DRAGE: Dual-Gaps Robust Alignment for Text-Based Person Search

Anonymous ACL submission

Abstract

Text-based Person Search (TBPS) faces two critical and intertwined challenges: the semantic gap caused by noisy image-text correspondences, and the modality gap stemming from the structural heterogeneity between dense visual features and sparse textual attributes. Crucially, we identify that this heterogeneity leads to intra-modal disorganization, where embeddings—particularly sparse text representations—lack internal structure, hindering robust alignment. To address these limitations, we propose Dual-Gaps Robust Aligned General Embedding (DRAGE) , a unified framework comprising two synergistic mechanisms. First, Semantic Embedding Reliability Division (SERD) dynamically partitions training data into reliable and noisy subsets using a Beta Mixture Model, providing clean supervision to mitigate the semantic gap. Second, Semantic Intra-Modality Regularization (SIMR) explicitly addresses intra-modal disorganization by enforcing semantic-aware distance constraints within each modality *before* cross-modal alignment. This transforms disorganized embeddings into coherent semantic clusters, establishing a stable geometric foundation for matching. Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid demonstrate that DRAGE significantly outperforms state-of-the-art methods, achieving **78.58%** Rank-1 accuracy on CUHK-PEDES and exhibiting superior robustness in cross-domain settings.

1 Introduction

Text-based Person Search (TBPS) aims to retrieve a specific person from a large image gallery using natural language descriptions (Lei et al., 2022; Miech et al., 2021; Sun et al., 2021). This task has critical applications ranging from video surveillance to social media analysis (Li et al., 2025; Zhao et al., 2024). Despite significant progress, deploying TBPS models in real-world scenarios remains constrained by two fundamental and intertwined challenges: the semantic gap and the modality gap, as illustrated in Figure 1.

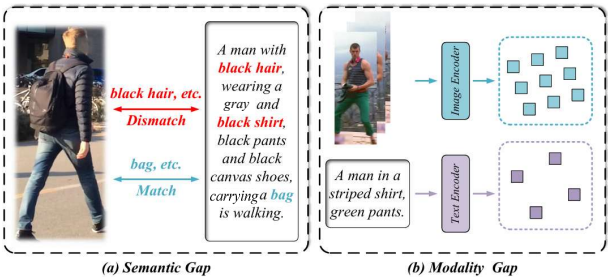


Figure 1: Illustration of the semantic gap and modality gap in Text-based Person Search. (a) **Semantic gap**: training image-text pairs often exhibit content inconsistencies due to annotation biases or visual ambiguities, resulting in noisy supervision. (b) **Modality gap**: the fundamental heterogeneity between dense visual features and sparse textual features, exacerbated by disorganized internal embedding spaces that hinder robust cross-modal alignment.

Recent research makes significant progress in TBPS. Many methods explore both global (Jiang and Ye, 2023b; Xie et al., 2025a) and local matching (Zuo et al., 2024; Park et al., 2024) strategies. Studies leverage large pre-trained models (Li et al., 2025; Liu et al., 2025b) by adapting vision-language models such as CLIP (Radford et al., 2021), or employ generative models for data augmentation through diffusion-based image synthesis (Song et al., 2024) and language-model-driven text enrichment (Tan et al., 2024; Xie et al., 2025b).Despite these advancements, existing TBPS methods remain fundamentally constrained by two critical yet rarely co-addressed challenges.

The first is the **semantic gap**, which arises from noisy and inconsistent image-text correspondences caused by real-world data variability such as pose variations, camera angles, and illumination changes.The second is the **modality gap**, stemming from a deeper structural disparity operating on two levels: (1) *cross-modal heterogeneity*—the

fundamental difference between **dense visual features** (characterized by redundant spatial information) and **sparse textual attributes** (composed of discrete, abstract tokens); and (2) *intra-modal disorganization*—the lack of structural organization within each modality's embedding space. Existing methods focus exclusively on cross-modal alignment while neglecting this internal organization. As a result, embeddings within a single modality (especially the sparse text modality) tend to be distributed chaotically, where semantically similar samples scatter and dissimilar ones cluster together. This internal disorder sets a low upper bound for the subsequent cross-modal alignment.

To address these limitations, we propose **D**ual-Gaps **R**obust **A**ligned **G**eneral **E**mbedding (DRAGE), a unified framework that systematically tackles both challenges through two synergistic mechanisms.

First, to address the *semantic gap*, we propose Semantic Embedding Reliability Division (SERD). Leveraging a Beta Mixture Model, SERD dynamically separates reliable image-text pairs from noisy ones. By prioritizing high-confidence data and suppressing noise, it ensures the model focuses on valid semantic correspondences.

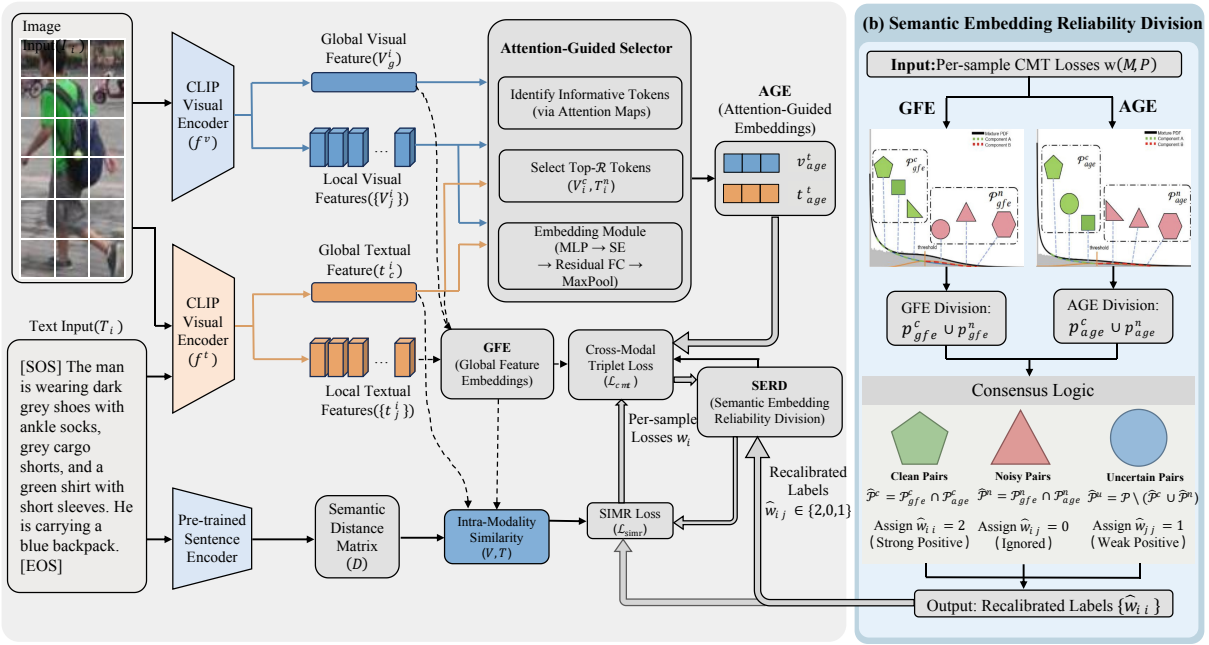
Second, to bridge the *modality gap*, DRAGE employs Semantic Intra-Modality Regularization (SIMR). This module addresses the overlooked issue of internal disorganization. Rather than immediately forcing cross-modal alignment, SIMR first organizes each modality's internal embedding structure by enforcing semantic-aware distance constraints: pulling similar samples closer while pushing dissimilar ones apart. This intra-modal organization is particularly crucial for sparse text representations, transforming disorganized embeddings into coherent semantic clusters and compensating for their inherent lack of structure. By establishing stable, well-organized foundations within each modality *before* cross-modal alignment, SIMR enables more effective and robust mapping into a unified space.

