Scale-TBPS: Unified Multi-Dataset Training with Noise-Aware Curation and Scalable Discriminative Identity Learning

Anonymous authors
Paper under double-blind review

Abstract

Text-Based Person Search (TBPS) has seen significant progress with vision–language models (VLMs), yet it remains constrained by limited training data and the fact that VLMs are not inherently pre-trained for pedestrian-centric recognition. Existing TBPS methods therefore rely on dataset-centric fine-tuning to handle distribution shift, resulting in multiple independently trained models for different datasets. While synthetic data can increase the scale needed to fine-tune VLMs, it still needs dataset-specific adaptation. This motivates a fundamental question: can we train a single unified TBPS model across multiple datasets? We show that naive joint training over all datasets remains sub-optimal because current training paradigms do not scale to a large number of unique person identities and are vulnerable to noisy image–text pairs. To address these challenges, we propose Scale-TBPS with two contributions: (i) a noise-aware unified dataset curation strategy that cohesively merges diverse TBPS datasets; and (ii) a scalable discriminative identity learning framework that remains effective under a large number of unique identities. Extensive experiments on CUHK-PEDES, ICFG-PEDES, RSTPReid, IIITD-20K, and UFine6926 demonstrate that a single Scale-TBPS model outperforms dataset-centric optimized models and naive joint training. The code is available at: https://anonymous.4open.science/r/Scale-TBPS-52D6/.

1 Introduction

Text-based Person Search (TBPS) is a fine-grained visual retrieval task to locate specific individuals in large image galleries. Since its introduction by Li et al. (2017) alongside the benchmark CUHK-PEDES dataset, the field has expanded through numerous benchmarks in recent years including ICFG-PEDES Ding et al. (2021), RSTPReid Zhu et al. (2021), IIITD-20K Subramanyam et al. (2023), and UFine6926 Zuo et al. (2024b). The inception of Vision-Language Models (VLMs) such as CLIP Radford et al. (2021) and ALBEF Li et al. (2021) has yielded notable performance gains. However, these models are primarily trained on broad online corpora Changpinyo et al. (2021); Thomee et al. (2016). Though they show notable zero shot results in multiple domains, the performance in zero shot fine-grained domains still remains unsatisfactory Wang et al. (2025). To tackle this issue, various TBPS Qin et al. (2024); Li et al. (2021); Cao et al. (2024); Jiang & Ye (2023); Qin et al. (2025) resort to dataset level fine-tuning and propose several loss formulations and local feature networks to improve fine-grained feature extraction.

Despite these advancements, the current paradigm remains heavily dataset-centric. Fine-tuning models on isolated benchmarks leads to poor cross-dataset generalization, as these models fail to account for the diverse distribution shifts caused by varying lighting conditions Li et al. (2025), and environmental contexts. A model trained on unified and curated data from multiple sources can leverage greater data diversity and scale, thereby achieving superior retrieval performance. This motivates a fundamental question: can we train a single unified TBPS model across multiple datasets?

Recent works Yang et al. (2023); Tan et al. (2024); Zuo et al. (2024a) leverage large volumes of unlabelled or synthetic pedestrian data to learn person-centric representations. Despite the scale of pretraining data,

these methods still require dataset-specific fine-tuning to achieve competitive performance, and the reliance on individual dataset-centric models persists even after pretraining. Similarly, naive joint training also fails to show robust performance across all distribution shifts despite having access to more data. This indicates that data scale alone does not resolve the unified training problem. We hypothesize that two bottlenecks explain this failure. First, due to annotation noise Sun et al. (2025); Qin et al. (2024) and cross-dataset distribution shifts, not all image–text pairs exhibit reliable semantic correspondence. Unreliable pairs corrupt alignment learning at scale in ways that do not surface when training on a single clean dataset. Second, under unified training the overall identity label space grows to a cardinality far exceeding that of any individual dataset, making naive ID supervision Zheng et al. (2020); Cao et al. (2024) increasingly ineffective as the identity space scales.

To address these challenges, we propose Scale-TBPS with two core components: a Noise-aware Dataset Curation (NDC) framework for cohesive multi-dataset merging, and a Discriminative ID Learning (DIL) framework for learning from a large number of unique identities in a shared embedding space. For NDC, we consider a set of diverse expert models trained on varied TBPS distributions and rely on relative similarity ranking between text–image pairs rather than a hard threshold. Combining consensus from diverse expert models enables more cohesive curation of the unified dataset. Crucially, this design enables scalability as additional datasets can be incorporated even when corresponding pretrained TBPS models are unavailable. NDC operates as a one-time preprocessing step, filtering the entire joint dataset of 169,810 text–image pairs in under ten minutes on a single GPU. For DIL, we adopt a geometry-oriented approach. The class features from different distribution shifts lie far apart from each other, thus we map features onto a common hypersphere with respect to identity-based prototypes and compute angular similarity between feature embeddings and class prototypes to capture directional alignment and discriminative features, enforcing stricter separation between identity classes as the identity space grows.

Our contributions are as follows:

- We identify annotation noise and identity-space explosion as the two bottlenecks of naive joint training for TBPS and show that solving both is necessary for a unified model to consistently outperform dataset-centric approaches.
- We propose Noise-aware Dataset Curation (NDC) and Discriminative ID Learning (DIL) to address each bottleneck respectively. NDC filters unreliable text–image pairs using consensus from diverse pretrained expert models via relative similarity ranking. DIL maps heterogeneous multi-dataset features onto a shared unit hypersphere and enforces angular separation between identity classes with multimodal angular identity loss (MA-ID).
- We train a single unified model that outperforms dataset-centric state-of-the-art models on CUHK-PEDES, RSTPReid, ICFG-PEDES, and IIITD-20K and demonstrates strong zero-shot generalization on UFine6926.

2 Related Works

With the advent of large-scale VLMs, TBPS methods have largely benefited from improved cross-modal alignment. In particular, CLIP has become a popular choice in TBPS due to its lightweight architecture and strong transferability, and has been widely adapted for this task Jiang & Ye (2023); Yan et al. (2023); Radford et al. (2021); Cao et al. (2024); Qin et al. (2024). ALBEF Chowdhury et al. (2024), although computationally heavier, provides stronger alignment capabilities and has been adopted in methods such as RaSa Bai et al. (2023).

CLIP-based methods. CFine Yan et al. (2023) was the first to adopt a CLIP-based baseline for TBPS. CFine also focused on fine-grained adaptation through local networks, as the distribution shift was still considerable. IRRA Jiang & Ye (2023) was among the first CLIP-based approaches to incorporate ID loss for learning identity-specific attributes. However, IRRA primarily focused on contrastive learning, local feature interaction, and masked language modeling (MLM). RDE Qin et al. (2024) adopted the IRRA architecture but replaced the contrastive objective with a ranking-based loss, demonstrating improved optimization

behavior. RDE also addressed the noisy nature of TBPS datasets through noisy correspondence learning. However, the correspondence refinement process is performed after every epoch, making training computationally expensive. TBPS-CLIP Cao et al. (2024) studies several augmentation strategies and loss functions without using sophisticated modules, while highlighting the limitations and usefulness of CLIP's performance in the TBPS field. However, individual TBPS datasets are relatively small, and these studies do not provide information on performance under joint learning where identity space grows considerably.

Non-CLIP-based methods. RaSA Bai et al. (2023) is built upon ALBEF, which is a heavier multimodal backbone compared to CLIP and consequently achieves stronger performance. RaSA models relational sensitivity between data samples and exploits these relations to improve contrastive learning. More recently, another research direction has focused on TBPS-specific pre-trained backbones Yang et al. (2023); Tan et al. (2024), where multimodal large language models (MLLMs) are used to generate dataset-level captions for large-scale pre-training. Although these approaches improve representation learning, they still exhibit limited zero-shot generalization and require subsequent dataset-specific fine-tuning to achieve competitive performance. Consequently, even after TBPS-specific pre-training, the dependence on separate dataset-specific models largely remains. This motivates the need to investigate whether TBPS can instead be scaled toward a single unified model capable of handling diverse distribution shifts. Notably, these methods are trained on more than 1M TBPS-specific pre-training samples, yet the noisy nature of the collected data still necessitates effective filtering mechanisms.

