RETHINKING FAIRNESS EVALUATION IN FEDERATED REID: BEYOND PERFORMANCE VARIANCE UNDER DOMAIN SKEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair federated ReID aims to achieve equitable identity retrieval across decentralized and heterogeneous ReID domains. However, when fairness is evaluated solely by performance parity among domains, smaller retrieval performance variance is regarded as stronger fairness. Used alone, this criterion overlooks a fundamental property of ReID scenarios: different domains inherently possess different retrieval difficulties caused by variations in cameras, viewpoints, backgrounds, illumination, and identity distributions. Consequently, equal raw performance does not necessarily imply equal reference-relative attainment, while low variance can coexist with uniformly poor performance. In this work, we rethink the fairness objective of fair federated ReID and argue that fair achievement should be evaluated relative to each domain's protocol-matched single-domain reference rather than raw performance disparity alone. To this end, we introduce **S**ingle-**D**omain **S**upervised **R**eference (**SDSR**), a reference calibration scheme that calibrates federated ReID performance using protocol-matched single-domain reference performance. Based on normalized domain achievement, we further propose **SDSR-HM**, which adopts harmonic aggregation to emphasize balanced reference-relative attainment across heterogeneous and challenging ReID domains. Extensive experiments demonstrate that raw dispersion alone can produce misleading fair-achievement rankings and fail to distinguish high attainment from uniform under-attainment. Our results reveal the necessity of considering reference-relative attainment in fair federated ReID and provide a more informative criterion for measuring fair achievement under domain skew.

1 INTRODUCTION

Person re-identification (ReID) aims to retrieve images of the same identity across different camera views and environments. With increasing privacy concerns, federated ReID enables collaborative training without exchanging raw identity data (McMahan et al., 2017; Li et al., 2020a; Wu & Gong, 2021; Zhuang et al., 2020). However, domain heterogeneity caused by variations in cameras, locations, identity distributions, and image conditions makes it difficult to achieve consistent retrieval capability across clients.

To address this issue, fair federated ReID has been investigated to reduce performance imbalance among clients (Li et al., 2020b; Mohri et al., 2019; Li et al., 2021b). Performance deviation is commonly reported in fairness evaluation, where smaller variations in mAP or Rank-1 scores indicate greater raw performance parity (Wang et al., 2021; Ezzeldin et al., 2023). Used alone, this criterion does not account for domain-specific reference performance. In practical ReID scenarios, different domains have different achievable performance due to variations in viewpoint, occlusion, background complexity, and identity ambiguity (Song et al., 2019; Dai et al., 2021). Therefore, raw retrieval comparison cannot separate unfair optimization from intrinsic domain difficulty.

Moreover, deviation-based metrics suffer from the **low-performance fairness illusion**, where uniformly poor models may obtain favorable fairness scores due to small inter-domain variance. Such evaluation reflects performance similarity rather than whether each domain benefits effectively from

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.

federated training. Therefore, fair federated ReID requires fair-achievement metrics that consider both domain-specific references and relative attainment.

As illustrated in Figure 1(a), Method B obtains lower retrieval performance on all domains after uniform degradation, while showing smaller cross-domain performance disparity. According to the STD metric, Method B may be considered fairer than Method A, although it provides consistently inferior ReID capability. This phenomenon reveals that using the STD metric alone can conflate raw performance equality with fair achievement. A fair federated ReID system should not pursue equality by reducing every domain to a similarly poor level in deployment scenarios.

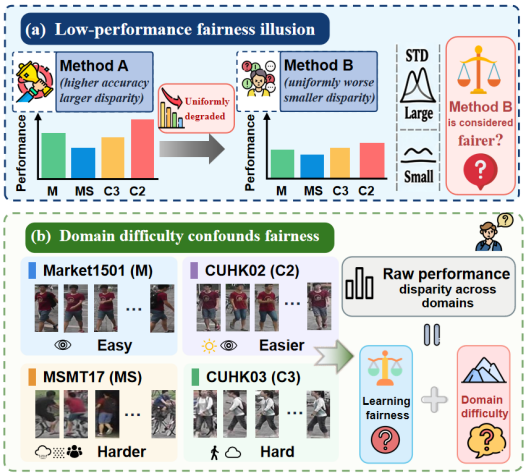


Figure 1: Illustration of fairness evaluation challenges in fair federated ReID. (a) Low-performance fairness illusion caused by disparity-based evaluation. (b) Domain difficulty confounding caused by heterogeneous ReID domains.

Beyond the low-performance fairness illusion, another fundamental issue is **domain difficulty confounding**. As shown in Figure 1(b), different ReID domains naturally exhibit different retrieval difficulties under the same protocol. Domains with larger viewpoint variations, severe occlusions, or complex backgrounds are inherently more challenging than relatively stable domains. Therefore, the observed performance gap among domains is determined by two coupled factors: the benefit obtained from federated optimization and the intrinsic difficulty of each domain. Raw-dispersion metrics directly measure performance discrepancy across domains without distinguishing these two sources of variation. Consequently, they may incorrectly attribute the performance gap caused by domain difficulty to unfair optimization, or overlook unfairness hidden behind naturally easy or difficult domains. This entanglement limits the use of raw dispersion alone for assessing balanced reference-relative attainment across heterogeneous ReID domains.