By integrating SERD and SIMR, DRAGE systematically addresses both gaps to achieve state-of-the-art performance. The main contributions of this work are:

- We propose DRAGE, a unified framework that addresses the semantic and modality gaps in TBPS. Unlike prior works that treat these challenges separately, DRAGE integrates SERD and SIMR to jointly mitigate noisy correspondences and structural heterogeneity.
- We introduce SERD, a dynamic reliability-aware mechanism. By leveraging a Beta Mixture Model to model per-sample loss distributions, SERD effectively filters noisy supervision signals caused by real-world data inconsistencies, ensuring the model learns from semantically reliable pairs.
- We propose SIMR to address the overlooked issue of intra-modal disorganization. By enforcing semantic-aware distance constraints, SIMR structures the embedding spaces of dense visual and sparse textual features *before* cross-modal alignment, providing a stable geometric foundation for matching.
- Extensive experiments on three benchmarks demonstrate that DRAGE achieves state-of-the-art performance. **78.58%** Rank-1 accuracy on CUHK-PEDES. Furthermore, the framework exhibits superior robustness in cross-domain settings, validating the effectiveness of our dual-gap mitigation strategy.

2 Related Work

Text-based Person Search. TBPS aims to match textual descriptions with images to identify specific individuals. Early TBPS approaches focus on extracting global features by mapping both image and text data into a shared embedding space (**?**Ge et al., 2019). However, these methods struggle to capture the fine-grained details necessary to differentiate similar-looking individuals. To improve accuracy, researchers develop local-matching methods (Chen et al., 2018; Guo et al., 2019; Ding et al., 2020) that explicitly associate specific regions in images with corresponding words in text descriptions. Although these methods yield more precise alignments, they incur significantly higher computational costs due to their complexity. Recent advances using pre-trained models such as CLIP (Radford et al., 2021) and ViT (Ge et al., 2019) further advance TBPS by leveraging large-scale multimodal learning to generate enhanced representations. Nevertheless, these models still encounter two key challenges: the semantic gap and the modality gap. Current methods have not fully resolved these challenges—the semantic gap results in misalignments even when text and images refer to the same person, while the



(a) Overview of our DRAGE framework

(b) Semantic Embedding Reliability Division

Figure 2: Overview of the DRAGE framework. DRAGE addresses the semantic gap through SERD, which identifies reliable image-text pairs for robust supervision, and bridges the modality gap through SIMR, which organizes each modality's internal embedding structure before cross-modal alignment. The framework integrates Attention-Guided Selector for discriminative features, Cross-Modal Triplet Loss for robust matching for alignment optimization.

modality gap persists due to fundamental differences between the two modalities. Consequently, more sophisticated techniques are required to systematically address these gaps and enhance TBPS performance in real-world applications.

Modality Gap. Early work such as CMPM (Zhang and Lu, 2018a) tackles this challenge by aligning global features across modalities. More recent efforts enhance this alignment through local attribute modeling (Yan et al., 2022a; Zuo et al., 2024) and refined loss functions (Jiang and Ye, 2023a; Qin et al., 2024), yielding better semantic correspondence across views. However, these methods typically treat each modality's embedding space as a black box—focusing only on cross-modal pairing while largely ignoring the internal structure of each modality. As a result, embeddings within the same modality can remain disorganized, leading to unstable representations that hinder alignment consistency, especially in challenging or noisy scenarios. In contrast, our study explicitly addresses this gap by proposing a dedicated intra-modal regularization strategy that enforces meaningful distance constraints within each modality, stabilizing the feature space for robust cross-modal alignment.

3 Method

3.1 Preliminary

The goal of TBPS is to retrieve pedestrian images matching a textual description. Let the image set be $\mathcal{V} = \{I_i\}_{i=1}^N$ and the text set be $\mathcal{T} = \{T_i\}_{i=1}^N$. We aim to learn a joint embedding space where matched pairs (I_i, T_i) are close. We employ CLIP (Radford et al., 2021) as the backbone, consisting of a visual encoder f^v and a textual encoder f^t. For an image I_i, f^v outputs features $\boldsymbol{V}_i = \{\boldsymbol{v}_g^i, \boldsymbol{v}_1^i, \dots\}^\top$, where $\boldsymbol{v}_g^i$ is the global $[CLS]$ token. For a text T_i, f^t outputs $\boldsymbol{T}_i = \{\boldsymbol{t}_s^i, \dots, \boldsymbol{t}_e^i\}^\top$, where $\boldsymbol{t}_e^i$ represents the global $[EOS]$ feature.

3.2 Attention-Guided Selector

Global features often miss fine-grained details. To address this, we use Attention-Guided Selector to extract discriminative local tokens. First, we select the top-$\mathcal{R}$ most informative local tokens from last attention layer, denoted as $\hat{\boldsymbol{V}}_i^a$ and $\hat{\boldsymbol{T}}_i^a$. Then, we project them into compact vectors, termed **A**ttention-**G**uided **E**mbedding (AGE) vectors:

$$\boldsymbol{v}_{\text{age}}^i = \text{MaxPool}(\text{SE}(\text{MLP}(\hat{\boldsymbol{V}}_i^a)) + \text{FC}(\hat{\boldsymbol{V}}_i^a)), \quad (1)$$

$$\boldsymbol{t}_{\text{age}}^i = \text{MaxPool}(\text{SE}(\text{MLP}(\hat{\boldsymbol{T}}_i^a)) + \text{FC}(\hat{\boldsymbol{T}}_i^a)). \quad (2)$$