One recent work, WoRA-TBPS Sun et al. (2025), performs data filtering using a pre-trained VLM; however, the filtering model itself is not specifically trained on a TBPS corpus. WoRA-TBPS adopts APTM Yang et al. (2023) as its baseline, where synthetic TBPS-specific data are additionally utilized during training.

Unified paradigm. Although VLMs demonstrate promising zero-shot capabilities, fine-grained tasks such as TBPS and vision-based person re-identification still require task-specific fine-tuning to achieve reliable performance. Recent research increasingly explores unified paradigms that aim to support multiple in-domain distribution shifts within a single model. For instance, ATReID Li et al. (2025) addresses illumination variation by enabling identity recognition across different times of day, while UniHCP Ci et al. (2023) trains a unified model across multiple human-centric perception tasks. In large-scale identity learning problems such as face recognition, angular-margin based objectives have also been extensively explored Deng et al. (2019). Similarly, the discriminative capability of angular objectives has recently been studied in positive-unlabeled learning settings Sevetlidis et al. (2025). However, these approaches are primarily designed for unimodal vision tasks and have not been effectively transferred to multimodal retrieval settings.

Although identity learning and contrastive objectives are widely adopted, these models are not explicitly designed to scale to extremely large identity spaces. Moreover, simply increasing the amount of training data does not necessarily improve performance, as existing methods fail to address the optimization and distributional challenges introduced by unified training. Motivated by these challenges, we propose a noise-aware dataset curation strategy for unified training together with a discriminative ID learning framework designed to scale effectively to large identity spaces.

3 Methodology

Scale-TBPS aims to learn a single unified model that can handle data scaling across all available TBPS datasets. Our method explicitly targets the cohesive merging of different data distributions while handling large volumes of unique person identities. We formulate our problem, followed by our noise-aware unified dataset curation strategy (NDC) in Section 3.1. Section 3.2 explains our Discriminative Identity Learning framework (DIL) designed to scale effectively with the number of person identities.

Problem Formulation Let there be n train-sets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$, each containing image–text pairs annotated with person identities. These datasets are merged to form a joint dataset:

$$\mathcal{D} = \bigcup_{j=1}^{n} \mathcal{D}_j = \{I_i, T_i, y_i\}_{i=1}^{N}, \tag{1}$$

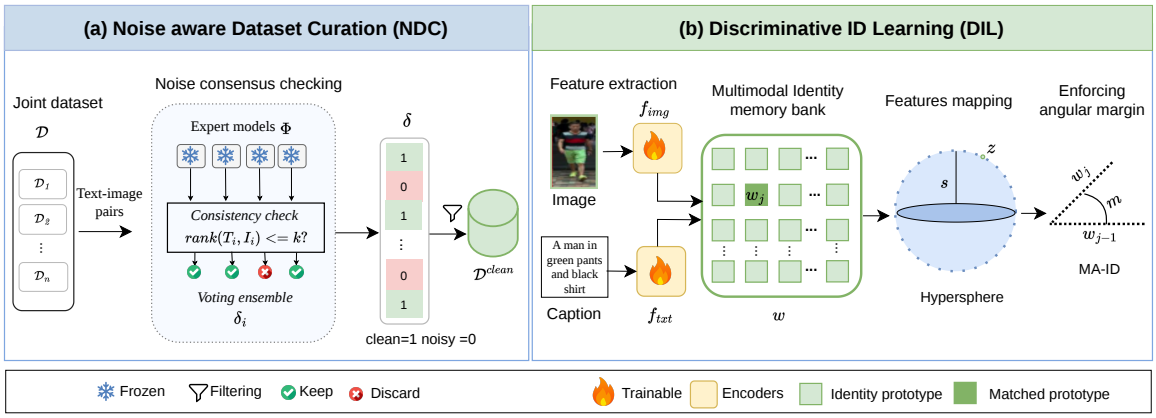


Figure 1: Overview of the proposed Scale-TBPS. (a) Noise-Aware Data Curation (NDC): Text–image pairs from the joint dataset ($\mathcal{D}$) are encoded using a set of pretrained and frozen models Φ. top-K retrieved samples are computed independently for each model. A pair is retained as a clean sample if it is ranked within the top-K results by at least one the experts and then their votes are ensembled. Selected pairs are highlighted in green. (b) Discriminative Identity Learning (DIL): Filtered text–image pairs from $\mathcal{D}^{\text{clean}}$ are encoded by the learnable image and text encoders and encoded features f^v and f^t are passed through a shared multimodal identity memory w to map the features on a hypersphere. MA-ID denotes the proposed multimodal angular identity loss while m denotes the margin

where I_i and T_i denote the i-th pedestrian image and its corresponding textual description, respectively, and y_i represents a unique person identity across the unified dataset. We employ image encoder ϕ_{img} and text encoder ϕ_{txt} initialized with CLIP weights. These encoders are jointly optimized during training to produce aligned global visual and textual representations.

Baseline. We adopt established dataset-centric TBPS model architectures as our baselines. To enable scalability, we design a training pipeline that can be applied on top of VLM baseline models, without modifying their core architectures.

Feature Extraction. For each image I_i, we extract a global visual representation f_i^v using an image encoder, where f_i^v corresponds to the `[CLS]` token or an equivalent global pooled feature. Similarly, for each text description T_i, a global textual representation f_i^t is obtained from a text encoder, using the `[EOS]` token.

3.1 Noise-Aware Unified Dataset Curation (NDC)

Due to annotation noise and cross-dataset distribution shifts, not all image-text pairs exhibit reliable semantic correspondence. Learning alignment confidence directly from untrained models is unreliable. Therefore, we adopt a noise-aware filtering strategy using a set of pretrained TBPS models. We perform the filtering only once and use the filtered dataset for training.

We consider a set of p pretrained person-specific models $\Phi = \{\phi_1, \phi_2, \ldots, \phi_p\}$. Individual experts may operate under different distributions and can be biased or overfit to specific pairs. As a result, a single expert may incorrectly discard a valid pair. By aggregating decisions among multiple experts, such errors can be mitigated to a certain extent. The image and text encoders of ϕ_l are frozen. The ensemble thus provides robust supervision for filtering unreliable text-image pairs.

Given $\mathcal{D}$ and ϕ_l, for every text query T_i, we examine whether its corresponding image I_i appears within the top-K retrieved images. Specifically, a text-to-image similarity matrix is computed over $\mathcal{D}$. The matrix is constructed using cosine similarity between ℓ_2-normalized global text and image feature embeddings, f_i^t and f_j^v, respectively. For T_i, similarity is computed against all images in $\mathcal{D}$:

$$\mathcal{S}_i = \left[S\left(f_i^t, f_1^v\right),\ S\left(f_i^t, f_2^v\right),\ \ldots,\ S\left(f_i^t, f_N^v\right)\right], \tag{2}$$

where $S(\cdot,\cdot)$ denotes cosine similarity.

The similarity scores $\mathcal{S}_i$ are sorted in descending order, inducing a ranking over all images. The retrieval rank for model ϕ_l, denoted as $\text{rank}_{\phi_l}(T_i, I_i)$, is defined as the position of the ground-truth image I_i in this sorted list.

