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

Anonymous ACL submission

Abstract

Text-based Person Search (TBPS), which retrieves individuals using natural language descriptions, faces two critical challenges. The first is the **semantic gap**, where noisy and inconsistent image-text pairs in training data severely corrupt the model's learning process. The second is the **modality gap**, which arises from the fundamental heterogeneity between dense visual features and sparse textual attributes, hindering reliable cross-modal alignment. To address these challenges, we propose the **D**ual-Gaps **R**obust **A**ligned **G**eneral **E**mbedding (DRAGE) framework. DRAGE introduces two synergistic mechanisms: (1) *Semantic Embedding Reliability Division* (SERD), which dynamically partitions training data into clean and noisy subsets to provide robust supervision, and (2) *Semantic Intra-Modality Regularization* (SIMR), which organizes the internal structure of each modality's embedding space *before* cross-modal alignment. Through semantic-aware calibration, SIMR transforms disorganized sparse text representations into well-structured embeddings, establishing a solid foundation for effective cross-modal matching. Extensive experiments demonstrate that DRAGE significantly outperforms state-of-the-art approaches, achieving 78.58% Rank-1 accuracy on CUHK-PEDES with strong cross-domain generalization, validating the effectiveness of our dual-gap strategy.

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 important applications ranging from video surveillance to social media analysis (Li et al., 2025; Zhao et al., 2024). Despite significant progress, existing TBPS models face two critical challenges in real-world scenarios, as illustrated in Figure 1.

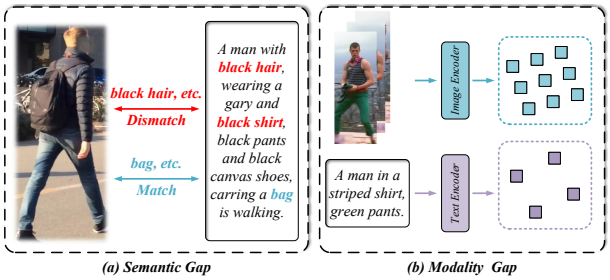

Figure 1: Illustration of the semantic gap and modality gap in Text-based Person Search. (a) **Semantic gap**: training image-text pairs often exhibit content inconsistencies due to annotation biases or visual ambiguities, resulting in noisy supervision. (b) **Modality gap**: the fundamental heterogeneity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

Recent research has made significant progress in TBPS. Many methods explore both global and local matching strategies. Global matching approaches (Jiang and Ye, 2023b; Xie et al., 2025a) focus on aligning overall features, while local matching techniques (Zuo et al., 2024; Park et al., 2024) associate fine-grained visual features, such as body parts, with corresponding textual attributes. More recently, studies leverage large pre-trained models (Li et al., 2025; Liu et al., 2025b) by adapting vision-language models like 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 significant progress, existing TBPS methods remain fundamentally constrained by two critical yet rarely co-addressed challenges: the **semantic gap** and the **modality gap**. The semantic gap arises from noisy and inconsistent image-text correspondences caused by real-world data variability such as pose variations, camera angles, and illu-