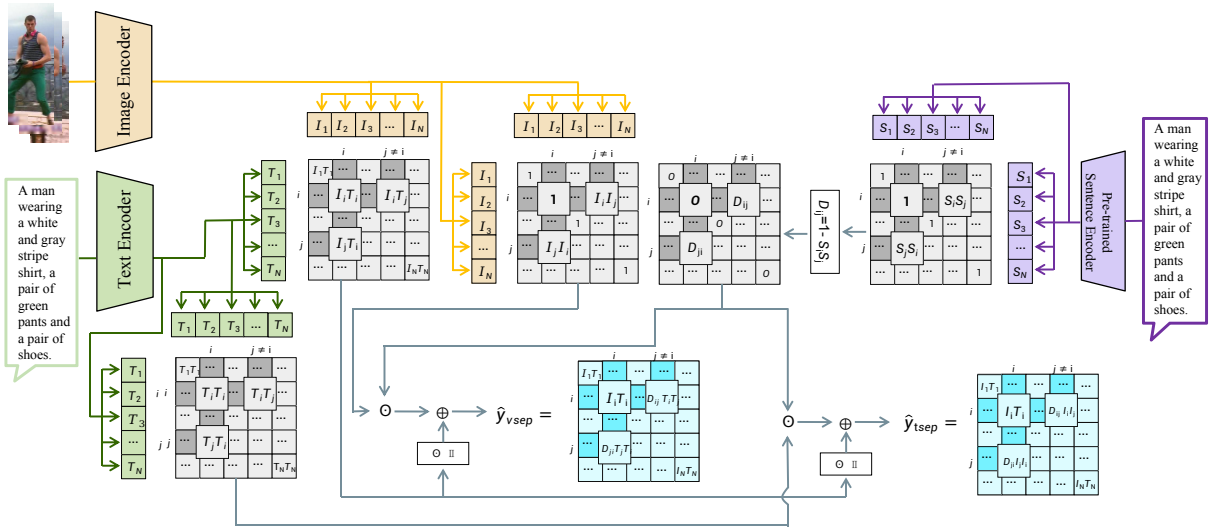


Figure 3: Overview of Semantic Intra-Modality Regularization (SIMR) method. This approach is implemented through vision-to-vision (v2v) and text-to-text (t2t) contrastive losses, adjusting embedding distances within each modality based on semantic distances between samples. Note: The input embeddings for SIMR are purely intra-modal (image-to-image or text-to-text).

The AGE similarity S^a_{ij} is the cosine similarity between these vectors, complementing the global similarity S^g_{ij}.

3.3 Semantic Embedding Reliability Division

Real-world data often contains noisy image-text pairs (e.g., mismatched descriptions), causing the *semantic gap*. SERD identifies and filters these pairs to ensure reliable supervision.

We leverage the "small-loss trick" (Arpit et al., 2017), observing that clean data exhibits lower loss values during early training. We first compute the per-sample loss $\ell_i = \mathcal{L}_{\text{cmt}}(I_i, T_i)$. To satisfy the input requirement of the Beta Mixture Model (BMM), we apply min-max normalization to map losses to $[0, 1]$:

$$\bar{\ell}_i = \frac{\ell_i - \min(\ell)}{\max(\ell) - \min(\ell)}. \tag{3}$$

We then fit a two-component BMM to $\bar{\ell}_i$. Based on the posterior probability $p(\text{clean}|\bar{\ell}_i)$ and a threshold $\delta = 0.6$, we first divide pairs into Reliable and Unreliable sets for both global and AGE views. By computing the consensus between these views, we assign a **recalibrated** reliability weight w_{ii} to each pair (I_i, T_i):

$$w_{ii} = \begin{cases} 2, & \text{if deemed Reliable (Strong Positive),} \\ 1, & \text{if deemed Uncertain (Weak Positive),} \\ 0, & \text{if deemed Unreliable (Ignored).} \end{cases} \tag{4}$$

For negative pairs $(i \neq j)$, we set $w_{ij} = 1$. This ensures the model learns only from semantically valid correspondences.

3.4 Cross-Modal Triplet Loss

To learn robust features against noisy correspondence, we introduce a reliability-aware Cross-Modal Triplet (CMT) loss. Unlike standard contrastive losses, we incorporate the reliability weights w_{ij} to prevent noisy positives from misguiding the gradient optimization.

Formally, the matching probability p_{ij} between image I_i and text T_j is defined as:

$$p_{ij} = \frac{w_{ij} \exp(S(I_i, T_j)/\tau)}{\sum_{k=1}^{N} w_{ik} \exp(S(I_i, T_k)/\tau)}, \tag{5}$$

where τ denotes the temperature parameter. Crucially, the inclusion of w_{ik} in the denominator ensures that p_{ij} remains a valid probability distribution over the batch.

Based on this, the reliability-aware CMT loss from image to text is formulated as:

$$\mathcal{L}_{\text{i2t}} = \frac{1}{N} \sum_{i=1}^{N} \Bigg[\alpha - \hat{S}_{i,i} + \tau \log \left(\sum_{j=1}^{N} w_{ij} \exp\left(\frac{S(I_i, T_j)}{\tau} \right) \right) \Bigg]_+, \tag{6}$$

where $[\cdot]_+ = \max(\cdot, 0)$ denotes the hinge function, and α is the margin hyperparameter. The

term $\hat{S}_{i,i} = \sum_j p_{ij} S(I_i, T_j)$ represents the expected similarity score of reliable positive pairs. This formulation encourages the model to maximize the similarity of reliable positive pairs while suppressing the scores of negative ones relative to the margin. The total CMT loss integrates both image-to-text and text-to-image directions across both Global and AGE feature views:

$$\mathcal{L}_{\text{cmt}} = \mathcal{L}_{\text{cmt}}^{\text{gfe}} + \mathcal{L}_{\text{cmt}}^{\text{age}}. \tag{7}$$

3.5 Semantic Intra-Modality Regularization

As illustrated in Figure 3, the *modality gap* is exacerbated by *intra-modal disorganization*. SIMR addresses this by enforcing semantic constraints *within* each modality to organize the embedding space before cross-modal alignment.

3.5.1 Target Topology Construction

We employ a pre-trained **Sentence Encoder** (e.g., MPNet) to define the ground-truth semantic topology, as shown in the purple blocks of Figure 3. Let $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^\top \in \mathbb{R}^{N \times d_s}$ denote the batch of sentence embeddings derived from the raw descriptions. We compute the target semantic distance matrix $\mathcal{D}$ by performing strictly row-wise normalization to ensure valid cosine distances:

$$\mathcal{S}_{ij}^s = \frac{\boldsymbol{s}_i \cdot \boldsymbol{s}_j}{\|\boldsymbol{s}_i\| \|\boldsymbol{s}_j\|}, \quad \mathcal{D}_{ij} = 1 - \mathcal{S}_{ij}^s. \tag{8}$$

3.5.2 Intra-Modal Alignment

Next, we align the internal structure of our trainable features to this target $\mathcal{D}$. To maintain consistency with the global features defined in Sec. 3.1, let $\boldsymbol{I} = [v_g^1, \ldots, v_g^N]^\top$ and $\boldsymbol{T} = [t_e^1, \ldots, t_e^N]^\top$ denote the batch matrices of global visual and textual embeddings, corresponding to the yellow and orange blocks in Figure 3, respectively.