We define an indicator variable $\delta_i \in \{0, 1\}$ to determine whether a pair (T_i, I_i) is valid under the ensemble of pretrained models. A pair is considered valid if at least v pretrained models rank the corresponding image within its top-K retrieval results:

$$\delta_i = \begin{cases} 1, & \text{if } \sum_{l=1}^{p} \mathbb{I}(\text{rank}_{\phi_l}(T_i, I_i) \leq K) \geq v, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Based on this criterion, we construct a curated noise-aware dataset which is subsequently used to train our unified model,

$$\mathcal{D}^{\text{clean}} = \{(T_i, I_i) \in \mathcal{D} \mid \delta_i = 1\}. \tag{4}$$

Our NDC model is inspired by the listwise ranking model of R. Duncan Luce Xia et al. (2019) and Robin L. Plackett Plackett (1975). We further validate this formulation and provide a probabilistic interpretation in the appendix.

3.2 Discriminative ID Learning framework (DIL)

Under unified training across n datasets with identity sets $\{\mathcal{Y}_k\}_{k=1}^{n}$, the overall label space becomes $y = \bigcup_{k=1}^{n} \mathcal{Y}_k$, with cardinality $C = |y|$. As C grows significantly larger than the identity space of any individual dataset, naive ID supervision becomes less effective. The class features from different distribution shifts lie far apart from each other. To mitigate this, our method geometrically maps the features on a common hypersphere Sevetlidis et al. (2025).

Let the normalized feature vectors $\hat{f}^v$ and $\hat{f}^t$ be obtained by the learnable image encoder ϕ_{img} and text encoder ϕ_{txt}, respectively. We employ a trainable multimodal identity memory bank $W = \{w_j\}_{j=1}^{C}$, where each $w_j \in \mathbb{R}^d$ represents the prototype corresponding to the j-th identity class. All prototypes are ℓ_2-normalized such that $\|w_j\|_2 = 1$. With this geometry W learns to map ℓ_2-normalized features onto a common unit hypersphere irrespective of its modality.

W is shared across both image and text modalities, allowing each prototype to accumulate identity-level information from multimodal features. Both f_v and f_t are projected with the same shared W to capture strictly class-identity specific features irrespective of the modality.

Regular classifier-based methods like instance loss Zheng et al. (2020) generally focus on capturing class-specific information, whereas our method focuses on capturing the discrimination between features. We maintain class prototypes w_j and model the directional difference between each class prototype and a given feature. The angular similarity between a feature and the j-th identity prototype is defined as:

$$\cos\theta_{i,j}^{v} = \hat{f}_i^v \cdot w_j, \qquad \cos\theta_{i,j}^{t} = \hat{f}_i^t \cdot w_j. \tag{5}$$

For the ground-truth identity y_i, θ_{i,y_i} denotes the angle between the feature and its corresponding prototype.

To further increase the discrimination, we use an additive angular margin m, which enforces stricter separation between identity classes. Specifically, the margin-adjusted similarity for the target class is defined as:

$$\gamma_{i,y_i} = \cos(\theta_{i,y_i} + m). \tag{6}$$

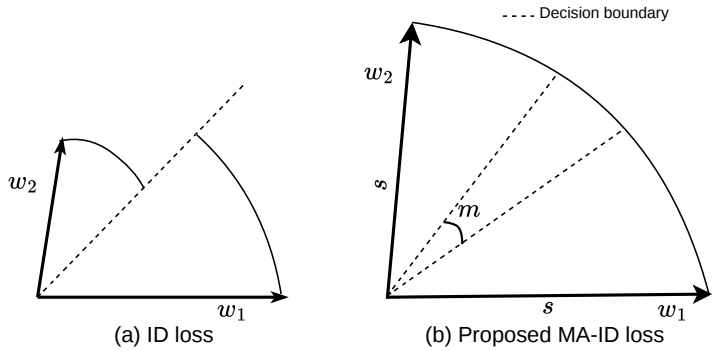


Figure 2: Geometric interpretation of MA-ID compared to ID loss. Compared to ID loss, our MA-ID applies angular separation between class prototypes while eliminating magnitude bias by scaling all weight vectors to s.

To ensure monotonicity of the angular function and maintain numerical stability, the margin-adjusted similarity is defined in a piecewise manner as

$$\gamma_{i,y_i} = \begin{cases} \cos(\theta_{i,y_i} + m), & \theta_{i,y_i} \leq \pi - m, \\ \cos(\theta_{i,y_i}) - m\sin(\pi - m), & \text{otherwise.} \end{cases} \tag{7}$$

The final logits used for training are defined as

$$z^v_{i,j} = \begin{cases} s\gamma_{i,y_i}, & j = y_i, \\ s\gamma_{i,j}, & j \neq y_i, \end{cases} \tag{8}$$

where s is a scaling factor that controls the magnitude of the otherwise normalized logits. Theoretically, we scale the features from a unit hypersphere to a hypersphere with radius s. The logits are then optimized using the standard cross-entropy loss:

$$\mathcal{L}_{img} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(z^v_{i,y_i})}{\sum_{j=1}^{C}\exp(z^v_{i,j})}. \tag{9}$$

Similarly, text-based identity loss is computed using the textual embeddings f^t and using the shared class weight w:

$$\mathcal{L}_{txt} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(z^t_{i,y_i})}{\sum_{j=1}^{C}\exp(z^t_{i,j})}. \tag{10}$$

Multimodal Angular Identity Loss. The final identity supervision jointly optimizes both modalities,

$$\mathcal{L}_{\text{MA-ID}} = \frac{1}{2}\left(\mathcal{L}_{\text{img}} + \mathcal{L}_{\text{txt}}\right). \tag{11}$$

Our proposed MA-ID loss achieves more discriminability compared to conventional identity-based objectives that lack explicit geometric regularization. We discuss this in detail.

3.3 Theoretical analysis: MA-ID loss

We analyze the behavior of the MA-ID loss at convergence to understand its geometric effect.

Lemma 3.1 (Uniform Angular Separation under MA-ID Loss)**.** *Let $\{\mathbf{f}_i\}_{i=1}^N$ denote ℓ_2-normalized feature embeddings and let $\theta_{ij} = \arccos(\mathbf{f}_i^\top \mathbf{f}_j)$ denote the geodesic angle between embeddings i and j on the unit hypersphere $\mathcal{S}^{d-1}$. Suppose the MA-ID loss enforces an additive angular margin $m > 0$ on positive pairs belonging to the same identity. Then at convergence, for any two embeddings $\mathbf{f}_i$, $\mathbf{f}_j$ belonging to distinct identities $y_i \neq y_j$, the inter-identity angle satisfies*

$$\theta_{ij} \geq m, \tag{12}$$

implying a lower bound on the cosine similarity gap between within-identity and cross-identity pairs of at least $\cos(0) - \cos(m) = 1 - \cos(m) > 0$.

Proof. Let w_c denote the ℓ_2-normalized weight vector for class c in the classifier. The MA-ID loss penalizes the logit of the ground-truth class y_i by inserting an additive margin m into the angle, replacing $\cos(\theta_{i,y_i})$ with $\cos(\theta_{i,y_i} + m)$. The training objective therefore requires

$$\cos(\theta_{i,y_i} + m) \geq \cos(\theta_{i,c}) \quad \forall\, c \neq y_i, \tag{13}$$

which, since $\cos(\cdot)$ is strictly decreasing on $[0, \pi]$, is equivalent to

$$\theta_{i,y_i} + m \leq \theta_{i,c} \quad \forall\, c \neq y_i. \tag{14}$$

Rearranging,

$$\theta_{i,c} - \theta_{i,y_i} \geq m \quad \forall\, c \neq y_i. \tag{15}$$

At convergence, intra-class angles $\theta_{i,y_i} \to 0$ as positive pairs are pulled toward their class prototype $\mathbf{w}_{y_i}$. Substituting $\theta_{i,y_i} \approx 0$ yields

$$\theta_{i,c} \geq m \quad \forall\, c \neq y_i, \tag{16}$$

establishing a uniform lower bound of m on all inter-identity angles. The cosine gap follows from the monotonicity of $\cos(\cdot)$:

$$\cos(\theta_{i,y_i}) - \cos(\theta_{i,c}) \geq \cos(0) - \cos(m) = 1 - \cos(m) > 0, \tag{17}$$

where the final inequality holds for any $m \in (0, \pi)$. □

Remark 3.1 (Discriminability via Uniform Separation)*.* By Lemma 2, every pair of cross-identity embeddings is separated by at least angle m, irrespective of class or sample difficulty. Conventional ID supervision provides no such geometric lower bound, leaving inter-class separation dependent on the relative scaling of logits rather than an explicit angular constraint.