These observations raise an important question: *What should fairness mean in fair federated ReID?*

Fair achievement in federated ReID cannot be characterized solely by reducing the performance discrepancy among clients (Li et al., 2020b; Mohri et al., 2019), since different ReID domains have different single-domain reference performances under a specified protocol. Therefore, our evaluation objective is not identical raw retrieval performance across domains, but comparable attainment relative to each domain's empirical reference performance.

We revisit fairness evaluation in fair federated ReID from a difficulty-aware perspective. Each ReID domain has a protocol-dependent single-domain reference performance influenced by its visual characteristics, training data, and training configuration. Consequently, our metric calibrates domain performance against these empirical references rather than directly comparing absolute retrieval scores.

To achieve this goal, we propose **S**ingle-**D**omain **S**upervised **R**eference (**SDSR**), a reference calibration scheme for fair federated ReID based on domain-specific reference performance. Specifically, for each ReID domain, we train a single-domain model using only the local data to obtain its protocol-matched reference performance. Such a domain-specific reference records the performance achieved without cross-domain federated optimization. By comparing the federated model with this reference, SDSR measures reference-relative attainment while reducing the influence of different domain-specific performance scales.

Based on this reference-aware calibration, SDSR transforms client performances into normalized achievement representations. Different from comparing raw retrieval scores, SDSR evaluates whether each domain attains a comparable proportion of its own empirical reference. Therefore, a domain with lower absolute performance does not necessarily have lower reference-relative attainment. Conversely, a domain with higher retrieval performance may still exhibit lower reference-relative attainment if it reaches a smaller proportion of its single-domain reference.

Furthermore, we introduce harmonic mean aggregation to obtain the final fairness metric, termed **SDSR-HM**. The harmonic aggregation emphasizes domains with lower achievement values and prevents high-performing domains from dominating the overall fairness evaluation. Consequently, SDSR-HM captures reference-relative fair achievement across heterogeneous domains and provides a more reliable fairness criterion for fair federated ReID.

Through this reference-aware evaluation paradigm, fair-achievement evaluation in federated ReID goes beyond minimizing raw performance disparity to measure reference-relative attainment and sensitivity to weak domains.Our contributions are summarized as follows:

- We study a challenging problem: reliable fairness evaluation in federated ReID under heterogeneous domain-specific performance scales. Our focus is on addressing two limitations of conventional deviation-based metrics: **low-performance fairness illusion** and **domain difficulty confounding**, which can make similar raw performance dispersion correspond to substantially different levels of domain-wise achievement.
- We propose SDSR-HM, a reference-relative fair-achievement metric based on protocol-matched single-domain calibration and harmonic aggregation. SDSR maps each domain's federated performance to a capped achievement ratio relative to its empirical reference, while harmonic aggregation increases the influence of weak domains and prevents above-reference performance from compensating for below-reference domains.
- We conduct comprehensive experiments on four ReID domains with ResNet50 and ViT backbones, covering general federated learning, federated ReID, and fairness-oriented methods. The results show that SDSR-HM distinguishes methods that raw dispersion may tie or rank inconsistently and provides a more informative assessment of reference-relative attainment and weak-domain performance under the evaluated protocol.

2 Related Work

2.1 Performance Fairness in Federated Learning

Statistical heterogeneity can produce inconsistent client objectives and updates. General federated optimizers address this problem through proximal regularization (FedProx) (Li et al., 2020a), control variates (SCAFFOLD) (Karimireddy et al., 2020), or representation alignment (MOON) (Li et al., 2021a). Performance-fair methods further protect under-performing clients: AFL considers adverse mixtures of client distributions (Mohri et al., 2019), q-FFL reweights clients according to their losses (Li et al., 2020b), and FedFV mitigates conflicting gradients (Wang et al., 2021). FedHEAL (Chen et al., 2024) and FedPW (Guan et al., 2025) additionally address parameter conflict and aggregation bias under domain heterogeneity. These methods modify training, whereas our work studies how the resulting performance should be evaluated.

2.2 Federated Person Re-identification

Federated ReID learns a shared representation from decentralized domains with disjoint identity spaces and different camera environments. FedPav established federated-by-dataset and federated-by-camera benchmarks with partial model averaging (Zhuang et al., 2020), while FedReID learned a generalizable embedding from private multi-domain labels (Wu & Gong, 2021). Subsequent methods improve domain generalization through meta-knowledge augmentation (MEDA) (Song et al., 2024), positive-style accumulation (SSCU) (Xu et al., 2025), domain-aware suppression and aggregation (FedSupWA) (Yu et al., 2026), and robust knowledge selection (FedARKS) (Xu et al., 2026). PEFL enriches pose diversity in the federated-by-camera setting (Oh et al., 2026).

These methods primarily optimize model performance. However, ReID domains differ in illumination, occlusion, camera layout, and query–gallery construction, making their raw scores not directly comparable as fairness outcomes. Our contribution instead concerns evaluation across such heterogeneous domains rather than further modifying the optimizer.

2.3 Fairness Evaluation under Heterogeneity

Fairness criteria may conflict when group-conditioned distributions differ (Kleinberg et al., 2017; Chouldechova, 2017; Pleiss et al., 2017). STD measures raw-score dispersion, while Jain's index measures outcome equality (Jain et al., 1984); neither establishes whether similar outcomes represent sufficient achievement. We therefore introduce SDSR-HM as a protocol-dependent, reference-relative complement to mAP, Rank-1, AVG, and STD.