We define calibrated logits $\hat{y}$ by combining the cross-modal matching confidence (Bias) with intra-modal topological constraints (Topology):

$$\hat{y}_v = s \cdot \left[\underbrace{\text{diag}(\mathcal{S}^{i2t}) \cdot \mathbf{1}^\top}_{\text{Alignment Bias}} + \underbrace{\mathcal{S}^{i2i} \odot (1 - \mathcal{D})}_{\text{Topology Constraint}} \right], \tag{9}$$

$$\hat{y}_t = s \cdot \left[\text{diag}(\mathcal{S}^{t2i}) \cdot \mathbf{1}^\top + \mathcal{S}^{t2t} \odot (1 - \mathcal{D}) \right], \tag{10}$$

where s is a scaling factor. Here, $\mathcal{S}^{i2i} = \bar{\boldsymbol{I}}\bar{\boldsymbol{I}}^\top$ and $\mathcal{S}^{t2t} = \bar{\boldsymbol{T}}\bar{\boldsymbol{T}}^\top$ represent the intra-modal similarity matrices (the $I \cdot I$ and $T \cdot T$ grids in Figure 3), while $\mathcal{S}^{i2t} = \bar{\boldsymbol{I}}\bar{\boldsymbol{T}}^\top$ is the cross-modal similarity matrix.

The term $\text{diag}(\mathcal{S}^{i2t}) \cdot \mathbf{1}^\top$ involves **broadcasting** the diagonal elements (scores of positive pairs) across the rows. This acts as an alignment-aware bias, ensuring that the intra-modal clustering remains consistent with the cross-modal matching objective. The second term, $\mathcal{S}^{i2i} \odot (1 - \mathcal{D})$, enforces the target topology: features with small semantic distance ($\mathcal{D}_{ij} \approx 0$) are encouraged to have high similarity scores.

Finally, the SIMR loss is formulated as a classification task using the standard Cross-Entropy loss $H(\cdot, \cdot)$:

$$\mathcal{L}_{\text{simr}} = H(\hat{y}_v, Y_{\text{id}}) + H(\hat{y}_t, Y_{\text{id}}), \tag{11}$$

where Y_{id} represents the ground-truth identity labels.

3.6 Overall Optimization Objective

Our framework integrates SERD, CMT, and SIMR to address the dual gaps systematically. The total optimization objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simr}} + \mathcal{L}_{\text{cmt}}. \tag{12}$$

By jointly optimizing these losses, DRAGE mitigates the semantic gap via reliable supervision (SERD) and bridges the modality gap through structured intra-modal regularization (SIMR).

4 Experiments

4.1 Datasets and Performance Measurements

Datasets. We evaluate our method on three widely-used TBPS datasets: CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021a), and RSTPReid (Zhu et al., 2021a). CUHK-PEDES is a dataset containing over 40,000 images of 13,003 identities. ICFG-PEDES includes over 54,000 images of 4,102 identities, and RSTPReid consists of over 20,000 images of 4,101 identities. A detailed statistical breakdown of each dataset is provided in the **supplementary material**.

Evaluation Metrics. In accordance with standard practices, search performance is assessed using the metrics Rank-k (with k set to 1, 5, and 10), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) (Ye et al., 2022). Higher values of these metrics signify enhanced search performance.

4.2 Implementation Details.

Consistent with previous studies (Jiang and Ye, 2023a; Li et al., 2023; Liu et al., 2025a), DRAGE

Table 1: Performance comparisons on the CUHK-PEDES dataset. The best results are in **bold**.

Methods	Ref.	Image Enc.	Text Enc.	R-1	R-5	R-10	mAP	mINP
Lapscore (Wu et al., 2021)	ICCV'21	RN50	BERT	63.40	-	87.80	-	-
AXM-Net (Farooq et al., 2022)	ACMMM'22	RN50	BERT	64.44	80.52	86.77	58.73	-
LGUR (Shao et al., 2022)	ACMMM'22	DeiT-Small	BERT	65.25	83.12	89.00	-	-
DCEL (Qin et al., 2022)	ACMMM'22	CLIP-ViT	CLIP-Xformer	71.36	88.11	92.48	64.25	48.26
CFine (Yan et al., 2022a)	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
IRRA (Jiang and Ye, 2023b)	CVPR'23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13	50.24
PBSL (Shen et al., 2023)	ACMMM'23	RN50	BERT	65.32	83.81	89.26	-	-
BEAT(Ma et al., 2023)	ACMMM'23	RN101	BERT	65.61	83.45	89.54	-	-
LCR^{2}S (Yan et al., 2023)	ACMMM'23	RN50	TextCNN	67.36	84.19	89.62	59.24	
DCEL (Li et al., 2023)	ACMMM'23	CLIP-ViT	CLIP-Xformer	75.02	90.89	94.52	-	-
APTM (Yang et al., 2023)	ACMMM'23	Swin-Transformer	Bert	76.53	90.04	94.15	66.91	
UniPT (Shao et al., 2023)	ICCV'23	CLIP-ViT	CLIP-Xformer	68.50	84.67	-	-	-
RaSa (Bai et al., 2023)	IJCAI'23	ALBEF	ALBEF	76.51	90.29	94.25	69.38	
SEN-XL (Liu et al., 2025a)	Neural Networks'25	CLIP-ViT	CLIP-Xformer	76.64	91.33	94.66	69.19	53.88
TBPS (Cao et al., 2024)	AAAI'24	CLIP-ViT	CLIP-Xformer	73.54	88.19	92.35	65.38	-
DP (Song et al., 2024)	AAAI'24	CLIP-ViT	CLIP-Xformer	75.66	90.59	94.07	66.58	-
IRRA+IRLT (Liu et al., 2024)	AAAI'24	CLIP-ViT	CLIP-Xformer	74.46	90.19	94.01	-	-
UASA (Zhao et al., 2024)	AAAI'24	CLIP-ViT	CLIP-Xformer	74.25	89.83	93.58	66.15	-
RDE (Qin et al., 2024)	CVPR'24	CLIP-ViT	CLIP-Xformer	75.94	90.14	94.12	67.56	51.44
CFAM (Zuo et al., 2024)	CVPR'24	CLIP-ViT	CLIP-Xformer	75.60	90.53	-	67.27	-
MGRL (Lv et al., 2024)	ICASSP'24	CLIP-ViT	CLIP-Xformer	73.91	90.68	-	67.28	-
DM-Adapter (Liu et al., 2025b)	AAAI'25	CLIP-ViT	CLIP-Xformer	72.17	88.74	92.85	64.33	-
OCDL (Li et al., 2025)	ICASSP'25	CLIP-ViT	CLIP-Xformer	75.10	89.43	-	68.18	-
HAM(1.0 M) (Jiang et al., 2025)	CVPR'25	CLIP-ViT	CLIP-Xformer	77.71	91.42	94.57	69.68	-
DRAGE (Ours)	-	CLIP-ViT	CLIP-Xformer	**78.58**	**91.68**	**95.19**	**70.68**	**53.89**

utilizes CLIP-ViT as the image encoder, the CLIP text transformer as the text encoder, and in addition, all-mpnet-base-v2 as the sentence encoder. Standard data augmentation techniques, including random horizontal flipping, random cropping, and random erasing, are applied. An Attention-Guided Selector with a sampling ratio of 0.4 is employed. The images are resized to 384×128 pixels, and text sequences are truncated to 77 tokens. Training is conducted for 60 epochs using the Adam optimizer, with an initial learning rate of 1×10^{-6}, which decays according to a cosine schedule after a 2-epoch warm-up. Furthermore, the α parameter in CMT is set to 0.1, and the temperature τ is set to 0.015. All experiments are carried out on a single NVIDIA RTX 4090 GPU.

