Cross-modal Collaborative Representation Learning for Text-to-Image Person Retrieval

Shuanglin Yan[1], **Jun Liu**[2], **Neng Dong**[1], **Liyan Zhang**[3*] and **Jinhui Tang**[1]
[1]Nanjing University of Science and Technology
[2]Lancaster University
[3]Nanjing University of Aeronautics and Astronautics
{shuanglinyan, neng.dong, jinhuitang}@njust.edu.cn, j.liu81@lancaster.ac.uk, zhangliyan@nuaa.edu.cn

Abstract

Text-to-image person retrieval (TIPR) aims to find images of the same identity that match a given text description. Current TIPR methods mainly focus on mining the association between images and texts, ignoring their potential complementarity. Besides, existing matching losses treat all positive pairs from the same identity equally, leading to noisy correspondences. In this paper, we propose CoRL: a cross-modal Collaborative Representation Learning framework designed to improve TIPR by effectively leveraging the complementarity between modalities. The text typically contains identity details with less noise, which helps distinguish visually similar pedestrians. This inspires us to integrate it into the corresponding image to emphasize identity-related and modality-shared visual information. However, corresponding text for each image is not always available, especially during inference. Accordingly, we introduce a Virtual-text Embedding Synthesizer that generates high-quality virtual-text features for cross-modal collaboration, eliminating the need for actual texts. We then design a Cross-Modal Collaboration learning process, incorporating a Cross-modal Relation Consistency loss to promote interaction and fusion between image and virtual-text features for mutual enhancement. Additionally, an Identity-bounded Matching loss is proposed to handle different types of image-text pairs distinctly, leading to more accurate cross-modal correspondences. Extensive experiments on multiple benchmarks demonstrate the superiority of CoRL over existing TIPR methods.

1 Introduction

Person re-identification (ReID) aims to retrieve a person-of-interest across different camera networks. ReID models [Gong *et al.*, 2022; Li *et al.*, 2023b; Li *et al.*, 2019a] trained on extensive labeled cross-camera image pairs have shown impressive retrieval capabilities. However, the close-set assumption of paired cross-camera images severely limits the practical deployment of these models. Recently, Text-to-image person retrieval (TIPR) [Li *et al.*, 2017] has been proposed to address situations where images are missing under certain cameras. As a realistic extension to ReID, TIPR assumes that text descriptions are available for the missing images, allowing the retrieval of target person images via these descriptions. The model must accurately discover identity information and establish correspondences between images and texts, making TIPR a challenging and meaningful task.

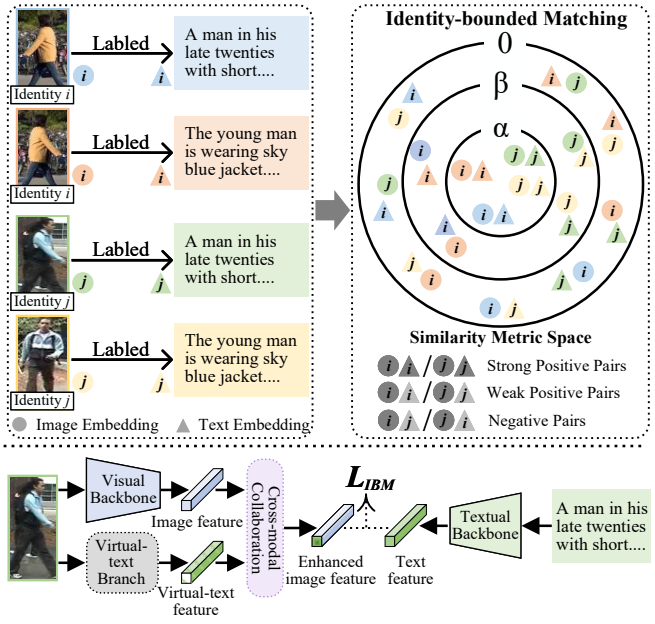

Figure 1: (1) Examples of different types of image-text pairs. Identity-bounded Matching (IBM) loss is designed to handle different types of image-text pairs distinctly by establishing different similarity boundaries, leading to more accurate cross-modal correspondences. (2) We propose a cross-modal collaborative representation learning framework that leverages the inter-modality complementarity to enhance image features without relying on actual texts.

Previous TIPR methods [Chen *et al.*, 2022; Yan *et al.*, 2023d] have relied on single-modal pre-trained models [Dosovitskiy *et al.*, 2021; Devlin *et al.*, 2019] as backbone networks, leveraging their robust initial representations to facilitate fine-tuning on training data. However, these

*Corresponding author

single-modal pre-trained models tend to overemphasize information from one modality, leading to significant modality gaps and alignment challenges. With the rise of vision-language pre-training (VLP), some works [Yan *et al.*, 2023c; Jiang and Ye, 2023] have adopted VLP models as backbone networks, allowing TIPR to benefit from rich multi-modal information. Although promising results have been achieved, existing methods mainly focus on modeling the association between images and texts, often overlooking their potential complementarity. In reality, the text primarily contains pedestrian identity details with less noise. Even when pedestrians are visually similar, their text descriptions are often distinct, making text information more discriminative and robust. This leads us to explore whether inter-modality complementary information can benefit the TIPR task. Inspired by this, we propose integrating text information into images to emphasize discriminative and modality-shared identity details, aiding in distinguishing fine-grained differences and enhancing robustness to background noise. However, this approach requires access to the corresponding text for each image, which is not always available, especially during inference.

