CO-EVO: Co-evolving Semantic Anchoring and Style Diversification for Federated DG-ReID

Fengchun Zhang[1], **Qiang Ma**[2], **Liuyu Xiang**[3], **Jinshan Lai**[1], **Tingxuan Huang**[4], **Jianwei Hu**[2,*]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China
[2]QiYuan Lab
[3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[4]School of Software, Tsinghua University

Abstract

Federated domain generalization for person re-identification (FedDG-ReID) aims to collaboratively train a pedestrian retrieval model across multiple decentralized source domains such that it can generalize to unseen target environments without compromising raw data privacy. However, this task is significantly challenged by the inherent stylistic gaps across decentralized clients. Without global supervision, models easily succumb to shortcut learning where representations overfit to domain specific camera biases rather than universal identity features. We propose CO-EVO, a novel federated framework that resolves this semantic-style conflict through a co-evolutionary mechanism. On the semantic side, Camera-Invariant Semantic Anchoring (CSA) learns identity prompts with cross-camera consistency to establish purified and domain-agnostic anchors that filter out local imaging noise. On the visual side, Global Style Diversification (GSD), powered by a Global Camera-Style Bank (GCSB), synthesizes realistic perturbations to expand the visual boundaries of training data. The core of CO-EVO is its co-evolutionary loop where purified anchors act as gravitational centers to guide the image encoder toward robust anatomical attributes amidst diverse style variations. Extensive experiments demonstrate that CO-EVO achieves state-of-the-art (SOTA) performance, proving that the synergy between semantic purification and style expansion is essential for robust cross-domain generalization. Our code is available at: https://github.com/NanYiyuzurn/ACL-LGPS-2026.

1 Introduction

Person Re-identification (ReID) is a pivotal technology in modern surveillance for cross-camera pedestrian tracking and public safety (Luo et al., 2019; Wang et al., 2022a; Gao et al., 2020, 2022). However, ReID models often face severe domain shifts in practical deployment due to heterogeneous camera characteristics and lighting conditions (Wang et al., 2022b; Ye et al., 2021). To address these challenges, domain generalization (DG) for ReID aims to learn a robust model from multiple sources that can generalize to unseen targets (Choi et al., 2021; Dai et al., 2021; Jin et al., 2020; Zhou et al., 2021; Nuriel et al., 2021). While traditional DG requires centralized data access, the demand for privacy has shifted focus toward Federated Learning (FL) as a paradigm for multi-domain collaboration without sharing raw data (McMahan et al., 2017). Consequently, Federated Domain Generalization for ReID (FedDG-ReID) has emerged to learn retrieval models from decentralized sources. Despite its potential, FedDG-ReID remains challenging due to open-set nature and client heterogeneity. Standard model aggregation often fails to achieve robust generalization as local models tend to internalize domain-specific biases.

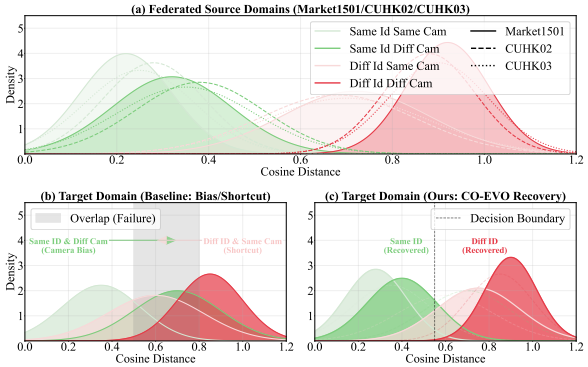

Figure 1: Cosine distance distributions illustrating the motivation of CO-EVO. (a) Source Training: Model learns identity discrimination under consistent source distributions. (b) Baseline Failure: On unseen target domains, camera bias and shortcut learning lead to distribution overlap. (c) CO-EVO Recovery: By coupling stable CSA with GSD, our framework restores the decision boundary.

The fundamental bottleneck of existing FedDG-

*Corresponding author: hjw17@tsinghua.org.cn.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32348–32359
July 2-7, 2026

ReID methods lies in a semantic-style conflict. During local optimization, the lack of a global semantic reference often leads the model toward shortcut learning. As illustrated in Figure 1(b), the network tends to exploit superficial cues such as background textures and camera related color footprints for identity discrimination. Although these cues are stable within a single client, they fail to generalize across the federation. This causes same-identity pairs from different cameras to drift apart while pulling different-identity pairs closer under similar imaging conditions. While vision-language models like CLIP show promise in anchoring visual features to stable semantic spaces, their application in FedDG-ReID is hindered by the lack of natural language names for ReID labels and high communication costs.

In parallel, local data diversification via style transfer has been adopted to mimic unseen domain shifts (Yang et al., 2024; Huang et al., 2023). However, existing strategies face a trade-off between privacy and efficiency, often relying on costly learning-based generators that scale poorly with the number of clients (Zhuang et al., 2020; Yan et al., 2020). These observations raise a fundamental question: How can we achieve robust FedDG-ReID by harmonizing stable semantic grounding with efficient style diversification?

To answer this, we propose CO-EVO, a framework for Co-evolving Semantic Anchoring and Style Diversification. Here, "co-evolving" does not mean that semantic anchors and style templates are both updated symmetrically at every step. Instead, it denotes a coupled training mechanism in which style diversification continuously expands the visual inputs seen by the encoder, while semantic anchoring continuously constrains those updates with stable identity-level targets. As shown in Figure 1(c), CO-EVO rectifies the distribution overlap by resolving the semantic-style conflict. First, we propose Camera-Invariant Semantic Anchoring (CSA), which equips each identity with learnable tokens to form textual descriptions. Unlike previous methods, CSA introduces cross-camera consistency to distill identity-specific features from local camera noise. By caching these as frozen identity-level textual prototypes, we provide stable and purified semantic anchors that prevent the model from drifting amid visual variations. Second, we introduce Global Style Diversification (GSD) powered by a lightweight Global Camera-Style Bank (GCSB). GCSB aggregates camera statistics from all clients to generate diverse and realistic perturbations without the need for expensive generators.