4.3 Comparison with State-of-the-Art

In this section, we compare our method with current sota approaches on widely-used datasets.

Performance Comparisons on CUHK-PEDES. Table 1 presents a comprehensive comparison of different methods on the CUHK-PEDES dataset, a well-established benchmark for Text-Based Person Search (TBPS). Our method, DRAGE, demonstrates superior performance across all key metrics, achieving **78.58%** in Rank-1 accuracy, 91.68% in Rank-5, and 95.19% in Rank-10. While several state-of-the-art methods, including DCEL, APTM, and RaSa, also yield competitive results, DRAGE's performance significantly surpasses theirs, underscoring its enhanced capability on this benchmark. Crucially, these results are achieved without employing post-processing techniques like re-ranking (e.g., ICL (Qin et al., 2025) and MTI (Xie et al., 2025b)) or leveraging external data sources, unlike competing methods such as APTM and HAM.

Performance Comparisons on ICFG-PEDES. We present the results on the ICFG-PEDES dataset in Table 4. Our method achieves a Rank-1 accuracy of 69.63%, a Rank-5 accuracy of 83.40%, and a mAP of 42.08%. Compared to the existing methods, our approach outperforms several recent models, including RaSa and TBPS, which show perfor-

mance gains of up to 4.56% on Rank-1 and 4.67% on mAP. Despite the relatively low mINP score of 8.97%, which reflects the challenge in identifying the hardest samples, our method demonstrates solid retrieval performance.

Performance Comparisons on RSTPReid. We report results on the RSTPReid dataset in Table 5. Our method achieves a Rank-1 accuracy of 67.55%, Rank-5 accuracy of 85.00%, and a mAP of 52.88%, significantly outperforming other methods, including TBPS and CFAM, by a considerable margin. In comparison to the global-matching method IVT (Shu et al., 2022), our approach shows an improvement of 13.5% on Rank-1, 11.3% on Rank-5, and 9.4% on Rank-10. Additionally, when compared to the local-matching method Cfine (Yan et al., 2022a), we achieve a notable improvement in performance with gains of up to 9.65 on Rank-1.

Table 2: Performance comparisons on the ICFG-PEDES dataset. The best results are in **bold**.

Methods	R-1	R-5	R-10	mAP	mINP
Dual Path (Zheng et al., 2020)	38.99	59.44	68.41	-	-
CMPM/C (Zhang and Lu, 2018b)	43.51	65.44	74.26	-	-
ViTAA (Wang et al., 2020)	50.98	68.79	75.78	-	-
SSAN (Ding et al., 2021b)	54.23	72.63	79.53	-	-
IVT (Shu et al., 2022)	56.04	73.60	80.22	-	-
ISANet (Yan et al., 2022b)	57.73	75.42	81.72	-	-
CFine (Yan et al., 2022a)	60.83	76.55	82.42	-	-
IRRA (Jiang and Ye, 2023b)	63.46	80.25	85.82	38.06	7.93
BiLMa (Fujii and Tarashima, 2023)	63.83	80.15	85.74	38.26	-
PBSL (Shen et al., 2023)	57.84	75.46	82.15	-	-
BEAT(Ma et al., 2023)	58.25	75.92	81.96	-	-
LCR^{2}S (Yan et al., 2023)	57.93	76.08	82.40	38.21	-
DCEL (Li et al., 2023)	64.88	81.34	86.72	-	-
UniPT (Shao et al., 2023)	60.09	76.19	-	-	-
RaSa (Bai et al., 2023)	65.28	80.40	85.12	41.29	-
TBPS (Cao et al., 2024)	65.05	80.34	85.47	39.83	-
CFAM (Zuo et al., 2024)	65.38	81.17	-	39.42	-
MGRL (Lv et al., 2024)	67.28	63.87	-	82.34	-
OCDL (Li et al., 2025)	64.53	80.23	-	40.76	-
DRAGE (Ours)	**69.63**	**83.40**	**88.04**	**42.08**	**8.97**

Table 3: Performance comparisons on the RSTPReid dataset. The best results are in **bold**.

Methods	R-1	R-5	R-10	mAP	mINP
DSSL (Zhu et al., 2021b)	39.05	62.60	73.95	-	-
SSAN (Ding et al., 2021b)	43.50	67.80	77.15	-	-
LBUL (Wang et al., 2022)	45.55	68.20	77.85	-	-
IVT (Shu et al., 2022)	46.70	70.00	78.80	-	-
CFine (Yan et al., 2022a)	50.55	72.50	81.60	-	-
IRRA (Jiang and Ye, 2023b)	60.20	81.30	88.20	47.17	25.28
BiLMA (Fujii and Tarashima, 2023)	61.20	81.50	88.80	48.51	-
PBSL (Shen et al., 2023)	47.80	71.40	79.90	-	-
BEAT(Ma et al., 2023)	48.10	73.10	81.30	-	-
LCR^{2}S (Yan et al., 2023)	54.95	76.65	84.70	40.92	-
DCEL (Li et al., 2023)	61.35	83.95	90.45	-	-
RaSa (Bai et al., 2023)	66.90	86.50	91.35	52.31	-
TBPS (Cao et al., 2024)	61.95	83.55	88.75	48.26	-
CFAM (Zuo et al., 2024)	62.45	83.55	-	49.50	-
OCDL (Li et al., 2025)	61.60	82.35	-	49.77	-
DRAGE (Ours)	**67.55**	**85.00**	**90.90**	**52.88**	**30.05**

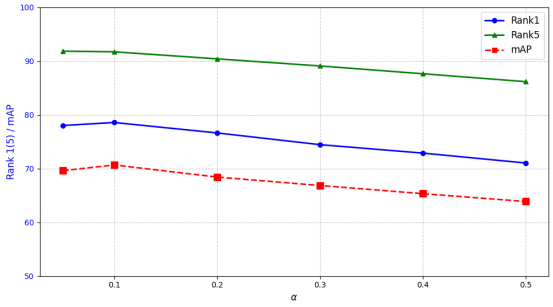


Figure 4: Performance variation with different values of the margin parameter α.

