Exploring the Potential of Large Vision-Language Models for Unsupervised Text-Based Person Retrieval

Zongyi Li[1], Li Jianbo[1], Yuxuan Shi[2*] Jiazhong Chen[1], Shijuan Huang[1], Linnan Tu[1], Fei Shen[3], Hefei Ling[1*]

[1]Department of Computer Science and Technology, Huazhong University of Science and Technology
[2]National Engineering Research Center of Educational Big Data and the Faculty of Artificial Intelligence in Education, Central China Normal University
[3]Nanjing University of Science and Technology
{zongyili, jianboli, jzchen,huang_shijuan,lntu,lhefei}@hust.edu.cn,shiyx@ccnu.edu.cn, feishen@njust.edu.cn

Abstract

The aim of text-based person retrieval is to identify pedestrians using natural language descriptions within a large-scale image gallery. Traditional methods rely heavily on manually annotated image-text pairs, which are resource-intensive to obtain. With the emergence of Large Vision-Language Models (LVLMs), the advanced capabilities of contemporary models in image understanding have led to the generation of highly accurate captions. Therefore, this paper explores the potential of employing Large Vision-Language Models for unsupervised text-based pedestrian image retrieval and proposes a Multi-grained Uncertainty Modeling and Alignment framework (MUMA). Initially, multiple Large Vision-Language Models are employed to generate diverse and hierarchically structured pedestrian descriptions across different styles and granularities. However, the generated captions inevitably introduce noise. To address this issue, an uncertainty-guided sample filtration module is proposed to estimate and filter out unreliable image-text pairs. Additionally, to simulate the diversity of styles and granularities in captions, a multi-grained uncertainty modeling approach is applied to model the distributions of captions, with each caption represented as a multivariate Gaussian distribution. Finally, a multi-level consistency distillation loss is employed to integrate and align the multi-grained captions, aiming to transfer knowledge across different granularities. Experimental evaluations conducted on three widely-used datasets demonstrate the significant advancements achieved by our approach.

Introduction

Text-based person retrieval aims to identify specific individuals based on a provided textual description (Li et al. 2017; Jiang and Ye 2023; Shen et al. 2023a). Diverging from traditional person re-identification methods, text-based person retrieval utilizes both image and text as input with the aim of integrating them into a unified space for precise retrieval. Consequently, recent researchers have endeavored to develop robust representations by employing robust loss functions and multi-grained feature extraction methodologies. Nevertheless, the efficacy of these endeavors relies heavily on the accessibility of annotated textual descriptors aligned with each image, which imposes substantial demands in terms of human labor and resources for labeling.

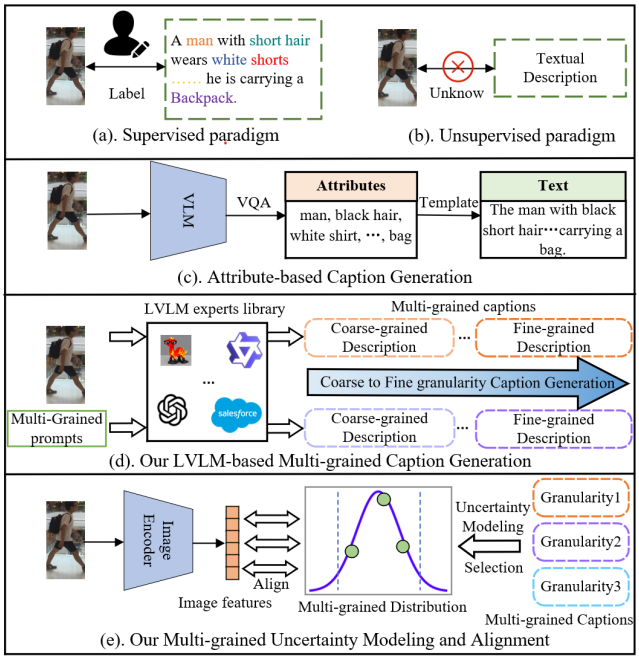

Figure 1: (a) and (b) depict the supervised and unsupervised Text-based person retrieval paradigms. (c) denotes the attribute-based caption generation method. (d) introduces our LVLM-based Multi-grained caption generation method. (e) represents our multi-grained uncertainty Modeling and Alignment framework