The core of CO-EVO lies in this coupled optimization loop, where purified semantic anchors act as gravitational centers. These anchors guide the image encoder to focus on robust anatomical attributes even as the input visuals undergo extreme style perturbations synthesized by GSD. Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to introduce language-guided semantic supervision into the FedDG-ReID task; our framework resolves the shortcut learning problem through the proposed CSA and its distilled, camera-invariant textual anchors.
- We propose a GSD mechanism utilizing a global camera-style bank, enabling efficient visual diversification to simulate unseen domain shifts without violating decentralization constraints. The bank is built once with negligible overhead and remains effective even when metadata are noisy or missing.
- We identify and resolve the semantic-style conflict through a coupled semantic–style optimization mechanism. Extensive experiments on multiple benchmarks demonstrate that CO-EVO achieves state-of-the-art (SOTA) performance and significantly enhances cross-domain generalization.

2 Related Work

2.1 Domain Generalization for Federated Person Re-ID

Domain Generalization (DG) for Re-ID aims to extract domain-invariant representations (Balaji et al., 2018; Liang et al., 2025). Previous studies have explored data diversification via style perturbations to mitigate domain shifts (Kang et al., 2022; Zhou et al., 2021). In the federated learning (FL) context, existing methods such as DACS (Yang et al., 2024) and SSCU (Xu et al., 2025) utilize a Style Transformation Model (STM) to achieve local diversification. However, these STM-based approaches require training an auxiliary generative network, which is computationally cumbersome and prone to instability during decentralized optimization. As visualized in Figure 3(c), STM-generated images frequently suffer from repetitive artifacts and unrealistic exposure. In contrast,

our CO-EVO addresses these limitations by introducing a lightweight diversification mechanism grounded in a global camera-style bank. Instead of relying on expensive generative models, we utilize template-based re-normalization of real-world style statistics to provide authentic variations with negligible overhead.

2.2 Vision–Language Learning for Re-ID

Vision–language models like CLIP (Radford et al., 2021) offer powerful semantic priors through contrastive pre-training. Since Re-ID datasets lack natural language descriptions, methods like CLIP-ReID (Li et al., 2023) and TF-CLIP (Yu et al., 2024) learn identity-specific prompts for semantic supervision. In federated settings, DiPrompT (Bai et al., 2024) explored disentangled prompt tuning for general DG tasks. However, the synergy between vision-language semantics and federated stylization remains underexplored. Existing VLM-based methods often fail to resolve the semantic-style conflict, where aggressive stylization distorts semantic grounding. Our CO-EVO bridges this gap by establishing a coupled training loop between stable semantic anchoring and dynamic style perturbations. By decoupling the learning process, we ensure that the model consistently aligns visual features with purified, camera-invariant semantic references, even under extreme input variations.

3 Methodology

3.1 Design Rationale: The Synergy of Semantics and Style

The core challenge in FedDG-ReID is the semantic-style conflict: purely visual supervision often succumbs to domain-specific shortcuts, while aggressive visual augmentation can corrupt identity-sensitive cues if not properly grounded. To resolve this, we propose a coupled interaction between stable semantic anchoring and dynamic style diversification. In our terminology, "co-evolution" specifically means that the visual distribution evolves through GSD while the encoder is repeatedly pulled back to fixed CSA anchors; it does not require simultaneous parameter updates for both branches. This mechanism compels the image encoder to map highly perturbed visual inputs back to a unified, domain-agnostic latent space.

As illustrated in Figure 2, we instantiate this rationale through two collaborative components: Camera-Invariant Semantic Anchoring (CSA) and Global Style Diversification (GSD). GSD forces the model to explore stylistic boundaries by synthesizing diverse camera effects grounded in global statistics, while CSA ensures that the learned representations remain anchored to intrinsic identity semantics, effectively neutralizing the impact of camera-related noise.

In our federated scenario with K clients, the non-overlapping identity sets necessitate a stable global reference to bridge domain gaps. We decouple the learning process into a stable anchoring phase and a dynamic diversification phase. By caching purified identity-level textual prototypes locally, we prevent semantic drift caused by biased local updates and eliminate the prohibitive communication cost of transmitting large-scale language models. These cached anchors serve as constant gravitational centers throughout the federated loop. As revealed in Figure 1, this coupled optimization prevents the collapse of identity distributions on unseen domains by penalizing camera-specific shortcuts while rewarding semantic-level consistency.

3.2 Phase I: Camera-invariant Semantic Anchoring (CSA)

Most FedDG-ReID methods suffer from shortcut learning due to the lack of explicit domain-agnostic guidance. CSA resolves this by learning camera-invariant identity prompts that serve as purified semantic anchors. Unlike standard visual-language alignment, CSA explicitly distills identity-specific features from local camera noise, ensuring the resulting textual prototypes are robust across heterogeneous environments.

For each identity y on client k, we introduce L learnable tokens $\{[X_\ell^y]\}_{\ell=1}^L$ inserted into a template: "a photo of a $[X_1^y]$...$[X_L^y]$ person". During this phase, we freeze the CLIP encoders, optimizing only the tokens.