mination changes. The modality gap stems from a deeper structural disparity operating on two levels: (1) *cross-modal heterogeneity*—the fundamental difference between dense, continuous visual features and sparse, abstract textual attributes; 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 internal organization, resulting in disorganized embeddings where semantically similar samples scatter and dissimilar ones cluster together. This internal chaos, particularly severe for sparse text features that inherently lack structural information, undermines stable cross-modal alignment. To address these limitations, we propose DRAGE (Dual-Gaps Robust Aligned General Embedding), a unified framework that systematically tackles both challenges through two synergistic mechanisms. First, to mitigate the semantic gap, DRAGE introduces Semantic Embedding Reliability Division (SERD), which dynamically identifies and separates noisy from reliable image-text pairs using a Beta Mixture Model. By providing reliability-based supervision—assigning higher confidence to clean pairs and lower weights to noisy ones—SERD enables robust training on real-world data. Second, to bridge the modality gap, DRAGE employs Semantic Intra-Modality Regularization (SIMR), which addresses the core 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 systematically addresses the semantic gap and modality gap in TBPS through two complementary mechanisms: SERD for handling noisy correspondences and SIMR for organizing internal embedding structures.
- We introduce Semantic Embedding Reliability Division (SERD), which identifies noisy image-text pairs and provides reliability-based supervision to mitigate the semantic gap caused by real-world data inconsistencies.
- We propose Semantic Intra-Modality Regularization (SIMR), which organizes each modality's internal embedding structure *before* cross-modal alignment, addressing the critical yet overlooked issue of intra-modal disorganization that exacerbates the modality gap.
- Extensive experiments demonstrate that DRAGE achieves state-of-the-art performance across multiple benchmarks, with 78.58% Rank-1 accuracy on CUHK-PEDES and strong cross-domain generalization, validating 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 (Bai et al., 2023a; 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 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 objective of Text-based Person Search (TBPS) is to retrieve pedestrian images from a gallery that correspond to a given textual description. Formally, we define the gallery set as $\mathcal{V} = \{I_i, y_i^p, y_i^v\}_{i=1}^{N_v}$ and the text set as $\mathcal{T} = \{T_i, y_i^v\}_{i=1}^{N_t}$, where N_v and N_t are the numbers of images and texts, respectively. Here, $y_i^p \in \mathcal{Y}_p = \{1, \ldots, C\}$ is the person identity label, and $y_i^v \in \mathcal{Y}_v = \{1, \ldots, N_v\}$ is the image-specific label. The training set is composed of image-text pairs $\mathcal{P} = \{(I_i, T_i), y_i^v, y_i^p\}_{i=1}^{N}$, where paired samples share the same labels. To learn the cross-modal correspondence, we employ the pre-trained CLIP model (Radford et al., 2021) as our backbone, which provides a visual encoder f^v and a textual encoder f^t. Given an input image I_i, the visual encoder f^v generates a sequence of features $\boldsymbol{V}_i = \{\boldsymbol{v}_g^i, \boldsymbol{v}_1^i, \ldots, \boldsymbol{v}_{N_\circ}^i\}^\top$, where $\boldsymbol{v}_g^i$ is the global feature from the [*CLS*] token and $\{\boldsymbol{v}_j^i\}$ are local patch features. Similarly, for a textual description T_i, the textual encoder f^t produces a sequence of representations $\boldsymbol{T}_i = \{\boldsymbol{t}_s^i, \boldsymbol{t}_1^i, \ldots, \boldsymbol{t}_{N_\diamond}^i, \boldsymbol{t}_e^i\}^\top$, where $\boldsymbol{t}_s^i$ and $\boldsymbol{t}_e^i$ are the features for the [*SOS*] and [*EOS*] tokens, while $\{\boldsymbol{t}_j^i\}$ are the word-level features. The $\boldsymbol{t}_e^i$ feature effectively serves as the global semantic representation of the text.

3.2 Attention-Guided Selector

While global features provide a baseline similarity using the [*CLS*] and [*EOS*] tokens, they often fail to capture the fine-grained details necessary for person retrieval. To address this limitation, we introduce an **Attention-Guided Selector** to create more discriminative fine-grained representations.

Our selector operates in three main steps. First, we leverage the self-attention maps from the final layer of the CLIP encoders to identify the most informative local tokens in both the image and text (i.e., those with the highest correlation to their global features). Second, we select the top-$\mathcal{R}$ proportion of these informative tokens, denoted as $\boldsymbol{V}_i^a$ and $\boldsymbol{T}_i^a$. Finally, these selected tokens are processed through a dedicated embedding module to produce compact and discriminative feature vectors, which we term the **A**ttention-**G**uided **E**mbedding (AGE) vectors, $\boldsymbol{v}_{\text{age}}^i$ and $\boldsymbol{t}_{\text{age}}^i$:

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

where $\hat{\boldsymbol{V}}_i^a$ and $\hat{\boldsymbol{T}}_i^a$ represent the normalized selected tokens. The resulting AGE similarity, S_{ij}^a, is the cosine similarity between these vectors. This similarity score, which focuses on salient local details, is used alongside the global similarity to provide a comprehensive cross-modal assessment.

3.3 Semantic Embedding Reliability Division

In this section, we introduce Semantic Embedding Reliability Division (SERD), our approach to addressing the semantic gap in TBPS. As mentioned in the introduction, the semantic gap is a major challenge when image-text pairs, despite depicting the same person, contain misaligned features or inconsistent details. SERD specifically addresses this semantic gap by assessing the reliability of image-text pairs.