Accordingly, we propose a cross-modal **Co**llaborative **R**epresentation **L**earning (**CoRL**) framework that leverages the inter-modality complementary information to enhance image features without relying on actual texts. Specifically, we introduce a Virtual-text Embedding Synthesizer (VES), which utilizes CLIP's vision-language alignment capabilities to produce high-quality virtual-text embeddings directly from images, eliminating the need for actual texts. These generated virtual-text embeddings can replace actual texts in subsequent cross-modal collaboration. To ensure consistency, we align the generated virtual-text embeddings with actual text embeddings at both the feature and semantic levels. To fully exploit the complementarity between images and texts, we design a dual-branch cross-modal collaboration learning process, incorporating a cross-modal relation consistency loss (CRCL). The visual backbone branch focuses on generating image features. The virtual-text branch uses VES to create virtual-text embeddings from images, which are then fed into an Adapter to produce virtual-text features adapted to the target domain. The CRCL loss enforces that image and virtual-text features maintain the same relationship with modality-specific prototypes. This encourages information exchange and collaboration between the two branches, allowing virtual textual information to emphasize the discriminative and modality-shared identity details in the images, thereby enhancing feature discriminability and reducing modality gaps.

Cross-modal matching loss is essential for learning accurate correspondences between modalities. However, existing losses [Zhang and Lu, 2018; Ding *et al.*, 2021; Jiang and Ye, 2023] treat all positive pairs from the same identity equally, resulting in noisy correspondences. Typically, each batch contains three types of image-text pairs: single-view strong positive pairs, cross-view weak positive pairs, and negative pairs, as illustrated in Figure 1. Due to significant appearance differences under the same identity caused by view variations, cross-view weak positive pairs may suffer from noisy correspondences. Consequently, the similarity between these three types of image-text pairs should decrease progressively. To address this, we propose an Identity-bounded Matching (IBM) loss, which defines distinct similarity boundaries for each category of image-text pairs, thereby leading to more precise cross-modal correspondences.

Here are the main contributions of our paper: (1) We propose a cross-modal collaborative representation learning framework, which is the first to leverage the inter-modality complementary information to improve the TIPR task without relying on actual texts. (2) Identity-bounded matching loss is proposed to learn precise cross-modal correspondences. (3) Extensive experiments verify the effectiveness of our method and achieve superior performance on multiple benchmarks.

2 Related Work

2.1 Text-to-Image Person Retrieval

TIPR extends ReID [Gong *et al.*, 2024; Dong *et al.*, 2024a; Dong *et al.*, 2024b] to a more realistic scenario. The TIPR model mainly contains two parts: the backbone network and the feature alignment network [Tang *et al.*, 2025]. A common practice is to use pre-trained backbones to leverage their strong initial representation capabilities, facilitating effective fine-tuning on TIPR data. Earlier methods [Shen *et al.*, 2023; Yan *et al.*, 2023b] employed single-modal pre-trained models such as ViT (pre-trained on ImageNet) and BERT. Recently, the success of vision-language pre-trained models (VLPs) has led to their widespread adoption in TIPR [Yan *et al.*, 2023c; Jiang and Ye, 2023], achieving promising results by exploiting their rich multi-modal knowledge. Notably, some recent works [Yang *et al.*, 2023; Tan *et al.*, 2024] have advanced further by retraining VLPs specifically for TIPR using large-scale datasets, yielding additional performance gains.

For feature alignment network, various strategies have been proposed to align images and texts. Early methods [Li *et al.*, 2017; Zhang and Lu, 2018] directly aligned the global features of images and texts. To achieve fine-grained correspondences, later methods [Chen *et al.*, 2022; Yan *et al.*, 2023d] introduced feature aggregation schemes to generate multiple local features, modeling fine-grained alignment through interaction or guidance between these local features. However, this approach increases storage costs and inference time, reducing practicality. To avoid explicitly generating local features, recent methods [Jiang and Ye, 2023; Li *et al.*, 2023a] propose to inject fine-grained information [Tang *et al.*, 2023; Yan *et al.*, 2023a] into global features by designing auxiliary tasks (e.g., masked language/region modeling) to model fine-grained matching.

Despite these advancements, existing methods mainly emphasize image-text associations but overlook their complementarity. In this study, we explore inter-modality complementarity to enhance image features with text information. Besides, existing matching losses treat image-text pairs of the same identity equally. Although RaSa [Bai *et al.*, 2023] incorporates a discriminator to differentiate between positive pairs, it still relies on conventional matching losses, with the discriminator serving only as a regularizer. In contrast, our IBM loss directly addresses this issue by establishing distinct boundaries for different types of image-text pairs.

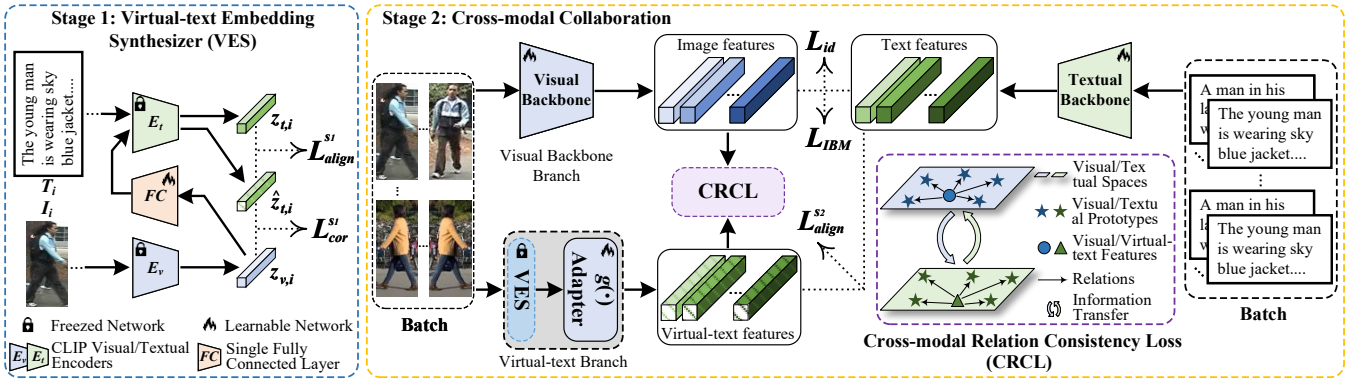

Figure 2: Overview of our CoRL. In the first stage, we introduce a Virtual-text Embedding Synthesizer to generate a virtual-text embedding from each image. The generated virtual-text embedding is then used to replace actual text in modeling multi-modal complementarity with the image. In the second stage, we propose a dual-branch Cross-Modal Collaboration learning process equipped with a Cross-modal Relation Consistency Loss, which encourages the exchange of information and mutual enhancement between image and virtual-text features.