3 Preliminaries

3.1 Federated ReID Formulation

We consider K decentralized clients $\mathcal{C} = \{c_k\}_{k=1}^{K}$. Client k corresponds to a ReID domain with disjoint training and test splits, denoted by $\mathcal{D}_k^{\mathrm{tr}}$ and $\mathcal{D}_k^{\mathrm{te}}$. A shared ReID model f_θ is trained without exchanging raw client images. For a retrieval metric $m \in \{\mathrm{mAP}, \mathrm{R1}\}$, its test performance on client k is $p_k^{(m)}$. Metrics are processed separately; mAP and Rank-1 are never mixed into one score.

3.2 Raw Cross-Domain Dispersion

Let $\bar{p}^{(m)} = K^{-1} \sum_{k=1}^{K} p_k^{(m)}$. We report the population standard deviation

$$D_{\mathrm{STD}}^{(m)} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(p_k^{(m)} - \bar{p}^{(m)} \right)^2}, \tag{1}$$

where lower values indicate smaller cross-domain dispersion. D_{STD} is invariant to a common additive change in all domain scores and therefore does not measure their absolute adequacy.

4 Methodology

4.1 Overview

The overall framework of SDSR-HM is illustrated in Figure 2. Different from existing federated ReID evaluation protocols that directly aggregate retrieval performances across domains, SDSR-HM evaluates the reference-relative fair achievement of federated models by considering protocol-matched reference performances across heterogeneous domains. Specifically, we first construct domain-specific reference models through independent single-domain training, and use their measured performances as calibration anchors. Then, the federated model performances are transformed into a unified reference-aware achievement space through Single-Domain Supervised Reference (SDSR) calibration. Finally, the calibrated achievements are aggregated by a fairness-aware harmonic mechanism to obtain the final evaluation score. The proposed framework does not alter the federated training process and can be directly applied to different federated ReID models.

4.2 Protocol-Matched References and SDSR Calibration

Protocol-matched single-domain references. For each client domain k, we independently train a reference model using only its official training split $\mathcal{D}_k^{\mathrm{tr}}$. The purpose of this model is not to define a theoretical optimum, but to provide an empirical estimate of the retrieval performance achievable under the same experimental protocol without cross-domain federated optimization. The reference parameters are obtained as

$$\theta_k^* = \mathrm{Train}(\mathcal{D}_k^{\mathrm{tr}}; \mathcal{L}_{\mathrm{ReID}}, \mathcal{P}), \tag{2}$$

where $\mathcal{L}_{\mathrm{ReID}}$ denotes the ReID training objective and $\mathcal{P}$ specifies the fixed training and checkpoint-selection protocol. Train returns the checkpoint selected by this protocol, not a guaranteed global minimizer. Once the reference model has been trained, both the single-domain model and the evaluated federated model are tested on the same domain-specific test split $\mathcal{D}_k^{\mathrm{te}}$ under retrieval metric m. Their performances are defined jointly as

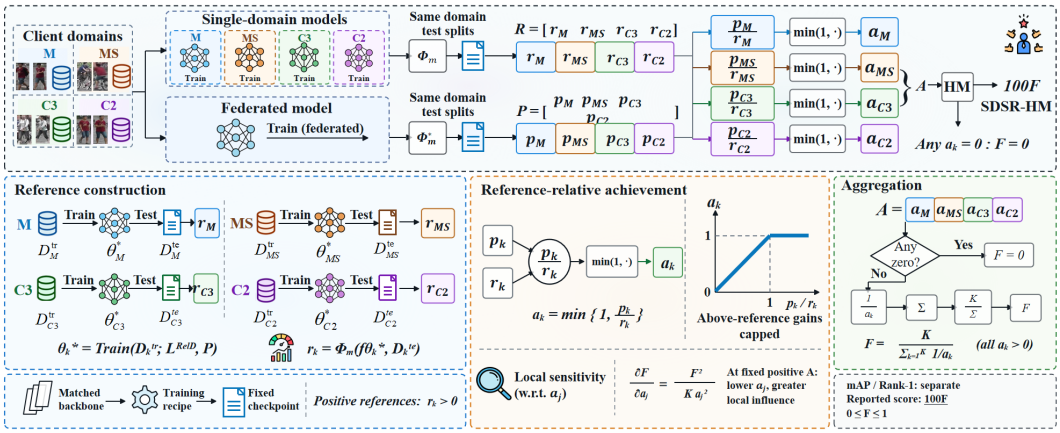


Figure 2: **Overview of the SDSR calibration and SDSR-HM aggregation pipeline**. For each domain, SDSR normalizes federated performance by a protocol-matched single-domain reference and caps the ratio at one (Eq. 4). SDSR-HM aggregates these domain achievements using the harmonic mean (Eq. 5), with the final score scaled to 0–100 (Eq. 7).

$$r_k^{(m)} = \Phi_m(f_{\theta_k^*}, \mathcal{D}_k^{\text{te}}), \qquad p_k^{(m)} = \Phi_m(f_\theta, \mathcal{D}_k^{\text{te}}), \qquad r_k^{(m)} > 0, \tag{3}$$

where $r_k^{(m)}$ denotes the single-domain reference performance and $p_k^{(m)}$ denotes the performance of the federated model on the same client domain. The condition $r_k^{(m)} > 0$ ensures that the subsequent ratio-based calibration is well defined.