To reduce the impact of semantic inconsistencies, we identify unreliable pairs in the training data using a two-component Beta Mixture Model (BMM). This approach leverages the observation that DNNs tend to exhibit smaller loss values for semantically consistent data during early training (Arpit et al., 2017). For a cross-modal model $\mathcal{M}$, we define the per-sample loss as:

$$\ell(\mathcal{M}, \mathcal{P}) = \{\ell_i\}_{i=1}^N = \{\mathcal{L}(I_i, T_i)\}_{i=1}^N, \quad (1)$$

where $\mathcal{L}$ is the loss function for pair $(I_i, T_i) \in \mathcal{P}$

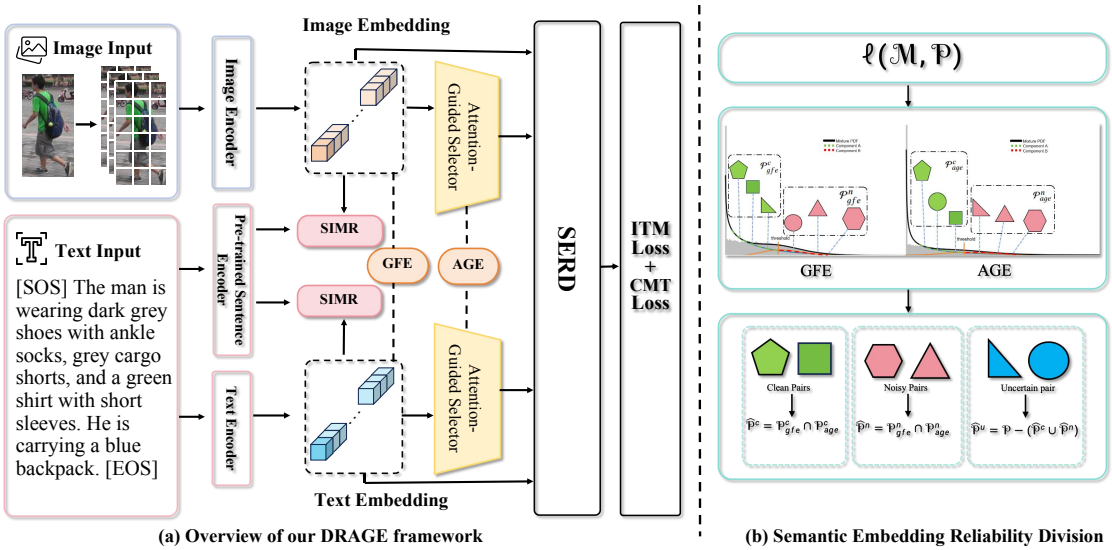

Figure 2: Overview of the DRAGE framework. DRAGE addresses the semantic gap through SERD, which identifies reliable image-text pairs for robust supervision, and bridges the modality gap through SIMR, which organizes each modality's internal embedding structure before cross-modal alignment. The framework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

that brings them closer in the shared latent space. In our method, $\mathcal{L}$ is the proposed $\mathcal{L}_{\text{cmt}}$ defined later.

The per-sample loss is then analyzed using BMM to separate semantically consistent and inconsistent data by assigning the Beta component with lower mean to the reliable set and the other to the unreliable set. Following (Li et al., 2020; Huang et al., 2021), we use the Expectation-Maximization (EM) algorithm to optimize the BMM and compute the posterior probability $p(k|\ell_i) = p(k)p(\ell_i|k)/p(\ell_i)$ for the i-th pair, indicating its reliability probability, where $k \in \{0, 1\}$ distinguishes between reliable and unreliable components. We then set a threshold $\delta = 0.6$ to $\{p(k = 0|\ell_i)\}_{i=1}^{N}$ to divide the data into reliable and unreliable sets. For GFE and AGE, the divisions are $\mathcal{P} = \mathcal{P}^c_{\text{gfe}} \cup \mathcal{P}^n_{\text{gfe}}$ and $\mathcal{P} = \mathcal{P}^c_{\text{age}} \cup \mathcal{P}^n_{\text{age}}$, respectively. To obtain the final reliable divisions, we exploit the consensus between the two divisions to identify the most semantically consistent pairs as $\hat{\mathcal{P}}^c = \mathcal{P}^c_{\text{gfe}} \cap \mathcal{P}^c_{\text{age}}$, the most unreliable pairs as $\hat{\mathcal{P}}^n = \mathcal{P}^n_{\text{gfe}} \cap \mathcal{P}^n_{\text{age}}$, and the uncertain pairs as $\hat{\mathcal{P}}^u = \mathcal{P} \setminus (\hat{\mathcal{P}}^c \cup \hat{\mathcal{P}}^n)$. Finally, we exploit these divisions to recalibrate the correspondence labels for each pair:

$$\hat{l}_{ii} = \begin{cases} 2, & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^c, \\ 0, & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^n, \\ 1, & \text{if } (I_i, T_i) \in \hat{\mathcal{P}}^u. \end{cases} \tag{2}$$

By distinguishing between semantically consistent and inconsistent image-text pairs, SERD provides more reliable supervision signals for robust representation learning, substantially alleviating the semantic gap in TBPS.

