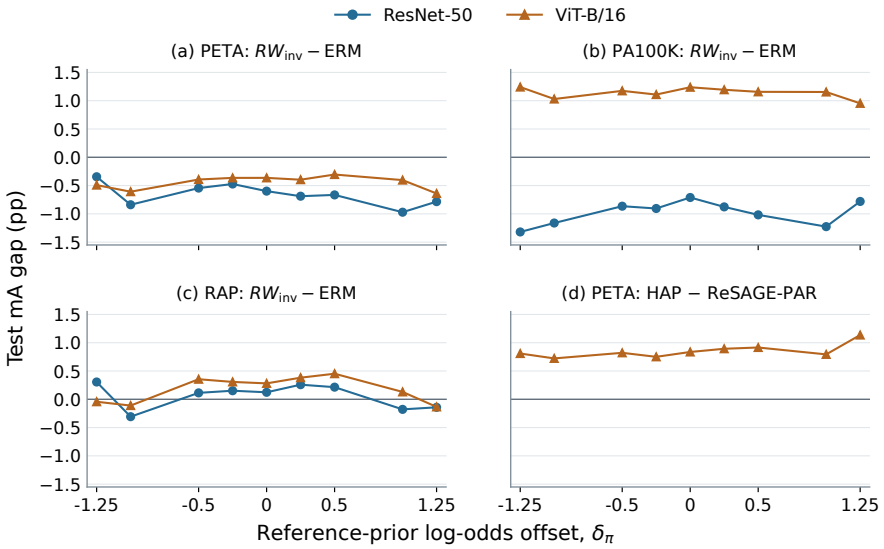


Figure 6: Sensitivity of Est. test mA gaps to reference-prior perturbations. A common log-odds offset δ_π is applied to the training priors before isotonic inversion; $\delta_\pi = 0$ recovers the default Est. evaluation. Markers show means across three seeds, with signed gaps following the method order in each panel.

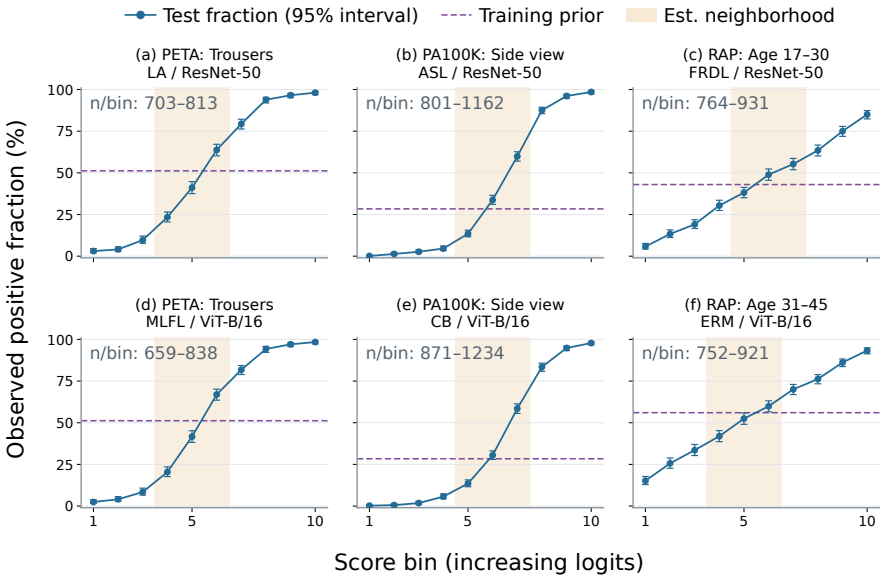


Figure 7: Illustrative raw score–label relations. The displayed cases exhibit nondecreasing ten-bin fractions across three seeds and consistent ordering near Est. with five and twenty bins. Panels show seed 1, with pointwise 95% Wilson intervals. Dashed lines mark training priors; shading marks the Est. neighborhood. Annotations give the range of test-image counts per bin.

attributes, seeds, and binning resolutions, each Est. neighborhood contains at least 514 test images per class. This ordered crossing illustrates how the mA reference can be expressed through a single score threshold, as required by the Est. construction.