2.2 Vision-Language Pre-Training

Vision-language pre-training involves training large-scale models on vast collections of image-text pairs, demonstrating exceptional capabilities in semantic understanding [Shen and Tang, 2024; Shen *et al.*, 2025], multi-modal alignment, and generalization. It has been widely used in diverse downstream tasks. The primary objective of pre-training is to achieve a deep understanding of both image and text semantics and their correspondences. Some studies [Li *et al.*, 2019b; Li *et al.*, 2021] focus on generating multi-modal representations by encoding interactions between images and texts with tasks such as mask language modeling and image captioning. However, these methods typically necessitate pairwise interactions for all image-text pairs, leading to inefficiency during training and inference, thus limiting their practicality in large-scale applications. Consequently, some studies [Radford *et al.*, 2021; Yao *et al.*, 2022] integrate contrastive representation learning into Vision-language pre-training. These methods encode images and texts separately into a joint space and learn modality-aligned representations by contrasting positive and negative pairs, such as the well-known CLIP [Radford *et al.*, 2021], trained on 400 million image-text pairs. The strong multi-modal alignment capability of CLIP prompts us to introduce it into the TIPR task.

3 Methods

This section presents our proposed CoRL framework, with an overview in Figure 2 and details in the following subsections.

3.1 Problem Formulation

The TIPR dataset contains image-text pairs of persons with multiple identities, where each identity has multiple images collected from different cameras with their annotated text descriptions. The goal is to accurately identify images that belong to the same identity as a given text description. This requires the TIPR model to extract identity-discriminative representations and accurately establish cross-modal correspondences. To achieve this, we first use a pre-trained CLIP as the backbone to provide high-quality initialization and facilitate fine-tuning on the TIPR dataset. Next, we propose a cross-modal collaborative representation learning framework that leverages inter-modality complementary information to enhance image features. Additionally, we design an identity-bounded matching loss to fully leverage identity information and precisely establish cross-modal correspondences.

3.2 Cross-modal Collaborative Representation Learning

Current TIPR methods mainly emphasize image-text associations but overlook their complementarity. Texts offer the following advantages over images: texts primarily contain pedestrian identity details with less noise. Even when pedestrians are visually similar, their text descriptions are often distinct, making text information more discriminative and robust. Complementing images with textual information enhances fine-grained discrimination and reduces modality gaps. However, this requires access to corresponding text for each image, which is not always available, especially during inference. In this paper, we tackle two key challenges: eliminating the dependency on actual text and effectively utilizing text information to enhance image representations. To achieve this, we propose a two-stage cross-modal collaborative representation learning strategy. In the first stage, we develop a Virtual-text Embedding Synthesizer (VES) to generate virtual-text embeddings directly from each image, eliminating the need for actual paired text. In the second stage, we implement a dual-branch cross-modal collaboration learning process to integrate the virtual-text information into images, thereby enhancing image features.

Virtual-text Embedding Synthesizer. To eliminate reliance on actual paired text, we propose a Virtual-text Embedding Synthesizer (VES) that generates a virtual-text embedding directly from each image, as illustrated in Figure 2. CLIP, trained on vast amounts of image-text data, can generate modality-aligned image-text features. We utilize its alignment capabilities to convert images into text embeddings. VES functions as a CLIP-based encoder-decoder model. Specifically, an image I_i is first processed by CLIP's

visual encoder E_v to produce an image embedding $z_{v,i}$. This embedding is then transformed into virtual-text tokens through a fully connected layer. Finally, CLIP's textual encoder E_t decodes these virtual-text tokens to generate the virtual-text embedding $\hat{z}_{t,i}$.

This generated virtual-text embedding $\hat{z}_{t,i}$ can serve as a substitute for the actual text embedding to enhance its image feature, given that the virtual-text and actual-text embeddings are consistent. To ensure this consistency, we introduce an alignment loss $\mathcal{L}^{s1}_{align}$ that aligns the generated virtual-text embedding with the actual text embedding $z_{t,i}$ at both feature and semantic levels, inheriting the advantages of the actual text T_i.

$$\mathcal{L}^{s1}_{align} = \mathcal{L}^{s1}_{nce} + \mathcal{L}^{s1}_{mse} \tag{1}$$

$$\begin{aligned}\mathcal{L}^{s1}_{nce} = &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(z^T_{t,i}\hat{z}_{t,i})}{\sum_{j=1}^{B} exp(z^T_{t,i}\hat{z}_{t,j})} \\ &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(\hat{z}^T_{t,i}z_{t,i})}{\sum_{j=1}^{B} exp(\hat{z}^T_{t,i}z_{t,j})}\end{aligned} \tag{2}$$

$$\mathcal{L}^{s1}_{mse} = \frac{1}{B}\sum_{i=1}^{B} \|\hat{z}_{t,i} - z_{t,i}\|^2 \tag{3}$$

where $z_{t,i}$ denotes the actual text embedding of text T_i, generated by CLIP's visual encoder. $\|\cdot\|^2$ denotes the L_2 distance, and τ_a denotes the temperature factor. To facilitate effective conversion between images and virtual-text embeddings, we impose a correlation loss $\mathcal{L}^{s1}_{cor}$ to ensure that valuable image information is retained as much as possible throughout the conversion process.

$$\begin{aligned}\mathcal{L}^{s1}_{cor} = &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(z^T_{v,i}\hat{z}_{t,i}/\tau_a)}{\sum_{j=1,l_j\neq l_i}^{B} exp(z^T_{v,i}\hat{z}_{t,j}/\tau_a)} \\ &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(\hat{z}^T_{t,i}z_{v,i}/\tau_a)}{\sum_{j=1,l_j\neq l_i}^{B} exp(\hat{z}^T_{t,i}z_{v,j}/\tau_a)}\end{aligned} \tag{4}$$

where l_i represents the identity label of I_i/T_i. The overall objective of our VES is calculated as: $\mathcal{L}^{s1}_{VES} = \mathcal{L}^{s1}_{align} + \mathcal{L}^{s1}_{cor}$. VES enables the generation of virtual-text embedding that is aligned with the actual text for each image, eliminating the need for actual text.