3.4 Cross-Modal Triplet Loss

To enhance the model's ability to extract fine-grained features, we introduce the **Cross-Modal Triplet Loss (CMT)**. The probability of a correct match is first computed using the softmax function:

$$p_{ij} = \frac{\hat{l}_{ij} \exp(S(I_i, T_j)/\tau)}{\sum_{k=1}^{N} \exp(S(I_i, T_k)/\tau)}, \tag{3}$$

where τ is a temperature hyperparameter that controls the sharpness of the probability distribution, and $S(I_i, T_j) \in \{S^g_{ij}, S^a_{ij}\}$ represents the similarity score between image I_i and text T_j. For an input pair (I_i, T_j) in a mini-batch, the CMT loss from image to text is defined as:

$$\mathcal{L}_{\text{i2t}} = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{S}(I_i, T_i) - \tau \log \left(\sum_{j=1}^{N} q_{ij} \exp \left(\frac{S(I_i, T_j)}{\tau} \right) \right) + \alpha \right]_+, \tag{4}$$

where $\hat{S}(I_i, T_i) = \sum_{j=1}^{N} p_{ij} S(I_i, T_j)$, $q_{ij} = 2 - \hat{l}_{ij}$, α is a margin hyperparameter, and $[x]_+ =$

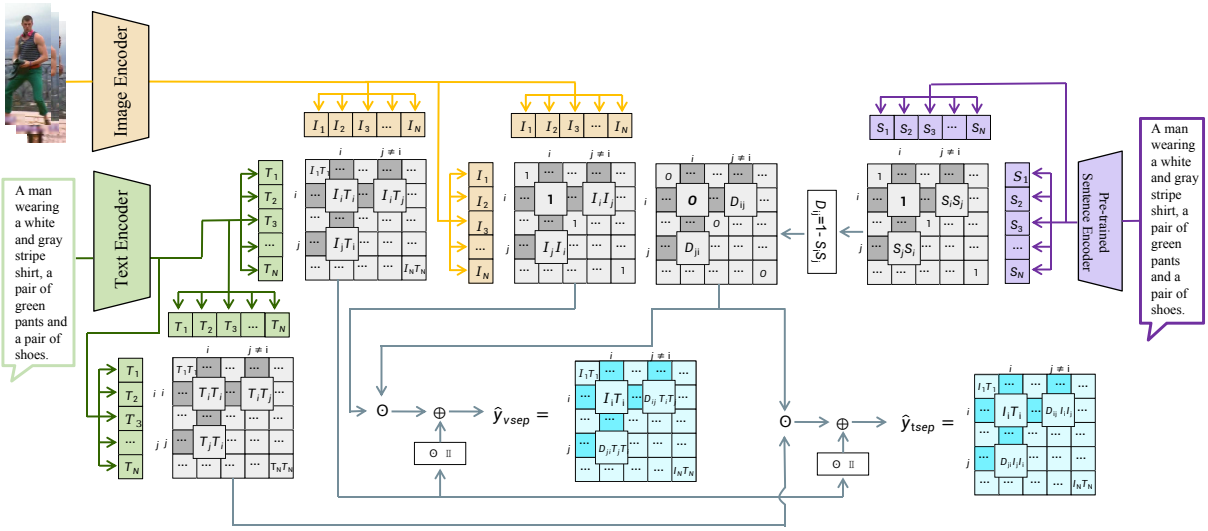

Figure 3: Overview of Semantic Intra-Modality Regularization (SIMR) method. This approach is implemented through vision-to-vision (v2v) and text-to-text (t2t) contrastive losses, adjusting embedding distances within each modality based on semantic distances between samples, ensuring that semantically similar samples maintain appropriate distances in the embedding space.

$\max(0, x)$ denotes the positive part of x. Similarly, the CMT loss $\mathcal{L}_{\text{t2i}}$ from text to image is obtained by swapping the roles of I_i and T_j in the formulas above. The same triplet concept is applied to the AGE features S^a_{ij}, leading to an additional AGE-based triplet loss $\mathcal{L}^{\text{age}}_{\text{cmt}}$. The bidirectional CMT loss, incorporating AGE features, is computed as:

$$\mathcal{L}_{\text{cmt}} = \mathcal{L}^{\text{gfe}}_{\text{i2t}} + \mathcal{L}^{\text{age}}_{\text{i2t}} + \mathcal{L}^{\text{gfe}}_{\text{t2i}} + \mathcal{L}^{\text{age}}_{\text{t2i}}. \quad (5)$$

3.5 Semantic Intra-Modality Regularization

As discussed in the introduction, the modality gap in TBPS stems from the structural disparity between dense visual features and sparse textual attributes. This textual sparsity poses a fundamental challenge: unlike dense visual representations that naturally form continuous distributions, sparse text embeddings lack sufficient structural information and are inherently prone to disorganization. As depicted in Figure 6(a), existing methods focus solely on cross-modal alignment while neglecting the internal structure within each modality. Consequently, embeddings become disorganized—especially for sparse text features—where semantically similar samples scatter while dissimilar ones cluster together, undermining stable cross-modal alignment.