In order to address these challenges, researchers have recently endeavored to alleviate the laborious and time-intensive process of caption annotation for pedestrian images through cross-domain training and automated labeling techniques. MAN (Jing et al. 2020) introduces a moment alignment network aimed at addressing the cross-domain text-based person search task, thereby mitigating the absence of pairwise text-image identity labels in the target domain. However, this methodology still requires the inclusion of textual data within the dataset. In contrast, GTR (Bai

*Corresponding authors.

et al. 2023c) proposes a generation-then-retrieval framework that operates without the need for parallel Image-Text data, thus liberating it from dependence on manually annotated text descriptions. This is achieved by labeling the images using attributes extracted from a Visual Question Answering (VQA) model. Furthermore, UniPT(Shao et al. 2023) suggests generating pseudo-textual descriptions by extracting attributes through the CLIP paradigm, employing a divide-and-conquer strategy. Nevertheless, both of these approaches focus on manually designing attributes, potentially overlooking finer-grained details and introducing noise. Additionally, captions generated by these methods tend to adopt a uniform style due to the utilization of standardized templates, deviating from the variability observed in real-world scenarios.

In light of recent advancements in Large Vision-Language Models (LVLMs), our objective is to leverage their notable capabilities in multi-modal comprehension to address the unsupervised text-based person retrieval task, which utilizes only image data to train a cross-modal retrieval model, as illustrated in Figure 1(b). Diverging from previous attribute-based caption generation approaches, shown in Figure 1(c), we aim to exploit the advantages of LVLMs equipped with multiple instructive prompts. Specifically, various LVLMs were employed to generate captions in distinct styles, simulating different annotation styles. The granularity of the generated captions was regulated by prompt instructions with defined constraints, ranging from coarse to fine. By utilizing a diverse array of LVLMs and prompts, we obtained a series of multi-grained descriptions for each image. Compared to attribute-based caption generation, this approach is capable of generating more diverse and multi-granular captions, which are characteristic features of real-world scenarios.

Although the generated captions provide a comprehensive description of person images, they inherently contain noise. To address this, we propose a Multi-grained Uncertainty Modeling and Alignment (MUMA) framework, which utilizes generated pseudo captions for training the retrieval model. We introduce an uncertainty-guided sample filtration module to select a clean subset, employing a Gaussian Mixture Model to characterize uncertainty distributions and eliminate noise. Additionally, we model the diverse textual representations with a multivariate Gaussian distribution to account for varying granularity and style, thereby capturing the reality that an image is typically described in different styles.

Furthermore, to achieve multi-granularity feature alignment and enable unified text feature extraction, we propose a multi-granularity consistency alignment module. This module integrates and aligns multi-granularity captions, facilitating knowledge transfer across different granularities. Specifically, we use integrated text-image similarity and matching scores as teacher signals to guide the learning of text features at each granularity. Through this mutual learning process, our method effectively aligns features across different granularities.

Our contribution can be summary as follows:

- We propose a methodology using Large Vision-Language Models to generate coarse-to-fine captions for each image, showcasing diverse styles and granularities.
- We introduce an uncertainty-guided sample filtration and modeling strategy to evaluate the reliability of generated text-image pairs. By modeling the inherent uncertainty in pseudo text, our approach mitigates noise and captures the multi-grained characteristic of text in real-world scenarios.
- We propose a multi-grained consistency alignment module designed to align the distribution of generated multi-grained captions with ensemble predictions. This module facilitates the extraction of unified text features across different granularities.

Related Work

Large Vision-Language Models

Large Vision-Language Models (LVLMs) have shown remarkable performance in various tasks by adapting Large Language Models (LLMs) for multimodal inputs and outputs, focusing primarily on aligning multimodal semantics. BLIP-2 (Li et al. 2023) introduces an efficient framework using a lightweight Q-Former to bridge modality gaps while utilizing frozen LLMs, enabling zero-shot image-to-text generation through natural language prompts. MiniGPT-4 (Zhu et al. 2023) aligns a pre-trained vision encoder with the LLM by training a single linear layer, effectively replicating GPT-4's capabilities (Achiam et al. 2023). Qwen-VL (Bai et al. 2023a) is a multilingual model that handles multiple image inputs during training and excels in object detection and localization tasks. LLaVA (Liu et al. 2024) pioneers Instruction Tuning in the multimodal domain, addressing data scarcity by introducing an open-source cross-modal instruction-following dataset developed with ChatGPT/GPT-4, and the LLaVA-Bench benchmark for performance evaluation. InstructBLIP (Dai et al. 2024) builds on BLIP-2 with updates to the Q-Former for Multi-modal Instruction Tuning, enhancing adaptable and varied feature extraction.