Given a mini-batch $\{(x_i, y_i, c_i)\}_{i=1}^B$, we adopt a bidirectional contrastive loss to align visual features v_i with their corresponding textual prototypes t_{y_i}:

$$L_{i2t}(i) = -\log \frac{\exp(s(v_i, t_{y_i})/\tau)}{\sum_{a=1}^{B} \exp(s(v_i, t_a)/\tau)}, \quad (1)$$

$$L_{t2i}(y) = -\frac{1}{|P(y)|} \sum_{p \in P(y)} \log \frac{\exp(s(v_p, t_y)/\tau)}{\sum_{a=1}^{B} \exp(s(v_a, t_y)/\tau)}, \quad (2)$$

where $P(y)$ denotes the set of sample indices with identity y, and $s(v,t) = \frac{v^\top t}{\|v\|\|t\|}$ denotes cosine similarity. To further distill the semantic information

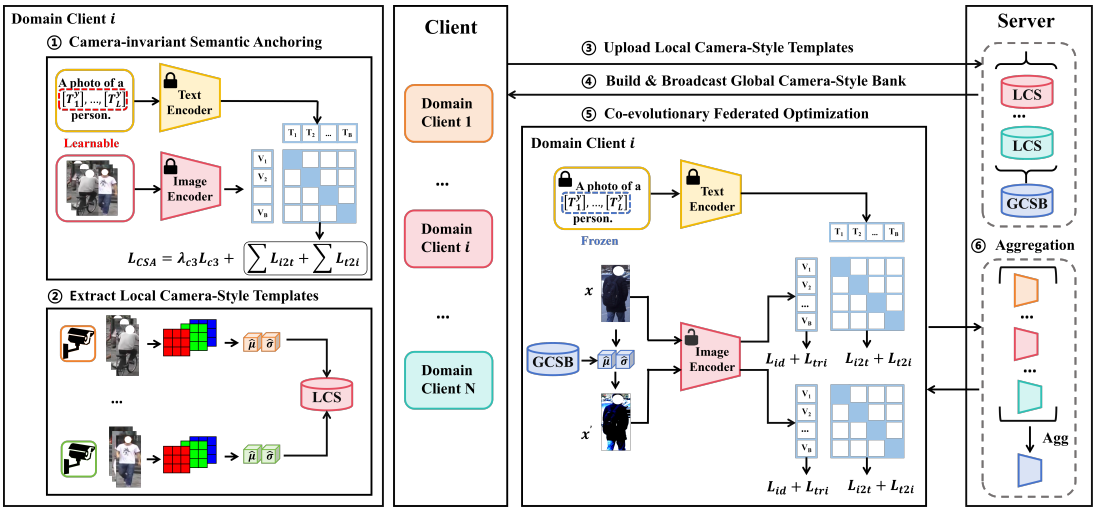

Figure 2: The overall architecture of CO-EVO for FedDG ReID. The coupled semantic–style procedure follows six key steps: ①–③ Clients establish purified semantic anchors via CSA and extract local camera-style templates; ④ The server constructs the GCSB to broadcast; ⑤ Clients perform local optimization by aligning both original and stylized views (x, x') with the fixed anchors; ⑥ The server aggregates local models to refine the global encoder.

from camera-related noise, we introduce a Cross-Camera Consistency (L_{c3}) regularization:

$$L_{c3} = \sum_{y \in Y_{batch}} \sum_{i,j \in P(y), c_i \neq c_j} \|s(v_i, t_y) - s(v_j, t_y)\|^2, \tag{3}$$

where c_i denotes the camera ID. This constraint compels the learnable tokens to ignore camera-specific visual shortcuts and focus on invariant pedestrian attributes. The overall CSA loss is $L_{CSA} = \sum L_{i2t} + \sum L_{t2i} + \lambda_{c3} L_{c3}$. After local optimization, we cache the resulting identity-level textual prototypes $T_k = \{t_y\}_{y \in Y_k}$ as purified semantic anchors for the federated loop:

$$T_k = \{t_y\}_{y \in Y_k} \in R^{|Y_k| \times D}. \tag{4}$$

3.3 Phase II: Global Style Diversification (GSD)

While CSA provides purified and camera-invariant semantic anchors, the image encoder must still encounter a vast spectrum of visual variations to achieve robust domain invariance. To this end, we propose Global Style Diversification (GSD), a lightweight mechanism to synthesize realistic cross-client domain shifts.

Central to GSD is the Global Camera-Style Bank (GCSB), which serves as a repository of camera-specific style statistics. We use channel-wise mean and variance because prior DG studies have shown that these first- and second-order feature statistics capture domain-specific appearance factors such as illumination, color tone, and texture while largely preserving semantic structure (Zhou et al., 2021; Tang et al., 2021). For each client k and camera $c \in C_k$, we extract a camera-style template by computing channel-wise statistics:

$$(\mu_{k,c}, \sigma^2_{k,c}) = Stat(\{x_i^k \mid c_i^k = c\}). \tag{5}$$

The server aggregates these templates into a global repository $\mathcal{B} = \cup_{k=1}^{K} \cup_{c \in C_k} \{(\mu_{k,c}, \sigma^2_{k,c})\}$. When camera IDs are unreliable or unavailable, we can replace c with pseudo-groups obtained from unsupervised clustering, allowing GSD to remain applicable without changing the training objective. During local optimization, we inject diverse camera effects via template-based re-normalization:

$$\hat{x} = \frac{x - \mu(x)}{\sqrt{\sigma^2(x) + \epsilon}}, \qquad x' = \hat{x} \odot \sqrt{\sigma_s^2} + \mu_s, \tag{6}$$

where $(\mu_s, \sigma_s^2) \sim \mathcal{B}$. This ensures that the augmented view x' is grounded in real camera distributions rather than arbitrary noise. In practice, the GCSB is constructed only once before federated optimization, takes about 4s per client on average, introduces no additional trainable parameters, and accounts for less than 0.1% of the full training time. We emphasize that GSD mainly targets photometric variation rather than explicit geometric transformation. Robustness to viewpoint and scale changes instead comes from the CLIP backbone's transferable priors and the cross-camera semantic constraint imposed by CSA.