As illustrated in Figure 2, for MA-ID loss class prototypes are being evenly distributed over a hypersphere with radius s providing balanced angular separation across all identities. In contrast, conventional identity (ID) loss primarily enforces alignment between features and their corresponding class prototypes by maximizing $w_y^\top \mathbf{f}_i$. While this ensures intra-class compactness, it does not explicitly model interactions between different class prototypes. As a result, inter-class relationships remain unconstrained, often leading to uneven spacing and suboptimal separation in the embedding space. Unlike MA-ID, which introduces a global repulsive structure through pairwise prototype interactions, standard ID loss operates locally and lacks an explicit mechanism to enforce uniform inter-class geometry. This is further shown with empirical evidence in analysis section.

4 Experiments

In this section, we conduct extensive experiments to comprehensively evaluate the effectiveness and scalability of the proposed Scale-TBPS framework.

Table 1: Comparison study of Scale-TBPS with existing methods. '*' denotes the results with test-time normalization.

Methods	Venue	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
CLIP based													
CFine Yan et al. (2023)	TIP	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
CLIP Radford et al. (2021)	PMLR	68.55	86.32	91.85	61.25	56.80	75.99	82.31	31.93	56.00	80.85	87.90	44.06
IRRA Jiang & Ye (2023)	CVPR	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
TBPS-CLIP Cao et al. (2024)	AAAI	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	62.10	81.90	87.75	48.00
CFAM Zuo et al. (2024b)	CVPR	75.60	90.53	94.36	67.27	65.38	81.17	86.35	39.42	62.45	83.55	91.10	49.50
ICL Qin et al. (2025)	CVPR	76.41	90.48	94.33	68.04	68.11	82.59	87.52	40.81	67.70	86.05	91.75	52.62
Non-CLIP based													
APTM Yang et al. (2023)	MM	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
RaSa Bai et al. (2023)	IJCAI	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	66.90	86.50	91.35	52.31
WoRA-TBPS Sun et al. (2025)	MM	76.38	89.72	93.49	67.22	68.35	83.10	87.53	42.60	66.85	86.45	91.10	52.49
RDE Qin et al. (2024)	CVPR	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
RDE*	-	76.32	90.95	94.43	68.13	66.73	82.87	87.63	39.42	65.30	85.90	90.90	51.63
Scale-TBPS	**Ours**	77.03	90.72	94.20	70.03	67.98	82.00	86.76	44.78	69.50	87.20	91.65	55.86
Scale-TBPS*	**Ours**	**77.60**	**91.29**	**94.54**	**70.65**	**68.11**	**82.86**	**87.60**	**45.43**	**70.40**	**87.60**	**92.30**	**57.88**

4.1 Experimental Details

For our unified dataset curation, we combine training set of four widely used text-based person search benchmarks: CUHK-PEDES Li et al. (2017), ICFG-PEDES Ding et al. (2021), RSTPReid Zhu et al. (2021) and IIITD-20K Subramanyam et al. (2023). We utilize pretrained models trained individually on CUHK-PEDES, ICFG-PEDES, and RSTPReid and IIITD-20K to filter pairs. Our unified dataset contains a total of 32,660 unique identities in the training set, compared to 11,003, 3,102, 3,701, 15000 training set identities in CUHK-PEDES, ICFG-PEDES, RSTPReid and IIITD-20K, respectively.

We set $K = 25$. Following NNN Chowdhury et al. (2024), we further employ a test-time normalization strategy for text-based image retrieval tasks using $\alpha = 0.75$ and $\mathcal{K} = 16$, without additional hyperparameter tuning for any experiment. Experiments utilizing NNN are denoted by *. We employ a discriminative identity classifier with a fixed scale of $s = 30$ and an angular margin of $m = 0.35$. We use RDE as our baseline. We follow the training configurations of RDE. We follow the standard test protocols used across all TBPS datasets. We evaluate retrieval performance using Rank-k accuracy at $k = \{1, 5, 10\}$, and mean Average Precision (mAP).

4.2 Main results

As shown in Table 1, we compare Scale-TBPS with state-of-the-art dataset-centric methods, where each model is trained independently on a single dataset. In contrast, Scale-TBPS learns a single unified model and outperforms these dataset-specific approaches in most of the metrics. We report results for Scale-TBPS with and without test-time similarity normalization NNN, where Scale-TBPS* denotes the variant using normalization. Scale-TBPS uses CLIP as the backbone, but we report discussion with both CLIP-based and non-CLIP-based approaches.

Comparison against CLIP-based methods. Scale-TBPS consistently outperforms RDE across datasets and metrics. In particular, we see a significant gain in terms of mAP. We see a gain of approximately 5% in mAP on both ICFG-PEDES and RSTPReid. In the transductive setting where test time normalization is applied, compared to RDE*, Scale-TBPS* shows a 2.52%, 6.01%, and 6.25% increase in mAP on CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively. On RSTPReid, our method shows a 7% improvement in Rank-1 accuracy compared to RDE.

The CLIP-based state-of-the-art method ICL Qin et al. (2025) relies on multimodal large language models (MLLMs) Wang et al. (2024) for data augmentation and uses backbone trained on TBPS specific data Qin et al. (2025). In contrast to that Scale-TBPS was initialized with vanilla CLIP weights. Despite not employing such auxiliary components, Scale-TBPS surpasses ICL in both Rank-1 and mAP. Overall, our single unified model outperforms all CLIP-based methods across metrics.

Table 2: Comparison results on IIITD-20K

Method	R-1	R-5	R-10	mAP
IRRA Jiang & Ye (2023)	77.08	–	–	–
RDE Qin et al. (2024)	80.90	94.48	97.02	87.04
Scale-TBPS (Ours)	83.58	95.48	97.70	88.97

Table 3: Zero-shot Comparison results on UFine6926

Method	R-1	R-5	R-10	mAP
IRRA Jiang & Ye (2023)	37.51	54.92	64.29	40.76
RDE Qin et al. (2024)	40.37	57.49	66.05	42.68
Scale-TBPS (Ours)	45.34	62.03	70.38	45.37

NNN addresses hubness which is when too many classes cluster together. To study the effect of NNN, we also report RDE*. We use the values of α and $\mathcal{K}$ as suggested in Chowdhury et al. (2024) and do not tune them either for Scale-TBPS or RDE. RDE* shows a gain of 0.38% in Rank-1 on CUHK-PEDES over RDE, whereas Scale-TBPS* shows a 0.57% boost on the same metric compared to Scale-TBPS. Further, Rank-1 accuracy decreases for ICFG-PEDES and RSTPReid. While NNN mitigates bias for Scale-TBPS, its benefit is less pronounced for RDE. This is because Scale-TBPS is already designed to promote uniformly separated clusters, which inherently reduces hubness, and test-time normalization further amplifies this effect.

Comparison against non-CLIP-based methods. Scale-TBPS outperforms RaSa Bai et al. (2023), which is ALBEF-based, even though ALBEF is generally considered a stronger backbone than CLIP. On ICFG-PEDES it shows a 2.7% gain in Rank-1, while on RSTPReid it shows a 2.6% gain. We also compare against APTM and WoRA-TBPS, whose backbones are pre-trained on pedestrian-specific data and thus exhibit strong capabilities. CLIP-based methods are much lighter in terms of parameter count. Compared to APTM Yang et al. (2023) and WoRA-TBPS, our method outperforms APTM on both CUHK-PEDES and RSTPReid. In contrast, the baseline RDE fails to surpass APTM and RaSa, whereas our method achieves consistent improvements over both. Both APTM and WoRA-TBPS are trained on 1.5 M pedestrian data and then finetuned on downstreaming datasets. Still our one unified model shows comparable performance against all three benchmarks. These results demonstrate that properly designed unified model can outperform different dataset-centric models. Our CLIP based method gives competing performance against stronger backbones too.