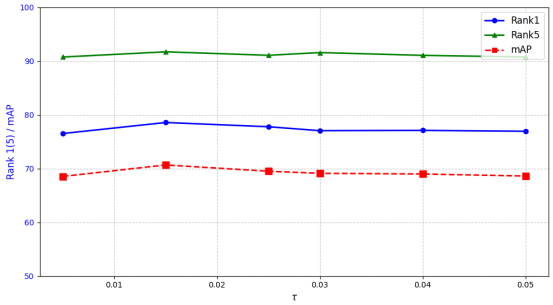


Figure 5: Performance variation with different values of the temperature parameter τ.

4.4 Ablation Study

We conduct ablation experiments on the CUHK-PEDES dataset to evaluate the contribution of each component in DRAGE. As shown in Table 5, removing any module leads to significant performance degradation, demonstrating that each component is essential. Specifically, removing SERD causes the largest drop (-6.40%), confirming its critical role in mitigating the semantic gap by identifying noisy pairs and providing reliable supervision. Without SIMR, performance decreases by 3.16%, validating its effectiveness in bridging the modality gap by organizing internal structures before cross-modal alignment. Eliminating AGE results in a 2.48% decline, showing that local features effectively complement global representations. The loss $\mathcal{L}_{\text{cmt}}$ also contributes to robust alignment. When removing both SERD and SIMR, performance drops sharply to 68.45%, exceeding the sum of their individual effects. This demonstrates strong complementarity: SERD provides reliable supervision that helps SIMR build stable internal structures, while SIMR's organized spaces help SERD better identify noisy pairs.

Table 4: Performance of methods and components in cross-domain tests. CUHK-PEDES is abbreviated as CUHK, ICFG-PEDES as ICFG, and RSTPReid as RSTP. The arrow → indicates that the method was trained on the dataset on the left and tested on the dataset on the right. The best results are in bold.

Method	Domain	R1	R5	mAP	Domain	R1	R5	mAP
IRRA (Jiang and Ye, 2023a)	CUHK → RSTP	53.30	77.15	39.63	CUHK → ICFG	42.42	62.11	21.77
	ICFG → RSTP	45.30	69.35	36.83	ICFG → CUHK	33.46	56.30	31.56
	RSTP → ICFG	32.30	49.69	20.54	RSTP → CUHK	32.80	55.25	30.29
SEN (Liu et al., 2025a)	CUHK → RSTP	55.50	77.85	45.29	CUHK → ICFG	45.34	63.45	23.26
	ICFG → RSTP	47.45	71.95	39.86	ICFG → CUHK	37.88	60.48	35.07
	RSTP → ICFG	36.23	53.31	22.32	RSTP → CUHK	35.40	57.71	33.41
DRAGE (Ours)	CUHK → RSTP	**62.10**	**81.40**	**47.31**	CUHK → ICFG	**52.45**	**69.61**	**28.46**
	ICFG → RSTP	**54.40**	**76.00**	**42.55**	ICFG → CUHK	**45.27**	**66.72**	**41.26**
	RSTP → ICFG	**43.31**	**58.94**	**27.18**	RSTP → CUHK	**42.58**	**63.35**	**38.84**
w/o $SIMR$	CUHK → RSTP	58.65	81.05	45.55	CUHK → ICFG	49.03	67.04	25.90
w/o $SERD$	CUHK → RSTP	50.25	72.40	38.44	CUHK → ICFG	39.15	56.92	21.57

Table 5: Ablation studies on the CUHK-PEDES dataset.

No.	Method	R-1	R-5	R-10	mAP	mINP
0	DRAGE (Full)	**78.58**	**91.68**	**95.19**	**70.68**	**53.89**
Single component ablation:						
1	w/o $\mathcal{L}_{\text{cmt}}$	77.24	91.11	94.96	68.98	52.67
2	w/o SIMR	75.42	89.86	93.51	67.35	51.22
3	w/o SERD	72.18	87.94	92.36	64.82	48.91
4	w/o AGE	76.10	90.81	94.40	67.91	51.60
Joint component ablation:						
5	w/o AGE + $\mathcal{L}_{\text{cmt}}$	75.94	90.69	94.54	67.87	51.51
6	w/o SIMR + SERD	68.45	84.72	90.18	61.29	46.08
7	w/o AGE + SERD	70.82	86.53	91.44	63.15	47.36

4.5 Parametric Analysis

To investigate the impact of hyperparameter selection on model performance, we conduct a systematic sensitivity analysis on the CUHK-PEDES dataset, focusing on two critical hyperparameters: the margin parameter α and the temperature parameter τ. The results are illustrated in Figure 4 and Figure 5. First, regarding the margin parameter α, Figure 4 reveals a clear optimal range. Setting α to excessively large values leads to significant performance degradation, particularly across the Rank-1, Rank-5, and mAP metrics. Based on these empirical observations, we set $\alpha = 0.1$ for all subsequent experiments. Second, for the temperature parameter τ (see Figure 5), we observe that extremely low values result in training instability. Conversely, while increasing τ initially improves Rank-1 and Rank-5 accuracy, excessively high values cause a gradual decline in performance, especially in mAP. Consequently, we determine that $\tau = 0.015$ yields the optimal trade-off between training stability and discriminative capability, achieving the best results.

4.6 Robustness Analysis

To evaluate our method's robustness, we conduct cross-domain tests by training the model on one dataset and testing it on another. As presented in Table 4, DRAGE demonstrates superior generalization capabilities, consistently outperforming prior state-of-the-art methods across all six settings. Notably, the improvements in Rank-1 accuracy reach up to 7.39% . This robust performance suggests that our strategy of addressing the fundamental semantic and modality gaps enables the model to learn more transferable representations. This conclusion is further validated by cross-domain ablation studies, where removing either the SERD or the SIMR module leads to a significant performance drop, confirming their crucial role in enhancing the model's robustness.

5 Conclusion

In this paper, we propose DRAGE, a unified framework that systematically addresses the semantic and modality gaps in TBPS. DRAGE introduces two synergistic mechanisms: SERD mitigates the semantic gap by identifying noisy pairs to provide robust supervision, while SIMR bridges the modality gap by organizing internal embedding structures. Notably, DRAGE consistently achieves superior performance, demonstrating a state-of-the-art 78.58% Rank-1 accuracy on CUHK-PEDES and significantly enhancing cross-domain generalization. By improving both internal organization and supervision reliability, DRAGE offers a robust solution for real-world person search applications.