Algorithm 1: CO-EVO: Coupled Federated Learning with Stable Anchors and Style Diversification

Input: Clients $\{D_k\}_{k=1}^K$; rounds R; local epochs E; weight λ_{c3}; weight λ.
Output: Global image encoder θ^R.

Phase I: Camera-Invariant Semantic Anchoring (CSA)
for $k = 1, \ldots, K$ *in parallel* **do**
  Optimize local identity tokens via L_{CSA} (Eq. 1-3);
  Cache purified prototypes $T_k = \{t_y\}_{y \in Y_k}$ (Eq. 4);
  Upload style templates $\{(\mu, \sigma^2)\}$ (Eq. 5) to server;
Phase II & III: Coupled Training Loop
Server constructs GCSB $\mathcal{B}$ and broadcasts $\mathcal{B}, \theta^0$ to all clients;
for $r = 1, \ldots, R$ **do**
  for $k = 1, \ldots, K$ *in parallel* **do**
    Set local model $\theta_k \leftarrow \theta^{r-1}$;
    for $e = 1, \ldots, E$ **do**
      Sample $(x, y) \sim D_k$ and synthesize $x' \sim GSD(\mathcal{B})$ (Eq. 6);
      Compute L_{loc} (Eq. 10) for original and stylized views;
      Update θ_k via backpropagation;
    Upload θ_k^r to server;
  $\theta^r \leftarrow \sum_{k=1}^K \frac{n_k}{N} \theta_k^r$ and broadcast θ^r to all clients;
return θ^R;

3.4 Coupled Federated Optimization

The core of CO-EVO lies in the joint optimization of semantic stability (via CSA) and visual diversity (via GSD) within the federated loop. During local training on client k, we sample a mini-batch $\{(x_i, y_i)\}$ and synthesize stylized counterparts x_i' using templates from the GCSB. This is the operational meaning of our "co-evolution" terminology: the input distribution evolves through sampled style templates, while the encoder parameters evolve under constant semantic anchors. To ensure discriminability, we apply identity loss L_{id} and triplet loss L_{tri} to both original (x) and stylized (x') views:

$$L_{id}(\tilde{x}) = -\frac{1}{B} \sum_{i=1}^{B} \log p_{i,y_i}(\tilde{x}), \tag{7}$$

$$L_{tri}(\tilde{x}) = \frac{1}{B} \sum_{i=1}^{B} \max(d_p^{(i)}(\tilde{x}) - d_n^{(i)}(\tilde{x}) + \alpha, 0), \tag{8}$$

where $\tilde{x} \in \{x, x'\}$. To prevent the model from exploiting domain-specific shortcuts, we utilize the purified textual prototypes T_k as fixed anchors. For each view $\tilde{x}$, the semantic alignment loss is defined as:

$$L_{align}(i; \tilde{x}) = -\log \frac{\exp(s(v_i(\tilde{x}), t_{y_i})/\tau)}{\sum_{y \in Y_k} \exp(s(v_i(\tilde{x}), t_y)/\tau)}. \tag{9}$$

By forcing both x and x' to align with the same camera-invariant anchor t_{y_i}, the image encoder is compelled to discard low-level stylistic noise. The total local objective is:

$$L_{loc} = \sum_{\tilde{x} \in \{x, x'\}} (L_{id}(\tilde{x}) + L_{tri}(\tilde{x}) + \lambda L_{align}(\tilde{x})). \tag{10}$$

This coupled process ensures that while GSD expands visual boundaries, the CSA anchors provide a consistent gravitational center that restores the decision boundary, as visualized in Figure 1. After E local epochs, the server performs weighted aggregation to update the global model $\theta^r = \sum_{k=1}^K \frac{n_k}{N} \theta_k^r$ and synchronizes the GCSB to incorporate evolving camera statistics. The complete procedure is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

In this section, we present our experimental setup, focusing on the datasets and representative baselines used for evaluation. Other detailed configurations, including the federated setup, backbone architectures, evaluation protocols, and specific implementation details, are provided in the Appendix.

Datasets. We conduct experiments on four large-scale person ReID benchmarks: CUHK02 (Li and Wang, 2013), CUHK03 (Li et al., 2014), MSMT17 (Wei et al., 2017), and Market1501 (Zheng et al., 2015). For clarity, these domains are denoted as C2, C3, MS, and M, respectively.

Baselines. We evaluate CO-EVO against representative methods from four categories: (i) Generic federated optimization algorithms including SCAFFOLD (Karimireddy et al., 2020), MOON (Li et al., 2021), and FedProx (Li et al., 2020), which address client drift and heterogeneity. (ii) Style-based domain generalization techniques such as MixStyle (Zhou et al., 2021) and CrossStyle (Tang et al., 2021) that diversify distributions via style statistics. (iii) Federated ReID frameworks like FedPav (Zhuang et al., 2020) and FedReID (Wu and Gong, 2021), which are tailored for open-set retrieval. (iv) Specialized DG-ReID and FedDG-ReID methods, including SNR (Jin et al., 2020), DACS (Yang

Table 1: Protocol I (leave-one-domain-out) results on FedDG-ReID. Each source domain is treated as a client. We report mAP and Rank-1 (%) on three held-out target domains and the average.