In Table 2 for IIITD-20K, We observe a 2.68% improvement in Rank-1 accuracy and a 1.93% improvement in mAP for Scale-TBPS compared to RDE. The results across multiple datasets demonstrate that a single unified model can outperform larger models trained on substantially more data.

Zeroshot results on UFine6926. As UFine6926 Zuo et al. (2024b) is not included in our training paradigm, we evaluate our model in a zero-shot setting on this dataset. In Table 3, we report results from prior methods that are trained exclusively on CUHK-PEDES, as this setting yields the strongest performance in their respective works. Our Scale-TBPS shows a 4.97% gain in Rank-1 and a 2.69% gain in mAP. These results indicate that the proposed method increases robustness to unseen domains and benefits from large-scale unified training.

4.3 Ablation Study

We analyze the contribution of the two core components of Scale-TBPS in Table 4: (i) noise-aware unified data curation (NDC), and (ii) discriminative identity learning framework (DIL) for scaling to a large number of unique identities.

For fair comparison, we include two strong naive joint training baselines, where IRRA and RDE are individually trained with the joint dataset. We keep RDE's original filtering protocols. As reported in Table 4, even though naive joint training leverages a larger volume of training data, it fails to yield better results across

Table 4: Ablation study for NDC and DIL. NDC+DIL is the Scale-TBPS method

Method	CUHK-PEDES		ICFG-PEDES		RSTPReid	
	R-1	mAP	R-1	mAP	R-1	mAP
Naive-IRRA	73.26	66.70	62.34	38.98	65.05	53.22
Naive-RDE	76.40	68.32	66.03	40.68	67.00	54.14
NDC	75.97	68.02	66.42	40.92	70.15	54.75
DIL	76.09	68.96	66.60	42.81	66.60	54.30
NDC+DIL	**77.03**	**70.03**	**67.98**	**44.78**	**69.50**	**55.86**

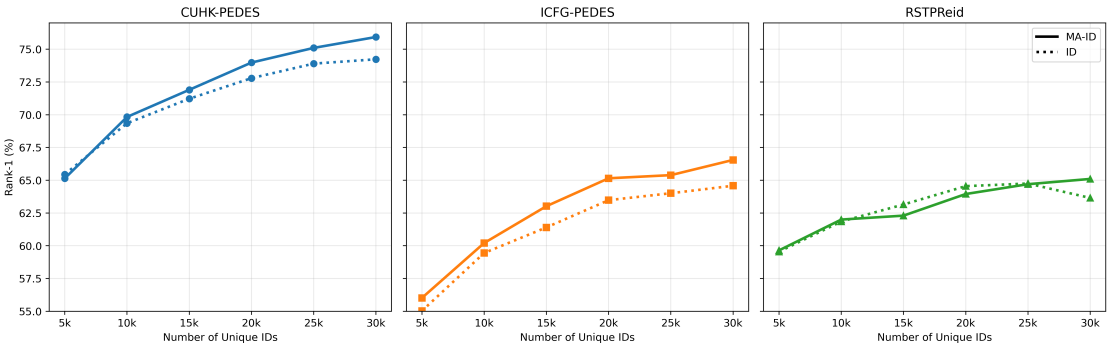
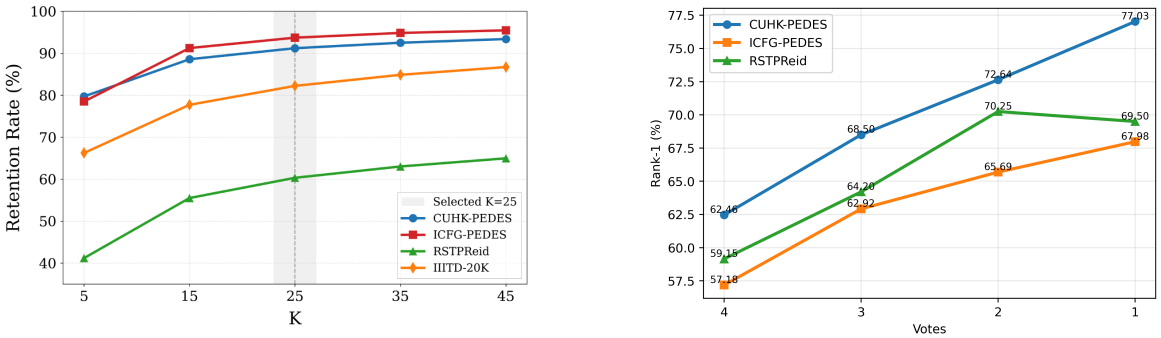


Figure 3: Rank-1 accuracy on three benchmarks as the number of unique identities included in unified training is progressively increased. MA-ID shows more consistent performance compared to standard ID loss as the joint identity space grows.

all datasets. In particular, it underperforms compared to RDE model on ICFG-PEDES. We also report the performance of our modules NDC and DIL independently.

Compared to naive joint training, NDC improves both Rank-1 and mAP on ICFG-PEDES and RSTPReid. On CUHK-PEDES, NDC shows a marginal drop. On the other hand, DIL shows significant improvement over the naive method on ICFG-PEDES. Our proposed method NDC combined with DIL outperforms naive method by 1% in CUHK-PEDES and nearly 2% in ICFG-PEDES and RSTPReid in Rank-1 accuracy. In terms of mAP, our method outperforms the naive approach by 1.73% on CUHK-PEDES, 4.48% on ICFG-PEDES, and 1.86% on RSTPReid. Finally, the combined model (NDC+DIL), referred to as Scale-TBPS, significantly outperforms all baselines across the three benchmarks. The two components are designed to address complementary challenges in unified training. NDC filters noisy pairs, while DIL ensures scalability to a large number of identities. When integrated, they yield a stronger model than either component individually, demonstrating the effectiveness of their complementary design. In addition, we conducted a thorough analysis of DIL and NDC to verify that each component aligns with our design motivations.

Analysis of DIL. In Figure 3, we analyze the impact of the proposed Multimodal Angular Identity (MA-ID) loss compared to naive supervision using the standard ID loss across three datasets. Specifically, we study the effect of increasing the number of unique identities obtained from combining multiple datasets. We randomly sample identities from the joint dataset and gradually increase the number of unique training IDs. We observe that models trained with fewer identities (e.g., 5k and 10k) achieve comparable Rank-1 accuracy for both MA-ID and standard ID loss on CUHK-PEDES. On RSTPReid, at the maximum number of identities, MA-ID outperforms ID loss. As the number of unique identities increases, the MA-ID-trained model demonstrates better adaptation to the increased data. Across all three datasets, ID loss fails to scale with increasing identity space. More visualization of embedding spaces using t-SNE diagram is given in the appendix.



(a) Effects of K on data retention rates

(b) Rank-1 vs vote consensus

Figure 4: Effects of K and expert votes of NDC on performance

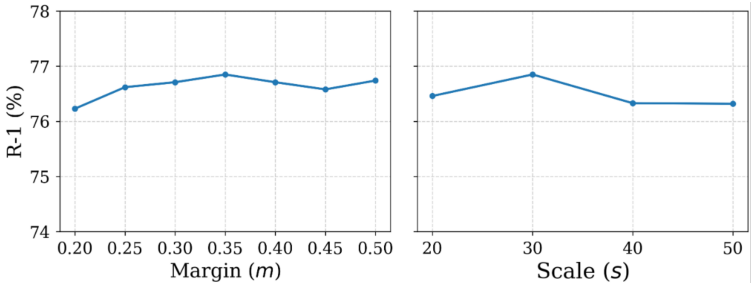


Figure 5: Hyperparameter Sensitivity Analysis on CUHK-PEDES

Analysis of NDC. In Figure 4-(a) we see for K=25, CUHK-PEDES and ICFG-PEDES retains approximately 90% of their data. In contrast, RSTPReid retains only 60.3% of the data. This is consistent with prior work, which identifies RSTPReid as one of the noisiest TBPS benchmarks.