6 Limitations

While DRAGE achieves state-of-the-art performance in TBPS, several limitations merit acknowledgment. First, the effectiveness of SIMR depends on the quality of the pre-trained sentence encoder used to construct semantic distance matrices. We adopt all-mpnet-base-v2, whose capability to capture fine-grained nuances in pedestrian descriptions inherently constrains SIMR's performance. Future work could explore domain-specific fine-tuning or alternative encoders for improved semantic structure modeling.Although DRAGE demonstrates strong cross-domain generalization, our evaluation is limited to standard TBPS benchmarks. These datasets may not fully capture real-world complexities such as extreme weather, heavy occlusion, or out-of-distribution language descriptions. Additionally, all experiments use English-language data; DRAGE's performance in multilingual or low-resource settings remains unexplored.

References

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023. Rasa: relation and sensitivity aware representation learning for text-based person search. In *IJCAI*, pages 555–563.

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. 2024. An empirical study of clip for text-based person search.

Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1879–1887. IEEE.

Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. 2020. Multi-task learning with coarse priors for robust part-aware person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1474–1488.

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

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

Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. 2022. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *AAAI*, pages 4477–4485.

Takuro Fujii and Shuhei Tarashima. 2023. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2786–2790.

Jing Ge, Guangyu Gao, and Zhen Liu. 2019. Visual-textual association with hardest and semi-hard negative pairs mining for person search. In *arXiv preprint arXiv:1912.03083*.

Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3642–3651. IEEE.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, IvorW. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Neural Information Processing Systems,Neural Information Processing Systems*.

D. Jiang and M. Ye. 2023a. 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*, pages 2787–2797.

Ding Jiang and Mang Ye. 2023b. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797.

Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. 2025. Modeling thousands of human annotators for generalizable text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9220–9230.

Jie Lei, Xinlei Chen, Ning Zhang, Mengjiao Wang, Mohit Bansal, Tamara L. Berg, and Licheng Yu. 2022. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *Preprint*, arXiv:2203.05465.

Haiwen Li, Delong Liu, Fei Su, and Zhicheng Zhao. 2025. Object-centric discriminative learning for text-based person retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. 2023. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5187–5196.

Delong Liu, Haiwen Li, Zhicheng Zhao, and Yuan Dong. 2025a. Text-guided image restoration and semantic enhancement for text-to-image person retrieval. *Neural Networks*, 184:107028.

Yating Liu, Zimo Liu, Xiangyuan Lan, Wenming Yang, Yaowei Li, and Qingmin Liao. 2025b. Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5703–5711.

Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14052–14060.

Tianle Lv, Shuang Li, Jiaxu Leng, and Xinbo Gao. 2024. Mgrl: Mutual-guidance representation learning for text-to-image person retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2895–2899. IEEE.

Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. 2023. Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4157–4168.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9821–9831.

Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. 2024. Plot: Text-based person search with part slot attention for corresponding part discovery. In *ECCV*, pages 474–490. Springer.

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

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184.

Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, pages 5566–5574.

Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. 2023. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931.

Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer.

Zifan Song, Guosheng Hu, and Cairong Zhao. 2024. Diverse person: Customize your own dataset for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4943–4951.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. *Preprint*, arXiv:2103.08784.

Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *CVPR*, pages 17127–17137.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 402–420. Springer.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *ACM MM*, pages 1984–1992.

Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. 2021. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633.

Zequn Xie, Haoming Ji, and Lingwei Meng. 2025a. Dynamic uncertainty learning with noisy correspondence for text-based person search. *arXiv preprint arXiv:2505.06566*.

Zequn Xie, Chuxin Wang, Sihang Cai, Yeqiang Wang, Shulei Wang, and Tao Jin. 2025b. Chat-driven text generation and interaction for person retrieval. *Preprint*, arXiv:2509.12662.

Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. 2023. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2022a. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*.

Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022b. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365*.

Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 2023 ACM on Multimedia Conference*.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. 2022. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 2872–2893.

Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. 2022. Progressive meta-learning with curriculum. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5916–5930.

Ying Zhang and Huchuan Lu. 2018a. Deep cross-modal projection learning for image-text matching. In *ECCV*.

Ying Zhang and Huchuan Lu. 2018b. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701.

Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. 2024. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7534–7542.

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

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021a. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*.

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

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

A Supplementary Material

In this supplementary material, we provide additional information for DRAGE.

A.1 Datasets

To comprehensively evaluate the performance of our proposed method under both clean and noisy supervision, we conduct experiments on three widely adopted benchmarks for text-to-image person retrieval: CUHK-PEDES (Zhang et al., 2022), ICFG-PEDES (Han et al., 2018), and RSTPReid (Zhu et al., 2021b). Each dataset presents unique characteristics in terms of scale, annotation style, and visual diversity.

CUHK-PEDES. CUHK-PEDES (Zhang et al., 2022) is the first and most well-established benchmark for the text-based person re-identification (ReID) task. It comprises a total of 40,206 pedestrian images corresponding to 13,003 unique identities. Each image is annotated with two independent natural language descriptions, collected via crowdsourcing, providing rich semantic diversity. Following the standard evaluation protocol, the dataset is divided into three subsets: a training set with 34,054 images from 11,003 identities, a validation set containing 3,078 images from 1,000 identities, and a testing set with 3,074 images from another 1,000 identities. The average caption length is approximately 23 words, encompassing a variety of appearance-related cues such as clothing, accessories, and actions.

ICFG-PEDES. ICFG-PEDES (Han et al., 2018) is a relatively newer but larger-scale dataset designed to reflect more complex and fine-grained textual descriptions. It contains 54,522 images corresponding to 4,102 person identities. Unlike CUHK-PEDES, each image in ICFG-PEDES is annotated with a single sentence, which on average contains around 37 words—making the descriptions more elaborate and detailed. The official data split consists of 34,674 image-text pairs from 3,102 identities for training, and 19,848 image-text pairs from the remaining 1,000 identities for testing. This dataset presents greater challenges due to longer and potentially more ambiguous captions.

RSTPReid. RSTPReid (Zhu et al., 2021b) is a large-scale and challenging benchmark constructed under real-world surveillance scenarios. It contains 20,505 images from 4,104 identities captured by 15 different cameras, introducing substantial variations in pose, lighting, and background. Each identity is associated with five images taken from distinct camera views. Importantly, each image is paired with two distinct textual descriptions, which significantly increases the textual diversity. According to the official splitting protocol, the training set includes image-text pairs from 3,701 identities. Both the validation and testing sets are composed of 200 identities each, ensuring consistent and reliable benchmarking. Compared to CUHK-PEDES and ICFG-PEDES, RSTPReid reflects more realistic retrieval conditions with strong cross-view challenges.