Cross-modal Collaboration. With the virtual-text embedding, we design a dual-branch cross-modal collaboration learning process to integrate it with the corresponding image, thereby enhancing the image feature. Specifically, for an image-text pair (I_i, T_i), the visual backbone branch generates the image feature v_i of image I_i. Simultaneously, the virtual-text branch first produces a virtual-text embedding $\hat{z}_{t,i}$ from image I_i using VES, which is then processed by an Adapter $g(\cdot)$ to yield a virtual-text feature $\hat{t}_i$ adapted to the target domain. Additionally, the text feature t_i for text T_i is generated via the textual backbone. We optimize the Adapter by aligning the virtual-text feature $\hat{t}_i$ with the text feature t_i through $\mathcal{L}^{s2}_{align} = \lambda_1\mathcal{L}^{s2}_{nce} + \mathcal{L}^{s2}_{mse}$.

$$\begin{aligned}\mathcal{L}^{s2}_{nce} = &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(sg(t_i)^T\hat{t}_i)}{\sum_{j=1}^{B} exp(sg(t_i)^T\hat{t}_j)} \\ &- \frac{1}{B}\sum_{i=1}^{B} log\frac{exp(\hat{t}_i^T sg(t_i))}{\sum_{j=1}^{B} exp(\hat{t}_i^T sg(t_j))}\end{aligned} \tag{5}$$

$$\mathcal{L}^{s2}_{mse} = \frac{1}{B}\sum_{i=1}^{B} \|\hat{t}_i - sg(t_i)\|^2 \tag{6}$$

Where $sg(\cdot)$ means stop-gradient operator, which ensures that only the Adapter is updated. With the image feature v_i and the virtual-text feature $\hat{t}_i$, we introduce a cross-modal relation consistency loss (CRCL). This loss enforces that the image and virtual-text features share the same relationship with modality-specific prototypes, promoting the two branches to exchange information and cooperate with each other. For each batch B, we cluster the image and text features based on identity labels to obtain visual and textual prototypes, p_v and p_t. We then compute the cosine similarity between the image feature v_i and the visual prototypes p_v to determine their relationship r^v_i. Similarly, we calculate the relationship r^t_i between the virtual-text feature $\hat{t}_i$ and these textual prototypes p_t. Thus, the CRCL is formulated as:

$$\mathcal{L}^{s2}_{CRCL} = \frac{1}{2|B|}\sum_{i=1}^{B}\left(D_{kl}(r^t_i||r^v_i) + D_{kl}(r^v_i||r^t_i)\right) \tag{7}$$

where D_{kl} is the Kullback-Leibler divergence. Mutual knowledge distillation on instance-to-prototype relations between the visual embedding v_i and virtual-text embedding $\hat{t}_i$ allows them to exchange information and benefit from one another. Incorporating virtual-text information emphasizes the discriminative and modality-shared identity details in the image I_i, boosting the discriminability and robustness of the image feature v_i, making it easier to differentiate visually similar pedestrians while effectively reducing the modality gap between the image I_i and the text T_i.

3.3 Identity-bounded Matching

When a batch contains multiple image-text pairs of the same identity, the model must perform pairwise matching between all images and texts in the batch. Due to view variations, significant differences among samples of the same identity can lead to noisy matching of cross-view image-text pairs of the same identity. Specifically, within the same identity, each image and its corresponding text form a single-view strong positive pair, while image-text pairs across different views are considered as cross-view weak positive pairs. Image-text pairs with different identities are categorized as negative pairs. Existing matching losses often treat strong and weak positive pairs equally, overlooking noisy correspondences of weak positive pairs. To address the issue, we propose an identity-bounded matching (IBM) loss to establish more precise cross-modal correspondences.

To fully leverage identity information, we use a PK sampling strategy to construct batch samples. For each batch,

we randomly sample K identities and then randomly select P images for each identity, with each image annotated with the corresponding text. Thus, each batch contains B = PK images and their corresponding texts. This results in PK strong positive pairs, $KP(P-1)$ weak positive pairs, and $P^2K(K-1)$ negative pairs. We compute the cosine similarity for these image-text pairs, denoted as $\{s_i^{sp}\}_{i=1}^{PK}$, $\{s_i^{wp}\}_{i=1}^{KP(P-1)}$, and $\{s_i^{n}\}_{i=1}^{P^2K(K-1)}$ respectively. To distinguish different image-text pairs, we propose an IBM loss that enforces the condition: $s^{sp}>s^{wp}>s^{n}$. This can be expressed equivalently as: $s^{sp}>\alpha, s^{n}<\beta, \beta<s^{wp}<\alpha$. We further reformulate this condition as follows:

$$(s^{sp}-\alpha)>0, -(s^{n}-\beta)>0, (s^{wp}-\beta)>0, -(s^{wp}-\alpha)>0 \tag{8}$$

where α and β are the upper and lower bounds for s^{sp} and s^{n}, with $\alpha>\beta$. We further implement IBM loss based on logistic loss as follows:

$$\begin{aligned}\mathcal{L}_{IBM} = \frac{1}{PK}\Bigg\{ &\sum_{i=1}^{PK} log\left[1+e^{-\tau_{sp}(s_i^{sp}-\alpha)}\right] \\ &+ \sum_{i=1}^{KP(P-1)} log\left[1+e^{-\tau_{wp}(s_i^{wp}-\beta)}\right] \\ &+ \sum_{i=1}^{KP(P-1)} log\left[1+e^{\tau_{wp}(s_i^{wp}-\alpha)}\right] \\ &+ \sum_{i=1}^{P^2K(K-1)} log\left[1+e^{\tau_{n}(s_i^{n}-\beta)}\right]\Bigg\}\end{aligned} \tag{9}$$

where τ_{sp}, τ_{wp} and τ_n are the temperature factors. Besides, we compute the cross entropy loss $\mathcal{L}_{id}$ on image and text features to classify them by identity. This loss forces the network to focus on the identity information, enabling it to correctly recognize pedestrians of the same identity while distinguishing between those of different identities.