Our NDC module selects a pair when the corresponding image is retrieved within the Top-K samples for a given caption. We observe that at $K = 25$, the retention curve plateaus, and we choose this value for our experiments, as below this point the data retention drops, while beyond it the increase is only marginal.

In Figure 4-(b), we study how the vote consensus threshold affects downstream retrieval accuracy. We increase the consensus requirement from $v = 1$ (any single expert must rank the pair in Top-K) to $v = 4$ (all experts must agree). Across all three datasets, performance degrades monotonically as v increases. This is expected as requiring stricter consensus prunes the retained set. This ends up removing valid pairs that only a subset of experts could correctly retrieve. The best performance is consistently achieved at $v = 1$, which retains a pair if at least one expert considers it a valid match. Since the expert models are trained on diverse distributions, validation by even one expert can indicate that the pair is meaningful. In contrast, requiring agreement from all experts may discard valid cross-domain pairs when some experts fail due to distribution differences.

Taken together, these results validate both hyperparameter choices in NDC: $K = 25$ captures the saturation point of the retention–quality tradeoff, and $v = 1$ preserves the diversity of the joint dataset while still filtering genuinely noisy correspondences.

4.4 Sensitivity Analysis.

We conduct a hyperparameter sensitivity analysis on the margin (m) and scale (s) used in our DIL. m is used to set the boundary between different identity classes, and s signifies the sharpness of logits in DIL. The hyperparameters are selected based on Rank-1 accuracy on CUHK-PEDES following previous TBPS

Figure 6: Qualitative results of rank wise Top-10 retrieved images by Naive joint baseline and our proposed Scale-TBPS

methods. In Figure 5 we show that the best results are achieved with $m = 0.35$ and $s = 30$. The proposed method exhibits robustness to hyperparameter variations, with Rank-1 accuracy fluctuating by less than 1% across a wide range of values.

5 Conclusion

In this work, we propose Scale-TBPS, a unified training paradigm that enables learning a single model across multiple text-based person search datasets. We introduce a noise-aware data curation strategy to cohesively merge multiple datasets, along with a discriminative ID learning framework designed to handle the rapid growth in unique identities that arises with dataset merging. Comprehensive experiments on CUHK-PEDES, RSTPReid, ICFG-PEDES, and IIITD-20K demonstrate that a single Scale-TBPS model consistently matches or surpasses dataset-specific state-of-the-art methods, as well as naive joint training, under diverse distribution shifts. Furthermore, we demonstrate zero-shot robustness on UFine6926, highlighting strong generalization to previously unseen data. Additionally, we observe that test time normalization significantly boosts the performance of Scale-TBPS compared to other methods.

References

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 555–563. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/62. URL https://doi.org/10.24963/ijcai.2023/62. Main Track.

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *AAAI*, volume 38, pp. 465–473, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22571–22582, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1257. URL https://aclanthology.org/2024.emnlp-main.1257/.

Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17840–17852, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2787–2797, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, volume 34, pp. 9694–9705, 2021.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pp. 1970–1979, 2017.

Xulin Li, Yan Lu, Bin Liu, Jiaze Li, Qinhong Yang, Tao Gong, Qi Chu, Mang Ye, and Nenghai Yu. Towards anytime retrieval: a benchmark for anytime person re-identification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1467–1475, 2025.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, pp. 27197–27206, 2024.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14390–14399, 2025.

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

Vasileios Sevetlidis, George Pavlidis, and Antonios Gasteratos. Angular regularization for positive-unlabeled learning on the hypersphere. *Transactions on Machine Learning Research*, 2025.

A Venkata Subramanyam, Vibhu Dubey, Niranjan Sundararajan, and Brejesh Lall. Dense captioning for text-image reid. In *ICVGIP*, pp. 1–8, 2023.

Jintao Sun, Hao Fei, Gangyi Ding, and Zhedong Zheng. From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search. In *Proceedings of the ACM on Web Conference 2025*, pp. 2341–2351, 2025.

Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17127–17137, 2024.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59 (2):64–73, 2016.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yue Wang, Shuai Xu, Xuelin Zhu, and Yicong Li. Msci: Addressing clip's inherent limitations for compositional zero-shot learning. *arXiv preprint arXiv:2505.10289*, 2025.

Tian Xia, Shaodan Zhai, and Shaojun Wang. Plackett-luce model for learning-to-rank task. *arXiv preprint arXiv:1909.06722*, 2019.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACM MM*, pp. 4492–4501, 2023.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 209–217, 2021.

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

Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22010–22019, 2024b.

A Datasets

We use total five datasets for conducting various experiments. CUHK-PEDES Li et al. (2017) is the most widely used benchmark for text-based person search. It contains 40,206 images paired with 80,440 textual descriptions from 13,003 identities. The dataset is split into 34,054 images and 68,126 texts from 11,003 identities for training, 3,078 images and 6,158 texts from 1,000 identities for validation, and 3,074 images and 6,156 texts from 1,000 identities for testing. The average caption length is 23 words.

ICFG-PEDES Ding et al. (2021) is a more recent dataset comprising 54,522 images from 4,102 identities. Each image is annotated with a single textual description. The training set contains 34,674 images from 3,102 identities, while the test set includes 19,848 images from 1,000 identities. The average text length is 37 words.

RSTPReid Zhu et al. (2021) is designed to better reflect real-world scenarios. It consists of 20,505 images from 4,101 identities, with each identity having five images captured from different camera views. Each image is annotated with two textual descriptions, and each description contains at least 23 words. The dataset is split into 3,701 identities for training, and 200 identities each for validation and testing.

IIITD-20K Subramanyam et al. (2023) consists of 15000 image–text pairs for training and 2500 pairs each for validation and testing, with no identity overlap across the splits. 20,000 total unique ids are there for with 15000 ids in training. Average caption length is 35.9 words.

UFine6926 Zuo et al. (2024b) is a recently introduced high-quality fine-grained text–image re-identification dataset containing 26,206 images from 6,926 identities. Each image is annotated with two highly detailed textual descriptions. The average caption length is 80.8 words, which is substantially longer than that of commonly used coarse-grained datasets. The official data split uses 4,926 identities for training and 2,000 identities for testing. We only use the test set for showing zero-shot results.

We combine the four training sets from these four datasets and ensure they have unique IDs. We do not change the test set configurations. All our results are evaluated on the test sets of these datasets individually.

B Implementation Details

We follow RDE's Qin et al. (2024) training configuration. We choose RDE models trained on CUHK-PEDES, ICFG-PEDES, and RSTPReid as our expert models. In order to perform zero-shot testing on UFine6926, following ICL Qin et al. (2025), we choose a context length of 168 to accommodate its longer text length. However, other datasets have a smaller text length and a context length of 77, like in RDE, is sufficient.

C Analysis

In this section we conduct additional experiments to verify effectiveness of our choices in method design.

C.1 Noise-aware Unified Dataset Construction (NDC)

Our NDC module operates using relative ranking rather than absolute similarity thresholds. While NDC does not explicitly optimize a Plackett–Luce (PL) objective, its design is conceptually aligned with the central principle of the PL model. Candidate relevance is determined relative to competing candidates rather than through independent score evaluations. Motivated by this connection, we provide a probabilistic interpretation of NDC through the lens of the Plackett–Luce (PL) model Zuo et al. (2024a); Luce (1959) and contrast it with hard similarity-based formulations.