To address this issue, we introduce Semantic Intra-Modality Regularization (SIMR), which organizes each modality's internal structure *before* cross-modal alignment. Rather than immediately aligning across modalities, SIMR first establishes well-structured intra-modal foundations by enforcing meaningful distance constraints: pulling semantically similar samples closer while pushing dissimilar ones apart. This is particularly crucial for sparse text representations, as SIMR transforms disorganized text embeddings into coherent semantic clusters, compensating for their inherent lack of structure. SIMR similarly organizes visual features into more discriminative structures. By establishing stable foundations within each modality, SIMR facilitates more effective cross-modal matching.

Conceptually, the SIMR objective can be formulated as a semantically-aware contrastive loss. For instance, the vision-to-vision loss ($\mathcal{L}_{v2v}$) and text-to-text loss ($\mathcal{L}_{t2t}$) are defined as:

$$\mathcal{L}_{v2v} = \frac{-1}{N} \sum_{i=1}^{N} \log \frac{\exp[(\boldsymbol{V}_i \cdot \boldsymbol{T}_i) \,/\, \tau]}{\sum_{\substack{j=1,\\ i\neq j}}^{N} \exp[(\mathcal{D}_{ij} \cdot \boldsymbol{V}_i \cdot \boldsymbol{V}_j) \,/\, \tau]}, \quad (6)$$

$$\mathcal{L}_{t2t} = \frac{-1}{N} \sum_{i=1}^{N} \log \frac{\exp[(\boldsymbol{T}_i \cdot \boldsymbol{V}_i) \,/\, \tau]}{\sum_{\substack{j=1,\\ i\neq j}}^{N} \exp[(\mathcal{D}_{ij} \cdot \boldsymbol{T}_i \cdot \boldsymbol{T}_j) \,/\, \tau]}, \quad (7)$$

where $\mathcal{D}_{ij}$ represents the semantic distance used to ensure that only semantically dissimilar pairs are pushed apart.

However, directly optimizing the complex log-sum-exp objectives in Eqs. 6 and 7 can be challenging. Therefore, we introduce a more direct and practical implementation of SIMR. Our approach involves re-scaling the similarity logits before applying a standard cross-entropy loss. First, we com-

pute the semantic distance matrix $\mathcal{D}$ required for this process. We obtain $\mathcal{D}$ by using a pre-trained sentence encoder to produce text embeddings E_s, from which we compute a similarity matrix S and the corresponding distance matrix $\mathcal{D} = 1 - S$. The raw intra-modal similarity matrices, $\mathcal{V}$ for vision and $\mathcal{T}$ for text, are computed as:

$$\mathcal{V} = E_v E_v^T, \quad \mathcal{T} = E_t E_t^T, \quad \text{where} \quad \mathcal{V}, \mathcal{T} \in \mathbb{R}^{b \times b}. \tag{8}$$

$$S = \frac{E_s E_s^T}{\|E_s\|^2}, \quad \mathcal{D} = 1 - S. \tag{9}$$

Next, we re-scale these raw similarities with the semantic distance $\mathcal{D}$ and combine them with the cross-modal similarities (from $\mathcal{M} = E_v E_t^T$) to produce the final logits for our loss function:

$$\hat{y}_v = \exp(\tau) \cdot [\text{diag}(\mathcal{M}) + \mathcal{V} \odot \mathcal{D}], \tag{10}$$

$$\hat{y}_t = \exp(\tau) \cdot \left[\text{diag}(\mathcal{M}^T) + \mathcal{T} \odot \mathcal{D}\right], \tag{11}$$

where $\odot$ denotes the element-wise product. Finally, as illustrated in Figure 7, the SIMR loss is defined as a simple cross-entropy function over these calibrated logits:

$$\mathcal{L}_{simr} = H(\hat{y}_v, Y) + H(\hat{y}_t, Y), \tag{12}$$

where H is the cross-entropy loss and Y is the ground-truth matrix. This implementation effectively achieves the goal of the conceptual losses (i.e., $\mathcal{L}_{v2v}$ and $\mathcal{L}_{t2t}$) in a more stable and efficient manner.

3.6 Overall Optimization Objective

To address both the semantic gap and the modality gap, our DRAGE framework integrates several complementary loss functions into a unified optimization objective.