Category	Methods	Reference	MS+C2+C3→M		M+C2+C3→MS		MS+C2+M→C3		Average	
			mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
Federated Learning	SCAFFOLD	ICML 2020	26.0	50.5	5.3	15.8	22.9	26.0	18.1	30.8
	MOON	CVPR 2021	26.8	51.1	4.8	14.5	20.9	22.5	17.5	29.4
	FedProx	MLSYS 2021	29.3	53.8	5.8	17.4	19.1	17.7	18.1	29.7
Domain Generalization	MixStyle	ICLR 2020	31.2	53.5	5.5	16.0	28.6	31.5	21.8	33.6
	CrossStyle	ICCV 2021	35.5	59.6	4.6	14.0	27.8	28.0	22.6	33.9
Federated-ReID	FedReID	AAAI 2021	30.1	53.7	4.5	13.7	26.4	26.5	20.3	31.3
	FedPav	MM 2020	25.4	49.4	5.2	15.5	22.5	24.3	17.7	29.7
DG-ReID	SNR	CVPR 2020	32.7	59.4	5.1	15.3	28.5	30.0	22.1	34.9
FedDG-ReID (RN50)	DACS	AAAI 2024	36.3	61.2	10.4	27.5	30.7	34.1	25.8	40.9
	SSCU	MM 2025	39.5	66.4	11.9	32.3	32.8	34.1	28.1	44.3
	CO-EVO	ours	**42.4**	**71.2**	**12.9**	**33.7**	**34.9**	**37.1**	**30.1**	**47.3**
FedDG-ReID (ViT)	FedPav (ViT)	MM 2020	37.4	62.6	14.6	33.7	23.7	25.0	25.2	40.4
	CrossStyle (ViT)	ICCV 2021	41.4	65.8	17.9	40.8	31.0	38.4	30.1	48.3
	DACS (ViT)	AAAI 2024	45.4	70.7	20.3	44.2	36.6	42.1	34.1	52.3
	CO-EVO	ours	**60.7**	**80.2**	**32.2**	**60.3**	**51.3**	**52.7**	**48.1**	**64.4**

Table 2: Protocol II results with a reduced number of source domains while keeping MS as a source client. We report mAP and Rank-1 (%) on targets M and C3 under different source combinations.

Methods	MS+C3→M		MS+C2→M		MS+C2+C3→M	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
FedPav	27.5	51.5	24.8	48.5	25.4	49.1
FedReID	31.0	55.0	28.1	52.4	30.1	53.7
DACS	33.2	58.1	30.3	56.3	36.3	61.2
SSCU	36.7	62.8	34.8	62.7	39.5	66.4
ours	**39.3**	**68.4**	**36.9**	**67.1**	**42.4**	**71.2**
Methods	MS+M→C3		MS+C2→C3		MS+C2+M→C3	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
FedPav	15.2	14.1	17.3	17.0	22.5	24.3
FedReID	16.1	15.3	21.8	20.4	26.4	26.5
DACS	18.2	17.7	22.9	23.5	30.7	34.1
SSCU	20.9	20.8	27.1	29.3	32.8	34.1
ours	**24.3**	**25.1**	**29.4**	**34.2**	**34.9**	**37.1**

Table 3: Protocol III (source-domain evaluation) results. Models are trained with clients {M, C2, C3} and evaluated on the test split of each source domain. We report mAP and Rank-1 (%).

Methods	M+C2+C3→M		M+C2+C3→C2		M+C2+C3→C3	
	mAP	rank-1	mAP	rank-1	mAP	rank-1
FedProx	61.0	80.4	66.8	65.5	24.2	23.9
FedPav	53.9	76.0	59.7	56.3	19.6	19.6
FedReID	71.8	87.6	82.9	82.8	44.0	44.9
DACS	72.1	88.2	84.5	83.4	47.4	50.1
SSCU	73.0	88.7	84.9	83.9	50.4	53.2
ours	**81.4**	**93.1**	**89.8**	**91.2**	**59.7**	**63.1**

et al., 2024), and SSCU (Xu et al., 2025). Among them, DACS and SSCU are the current SOTA for FedDG-ReID. For a fair comparison, all baselines are trained under the same federated setup (domain-as-client), backbone, and communication rounds.

4.2 Comparison under Three Evaluation Protocols

Protocol I: Leave-One-Domain-Out. Table 1 summarizes the generalization results under the most rigorous DG setting. CO-EVO achieves SOTA performance across all benchmarks, outperforming the strongest CNN-based baseline by an average of +2.0% mAP and +3.0% Rank-1 (28.1/44.3→30.1/47.3). Notably, on the most challenging MSMT17 domain, our method obtains the best results (12.9/33.7). We attribute this to the coupled effect of CSA and GSD: while CSA distills purified semantic anchors via cross-camera consistency, GSD forces the model to maintain these anchors amid extreme stylistic shifts. Consistent gains under Transformer backbones further verify the architectural robustness of our design.

Protocol II: Scaling with Source Domains. Table 2 evaluates generalization under varying numbers of source clients. While performance generally improves as more domains participate, CO-EVO consistently maintains its lead under all source combinations. For instance, on Market-1501, the performance scales from 39.3% mAP (MS+C3) to 42.4% mAP (MS+C2+C3). This suggests that even with limited source diversity, our camera-invariant semantic anchors provide essential structural constraints that prevent the model from overfitting to specific local camera styles, a common pitfall in standard federated ReID.

Table 4: Robustness of CO-EVO under imperfect camera metadata. The clean SSCU baseline is included for reference.