For a valid comparison, the reference model and federated model must use the same backbone architecture, initialization family, preprocessing pipeline, retrieval protocol, evaluation metric, and reported training budget. Differences in these settings would change the empirical reference and make the resulting achievement ratios incomparable. Under the matched conditions, $r_k^{(m)}$ serves as a protocol-dependent calibration anchor rather than an intrinsic property or theoretical upper bound of domain k. References are constructed separately for each backbone and retrieval metric.

Single-Domain Supervised Reference calibration. Directly comparing raw retrieval scores across heterogeneous domains can be misleading because the same numerical performance may represent different levels of achievement relative to the corresponding single-domain models. SDSR therefore evaluates each client relative to its own empirical reference. For every domain, we compute the ratio between the federated performance $p_k^{(m)}$ and its matched reference $r_k^{(m)}$.

Since a domain that exceeds its reference should not compensate for another domain that remains substantially below its own reference, the ratio is capped at one. The resulting domain achievement and achievement vector are defined as

$$a_k^{(m)} = \min\left\{1, \frac{p_k^{(m)}}{r_k^{(m)}}\right\}, \qquad \mathbf{A}^{(m)} = [a_1^{(m)}, \ldots, a_K^{(m)}]. \tag{4}$$

Each $a_k^{(m)} \in [0, 1]$ represents the fraction of the protocol-matched reference attained by the federated model on domain k. A value close to one indicates that the federated model approaches or exceeds the corresponding single-domain reference, whereas a small value indicates substantial under-attainment. Consequently, two domains with different absolute retrieval scores can obtain similar achievement ratios if both approach their respective references to a comparable degree.

This reference-relative representation reduces the influence of different domain-specific performance scales and enables cross-domain comparison in a common achievement space. The ratios are calculated separately for different retrieval metrics, such as mAP and Rank-1, because these metrics have different semantics and reference values.

The upper cap in Eq. 4 is essential to the intended interpretation of SDSR-HM. Without the cap, a domain satisfying $p_k^{(m)} > r_k^{(m)}$ would obtain an achievement ratio larger than one and could numerically compensate for another domain that remains below its reference. Restricting $a_k^{(m)}$ to $[0, 1]$ prevents this compensation. Performance beyond the empirical reference is still retained in the raw retrieval results, but it does not artificially increase the final fair-achievement score.

4.3 Harmonic Aggregation and Metric Properties

After SDSR calibration, all client domains are represented in the same reference-relative achievement space. A simple arithmetic average can still understate weak-domain behavior because high-achievement domains can offset domains with substantially lower achievement. To increase the influence of these low-achievement domains, we assign equal importance to all clients and aggregate their capped achievements using the harmonic mean:

$$F_{\text{SDSR-HM}}^{(m)} = \begin{cases} 0, & \min_k a_k^{(m)} = 0, \\ \dfrac{K}{\sum_{k=1}^{K} 1/a_k^{(m)}}, & \text{otherwise.} \end{cases} \tag{5}$$

The explicit zero case avoids division by zero and preserves the intended weak-domain sensitivity: if one domain obtains no reference achievement, the overall score also becomes zero. For positive achievements, the harmonic mean is no greater than the corresponding arithmetic mean (Bullen, 2003). This bounds the harmonic score by average attainment. Its sensitivity to weak domains is characterized separately by the marginal derivative with respect to the achievement of domain j:

$$\frac{\partial F_{\text{SDSR-HM}}^{(m)}}{\partial a_j^{(m)}} = \frac{\left(F_{\text{SDSR-HM}}^{(m)}\right)^2}{K\left(a_j^{(m)}\right)^2}. \tag{6}$$

For a fixed positive achievement vector, the derivatives share the same numerator and are inversely proportional to $\left(a_j^{(m)}\right)^2$. Thus, a domain with lower achievement has greater local influence on the score. SDSR-HM therefore emphasizes bottleneck domains instead of allowing well-performing domains to dominate the evaluation. At the same time, every domain receives the same explicit weight; the difference in influence arises from harmonic aggregation rather than from a manually selected client weight.

The normalized score $F_{\text{SDSR-HM}}^{(m)}$ satisfies several useful properties. First, it is bounded in $[0, 1]$ because every capped achievement belongs to $[0, 1]$. Second, it is permutation-invariant, so changing the order of the domains does not affect the result. Third, it is nondecreasing in every $a_k^{(m)}$, meaning that improving one domain while holding all others fixed cannot reduce the score.

SDSR-HM preserves absolute reference-relative attainment under uniform performance changes. If all domains attain the same fraction of their references, then the uniform-achievement property and the reported percentage-scale score can be written jointly as

$$\begin{gathered} p_k^{(m)} = c\, r_k^{(m)} \quad (\forall k), \quad 0 \le c \le 1 \Longrightarrow a_k^{(m)} = c, \quad F_{\text{SDSR-HM}}^{(m)} = c, \\ \text{SDSR-HM}^{(m)} = 100 F_{\text{SDSR-HM}}^{(m)}. \end{gathered} \tag{7}$$

Thus, in the uniform-achievement case above, lowering c lowers SDSR-HM even though the achievement ratios remain equal. More generally, performance degradation cannot increase the score when references are fixed; the decrease is strict if the initial achievements are all positive and at least one capped achievement decreases. A zero-achievement domain produces a zero score, while the cap in Eq. 4 prevents above-reference gains from compensating for a below-reference domain. Algorithm 1 summarizes the complete reference construction, calibration, and aggregation procedure.