For cross-modal alignment, we employ the Image-Text Matching (ITM) loss (Li et al., 2021), which maximizes agreement for matched pairs while minimizing it for unmatched pairs. The ITM loss is computed over both global features and AGE features:

$$\mathcal{L}_{\text{itm}} = \mathcal{L}_{\text{itm}}^{\text{gfe}} + \mathcal{L}_{\text{itm}}^{\text{age}}. \tag{13}$$

To bridge the modality gap, we employ the SIMR loss ($\mathcal{L}_{\text{simr}}$) to organize the internal structure within each modality, and the Cross-Modal Triplet (CMT) loss ($\mathcal{L}_{\text{cmt}}$) to enhance discriminative power across modalities. Our complete optimization objective integrates these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simr}} + \mathcal{L}_{\text{cmt}} + \mathcal{L}_{\text{itm}}. \tag{14}$$

By jointly optimizing these complementary losses, DRAGE effectively addresses both challenges: the semantic gap is mitigated through SERD's reliability-based data division and CMT's robust triplet constraints, while the modality gap is bridged through SIMR's intra-modal organization and ITM's cross-modal alignment. This unified framework yields robust representations that are resilient to noisy correspondences and structural disparities.

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, including the training, and test splits, 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 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.

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
ViTAA (Wang et al., 2020)	ECCV'20	RN50	LSTM	54.92	75.18	82.90	51.60	-
Lapscore (Wu et al., 2021)	ICCV'21	RN50	BERT	63.40	-	87.80	-	-
LBUL (Wang et al., 2022b)	ACMMM'22	RN50	BERT	64.04	82.66	87.22	-	-
CAIBC (Wang et al., 2022a)	ACMMM'22	RN50	BERT	64.43	82.87	88.37	-	-
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
IVT (Shu et al., 2022)	ECCVW'22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine (Yan et al., 2023b)	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., 2023a)	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., 2023b)	IJCAI'23	ALBEFF	ALBEFF	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**

4.3 Comparison with State-of-the-Art

In this section, we compare our method with current state-of-the-art approaches on three 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 and RSTPReid. Our method's strong performance extends to the ICFG-PEDES and RSTPReid datasets. On ICFG-PEDES, DRAGE achieves a Rank-1 accuracy of **69.63%** and a Rank-5 accuracy of 83.40%, marking a significant improvement over previous methods. This success is mirrored on the RSTPReid dataset, where our approach sets a new state-of-the-art with **67.55%** in Rank-1 accuracy and 52.88% in mAP. In conclusion, DRAGE consistently outperforms state-of-the-art methods across all three benchmarks, validating its superior performance and generalization capabilities. Detailed results for both datasets are available in the **supplementary material**.

4.4 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 5, DRAGE demonstrates superior generalization capabilities, consistently outperforming prior state-of-the-art methods across all six settings. No-

Table 2: 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 3: 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 $\mathcal{L}_{\text{itm}}$	77.76	91.36	94.84	69.57	53.46
4	w/o SERD	72.18	87.94	92.36	64.82	48.91
5	w/o AGE	76.10	90.81	94.40	67.91	51.60
Joint component ablation:						
6	w/o AGE + $\mathcal{L}_{\text{cmt}}$	75.94	90.69	94.54	67.87	51.51
7	w/o SIMR + SERD	68.45	84.72	90.18	61.29	46.08
8	w/o AGE + SERD	70.82	86.53	91.44	63.15	47.36
9	w/o AGE + $\mathcal{L}_{\text{itm}}$	75.89	90.72	94.56	67.86	51.51

tably, 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.

4.5 Ablation Study

We conduct ablation experiments on the CUHK-PEDES dataset to evaluate the contribution of each component in DRAGE. As shown in Table 3, 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 losses $\mathcal{L}_{\text{itm}}$ and $\mathcal{L}_{\text{cmt}}$ also contribute substantially 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.

5 Conclusion

In this paper, we propose DRAGE, a unified framework that systematically addresses two critical challenges in TBPS: the semantic gap and the modality gap. DRAGE introduces two synergistic mechanisms: SERD mitigates the semantic gap by identifying and handling noisy image-text pairs to provide robust supervision, while SIMR bridges the modality gap by organizing each modality's internal embedding structure before cross-modal alignment. Extensive experiments demonstrate that DRAGE significantly outperforms state-of-the-art methods across multiple benchmark datasets, achieving 78.58% Rank-1 accuracy on CUHK-PEDES with strong cross-domain generalization. As a general framework, DRAGE can be integrated with existing methods to boost performance, offering a robust solution for real-world 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.

Second, 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.