PL model defines a distribution over permutations based on latent utility scores. Given a set of items with utilities $w = [w_1, \ldots, w_N]$, the probability of an ordered list $(i_1, \ldots, i_N)$ is:

$$p((i_1, \ldots, i_N); w) = \prod_{j=1}^{N} \frac{\exp(w_{i_j})}{\sum_{l=j}^{N} \exp(w_{i_l})}. \tag{18}$$

Method	CUHK-PEDES				RSTPReid				ICFG-PEDES			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Image-to-Text												
NDC (I2T)	75.65	90.24	94.23	68.02	70.20	87.95	91.60	54.81	67.14	81.77	86.80	41.03
NDC + DIL	76.67	90.35	93.97	70.03	66.90	85.60	90.80	55.12	68.13	82.18	86.94	45.26
NDC + DIL + NNN	77.40	90.72	94.15	70.75	68.90	86.30	91.00	56.82	**69.05**	82.86	87.39	**46.60**
Text-to-Image												
NDC (T2I)	75.97	90.27	94.38	68.02	70.15	87.25	92.10	54.75	66.42	81.74	86.71	40.92
NDC + DIL	76.85	90.22	94.23	70.05	68.90	87.00	91.65	56.00	68.22	82.09	86.80	45.16
NDC + DIL + NNN	**77.91**	**91.07**	**94.49**	**70.88**	**71.70**	**87.40**	**91.95**	**58.38**	68.24	**83.00**	**87.62**	45.68

Table 5: Ablation study for filtering strategies of NDC module . T2I variants demonstrate better performance than I2T variants

The PL model maps scores to a probabilistic distribution over utilities, capturing relative relationships among items rather than relying on independent, absolute scores. In Equation 3, we showed that our descending ranking system leverages global similarity across all candidates. This can be formally expressed in probabilistic terms as follows.

We model the probability over all candidate pairs for a given query i and similarity s as:

$$P(i, j \mid s) = \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})}. \tag{19}$$

The probability of the correct match is given by $P(i, i \mid s)$. This formulation induces a distribution over all candidates $\{j\}$, where ranking is obtained by sorting $P(i, j \mid s)$ in descending order.

Let π denote the permutation such that:

$$P(i, \pi_1 \mid s) \geq P(i, \pi_2 \mid s) \geq \cdots \geq P(i, \pi_N \mid s). \tag{20}$$

The Top-K set is defined as $\{\pi_1, \ldots, \pi_K\}$. Our objective is to ensure $P(i, i \mid s)$ relative to all competing $P(i, j \mid s)$, and verifying the ground-truth pair (i, i) is in the given set or not:

$$(i, i) \in \{(i, \pi_1), \ldots, (i, \pi_K)\}. \tag{21}$$

PL method has been empirically validated across other methods to be more preferrable than hard thresholding methods Xia et al. (2019). However, a limitation of the PL-based approach is that it can become computationally intensive with large-scale data. To address this, we employ it as a principle for data filtering during preprocessing only. .

We analyze the impact of the Noise-Aware Unified Dataset Curation (NDC) module across four major benchmarks: CUHK-PEDES, ICFG-PEDES, RSTPReid, and IIITD-20K. For all filtering experiments, we set the retrieval threshold to $K = 25$ using an ensemble of $p = 3$ pretrained models. Our unified curated dataset has 30,957 IDs and 1,41,631 text-to-image pairs.

We experiment with Image-to-Text (I2T) ranking and Text-to-Image (T2I) ranking based filtering. In I2T ranking, each image is used as a query to retrieve the most relevant textual descriptions, and a pair is retained if its corresponding caption appears within the Top-K retrieved texts. Conversely, in T2I ranking, each text serves as a query to retrieve the most relevant images, and a pair is retained if its corresponding image is ranked within the Top-K results. These two strategies offer complementary perspectives on cross-modal alignment quality and allow us to assess how directional retrieval affects noise-aware dataset curation. Table 5 indicates that the Text-to-Image (T2I) filtering strategy is better. On CUHK-PEDES and RSTPReid T2I filtering shows better results. Therefore, we choose the T2I direction.

Figures 7 and 8 illustrate the data retention rates and the distribution of model consensus or votes. As shown in Figures 7(a) and 8(a) both methods effectively identify easy or high-confidence pairs in CUHK-PEDES

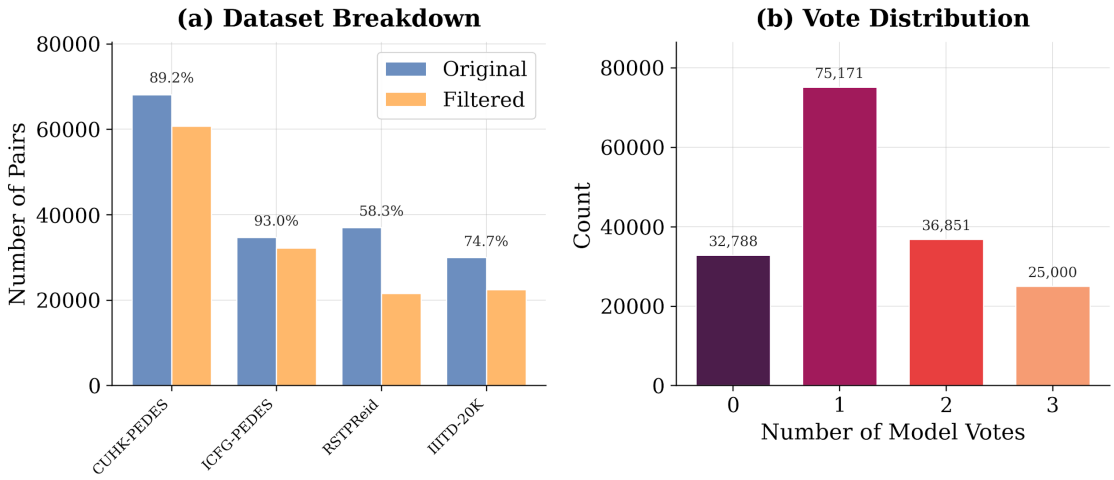


Figure 7: Image-to-Text (I2T) Filtering Statistics ($K = 25$).

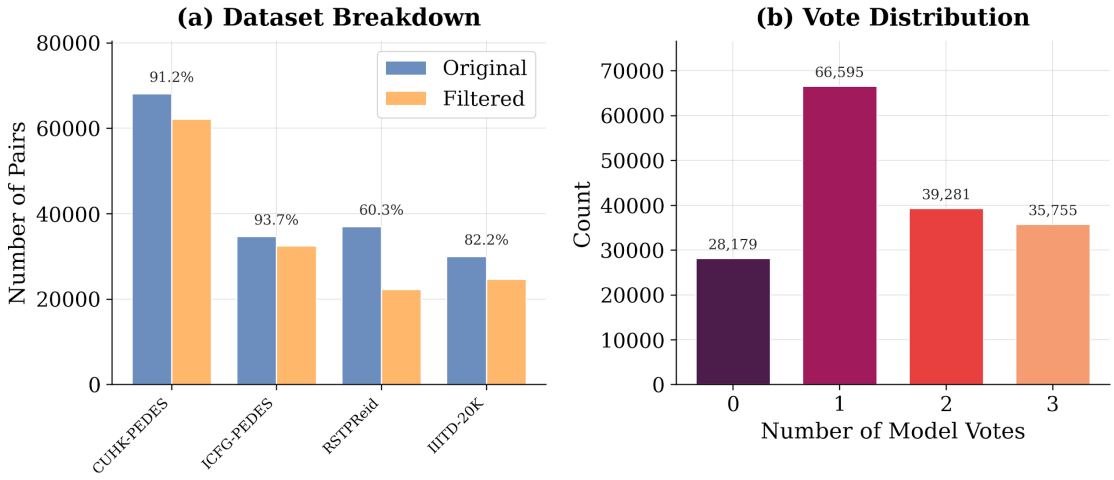


Figure 8: Text-to-Image (T2I) Filtering Statistics ($K = 25$).

and ICFG-PEDES, retaining over 89% of the data. However, on more challenging or noisier datasets like RSTPReid, filtration becomes more selective. Specifically, the T2I direction appears more robust, retaining a higher proportion of the original pairs compared to I2T.