The resulting score jointly reflects the overall reference-achievement level and the influence of weak domains. It should therefore be interpreted as a reference-relative fair-achievement criterion rather than a parity index independent of retrieval quality.

Algorithm 1 SDSR-HM Computation for Retrieval Metric m

Input: splits $\{\mathcal{D}_k^{\mathrm{tr}}, \mathcal{D}_k^{\mathrm{te}}\}_{k=1}^{K}$, federated model f_θ, metric m, and fixed protocol $\mathcal{P}$
Output: $F_{\text{SDSR-HM}}^{(m)}$
for $k = 1, \ldots, K$ **do**
  $\theta_k^* \leftarrow \mathrm{Train}(\mathcal{D}_k^{\mathrm{tr}}; \mathcal{L}_{\mathrm{ReID}}, \mathcal{P})$ // Train reference model
  $r_k^{(m)} \leftarrow \Phi_m(f_{\theta_k^*}, \mathcal{D}_k^{\mathrm{te}})$ // Obtain reference performance
  assert $r_k^{(m)} > 0$ // Ensure valid reference
  $p_k^{(m)} \leftarrow \Phi_m(f_\theta, \mathcal{D}_k^{\mathrm{te}})$ // Evaluate federated model
  $a_k^{(m)} \leftarrow \min\{1, p_k^{(m)}/r_k^{(m)}\}$ // Compute SDSR achievement
end for
if $\min_k a_k^{(m)} = 0$ **then**
  $F_{\text{SDSR-HM}}^{(m)} \leftarrow 0$ // Handle zero achievement
else
  $F_{\text{SDSR-HM}}^{(m)} \leftarrow K \left(\sum_{k=1}^{K} 1/a_k^{(m)}\right)^{-1}$ // Harmonic aggregation
end if
return $F_{\text{SDSR-HM}}^{(m)}$

5 Experiments

5.1 Experimental Setup

Evaluation protocol. We use Market-1501 (Zheng et al., 2015), MSMT17 (Wei et al., 2018), CUHK03 (Li et al., 2014), and CUHK02 (Li & Wang, 2013) as four client domains, denoted by M, MS, C3, and C2, respectively. We evaluate representative general federated learning, federated ReID, and fairness-oriented methods under a unified source-domain protocol for consistent comparison. Methods originally designed for other federated ReID settings are adapted to this protocol instead of using their published results. We report mAP and Rank-1 on the official test split of each domain. We also evaluate FedHRA, a separately developed federated ReID training method submitted to AAAI. Observations made during its development motivated this evaluation study; FedHRA is not a training contribution of the present paper.

References and reported metrics. For each backbone, we use the seed-42 single-domain runs trained for 80 epochs and select the last checkpoint according to the protocol. In the order M, MS, C3, and C2, the ResNet50 references are $(75.29, 37.92, 53.74, 67.28)$ for mAP and $(89.49, 64.93, 56.07, 67.36)$ for Rank-1. The ViT references are $(79.65, 46.10, 59.60, 74.72)$ and $(90.62, 69.60, 61.86, 76.78)$, respectively. These references are displayed to two decimal places; reference-based calculations use the unrounded values. AVG and STD denote the unweighted mean and population standard deviation of the four domain scores. SDSR-HM is computed for mAP and Rank-1 using Eqs. 4–7. All aggregate values in Tables 1 and 2 are computed from the displayed domain results, using the unrounded references where required.

5.2 Experimental Results

Table 1 compares federated ReID methods from both retrieval and fairness perspectives. A clear discrepancy can be observed between raw cross-domain dispersion and reference-relative fair achievement. Under ResNet50, SCAFFOLD obtains the lowest mAP STD of 14.6, even smaller than FedHRA's 18.7. However, SCAFFOLD achieves only 32.6 average mAP and 49.1 SDSR-HM, whereas FedHRA reaches 60.1 average mAP and 95.0 SDSR-HM. This substantial gap shows that a small STD can result from uniformly weak performance rather than high reference-relative attainment. An even stronger example appears for Rank-1: q-FFL and FedHRA have the same reported STD of 15.0 to one decimal place, but their SDSR-HM scores differ markedly, from 69.1 for q-FFL to 95.8 for FedHRA. Therefore, similar raw dispersion can correspond to very different levels of fair achievement once domain-specific reference performance is taken into account.

The same phenomenon consistently appears with the ViT backbone. FedPav achieves a smaller mAP STD than FedHRA (15.4 vs. 16.6), yet its SDSR-HM is only 66.0 compared with 96.9 for FedHRA, together with a much lower average mAP of 45.2 versus 66.3. For Rank-1, FedReID obtains the smallest STD of 12.5, slightly below FedHRA's 13.0, but its SDSR-HM remains substantially lower

Table 1: **Source-domain federated ReID comparison with ResNet50 and ViT.** M, MS, C3, and C2 denote Market-1501, MSMT17, CUHK03, and CUHK02; R1 denotes Rank-1 accuracy. AVG and STD report the unweighted mean and population standard deviation across domains. SRM abbreviates SDSR-HM (Eq. 5), computed by normalizing domain scores against backbone-matched single-domain references, capping ratios at one, and taking their harmonic mean on a 0–100 scale.