Metadata Setting	Method / Scope	MS+C2+C3→M mAP	MS+C2+C3→M R1	M+C2+C3→MS mAP	M+C2+C3→MS R1	MS+C2+M→C3 mAP	MS+C2+M→C3 R1
Clean	CO-EVO	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**
Noisy	30% Camera ID Noise	40.4	68.7	11.7	32.4	34.6	36.2
Missing	K-means Pseudo-Groups	41.1	69.8	12.2	33.1	35.8	37.5
Reference	SSCU (clean labels)	39.5	66.4	11.9	32.3	32.8	34.1

Protocol III: Source-Domain Discriminability. Table 3 reports the testing results on participating source domains. CO-EVO achieves the best performance across all clients, indicating that the pursuit of domain generalization does not compromise the model's discriminative power on local data. By resolving the semantic-style conflict, the learned representations successfully internalize robust identity cues that are both domain-agnostic for unseen targets and highly discriminative for participating clients. This dual-advantage confirms that our purified semantic grounding effectively captures intrinsic biometric features rather than superficial imaging shortcuts.

4.3 Robustness to Imperfect or Missing Metadata

Noisy, Missing, and Unreliable Metadata. Table 4 shows that CO-EVO remains effective even when the metadata used to build the GCSB are imperfect. Under 30% camera-ID corruption, the performance drops are moderate, indicating that the shared style bank is not overly brittle to annotation noise. In the zero-metadata case, we replace camera IDs with K-means pseudo-groups and still outperform SSCU trained with clean labels on all three transfers. These results suggest that GSD benefits from coarse grouping structure rather than perfect camera annotations: as long as the pseudo-groups preserve dominant appearance patterns, the resulting templates remain useful for expanding the visual support seen during training. This robustness is particularly important in federated deployments, where camera identifiers may be incomplete, noisy, or inconsistent across institutions.

4.4 Hyperparameter Analysis

Impact of Token Length (L). We evaluate the sensitivity of token length $L \in \{1, 4, 8, 16\}$ in Table 5. Performance consistently peaks at $L = 4$ across all transfer tasks. While $L = 1$ is too generic to capture identity nuances, larger values ($L \geq 8$) introduce redundant parameters that overfit to local camera-specific noise, leading to semantic drift. A compact set of 4 tokens provides the optimal balance between discriminative power and anchoring stability for the coupled semantic–style optimization.

Table 5: Impact of token length L in CSA. Results follow Protocol I (Leave-one-domain-out).

L	MS+C2+C3→M mAP	MS+C2+C3→M R1	M+C2+C3→MS mAP	M+C2+C3→MS R1	MS+C2+M→C3 mAP	MS+C2+M→C3 R1
1	40.8	69.4	10.3	29.5	35.5	36.8
4	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**
8	41.9	70.7	12.1	32.8	36.6	38.2
16	41.5	70.2	11.5	31.6	36.2	37.7

Sensitivity of Cross-Camera Consistency (λ_{c3}). We investigate the impact of the consistency weight λ_{c3} in Table 6. When $\lambda_{c3} = 0$, the textual prototypes are prone to capturing local camera footprints, leading to sub-optimal generalization on unseen domains. As λ_{c3} increases to 0.1, the performance improves significantly, particularly on the complex MSMT17 domain (+1.4% mAP). This gain validates that $\mathcal{L}_{c3}$ effectively "purifies" the semantic anchors by filtering out camera-specific noise during the anchoring phase. When λ_{c3} becomes too large, however, the constraint starts to over-suppress subtle cross-camera appearance differences that are still useful for fine-grained identity discrimination, which explains the mild degradation beyond 0.1.

Table 6: Sensitivity analysis of the cross-camera consistency weight λ_{c3} in CSA.

λ_{c3}	MS+C2+C3→M mAP	MS+C2+C3→M R1	M+C2+C3→MS mAP	M+C2+C3→MS R1	MS+C2+M→C3 mAP	MS+C2+M→C3 R1
0.0	41.2	69.8	11.5	31.8	35.8	37.5
0.1	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**
0.2	42.1	70.9	12.4	33.1	36.7	38.3
0.5	41.5	70.2	11.8	32.0	36.2	37.8

4.5 Ablation Study

We conduct systematic ablations under the CO-EVO framework to quantify the individual and synergistic gains of its core modules.

Table 7: Ablation of key components: CSA (Phase I) and GSD (Phase II).

Components CSA	Components GSD	MS+C2+C3→M mAP	MS+C2+C3→M R1	M+C2+C3→MS mAP	M+C2+C3→MS R1	MS+C2+M→C3 mAP	MS+C2+M→C3 R1
-	-	25.4	49.4	5.2	15.5	22.5	24.3
✓	-	38.7	66.7	9.1	30.7	33.8	35.2
-	✓	39.8	67.6	10.8	30.3	34.1	35.9
✓	✓	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**

Synergy of Semantics and Style. Table 7 demonstrates that the vanilla federated baseline fails significantly under severe domain shifts (5.2% mAP on MSMT17). (1) CSA Stability: CSA provides stable, purified semantic anchors, preventing identity features from drifting across heterogeneous clients. (2) GSD Diversity: GSD mitigates camera bias by diversifying training distributions via global statistics, breaking visual shortcuts. (3) Full Synergy: CO-EVO achieves optimal performance, confirming that CSA ensures semantic grounding while GSD explores diverse style boundaries.

Table 8: Ablation of CSA anchoring strategy. Stat.: Static Caching (Ours).

Strategy	MS+C2+C3→M		M+C2+C3→MS		MS+C2+M→C3	
	mAP	R1	mAP	R1	mAP	R1
Baseline	25.4	49.4	5.2	15.5	22.5	24.3
CSA (Dyn.)	39.1	67.8	10.4	30.1	34.2	35.6
CSA (Stat.)	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**

Semantic Anchoring Strategy. Table 8 investigates prototype stability. Compared to dynamic updates, our *Static Caching* achieved higher accuracy, confirming that dynamic prompts are prone to inheriting local camera noise, whereas frozen, purified anchors provide a reliable "gravitational center" for cross-modal alignment.

Table 9: Ablation of GSD sampling scope. Glob.: Global-Bank (Ours).