3.4 Optimization and Inference

The optimization of CoRL involves two stages. In the first stage, we train a virtual-text embedding synthesizer, where CLIP's visual and textual encoders are frozen and only a fully connected layer is optimized via $\mathcal{L}_{VES}^{s1}$ to convert image embeddings to virtual-text tokens. We pre-extract and save features from the dataset, allowing us to focus on optimizing this layer, reducing computational cost. In the second stage, VES is frozen, and the backbone network and Adapter are optimized. The objective of the second stage is as follows:

$$\mathcal{L}^{s2} = \mathcal{L}_{IBM} + \mathcal{L}_{id} + \mathcal{L}_{align}^{s2} + \mathcal{L}_{CRCL}^{s2} \tag{10}$$

During inference, both visual and virtual-text features of gallery images are extracted, and calculate their similarities to the query text features. The final retrieval score is obtained by summing these similarities without weighting.

4 Experiments

4.1 Experiment Settings

Datasets and Metrics: The evaluations are conducted on three TIPR datasets. **CUHK-PEDES** [Li *et al.*, 2017] has 40,206 images and 80,412 descriptions of 13,003 people. Each image has 2 descriptions, averaging 23 words. The dataset is split into 34,054 images for training, 3,078 for validation, and 3,074 for testing. **ICFG-PEDES** [Ding *et al.*, 2021] consists of 54,522 image-text pairs of 4,102 persons, with descriptions averaging 37 words. Training uses 34,674 pairs from 3,102 people, with the remaining 1,000 people reserved for evaluation. **RSTPReid** [Zhu *et al.*, 2021] includes 20,505 images of 4,101 people, each with 2 descriptions averaging 23 words. Training includes 3,701 people, while validation and testing include 200 people each. Performance is evaluated using Rank-k accuracy (R@k, k=1, 5, 10).

Implementation Details: Images are resized to 384×128 and augmented with random horizontal flipping, cropping with padding, and random erasing. The maximum length of the text sequence is set to 77, and random masking is employed for text augmentation. We use CLIP-ViT-B/16 as the backbone. Temperature factors are set to τ_a = 0.02, τ_{sp} = 10, τ_{wp} = 5, and τ_n = 40. Loss weight λ_1 is 0.1, and the boundaries α and β in IBM loss are 0.6 and 0.4. Each mini-batch comprises $B = P \times K$ images, , with $P = 32$ identities and $K = 4$ images per identity. In the first stage, only a fully connected layer is optimized for 60 epochs using a cosine learning rate schedule, starting at 1×10^{-4}. In the second stage, we fine-tune the visual/textual backbones with an initial learning rate of 1×10^{-5} and the Adapter with 5×10^{-5}, also using a cosine schedule and trained for 60 epochs. Both stages adopt the Adam optimizer with a linear warm-up over the first 5 epochs. Experiments are implemented using the PyTorch library on a single NVIDIA RTX 3090 (24GB) GPU.

4.2 Comparisons with State-of-the-art Models

Table 1 compares our CoRL with current state-of-the-art methods across three TIPR benchmarks: CUHK-PEDES, ICFG-PEDES, and RSTPReid. Our CoRL achieves leading performance on these benchmarks, underscoring its effectiveness and advantages. On the CUHK-PEDES dataset, CoRL achieves an R@1 accuracy of 78.15%, surpassing the second-best AUL method by 0.92%. For the ICFG-PEDES dataset, our CoRL sets a new state-of-the-art with R@1 and R@5 accuracies of 69.50% and 85.63%, respectively. On the RSTPReid dataset, CoRL delivers impressive results with 69.10%, 87.30%, and 92.90% on R@1, R@5, and R@10. These results highlight the robustness and versatility of CoRL across different scenarios. The superior performance of our method is attributed to its innovative approach in leveraging cross-modal complementarity and establishing precise cross-modal correspondences. CoRL effectively integrates textual information into images to enhance their discriminability and bridge the modality gap, and employs identity-bounded matching loss to address noisy correspondences among different types of image-text pairs.

4.3 Ablation Studies and Analysis

Effectiveness of different components: We conduct an ablation study to assess the effectiveness of various components on CUHK-PEDES in Table 2. 0# represents the result of Baseline, which involves only the backbone network and is trained using SDM and cross-entropy losses. 5# denotes