Category	Methods	Venue	M		MS		C3		C2		mAP			Rank-1		
			mAP	R1	mAP	R1	mAP	R1	mAP	R1	AVG ↑	STD ↓	**SRM ↑**	AVG ↑	STD ↓	**SRM ↑**
Backbone: ResNet50																
Federated Learning	SCAFFOLD	ICML'20	32.4	57.6	24.1	46.4	17.6	17.8	56.2	55.7	32.6	**14.6**	49.1	44.4	15.9	54.7
	MOON	CVPR'21	34.0	58.4	24.3	48.0	18.1	17.4	58.9	56.5	33.8	15.5	50.7	45.1	16.5	54.8
	FedProx	MLSys'20	30.4	56.8	24.7	48.8	18.0	17.5	58.5	57.2	32.9	15.4	49.1	45.1	16.3	54.9
Fed ReID	FedPav	MM'20	30.7	58.4	17.3	43.3	34.6	27.1	71.6	73.8	38.6	20.1	55.6	50.7	17.4	65.6
	FedReID	AAAI'21	61.9	82.2	30.9	60.1	37.7	38.1	76.6	76.4	51.8	18.4	82.2	64.2	17.1	86.2
	MEDA	ICASSP'24	65.4	82.9	23.2	46.7	49.0	50.6	81.9	79.9	54.9	21.7	81.9	65.0	16.5	87.4
	PEFL	CVPR'26	66.9	85.0	25.4	50.4	49.2	51.1	81.0	81.7	55.6	20.8	84.9	67.1	16.3	90.1
	SSCU	MM'25	57.9	79.1	23.8	51.8	38.6	42.2	75.7	77.5	49.0	19.6	75.7	62.7	16.0	84.9
	FedSupWA	AAAI'26	59.0	79.9	24.6	52.8	39.8	43.6	76.8	79.0	50.1	19.7	77.4	63.8	16.0	86.3
	FedARKS	AAAI'26	59.8	80.4	25.2	53.5	40.9	44.7	77.9	79.9	51.0	19.8	78.8	64.6	15.8	87.3
Fair Federated Learning	q-FFL	ICLR'20	56.4	67.3	22.5	47.0	25.4	28.4	62.8	61.2	41.8	18.0	64.4	51.0	15.0	69.1
	FedFV	IJCAI'21	65.2	83.5	30.6	53.5	50.3	54.0	79.1	81.6	56.3	18.0	89.6	68.2	**14.4**	92.5
	FedHEAL	CVPR'24	67.8	84.8	22.0	45.0	37.5	39.2	74.7	74.8	50.5	21.6	75.9	61.0	19.3	81.2
	FedPW	NeurIPS'25	64.3	76.3	28.6	54.2	32.4	32.9	72.9	72.6	49.6	19.3	77.6	59.0	17.2	78.8
Fair Federated ReID	**FedHRA**	**Submitted**	**70.2**	**87.3**	**34.7**	**59.1**	**51.4**	**53.4**	**84.2**	**84.5**	**60.1**	18.7	**95.0**	**71.1**	15.0	**95.8**
Backbone: ViT																
Fed ReID	FedPav	MM'20	47.8	70.0	27.1	53.9	37.1	36.2	68.7	66.1	45.2	**15.4**	66.0	56.6	13.2	73.3
	FedReID	AAAI'21	64.6	79.2	38.4	64.2	43.2	46.1	76.8	73.2	55.8	15.7	83.1	65.7	**12.5**	86.6
	MEDA	ICASSP'24	70.2	87.8	27.1	53.0	51.3	54.1	84.6	85.6	58.3	21.5	80.0	70.1	16.6	89.1
	PEFL	CVPR'26	68.4	86.6	26.6	52.7	50.5	52.5	85.3	86.2	57.7	21.8	78.8	69.5	16.9	88.0
	SSCU	MM'25	65.9	84.1	30.8	57.8	47.6	49.2	83.2	83.0	56.9	19.6	80.7	68.5	15.3	88.1
	FedSupWA	AAAI'26	67.0	84.9	31.6	58.8	48.8	50.6	84.3	84.5	57.9	19.7	82.2	69.7	15.3	89.4
	FedARKS	AAAI'26	67.8	85.4	32.2	59.5	49.9	51.7	85.4	85.4	58.8	19.8	83.3	70.5	15.2	90.3
Fair Federated Learning	q-FFL	ICLR'20	59.7	77.4	31.2	55.0	38.0	43.6	71.4	72.7	50.1	16.2	73.7	62.2	13.6	81.4
	FedFV	IJCAI'21	70.0	84.4	34.2	58.8	48.9	48.7	83.4	82.3	59.1	18.9	85.0	68.6	15.2	88.3
	FedHEAL	CVPR'24	65.5	83.3	37.3	62.1	39.5	39.1	76.3	76.4	54.7	16.7	80.6	65.2	16.9	83.4
	FedPW	NeurIPS'25	66.5	84.7	35.6	60.4	38.6	38.5	75.8	73.7	54.1	17.4	79.4	64.3	17.2	82.1
Fair Federated ReID	**FedHRA**	**Submitted**	**75.8**	**89.7**	**44.4**	**68.4**	**57.4**	**58.1**	**87.6**	**86.4**	**66.3**	16.6	**96.9**	**75.7**	13.0	**97.7**