Finally, DRAGE's architectural complexity—integrating multiple encoders, alignment modules, and loss objectives—incurs substantial computational cost during training, posing challenges for resource-constrained or real-time deployment scenarios.

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. 2023a. Rasa: Relation and sensitivity aware representation learning for text-based person search. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pages 555–563.

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023b. 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.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419.

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.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

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. 2022a. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *ACM MM*, pages 5314–5322.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022b. 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. 2023a. 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, Neng Dong, Liyan Zhang, and Jinhui Tang. 2022b. Clip-driven fine-grained text-image person re-identification. *arXiv:2210.10276*.

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

Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022c. 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.

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.

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);
    $\Theta = \text{Optimizer}(\Theta, \mathcal{L}_{total})$;
  end for
end for
Output: The optimized parameters $\hat{\Theta}$;

C 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 performance 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.

Table 4: 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., 2022c)	57.73	75.42	81.72	-	-
CFine (Yan et al., 2022b)	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., 2023a)	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., 2023b)	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**

D 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., 2022b), we achieve a notable improvement in performance with gains of up to 9.65 on Rank-1.

Table 5: 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., 2022b)	45.55	68.20	77.85	-	-
IVT (Shu et al., 2022)	46.70	70.00	78.80	-	-
CFine (Yan et al., 2022b)	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., 2023a)	54.95	76.65	84.70	40.92	-
DCEL (Li et al., 2023)	61.35	83.95	90.45	-	-
RaSa (Bai et al., 2023b)	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**

E Parametric Analysis

To investigate the impact of hyperparameter selection on model performance, we conducted a

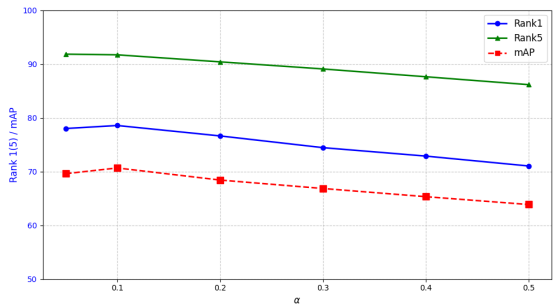

Figure 4: Variation of performance with different α . Three Paradigms of Prompt Tuning in the Continual Learning Scenario.

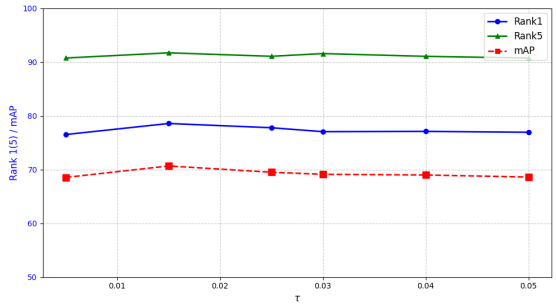

Figure 5: Variation of performance with different τ Three Paradigms of Prompt Tuning in the Continual Learning Scenario.

thorough sensitivity analysis for two critical parameters, α and τ, using the CUHK-PEDES dataset. As shown in fig:5,fig:6, our findings reveal the following insights:

(1) The parameter α has an optimal range, with excessively large values leading to a significant decline in performance, particularly affecting Rank1, Rank5, and mAP. Based on these results, we selected $\alpha = 0.1$ for all experiments.

(2) For the temperature parameter τ, values that are too small resulted in training instability. As τ increased, the performance of Rank1 and Rank5 improved, but beyond a certain point, further increases in τ led to a gradual decline in performance, especially in mAP. The most optimal performance was achieved when $\tau = 0.015$, which balanced both stability and discriminative ability, resulting in the best overall results in our experiments.

E.1 Visualization of Modality Gap Reduction

To visualize the modality gap reduction, we project the learned image and text embeddings from the CUHK-PEDES dataset into a 2D space using Principal Component Analysis (PCA). As illustrated in Figure 6, this visualization shows the progressive alignment of image and text embeddings throughout the training process. Initially, at epoch 1, the embeddings form two distinct clusters. These clusters begin to merge by epoch 30 and become nearly inseparable by epoch 60, indicating a successful alignment and an effective bridging of the modality gap.

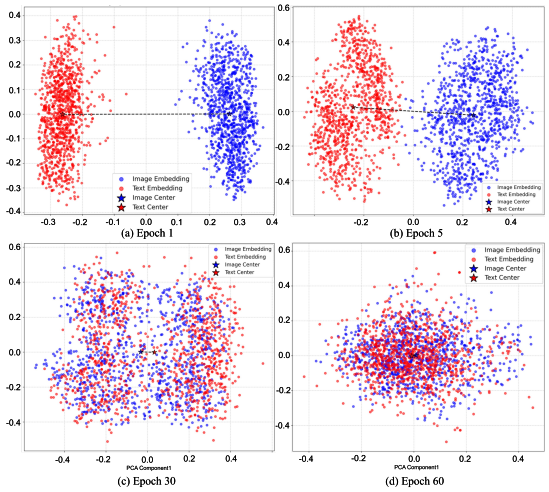

Figure 6: Visualization of modality gap reduction between image and text embeddings in CUHK-PEDE dataset .