Text-based Person Retrieval

Li *et al.* (Li et al. 2017) initially introduced the text-based person retrieval task and presented the cross-modal dataset CUHK-PEDES, which comprises image-text pairs. Zhang *et al.* (Zhang and Lu 2018) proposes two key losses, the cross-modal projection matching (CMPM) loss and the cross-modal projection classification (CMPC) loss, to accurately gauge the similarity between visual and textual inputs. IRRA (Jiang and Ye 2023) introduces a novel cross-modal matching loss based on the CMPM loss and incorporates fine-grained interaction to enhance global alignment without the need for additional supervision. Conversely, DSSL (Zhu et al. 2021) and ViTAA (Wang et al. 2020) focus on foreground-background separation, leveraging the pedestrian foreground to supervise network training and mitigate the impact of background noise. Additionally, some methods (Zhao et al. 2024; Shen et al. 2023b) employ uncertainty modeling to align image and text distributions.

In weakly supervised methods, MAN (Jing et al. 2020) first explored cross-domain text-based person retrieval,

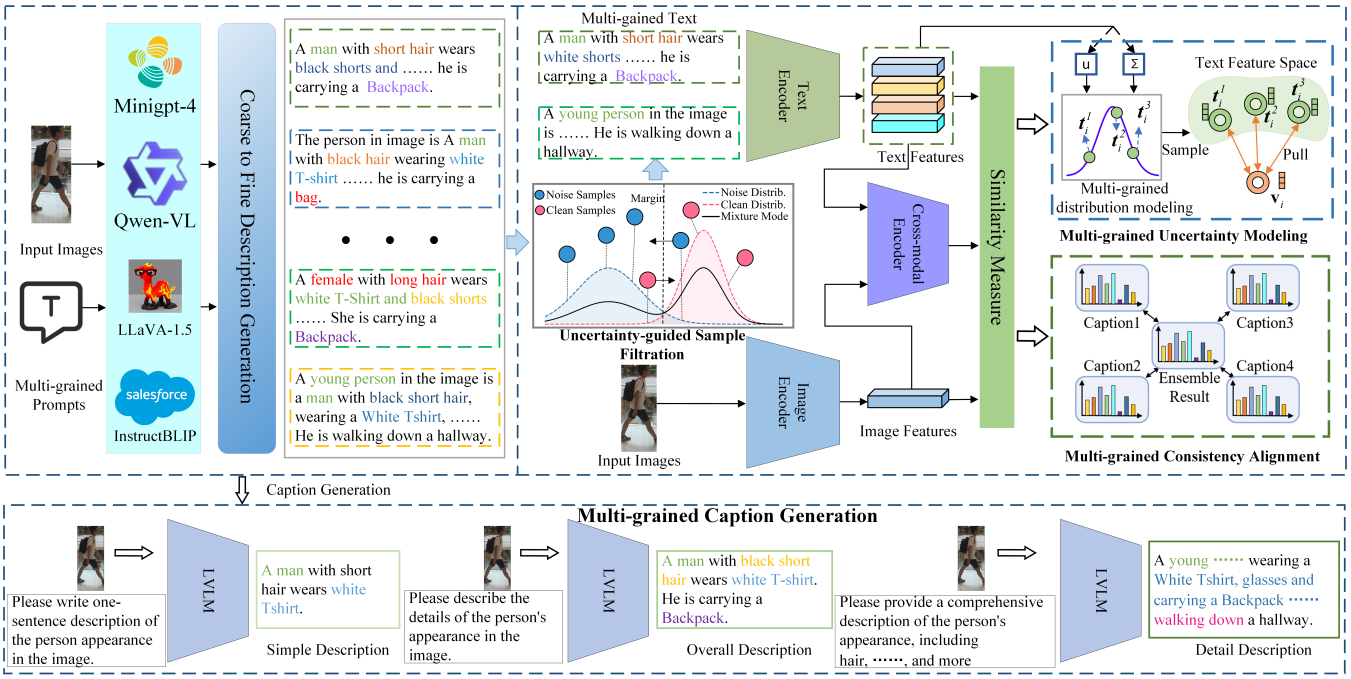

Figure 2: An overview of our proposed Unsupervised Text-based Person Retrieval pipeline. We employ multiple large Vision-Language models with multi-grained prompts to generate diverse captions. An uncertainty modeling and alignment training framework designed to elucidate the relevance of generated captions for the training of text-based person retrieval models.