Methods	Reference	CUHK-PEDES			ICFG-PEDES			RSTPReid		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CFine [Yan *et al.*, 2023c]	TIP'23	69.57	85.93	91.15	60.83	76.55	82.42	50.55	72.50	81.60
VGSG [He *et al.*, 2023]	TIP'23	71.38	86.75	91.86	63.05	78.43	84.36	-	-	-
IRRA [Jiang and Ye, 2023]	CVPR'23	73.38	89.93	93.71	63.46	80.25	85.82	60.20	81.30	88.20
TCB [Zang *et al.*, 2023]	MM'23	74.45	90.07	94.66	61.60	76.33	81.90	65.80	82.85	88.20
DCEL [Li *et al.*, 2023a]	MM'23	75.02	90.89	94.52	64.88	81.34	86.72	61.35	83.95	90.45
RaSa [Bai *et al.*, 2023]	IJCAI'23	76.51	90.29	94.25	65.28	80.40	85.12	66.90	86.50	91.35
CFAM [Zuo *et al.*, 2024]	CVPR'24	72.87	88.61	92.87	62.17	79.57	85.32	59.40	81.35	88.50
MACF [Sun *et al.*, 2024]	IJCV'24	73.33	88.57	93.02	62.95	79.93	85.04	-	-	-
TBPS-CLIP [Cao *et al.*, 2024]	AAAI'24	73.54	88.19	92.35	65.05	80.34	85.47	61.95	83.55	88.75
SAMC [Lu *et al.*, 2024]	TIFS'24	74.03	89.18	93.31	63.68	79.69	85.21	60.80	82.35	89.00
UMSA [Zhao *et al.*, 2024b]	AAAI'24	74.25	89.83	93.58	65.62	80.54	85.83	63.40	83.30	90.30
LSPM [Li *et al.*, 2024a]	TMM'24	74.38	89.51	93.42	64.40	79.96	85.41	-	-	-
IRLT [Liu *et al.*, 2024]	AAAI'24	74.46	90.19	94.01	64.72	81.35	86.31	61.49	82.26	89.23
DCGA [Zhao *et al.*, 2024a]	ICASSP'24	74.54	90.22	95.10	66.56	85.04	86.98	60.49	80.60	89.32
MDRL [Yang *et al.*, 2024]	AAAI'24	74.56	**92.56**	**96.30**	65.88	85.25	**90.38**	-	-	-
Propot [Yan *et al.*, 2024]	MM'24	74.89	89.90	94.17	65.12	81.57	86.97	61.87	83.63	89.70
DP [Song *et al.*, 2024]	AAAI'24	75.66	90.59	94.07	65.61	81.73	86.95	62.48	83.77	89.93
RDE [Qin *et al.*, 2024]	CVPR'24	75.94	90.63	94.04	67.60	82.47	87.17	65.00	84.75	90.60
FSRL [Wang *et al.*, 2024]	ICMR'24	74.86	89.97	94.14	64.93	80.71	86.19	60.65	83.05	89.60
APTM [Yang *et al.*, 2023]	MM'23	76.17	89.47	93.57	68.22	82.87	87.50	66.45	85.60	90.60
MLLM [Tan *et al.*, 2024]	CVPR'24	76.82	91.16	94.46	67.05	82.16	87.33	68.50	87.15	92.10
AUL [Li *et al.*, 2024b]	AAAI'24	77.23	90.43	94.41	69.16	83.32	88.37	**71.65**	**87.55**	92.05
CoRL (Ours)	IJCAI'25	**78.15**	92.16	95.57	**69.50**	**85.63**	88.86	69.10	87.30	**92.90**

Table 1: Performance comparison with state-of-the-art methods on three TIPR benchmarks. The first part lists non-pretrained methods, while the following section presents pretrained methods. R@1, R@5, and R@10 are listed.

No.	IBM	VES	CMC	Pre	R@1	R@5	R@10
0#					70.63	87.67	92.27
1#	✓				74.66	89.46	93.47
2#		✓			73.58	89.08	93.29
3#		✓	✓		74.24	89.70	93.54
4#	✓	✓	✓		75.48	90.45	94.22
5#				✓	75.02	90.21	93.92
6#	✓			✓	77.32	91.76	94.81
7#		✓		✓	75.89	90.33	94.04
8#		✓	✓	✓	76.41	90.77	94.51
9#	✓	✓	✓	✓	78.15	92.16	95.57

Table 2: Effectiveness of different components on CUHK-PEDES.

$\mathcal{L}_{mse}^{s1}$	$\mathcal{L}_{nce}^{s1}$	$\mathcal{L}_{cor}^{s1}$	R@1	R@5	R@10
✓			73.31	89.31	93.29
	✓		73.18	89.39	93.58
✓	✓		73.45	89.46	93.50
✓		✓	73.78	89.43	93.44
	✓	✓	73.64	89.38	93.36
✓	✓	✓	74.24	89.70	93.54

Table 3: Effectiveness of different losses in VES on CUHK-PEDES.

Baseline with additional pre-training (Pre) on LUPerson-MLLM [Tan *et al.*, 2024]. The SDM [Jiang and Ye, 2023] loss treats image-text pairs of the same identity equally, leading to noisy correspondences and poor performance. Comparing 1# (6#) and 0# (5#), our IBM loss creates distinct boundaries for differentiating various categories of image-text pairs and establishes more precise cross-modal correspondences, achieving a 4.03% (2.30%) R@1 improvement. Comparing 2# (7#) and 0# (5#), VES generates a virtual-text feature for each image to enhance its image feature, resulting in a 2.95% (0.87%) R@1 improvement. Besides, CMC facilitates information exchange between image and virtual-text features, allowing them to benefit from each other and resulting in an additional 0.66% (0.52%) improvement. The combination of these components achieves an R@1 accuracy of 75.48% (78.15%), surpassing all methods listed in Table 1 across various settings. These results underscore the positive contributions of each component to the overall performance.

Effectiveness of different losses in VES. Table 3 summarizes the impact of different losses in VES, leading to the following conclusions: (1) $\mathcal{L}_{mse}^{s1}$ and $\mathcal{L}_{nce}^{s1}$ jointly align virtual and actual text embeddings at both feature and semantic levels, which is crucial for ensuring that VES generates virtual-text embeddings aligned with actual text. (2) The introduction of $\mathcal{L}_{cor}^{s1}$ evidently improves performance, highlighting its importance in preserving information during conversion. (3) combining all losses effectively ensures the generation of high-quality virtual-text embeddings.