(86.6 vs. 97.7). These comparisons demonstrate that STD mainly reflects the similarity of raw client performances, while SDSR-HM further evaluates whether each domain attains its own reference performance, with greater sensitivity to weak domains. Across both backbones and both retrieval metrics, FedHRA achieves the highest SDSR-HM scores, indicating the highest reference-relative fair achievement under this criterion among the compared methods.

5.3 Complementary Diagnostics

Analytic behavior. Three cases isolate the behavior of the normalized SDSR-HM score. If $\mathbf{A} = (1, 1, 1, 1)$, then $Q = J = F_{\text{SDSR-HM}} = 1$. If $\mathbf{A} = (c, c, c, c)$ for $0 < c \leq 1$, then $J = 1$ for every c, whereas $Q = F_{\text{SDSR-HM}} = c$. Thus, a scale-invariant parity index cannot detect uniform degradation, while the normalized score decreases with c in this uniform-achievement case. Finally, if $\mathbf{A} = (1, 1, 1, 0.25)$, then $Q = 0.8125$, $J \approx 0.862$, and $F_{\text{SDSR-HM}} = 4/(1 + 1 + 1 + 4) = 4/7 \approx 0.571$. The reciprocal aggregation therefore makes one severely under-attained domain visible even when the arithmetic mean and parity index remain comparatively high.

Attainment and parity. We complement SDSR-HM with the mean capped achievement Q and Jain's parity index J (Jain et al., 1984):

$$Q^{(m)} = \frac{1}{K} \sum_{k=1}^{K} a_k^{(m)}, \qquad J^{(m)} = \frac{\left(\sum_{k=1}^{K} a_k^{(m)} \right)^2}{K \sum_{k=1}^{K} \left(a_k^{(m)} \right)^2}. \tag{8}$$

Here, Q measures average reference-relative attainment, whereas J measures parity and is defined only when $\sum_k a_k^{(m)} > 0$. For $\mathbf{A} = (c, c, c, c)$ with $0 < c \leq 1$, $J = 1$ regardless of c, while $Q = F_{\text{SDSR-HM}} = c$; thus, parity alone cannot detect uniform under-attainment. For $\mathbf{A} = (1, 1, 1, 0.25)$, $Q = 0.8125$ and $J \approx 0.862$, but $F_{\text{SDSR-HM}} = 4/7 \approx 0.571$, making the weak domain more visible.

Table 2: **Complementary mAP diagnostics.** Q, J, and SDSR-HM are reported on a 0–100 scale, using tabulated domain scores and unrounded references.

Method	ResNet50			ViT		
	$Q\uparrow$	$J\uparrow$	SDSR-HM↑	$Q\uparrow$	$J\uparrow$	SDSR-HM↑
FedPav	62.7	87.9	55.6	68.2	96.1	66.0
q-FFL	68.7	94.1	64.4	75.5	97.4	73.7
FedFV	90.2	99.4	89.6	86.0	98.8	85.0
FedHRA	**95.1**	**99.9**	**95.0**	**97.0**	**100.0**	**96.9**

Table 2 confirms this distinction. On ViT, q-FFL and FedHRA have similarly high parity ($100J = 97.4$ and 100.0), yet their SDSR-HM scores differ by 23.2 points (73.7 versus 96.9). On ResNet50, the gap between $100Q$ and SDSR-HM decreases from 7.1 points for FedPav to only 0.1 for FedHRA. Thus, SDSR-HM remains close to average attainment when domains are balanced, but penalizes weak domains that Q or J alone may obscure.

Figure 3a examines single-domain degradation for FedPav and FedHRA. Only M or MS is degraded, while other domain scores and references remain fixed throughout all four constructed degradation intervals. Colors denote relative mAP reduction intervals; boxes summarize constructed perturbations, not repeated training runs. For FedHRA on M, median STD initially decreases from 18.32 to 17.85, despite worsening retrieval performance, whereas median SDSR-HM decreases from 93.74 to 90.91. Across conditions, SDSR-HM decreases with greater degradation, while STD can respond nonmonotonically. Thus, smaller STD need not indicate better reference-relative attainment. Panel-specific axis ranges require numerical comparisons across methods.

Figure 3b shows that SDSR-HM decreases consistently as the constructed heterogeneity becomes stronger. Starting from 95.0, it falls to 59.9, 56.0, and 51.3 under the strongest identity-skew, data-imbalance, and camera-shift settings, respectively. The corresponding F_{STD} values decrease from 81.3 to 70.9, 72.5, and 74.6. This behavior is consistent with SDSR-HM's emphasis on low achievement ratios across all three shifts. Since the curves use inferred degradation factors, they demonstrate metric sensitivity rather than empirical robustness.