while CMMT (Zhao et al. 2021) addressed missing ID labels through clustering-based pseudo-labels. However, both methods require textual data. To address this limitation, GTR (Bai et al. 2023c) generated text descriptions for person images using VQA inspired techniques. Recent researchers (Shao et al. 2023; Yang et al. 2023b; Zuo et al. 2024; Tan et al. 2024; Jing et al. 2023; Shen et al. 2024a) focus on automatically constructing large-scale image-text datasets for text-based person search model pretraining.

Learning with Noisy Labels

Cross-modal noise, including false positive and negative instances, can lead to model generalization deterioration. To address this issue, Qin *et al.* (Qin et al. 2024b) proposed Active Complementary Loss and Self-refining Correspondence Correction (SCC) for stable soft correspondences. Hu *et al.* (Hu et al. 2023) tackled partially mismatched pairs through an unbiased estimator of cross-modal retrieval risk. Yang *et al.* (Yang et al. 2023a) introduced Bidirectional Cross-modal similarity consistency for label correction, based on the assumption that similar images should share similar textual descriptions. Recent works (Chen et al. 2022; Ji et al. 2023; Li et al. 2022b; Cui et al. 2024; Shen et al. 2024b) have focused on modeling noise uncertainty to learn robust features for multi-modal retrieval.

Methodology

In contrast to supervised text-based person retrieval, which depends on the availability of image-text pairs along with their associated identity labels $\{T, I, Y\}$, unsupervised text-based person retrieval centers on the development of a robust cross-modal retrieval model θ solely through the utilization of image data $I = \{I_1, I_2, I_3, ..., I_N\}$, where N denotes the total number of images within the dataset.

Therefore, the initial step in the unsupervised text-based person search framework involves the generation of pseudo captions for each image. To promote diversity in captions, a variety of Large Vision-Language Models with distinct prompts are utilized to produce captions ranging from coarse to fine granularity. Subsequently, in order to efficiently leverage the generated image-text pairs, we propose a Multi-grained Uncertainty Modeling and Alignment framework (MUMA) to train a robust text-based retrieval model capable of managing noisy image-text captions and extracting unified-grained text features. Our overall framework is illustrated in Figure 2.

Multi-grained Caption Generation

Given the absence of textual descriptions for images in unsupervised-based person search, it is imperative to generate captions for each image and subsequently fine-tune the cross-modal retrieval model. Previous approaches have relied on the Attribute VQA(Bai et al. 2023c) and attribute classification(Shao et al. 2023) methods to complete the attribute template for caption generation. However, these methodologies necessitate pre-defining the attribute space, potentially overlooking finer attributes. Furthermore, the inferred attributes often lack accuracy, leading to increased noise in the generated captions.

To enhance the diversity of caption styles for person images, a range of open-source Large Vision-Language Models (LVLMs) is employed. These models consist of Qwen-VL (Bai et al. 2023a), LLaVA-1.5(Liu et al. 2024), MiniGPT-4(Zhu et al. 2023), and instruct-BLIP(Dai et al. 2024). These multi-modal LVLMs are utilized to process person images in conjunction with prompts, resulting in descriptive outputs. Specifically, to generate captions with diverse stylistic variations, prompts with varying instructions are employed, spanning from coarse to fine granularity. As depicted in Figure 2, employing the prompt "Please write a one-sentence description of the person's appearance in the image." may elicit concise sentences from the multi-modal LVLM. Though such brief captions may not encapsulate the entirety of details within the image, they effectively capture its most salient attributes. In contrast, prompts like "Please describe the intricate details of the person's appearance in the image" prompt the LVLM to produce more detailed captions. However, while offering more comprehensive descriptions, these captions may also introduce additional noise. Consequently, different levels of granularity exhibit distinct characteristics, prompting the construction of varied captions for person descriptions.