F Qualitative Results

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

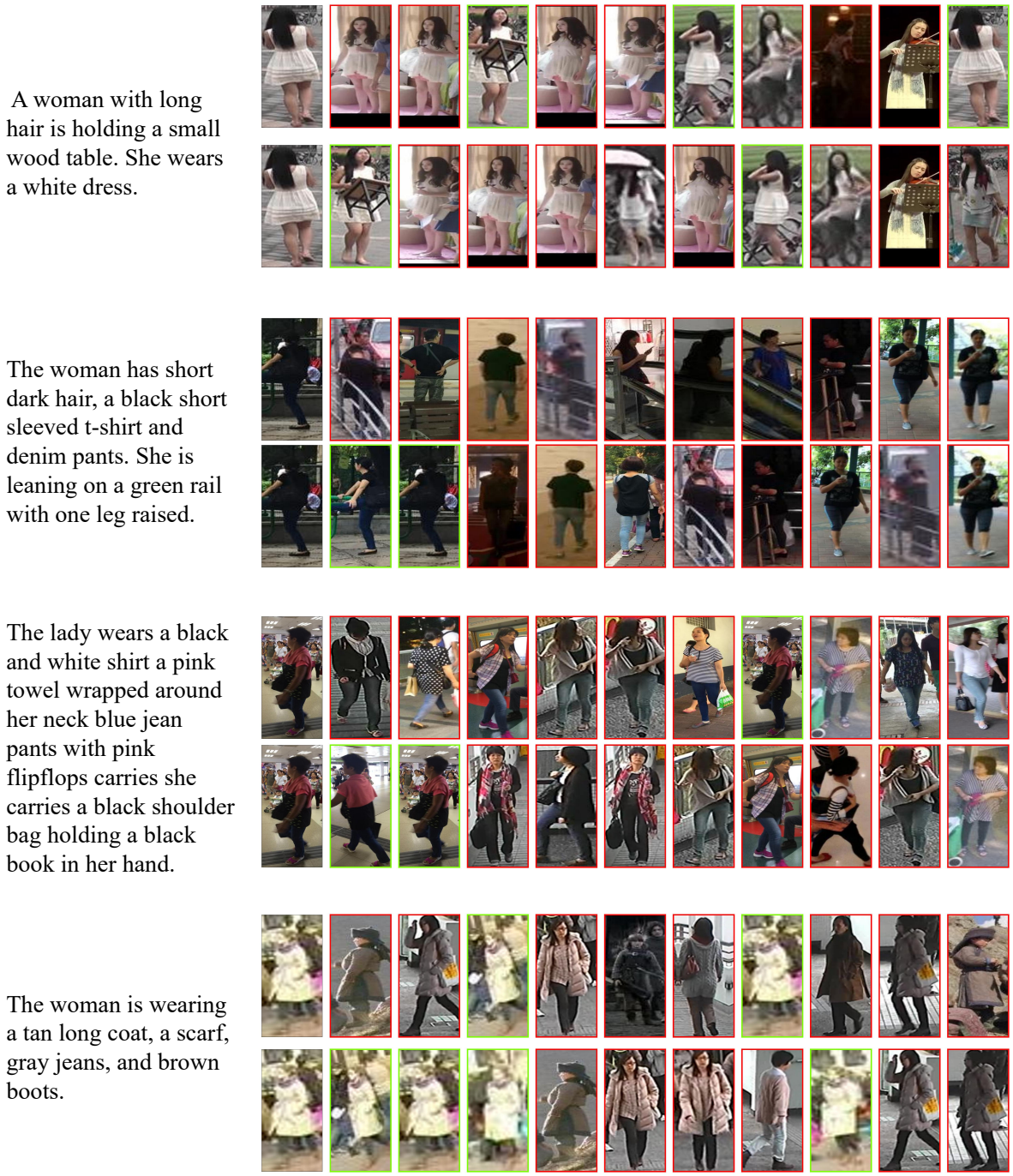

Figure 7: A comparison of the top-10 retrieved results on the CUHK-PEDES dataset between the baseline IRRA (first row) and our DRAGE (second row) is presented for each text query. The matched and mismatched person images are highlighted with red and green rectangles, respectively. Additionally, the first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