B Training Process

Algorithm 1 The training process of our model

Input: The training data $\mathcal{P}$ with N image-text pairs, maximal epoch N_e, the model $\mathcal{M}(\Theta)$, and the hyper-parameters $\mathcal{R}$, α, τ
Initialize the backbones with the weights of the pretrained model CLIP;
Initialize the Pre-trained Sentence Encoder: sbert = SBERT($sentence-transformers/all-mpnet-base-v2$);
for $e = 1, 2, \cdots, N_e$ **do**
    Calculate the per-sample loss $\ell(\mathcal{M}, \mathcal{P})$ with Equation (3) ;
    Divide the training data with the predictions of GFE and AGE using BMM;
    SERD: Deriving and Recalibrating the Labels $\{\ell_i\}_{i=1}^{N}$;
    for each x in mini-batches $\{x_m\}_{m=1}^{M}$ of x **do**
        Extract the GFE & AGE features of x;
        Compute the similarities between K image-text pairs in x with above features;
        Calculate the $\mathcal{L}_{v2v}$ for vision-to-vision and $\mathcal{L}_{t2t}$ for text-to-text using SIMR;
        Calculate $\mathcal{L}_{total}$ with Equation (17);
        Θ = Optimizer($\Theta, \mathcal{L}_{total}$);
    end for
end for
Output: The optimized parameters $\hat{\Theta}$;

C Visualization of Modality Gap Reduction

To visualize the modality gap reduction, we project the learned image and text embeddings from CUHK-PEDES into a 2D space using Principal Component Analysis. Figure 6 illustrates the pro-

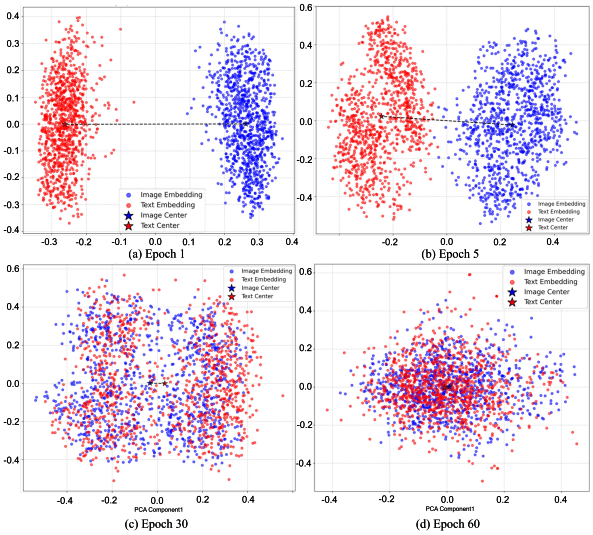


Figure 6: Visualization of modality gap reduction between image and text embeddings in CUHK-PEDES dataset .

gressive alignment throughout the training process. At epoch 1, the embeddings form distinct clusters. These clusters merge by epoch 30 and become indistinguishable by epoch 60, demonstrating the effective bridging of the modality gap.

D Qualitative Results

To illustrate the advantages of our DRAGE, some retrieved examples for TBPS are presented in Figure 7.

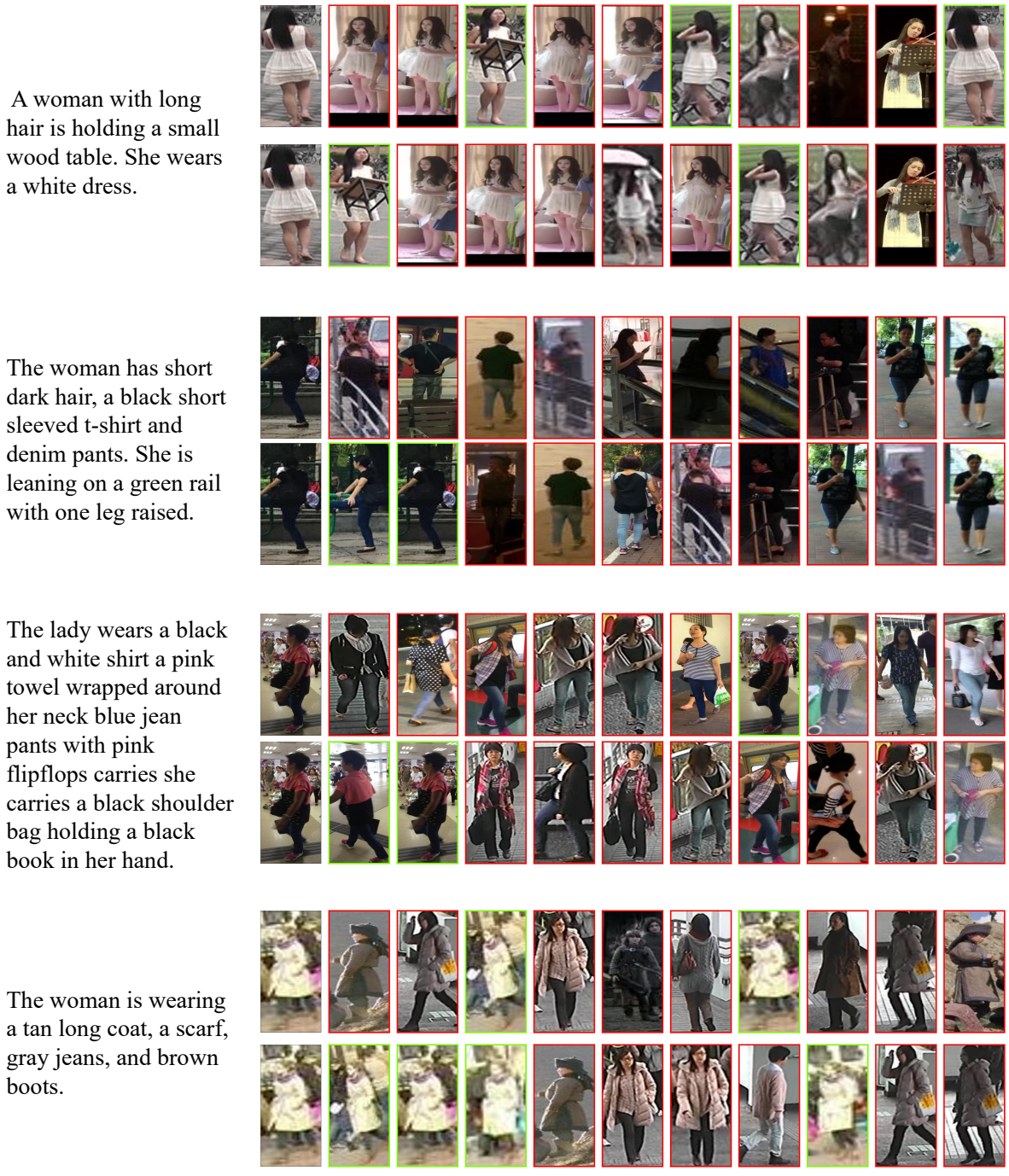


Figure 7: Qualitative comparison of the top-10 retrieved results on the CUHK-PEDES dataset. For each text query, we compare the baseline IRRA (first row) with our DRAGE (second row). The **matched** and **mismatched** person images are highlighted with **green** and **red** rectangles, respectively. The first column of each row represents the ground truth.