By employing n LVLMs and g multi-grained prompts, we generate a total of $n \times g$ descriptions associated with each image in the pseudo dataset $\{I_i, \mathbf{T}_i\}$, where $\mathbf{T}_i = \{T_i^1, T_i^2, ..., T_i^{n\times g}\}$. These diverse captions generated by LVLMs play a crucial role in training a robust cross-modal person retrieval model. However, due to inherent uncertainties within LVLMs, mismatches between pseudo image-text pairs may occur, potentially compromising the performance of the retrieval model.

To address these challenges, we introduce a Multi-grained Uncertainty Modeling and Alignment (MUMA) framework aimed at handling scenarios characterized by noisy captions and multiple granularities. As shown in figure 2, the proposed MUMA framework consists of three sub-modules: Uncertrainty-guided Sample Filtration, Multi-grained Uncertainty Modeling and Multi-grained Consistency Alignment.

Uncertainty-guided Sample Filtration

As descriptions are generated by the large Vision-language models, noise inevitably appears in the textual content, leading to potential inaccuracies in correspondence with the associated image. To identify and select clean image-text pairs suitable for training, we observe that clean pairs typically exhibit higher similarities and matching scores compared to their noise counterparts. Consequently, we utilize the distribution of cosine similarities and matching scores to distinguish noisy samples from clean ones, subsequently selecting reliable pairs for training purposes. Following previous noise-correspondence methodologies (Yang et al. 2023a; Qin et al. 2023), the similarity distribution of clean and noise pairs can be modeled by the Gaussian Mixture Model (GMM). In this manner, we can infer the probability of a given pair being noisy or clean by fitting the GMM to the distribution of pair similarity and matching score. For instance, considering the pair similarity $s_i^{itc} = cos(v_i, t_i)$, we utilize a Gaussian Mixture Model to characterize the similarity distribution of image-text pairs within the training dataset:

$$p(s_i^{itc} \mid \theta) = \sum_{k=1}^{K} \alpha_k \phi\left(s_i^{itc} \mid \theta_k\right), \tag{1}$$

where α_k denotes the mixture coefficient, and $\phi\left(s_i^{itc} \mid \theta_k\right)$ represents the probability density of the k-th component. s_i^{itc} denotes the cosine similarity between image I_i and text T_i. We posit that pairs exhibiting higher similarity and matching scores correspond to clean pairs, while others constitute noise pairs. The Expectation-Maximization algorithm is employed to optimize the GMM. The probability of cleanliness for the i-th pairs can be computed using the posterior probability as follows:

$$w_i^{itc} = p\left(\theta_k \mid s_i^{itc}\right) = p\left(\theta_k\right) p\left(s_i^{itc} \mid \theta_k\right) / P\left(s_i^{itc}\right), \tag{2}$$

where θ_k refers to the Gaussian component with a higher mean. Similarly, we can obtain the cleanliness probabilities w_i^{itm} based on matching score by replacing s_i^{itc} with m_i^{itm} in Eq.1 and Eq.2, where m_i^{itm} is the i-th matching score obtained by the cross-modal encoder. We denoted samples with both $w_i^{itc} > th$ and $w_i^{itm} > th$ as clean pairs, which can be expressed as :

$$\mathcal{D}_{train}^{c} = \left\{(I_i, T_i) \mid p(\theta_{k=1} \mid s_i^{itc}) > th,\ p(\theta_{k=1} \mid s_i^{itm}) > th\right\}, \tag{3}$$

Here, $\mathcal{D}_{train}^{c}$ denotes the clean training set. The mean of these two cleanliness probabilities can be interpreted as the reliability of the sample pair:

$$w_i^{clean} = \left(p\left(\theta_{k=1} \mid s_i^{itc}\right) + p\left(\theta_{k=1} \mid s_i^{itm}\right)\right)/2. \tag{4}$$

Moreover, in each epoch, instead of utilizing all $n \times g$ generated texts for each image, we sample K reliable texts for image I_i from the set $\{T_i^1, T_i^2, ..., T_i^{n\times g}\}$ based on the probability $\mathbf{w}_i^{clean}$, where texts with higher w_i^{clean} for the image are more likely to be selected.