Impact of boundaries α and β in IBM: Figure 3 illustrates the effects of α and β. α defines the boundary between strong and weak positive pairs. A large α may cause the model to ignore weak positives, while a small α may disrupt the cross-view matching for weak positive pairs. We set $\alpha = 0.6$. β denotes the boundary between weak positive and negative pairs. An improper β can introduce noisy correspondences—too large makes the model overly tolerant to negatives, while too small blurs the distinction with weak positives. We set $\beta = 0.4$ to balance these trade-offs.

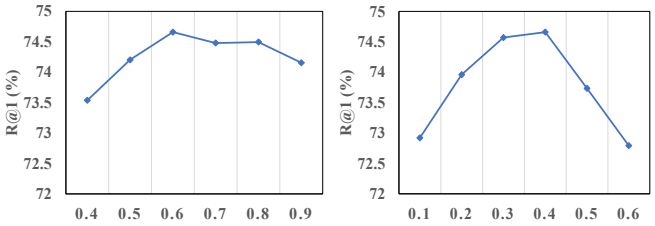

Figure 3: Effects of boundaries α (left) and β (right) of IBM loss on CUHK-PEDES.

Method	R@1	R@5	R@10
SDM	70.63	87.67	92.27
IBM*	73.03	87.75	92.66
IBM (Ours)	74.66	89.46	93.47

Table 4: Effectiveness of IBM on CUHK-PEDES.

Effectiveness of IBM. As shown in Table 4, the SDM loss treats strong and weak positive pairs equally, and its performance heavily relies on the number of negative pairs within a batch. When using the PK sampling strategy to construct batches, performance deteriorates significantly due to the reduced number of negative pairs. We modified IBM by removing the middle two lines of Equation 9 (IBM*) to treat strong and weak positive pairs equally, which led to a notable performance drop (-1.63%). This indicates the importance of distinguishing between strong and weak positive pairs for learning accurate cross-modal correspondence. Despite this drop, IBM* still outperforms SDM significantly and is less affected by the number of negative pairs. Our IBM loss offers greater robustness for cross-modal retrieval tasks.

Computational Complexity: We compare the computational cost and inference time of CoRL with classic methods in Table 5. TIPCB and CFine build fine-grained correspondences by learning local features, reducing retrieval efficiency due to pairwise similarity calculations. IRRA and Propot enhance global features with auxiliary fine-grained tasks, improving retrieval efficiency but increasing parameters and storage. In contrast, CoRL avoids fine-grained tasks, reducing parameters. The virtual-text feature in CoRL adds minimal inference time, balancing accuracy and efficiency.

Qualitative Results: We qualitatively assess the effectiveness of our CoRL in Figure 4, showcasing the Top-10 retrieved images for each query text using both Baseline and CoRL. The comparison reveals that our CoRL outperforms Baseline even in cases where Baseline fails, ensuring that images with the same identity as the given query text are ranked highly. This success is attributed to our method's ability to effectively exploit cross-modal complementarity to learn discriminative and modality-shared feature representations, and establish precise cross-modal correspondences.

Method	Parms	FLOPs	Time	R@1
Baseline	155.26	20.27	18.7s	70.63
TIPCB [Chen *et al.*, 2022]	184.75	43.86	25.1s	64.26
CFine [Yan *et al.*, 2023c]	204.74	27.69	37.2s	69.57
IRRA [Jiang and Ye, 2023]	194.54	26.36	18.7s	73.38
Propot [Yan *et al.*, 2024]	245.91	37.35	18.7s	74.89
CoRL (Ours)	155.66	37.60	20.8s	78.15

Table 5: Computational complexity comparison with several state-of-the-art methods on CUHK-PEDES.

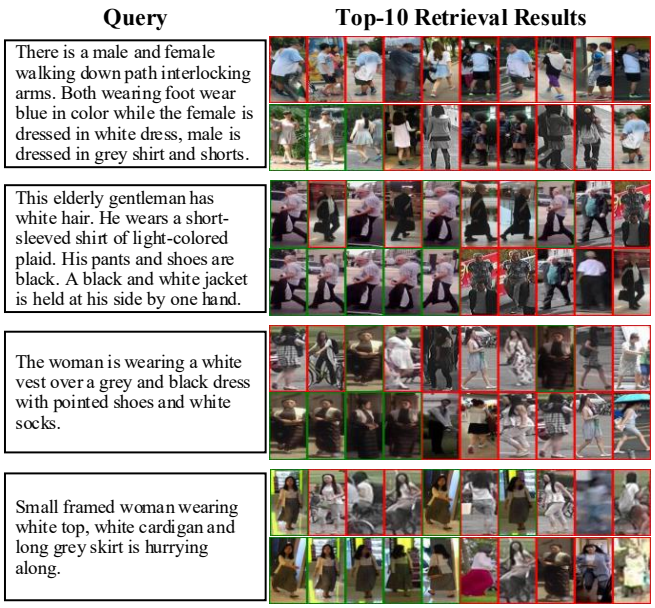

Figure 4: Retrieval results of Baseline (the 1st row) and CoRL (the 2nd row) on CUHK-PEDES. The matched and mismatched person images are marked with green and red rectangles, respectively.

5 Conclusion

In this paper, we propose leveraging multi-modal complementarity to improve the TIPR task. Specifically, we introduce a virtual-text embedding synthesizer that generates high-quality virtual-text embeddings aligned with actual text from images. These virtual-text embeddings replace actual texts, allowing us to explore multimodal complementarity with images and eliminating the dependency on actual texts. Meanwhile, we design a dual-branch cross-modal collaborative learning process equipped with a cross-modal relation consistency loss, which enforces the exchange of information and mutual enhancement between image and virtual-text features, resulting in more discriminative and robust image features for retrieval. Additionally, we propose an identity-bounded matching loss to distinguish between different types of image-text pairs, establishing more accurate cross-modal correspondences. The superior performance of CoRL across multiple TIPR benchmarks underscores its effectiveness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172212 and Grant 62427819, the Natural Science Foundation of Jiangsu Province under Grant BK20230031, the Jiangsu Provincial Science and Technology Major Project under Grant BG2024042.