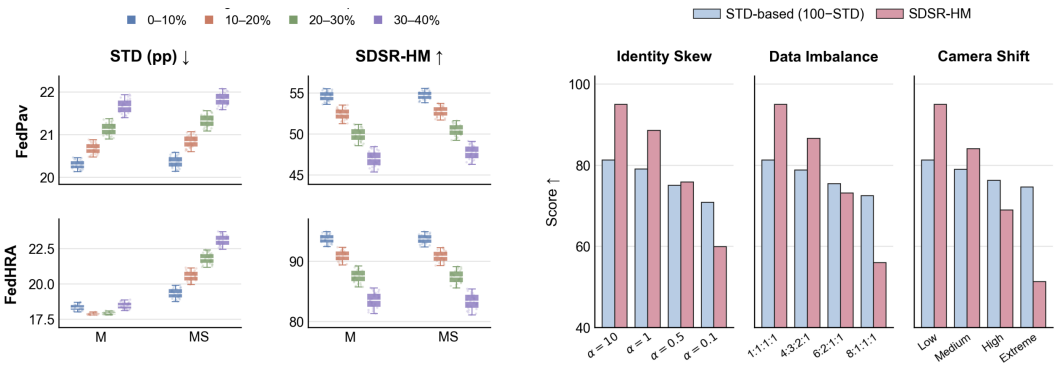


Figure 3: **Complementary diagnostics of SDSR-HM**: (a) controlled single-domain degradation and (b) sensitivity to constructed heterogeneity.

6 Conclusion

In this paper, we revisit fairness evaluation in fair federated ReID and propose SDSR-HM, a reference-aware fairness metric based on **S**ingle-**D**omain **S**upervised **R**eference (**SDSR**) calibration. SDSR measures each domain by its capped reference-relative achievement, while harmonic aggregation emphasizes weak-achievement domains and prevents above-reference gains from compensating for disadvantaged clients. Experiments across four ReID domains with ResNet50 and ViT demonstrate that SDSR-HM provides a more informative fairness assessment than raw performance dispersion. We hope this work offers a new perspective for fairness evaluation in federated ReID.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to assist with language editing, LaTeX formatting, citation style consistency, and the clarity and presentation of the metric formulation. In particular, these tools were used to improve grammar, readability, and organization, as well as to suggest clearer ways of expressing selected technical descriptions and mathematical notation. AI-generated outputs were treated only as editorial and drafting suggestions and were not accepted without review.

The authors carefully examined, revised, and verified all AI-assisted material. Responsibility for the study design, analytical reasoning, mathematical derivations, interpretation of results, factual claims, citation accuracy, and final presentation remains entirely with the authors. The authors also take full responsibility for any errors or omissions in the final manuscript and associated artifacts.

REFERENCES

Peter S. Bullen. *Handbook of Means and Their Inequalities*, volume 560 of *Mathematics and Its Applications*. Springer, 2 edition, 2003. doi: 10.1007/978-94-017-0399-4.

Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12077–12086, 2024. doi: 10.1109/CVPR52733.2024.01148.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16145–16154, 2021.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7494–7502, 2023. doi: 10.1609/aaai.v37i6.25911.

Kaiqi Guan, Wenke Huang, Xianda Guo, Yueyang Yuan, Bin Yang, and Mang Ye. Rethinking fair federated learning from parameter and client view. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Raj Jain, Dah-Ming Chiu, and William R. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report DEC-TR-301, Digital Equipment Corporation, 1984.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021a. doi: 10.1109/CVPR46437.2021.01057.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020a.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR, 2021b.

Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3594–3601, 2013. doi: 10.1109/CVPR.2013.461. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Li_Locally_Aligned_Feature_2013_CVPR_paper.html.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159, 2014.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

JooHyung Oh, Minyoung Oh, Sung Whan Yoon, and Jae-Young Sim. Pose-guided enriched feature learning for federated-by-camera person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40458–40467, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Oh_Pose-guided_Enriched_Feature_Learning_for_Federated-by-camera_Person_Re-identification_CVPR_2026_paper.html.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chunli Song, Xiaohua Chen, Wenqiu Zhu, Yucan Zhou, Xiaoyan Gu, and Bo Li. Meta-knowledge enhanced data augmentation for federated person re-identification. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8901–8905. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447543.

Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 719–728, 2019.

Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1615–1623, 2021. doi: 10.24963/ijcai.2021/223.

Long Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 79–88, 2018.

Guile Wu and Shaogang Gong. Decentralised learning from independent multi-domain labels for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2898–2906, 2021.

Xin Xu, Chaoyue Ren, Wei Liu, Wenke Huang, Bin Yang, Zhixi Yu, and Kui Jiang. Positive style accumulation: A style screening and continuous utilization framework for federated DG-ReID. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8527–8536, 2025. doi: 10.1145/3746027.3755549.

Xin Xu, Binchang Ma, Zhixi Yu, and Wei Liu. FedARKS: Federated aggregation via robust and discriminative knowledge selection and integration for person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11424–11432, 2026. doi: 10.1609/aaai.v40i14.38124.

Zhixi Yu, Wei Liu, Wenke Huang, Bin Yang, Qian Bie, Guancheng Wan, and Xin Xu. Domain-aware suppression and aggregation for federated DG ReID. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):12231–12239, 2026. doi: 10.1609/aaai.v40i14.38214.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 955–963, 2020.