Figures 7(b) and 8(b) show the ensemble consensus or votes. For a text-image pair, vote signifies the number of models that retrieve the text-image pair to remain in the filtered dataset. We observe that T2I filtration results in fewer rejected pairs (approximately 28,179 pairs with 0 votes) compared to I2T filtering (approximately 32,788 pairs with 0 votes). Based on performance on gains and data retention rates we choose T2I filtration for NDC.

C.2 Discriminative ID learning Framework (DIL)

The DIL module employs Multimodal Angular Identity Loss (MA-ID) to enhance the discriminative power of person identities across the unified dataset. In this loss, the margin (m) is a key hyperparameter. We investigate the sensitivity of this margin parameter in Table 6, reporting results across all metrics on CUHK-PEDES.

We track Rank-1 accuracy and observe the best performance at $m = 0.35$. Across the margin range from 0.20 to 0.50, Rank-1 accuracy remains stable. A similar trend is observed for Rank-5 and Rank-10. These results indicate that our method is robust to different choices of the margin parameter m.

Table 6: Sensitivity analysis of m on CUHK-PEDES.

Margin (m)	R-1	R-5	R-10	mAP
0.20	76.23	90.45	94.18	69.15
0.25	76.62	90.12	94.23	69.55
0.30	76.71	90.50	94.23	69.77
0.35	**76.85**	90.22	94.23	70.05
0.40	76.71	90.25	94.10	70.05
0.45	76.58	90.24	94.27	70.29
0.50	76.74	90.22	**94.35**	**70.53**

C.3 Qualitative Analysis of Filtered Noisy Samples

The effectiveness of the Noise-Aware Unified Dataset Curation (NDC) module is evidenced by the specific types of samples it identifies as noisy and removes from the training set. Figure 9 shows representative rejected pairs, where red text highlights specific inconsistencies. As shown in the figure, the noise identified by our ensemble of pretrained experts falls into three distinct categories:

- **Attribute Contradictions:** The dataset contains significant semantic conflicts where the caption describes attributes diametrically opposed to the visual evidence. For example, the top-left and middle-right pairs describe the subject as “A woman," despite the images clearly depicting male subjects. Similarly, colors are often mislabeled, such as a “purple jacket" described for a subject wearing blue/gray.
- **Textual Corruption and Typos:** Large-scale web-scraped or crowd-sourced annotations often contain typographical errors that confuse semantic alignment. A clear example is observed in the middle-left pair, where the caption describes “gay shoes" a likely typographical error for “gray shoes" which creates a token that lacks meaningful visual grounding.
- **Subjectivity and Uncertainty:** Some captions fail to describe visual attributes entirely. The bottom-right example explicitly states the annotator is “unsure if this person is a man or a woman," while the bottom-left caption focuses on a narrative about a “crowded mall" and “safety" rather than the person’s appearance.

By systematically filtering these semantically unreliable pairs, Scale-TBPS prevents the model from overfitting to false correspondences, leading to more robust feature representations.

C.4 t-SNE Visualization

To qualitatively evaluate the impact of our proposed method on representation learning, we visualize the joint embedding space using t-SNE.

In Figure 10, we visualize the learned feature distributions using t-SNE Maaten & Hinton (2008).The MA-ID representation exhibits more uniformly separated clusters compared to standard ID loss. While ID loss produces compact intra-class groupings, several identity clusters remain close to one another, which can become problematic as the number of identities increases. In contrast, MA-ID encourages stronger inter-class separation, resulting in better-structured feature distributions. Consistent with this observation, MA-ID achieves a silhouette score of 0.0762 compared to 0.0688 for ID loss, representing an approximately 10% relative improvement. These findings provide qualitative and quantitative evidence that MA-ID scales more effectively to large identity spaces.

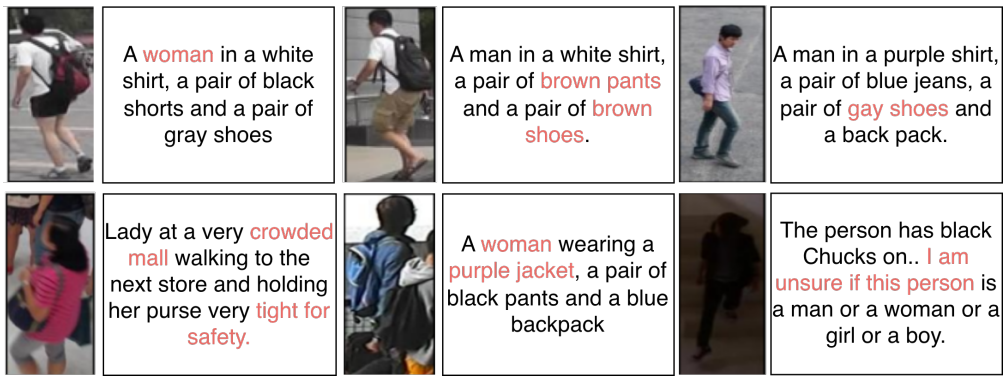


Figure 9: Qualitative examples of noisy image-caption pairs identified and removed by our NDC module. The samples highlight three common noise types: gender(top-left)/attribute contradictions(middle row) , typographical errors (top-right), and non-visual subjective uncertainty (bottom right). Text highlighted in red indicates the specific source of noise.

C.5 t-SNE visualization for MA-ID vs ID loss

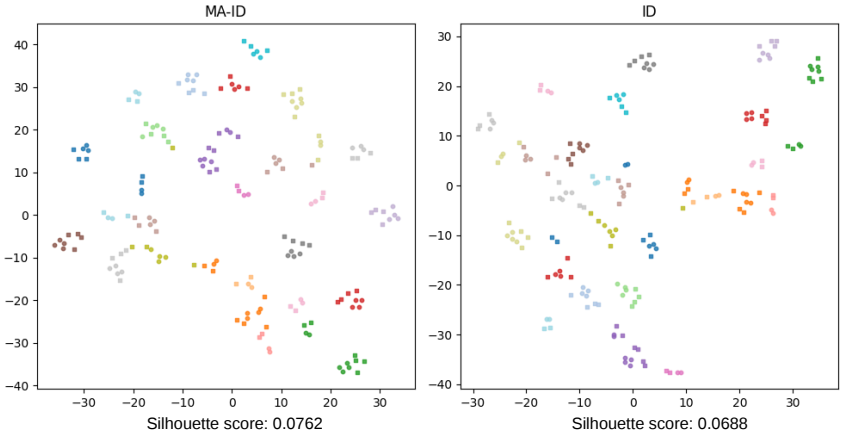


Figure 10: t-SNE visualization comparing naive ID supervised training and the proposed DIL. Different color signifies different IDs. Our method exhibits tighter intra-class clustering and improved inter-class separation

C.6 Qualitative Analysis

In Figure 11 We provide additional qualitative retrieval examples. In the first example, the naive joint baseline fails to capture all attributes in the query and incorrectly retrieves a person wearing blue pants despite the caption explicitly specifying black clothing. In the second and third examples, both models correctly attend to salient contextual cues such as bicycle and with a man. However, the naive baseline struggles to distinguish finer-grained identity attributes, confusing a child with an adult male. In the fourth example, the baseline incorrectly associates a scooter with a bicycle, indicating weaker attribute-level discrimination. These observations suggest that the proposed unified training strategy learns more discriminative identity-aware representations, leading to retrievals that are more faithful to the complete query description.

Figure 11: More Qualitative results