References

[Bai *et al.*, 2023] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2023.

[Cao *et al.*, 2024] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Chen *et al.*, 2022] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *The North American Chapter of the Association for Computational Linguistics, NAACL*, 2019.

[Ding *et al.*, 2021] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification, 2021.

[Dong *et al.*, 2024a] Neng Dong, Shuanglin Yan, Hao Tang, Jinhui Tang, and Liyan Zhang. Multi-view information integration and propagation for occluded person re-identification. *Information Fusion*, 104:102201, 2024.

[Dong *et al.*, 2024b] Neng Dong, Liyan Zhang, Shuanglin Yan, Hao Tang, and Jinhui Tang. Erasing, transforming, and noising defense network for occluded person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4458–4472, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, ICLR*, 2021.

[Gong *et al.*, 2022] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *IEEE conference on computer vision and pattern recognition, CVPR*, 2022.

[Gong *et al.*, 2024] Yunpeng Gong, Zhun Zhong, Yansong Qu, Zhiming Luo, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. In *Advances in neural information processing systems, NeurIPS*, 2024.

[He *et al.*, 2023] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding. Vgsg: Vision-guided semantic-group network for text-based person search. *IEEE Transactions on Image Processing*, 2023.

[Jiang and Ye, 2023] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Li *et al.*, 2017] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[Li *et al.*, 2019a] Huafeng Li, Shuanglin Yan, Zhengtao Yu, and Dapeng Tao. Attribute-identity embedding and self-supervised learning for scalable person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3472–3485, 2019.

[Li *et al.*, 2019b] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems, NeruIPS*, 2021.

[Li *et al.*, 2023a] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *ACM International Conference on Multimedia, MM*, 2023.

[Li *et al.*, 2023b] Shuang Li, Fan Li, Jinxing Li, Huafeng Li, Bob Zhang, Dapeng Tao, and Xinbo Gao. Logical relation inference and multiview information interaction for domain adaptation person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Li *et al.*, 2024a] Jiayi Li, Min Jiang, Jun Kong, Xuefeng Tao, and Xi Luo. Learning semantic polymorphic mapping for text-based person retrieval. *IEEE Transactions on Multimedia*, pages 1–14, 2024.

[Li *et al.*, 2024b] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Liu *et al.*, 2024] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-inspired invariant representation learning for text-based person retrieval. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Lu *et al.*, 2024] Zefeng Lu, Ronghao Lin, and Haifeng Hu. Mind the inconsistent semantics in positive pairs: Semantic aligning and multimodal contrastive learning for text-based pedestrian search. *IEEE Transactions on Information Forensics and Security*, 19:6409–6424, 2024.

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

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning, ICML*, 2021.

[Shen and Tang, 2024] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. In *Advances in neural information processing systems, NeurIPS*, 2024.

[Shen *et al.*, 2023] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *ACM International Conference on Multimedia, MM*, 2023.

[Shen *et al.*, 2025] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *AAAI Conference on Artificial Intelligence, AAAI*, 2025.

[Song *et al.*, 2024] Zifan Song, Guosheng Hu, and Cairong Zhao. Diverse person: Customize your own dataset for text-based person search. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Sun *et al.*, 2024] Mengyang Sun, Wei Suo, Peng Wang, Kai Niu, Le Liu, Guosheng Lin, Yanning Zhang, and Qi Wu. An adaptive correlation filtering method for text-based person search. *International Journal of Computer Vision*, pages 1–16, 2024.

[Tan *et al.*, 2024] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

[Tang *et al.*, 2023] Hao Tang, Jun Liu, Shuanglin Yan, Rui Yan, Zechao Li, and Jinhui Tang. M3net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *ACM International Conference on Multimedia, MM*, 2023.

[Tang *et al.*, 2025] Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1958–1974, 2025.

[Wang *et al.*, 2024] Di Wang, Feng Yan, Yifeng Wang, Lin Zhao, Xiao Liang, Haodi Zhong, and Ronghua Zhang. Fine-grained semantics-aware representation learning for text-based person retrieval. In *International Conference on Multimedia Retrieval, ICMR*, 2024.

[Yan *et al.*, 2023a] Rui Yan, Lingxi Xie, Xiangbo Shu, Liyan Zhang, and Jinhui Tang. Progressive instance-aware feature learning for compositional action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10317–10330, 2023.

[Yan *et al.*, 2023b] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *ACM international conference on Multimedia, MM*, 2023.

[Yan *et al.*, 2023c] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, pages 1–14, 2023.

[Yan *et al.*, 2023d] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023.

[Yan *et al.*, 2024] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. Prototypical prompting for text-to-image person re-identification. In *ACM International Conference on Multimedia, MM*, 2024.

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

[Yang *et al.*, 2024] Fan Yang, Wei Li, Menglong Yang, Binbin Liang, and Jianwei Zhang. Multi-modal disordered representation learning network for description-based person search. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Yao *et al.*, 2022] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *International Conference on Learning Representations, ICLR*, 2022.

[Zang *et al.*, 2023] Xianghao Zang, Wei Gao, Ge Li, Han Fang, Chao Ban, Zhongjiang He, and Hao Sun. A baseline investigation: Transformer-based cross-view baseline for text-based person search. In *ACM International Conference on Multimedia, MM*, 2023.

[Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *European Conference on Computer Vision, ECCV*, 2018.

[Zhao *et al.*, 2024a] Weichen Zhao, Yuxing Lu, Ge Jiao, and Yuan Yang. Dual-color granularity alignment for text-based person search. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2024.

[Zhao *et al.*, 2024b] Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *AAAI Conference on Artificial Intelligence, AAAI*, 2024.

[Zhu *et al.*, 2021] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. DSSL: deep surroundings-person separation learning for text-based person retrieval. In *ACM International Conference on Multimedia, MM*, 2021.

[Zuo *et al.*, 2024] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.