Scope	MS+C2+C3→M		M+C2+C3→MS		MS+C2+M→C3	
	mAP	R1	mAP	R1	mAP	R1
Baseline	38.7	66.7	9.1	30.7	33.8	35.2
Random-Stat	39.1	67.2	9.7	30.9	34.5	35.6
GSD (Loc.)	40.2	68.5	10.5	31.4	35.1	36.3
GSD (Glob.)	**42.4**	**71.2**	**12.9**	**33.7**	**37.1**	**38.9**

GSD Sampling Scope. Table 9 compares stylization scopes. Randomly perturbing feature statistics yields only marginal gains over the baseline, which suggests that naive noise injection is insufficient to approximate realistic deployment shifts. In contrast, the Global-Bank strategy yields significantly better results, validating that sharing low-dimensional camera statistics effectively constructs a proxy distribution of unseen domains and that the improvement of GSD comes from real camera-driven templates rather than arbitrary perturbations.

4.6 Visual and Statistical Insights

To delve deeper into the mechanism of CO-EVO, we provide comprehensive visualizations and statistical evidence in Figure 3. Quantitatively, Figure 3(a) highlights the restoration of the discriminative margin. By leveraging CSA anchors, CO-EVO rectifies the severe camera bias, reducing the mean cosine distance for same-identity pairs to 0.35, which represents a 48.5% reduction compared with SSCU. Simultaneously, it expands the margin for different-identity pairs to 0.78, an 85.7% improvement. This shift ensures that intrinsic identity cues override stylistic shortcuts. The representation capability in the feature space is further elucidated via the t-SNE visualization in Figure 3(b), which displays the sample distribution on the unseen MS target domain under the M+C2+C3 to MS setting. In the DACS and SSCU baseline models, identity clusters are scattered and heavily interleaved, indicating that the encoder fails to distinguish between different individuals under significant domain shifts. In contrast, CO-EVO successfully reconciles these scattered samples into compact and well-separated clusters as indicated by the arrows. This qualitative evidence confirms that our CSA anchors act as stable gravitational centers, pulling stylistically diverse inputs toward domain-agnostic semantic centers and thereby establishing a robust decision boundary on novel domains. Qualitatively, Figure 3(c) compares our GSD with the standard Style Transformation Model (STM) used in baselines. Notably, GSD maintains superior stylization diversity and image quality across both early and late training stages. While STM-generated images frequently suffer from repetitive artifacts or unrealistic over-exposure in the late stage, our GCSB-based GSD module synthesizes diverse yet realistic camera effects such as varying illumination and color tones. This stability throughout the entire training process ensures that the model consistently learns from high-fidelity proxies of unseen imaging pipelines, effectively enhancing the generalization to novel domains.

5 Conclusion

In this paper, we identify and resolve the semantic-style conflict in FedDG-ReID. We propose CO-EVO, a novel framework that bridges the gap between semantic stability and visual diversity through a coupled semantic–style optimization mechanism. By introducing CSA, we establish purified, domain-agnostic identity prototypes that effectively filter out local camera noise via cross-camera consistency. Simultaneously, our GSD module utilizes a privacy-preserving camera-style

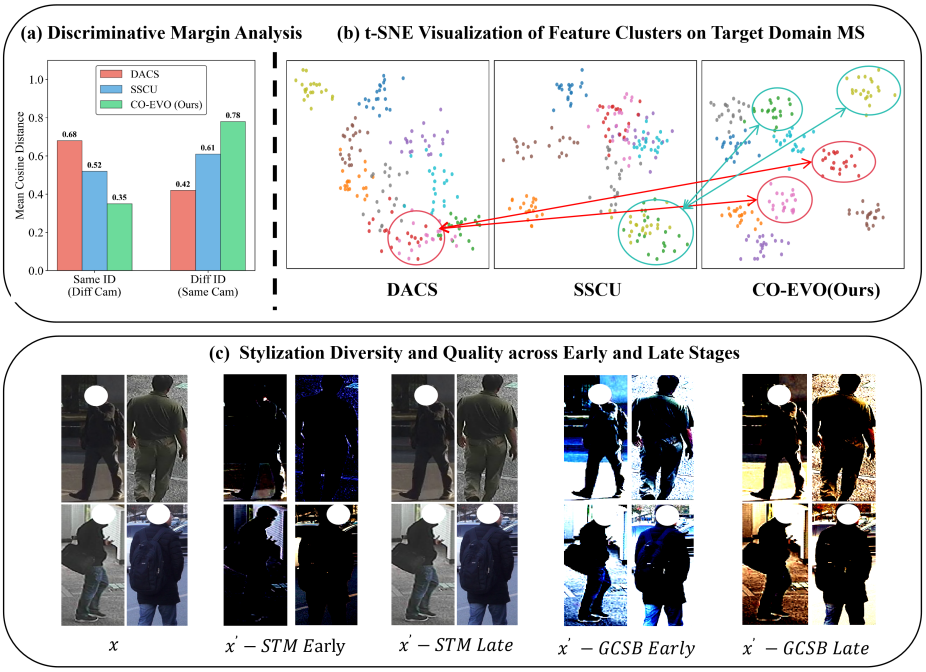

Figure 3: Comprehensive evaluation of the proposed CO-EVO. (a) Discriminative Margin Analysis: Quantitative comparison of mean cosine distances for same-identity and different-identity pairs. (b) t-SNE Performance on Target Domain MS: Visualization of feature clusters on the unseen MS domain (models trained on M+C2+C3), showing the evolution from DACS and SSCU to our CO-EVO. (c) Stylization Diversity and Quality across Early and Late Stages: Qualitative comparison between the baseline STM and our GCSB-based GSD module during the training process.

bank to synthesize realistic domain shifts, expanding the visual boundaries of training data with negligible overhead. Extensive experiments, including analyses with noisy and missing metadata, demonstrate that CO-EVO achieves SOTA performance across multiple benchmarks. These results show that stable semantic grounding and realistic style expansion are jointly essential for robust cross-domain generalization in federated environments.