Multi-grained Uncertainty Modeling

In addition to sampling clean image-text pairs for robust model training, we also model the uncertainty in the feature space and generate more diverse features by representing the features of generated multi-grained captions as a multivariate Gaussian distribution. As illustrated in Figure 2, we initially calculate the mean and variance features of pseudo captions for each image I_i. Denoting K as the number of multi-grained captions for image I_i, their mean and standard deviation are calculated as follows:

$$\mu_i^T = \frac{1}{\sum_{j=1}^{K} w_i^k} \sum_{k=1}^{K} w_i^k E_{text}(T_i^k), \tag{5}$$

$$\Sigma_i^T = \sqrt{\frac{1}{K} \sum_{j=1}^{K} \left(E_{text}(T_i^j) - \mu_i^T\right)^2}, \tag{6}$$

where $E_{text}(\cdot)$ represents the text encoder. After obtaining the mean and variance of the text, the caption for each

image can be modeled as a multivariate Gaussian distribution $p(T|I_i) \sim \mathcal{N}\left(\mu_i^T, (\Sigma_i^T)^2\right)$. This distribution for multi-grained captions can be expressed as:

$$\mathbf{t}_i^{(m)} = \epsilon^m \cdot \Sigma_i^T + \mu_i^T, \tag{7}$$

where $\epsilon^{(m)} \sim \mathcal{N}(0, I)$ is random sampled from the normal distribution. Given the image I_i, we obtain $K + M$ positive text features, consisting of K features from LVLM-generated captions and M sampled features from the monitored uncertainty distribution. Consequently, we compute the contrastive loss between the image and the text feature. For each image, the uncertainty-aware contrastive loss is calculated among these generated and sampled positive text features, which can be expressed as:

$$\mathcal{L}_{\text{uitc}}^{i2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\sum_{\mathbf{t}_i\in\mathcal{P}_i\cup\mathcal{P}_i^s} w_{t_i}^{clean}\exp\left(s\left(\mathbf{v}_i,\mathbf{t}_i\right)\right)}{\sum_{\mathbf{t}_j\in\left\{\mathcal{P}_i\cup\mathcal{P}_i^s\cup\widetilde{\mathcal{P}}_i\right\}}\exp\left(s\left(\mathbf{v}_i,\mathbf{t}_j\right)\right)}, \tag{8}$$

where $\mathcal{P}_i$ and $\mathcal{P}_i^s$ denote the set of generated K text and sampled M text features, respectively, while $\widetilde{\mathcal{P}}_i$ represents the set of negative text features. Similarly, the uncertainty-aware text-to-image contrastive loss can be expressed as:

$$\mathcal{L}_{\text{uitc}}^{t2i} = -\frac{1}{B}\sum_{i}^{B}\sum_{\mathbf{t}_i\in\mathcal{P}_i\cup\mathcal{P}_i^s} w_{t_i}^{clean}\log\frac{\exp\left(s\left(\mathbf{t}_i,\mathbf{v}_i\right)\right)}{\sum_j^B\exp\left(s\left(\mathbf{t}_i,\mathbf{v}_j\right)\right)}, \tag{9}$$

where $w_{t_i}^{clean}$ is the reliability scores for the text-image pair estimated in section 3.2. In this manner, the image features are pulled with multi-grained text, enhancing the ability to handle various scenarios with different granularity. Therefore, the multi-grained uncertainty-guided contrastive loss can be expressed as:

$$\mathcal{L}_{\text{uitc}} = L_{\text{uitc}}^{i2t} + L_{\text{uitc}}^{t2i}. \tag{10}$$

Additionally, we employ an uncertainty-guided image-text matching loss to learn the multi-grained image-text matching relations. We designate all multi-grained positive texts as positive image-text pairs, forming the set $\mathcal{P}$. For each image and text pair, we sample negative pairs based on their similarities, resulting in the negative set $\widetilde{\mathcal{P}}$. Reliability scores are also applied to image-text pairs to mitigate the influence of noise samples:

$$\mathcal{L}_{\text{uitm}} = \mathbb{E}_{(I,T)\sim\mathcal{P}\cup\widetilde{\mathcal{P}}}\boldsymbol{w}^{clean}\mathrm{H}\left(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T)\right). \tag{11}$$

Multi-grained Consistency Alignment

Considering the varied granularity and stylistic characteristics found in textual descriptions within real-world scenarios, we utilize the mean features of our pseudo multi-grained captions, denoted as μ, to facilitate the establishment of unified granularity text features via a multi-grained ensemble methodology. As a result, the computed similarity between the mean multi-grained caption features and image features can be construed as the measure of unified granularity similarities and can be expressed as follows:

$$\bar{p}_{i,j}^{\text{i2t}}(I) = \frac{\exp\left(s\left(v_i,\mu_i^t\right)/\tau\right)}{\sum_{m=1}^{B}\exp\left(s\left(v_i,\mu_m^t\right)/\tau\right)}, \tag{12}$$