6 Limitations

Despite the performance gains achieved by CO-EVO, certain limitations remain to be addressed in future research. First, while CSA provides robust semantic grounding by distilling identity-specific features, these anchors are primarily derived from source-domain identities. Under extreme conditions such as severe occlusion, significantly low resolution, or out-of-distribution appearances not covered by the source distribution, the efficacy of semantic anchoring as a regularizer may diminish. Second, the GCSB represents domain shifts using channel-wise statistics. Although this mechanism is lightweight and controllable compared to learning-based generators, it may not fully capture complex geometric transformations, background structural variations, or intricate occlusion patterns inherent in certain imaging pipelines. Third, the construction of style templates relies on the availability of camera identifiers or reliable grouping metadata to summarize statistics. In scenarios where such metadata is missing or highly noisy, the fidelity and diversity of the synthesized style bank can degrade, potentially leading to suboptimal generalization. Finally, although CO-EVO avoids the heavy computational burden of training generative networks and adheres to the decentralized nature of federated learning, it still introduces extra local computation for generating stylized views and optimizing textual prototypes. Furthermore, while sharing low-dimensional statistics minimizes privacy leakage, we do not provide a formal privacy guarantee such as differential privacy or secure aggregation for the shared camera footprint. Exploring more comprehensive style representations and integrating formal privacy-preserving mechanisms remain promising directions for future work.

References

Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. 2024. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27284–27293.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31.

Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. 2021. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3425–3435.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16145–16154.

Zan Gao, Lishuai Gao, Hua Zhang, Zhiyong Cheng, Richang Hong, and Shengyong Chen. 2020. Dcr: A unified framework for holistic/partial person reid. *IEEE Transactions on Multimedia*, 23:3332–3345.

Zan Gao, Hongwei Wei, Weili Guan, Weizhi Nie, Meng Liu, and Meng Wang. 2022. Multigranular visual-semantic embedding for cloth-changing person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3703–3711.

Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. 2023. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):712–728.

Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152.

Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. 2022. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR.

Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722.

Siyuan Li, Li Sun, and Qingli Li. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1405–1413.

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

Wei Li and Xiaogang Wang. 2013. Locally aligned feature transforms across views. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601.

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

Ke Liang, Lingyuan Meng, Hao Li, Jun Wang, Long Lan, Miaomiao Li, Xinwang Liu, and Huaimin Wang. 2025. From concrete to abstract: multi-view clustering on relational knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Oren Nuriel, Sagie Benaim, and Lior Wolf. 2021. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9482–9491.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. 2021. Crossnorm and selfnorm for generalization under distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 52–61.

Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. 2022a. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7297–7307.

Yang Wang, Jinjia Peng, Huibing Wang, and Meng Wang. 2022b. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Science China Information Sciences*, 65(6):160103.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2017. Person transfer gan to bridge domain gap for person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88.

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

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

Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. 2020. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1445–1451.

Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yifan He, Shaozi Li, and Nicu Sebe. 2024. Diversity-authenticity co-constrained stylization for federated domain generalization in person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6477–6485.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893.

Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. 2024. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6764–6772.

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

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*.

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

A Appendix

A.1 A. Implementation Details and Evaluation Protocols

Evaluation Protocols. To comprehensively evaluate both source-domain performance and generalization to unseen target domains, we adopt three evaluation protocols. (i) Protocol I (Leave-one-out): We follow a leave-one-domain-out scheme over C2, C3, MS, M. In each run, one dataset is held out for testing and the remaining three datasets serve as source domains for federated training. (ii) Protocol II (Robustness): As a complement to Protocol I, we further examine the robustness of our method under a smaller number of participating clients by removing one additional source dataset when MS is included in training. (iii) Protocol III (Source-domain evaluation): We evaluate the learned model on source domains by testing on their corresponding test splits, which reflects performance in the federated source-domain setting. For all protocols, performance is measured by Mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-1 (R1).

Federated Setup and Hyperparameters. We treat each source domain as an individual client and train the shared ReID model under the standard federated learning paradigm. The training process consists of two stages. First, we conduct a semantic learning phase for 120 rounds to initialize stable identity anchors. Subsequently, we run 60 communication rounds for the federated optimization phase. In each round, the server broadcasts the global parameters and each client performs local optimization for E=1 epoch before uploading model updates for aggregation. Based on our hyperparameter analysis, the token length L in CSA is set to 4 to provide the optimal balance between discriminative power and anchoring stability. The cross-camera consistency weight λ_{c3} is set to 0.1 to effectively purify semantic anchors by filtering out local camera specific noise. The temperature parameter τ used in the semantic alignment loss is set to 0.07.

Implementation Details. During federated training, the mini-batch size is set to 64. We use the SGD optimizer with an initial learning rate of 1e-3, weight decay of 5e-4, and momentum of 0.9. The learning rate is scheduled by MultiStepLR with decay factors at specific communication rounds. Our framework is compatible with various backbones, including ResNet-50 (RN50) and Vision Transformer (ViT). For the vision-language component, we employ the CLIP model with a ViT-B/16 image encoder as the base for semantic anchoring. For GSD, the camera-style templates are extracted once at the beginning of Phase II to construct the GCSB, which takes about 4s per client on average and contributes less than 0.1% of the full training time. This ensures that the diversification process is grounded in real camera distributions without requiring expensive generators or incurring additional communication overhead in subsequent training rounds.