Denoting the similarity between the mean caption features and image features within a batch as $\bar{\boldsymbol{p}}^{i2t}$ and $\bar{\boldsymbol{p}}^{t2i}$ respectively, we can define the ensemble labels as follows:

$$\bar{\boldsymbol{y}}^{i2t} = (1-\alpha)\boldsymbol{y}^{i2t} + \alpha_1\bar{\boldsymbol{p}}^{\text{i2t}}(I) \tag{13}$$

Then we utilize the Kullback-Leibler (KL) divergence to align the distribution of multi-grained text-image similarities with the unified image-text distribution. This regularization aids in transferring knowledge between different granularities. Consequently, the multi-grained consistency alignment loss can be represented as:

$$\mathcal{L}_{\text{mcl}} = \frac{1}{K}\Big(\sum_{i=1}^{K}\mathrm{KL}(\bar{\boldsymbol{y}}^{i2t}\|\boldsymbol{p}_k^{i2t}(I)) + \sum_{i=1}^{K}\mathrm{KL}(\bar{\boldsymbol{y}}^{t2i}\|\boldsymbol{p}_k^{t2i}(I))\Big), \tag{14}$$

where K represents the number of generated descriptions, and $\boldsymbol{p}_k^{t2i}(I)$ and $\boldsymbol{p}_k^{i2t}(I)$ denote the relationship between the image and the k-th style's caption. Through the utilization of this multi-grained consistency loss, our approach effectively transfers knowledge from each granularity level, thereby enhancing the robustness of features.

Objective Functions

We incorporate the multi-grained uncertainty loss and the multi-level consistency alignment loss as our ultimate objective functions:

$$\mathcal{L}_{\text{uitc}} = \mathcal{L}_{\text{uitc}} + \mathcal{L}_{\text{uitm}} + \mathcal{L}_{\text{mcl}} \tag{15}$$

During the evaluation, our model only comprises the image, text, and cross-modal encoders for similarity and matching score computation, similar to BLIP(Li et al. 2022a) without the need for additional steps.

Experiments

Datasets and protocol

We evaluate our approach using three Text-based Person Retrieval datasets: CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021). Our training solely utilizes image data, devoid of any dependency on manually annotated text data. During the testing phase, captions from the dataset are leveraged for retrieval.

The Rank-K metric (where k = 1, 5, 10) is a commonly utilized evaluation measure fundamental to text-based person retrieval. It entails the identification of the most pertinent one/five/ten image(s) by assessing the similarity between textual descriptions and images. Additionally, to provide a comprehensive evaluation, we integrate the mean average precision (mAP) as an additional criterion for assessing the overall retrieval performance.

Implementation Details

For caption generation, we utilize three Vision-Language Models (VLMs) with three prompts of varying granularity, resulting in the generation of nine captions per image. These VLMs include Instruct-BLIP, LLava1.5, and Qwen-VL. The multi-grained prompts employed in our framework are illustrated in Figure 2. The parameters of the image encoder, text

Methods	Ref	CUHK-PEDES			
		R-1	R-5	R-10	mAP
Fully Supervised					
GNA-RNN (Li et al. 2017)	CVPR17	19.05	-	53.64	-
Dual Path (Zheng et al. 2020)	TOMM20	44.40	66.26	75.07	-
CMPM/C (Zhang and Lu 2018)	ECCV18	49.37	-	79.27	-
ViTAA (Wang et al. 2020)	ECCV20	55.97	75.84	83.52	51.60
LBUL (Wang et al. 2022)	MM22	64.04	82.66	87.22	-
LGUR (Shao et al. 2022)	MM22	65.25	83.12	89.00	-
BLIP (Li et al. 2022a)	ICML22	65.61	82.84	88.65	58.02
CFine (Yan et al. 2022)	arXiv22	69.57	85.93	91.15	-
LCR^2S (Yan et al. 2023)	MM23	67.36	84.19	89.62	59.24
IRRA (Jiang and Ye 2023)	CVPR23	73.38	89.93	93.71	66.13
RDE (Qin et al. 2024a)	CVPR24	75.94	90.14	94.12	67.56
RaSa (Bai et al. 2023b)	IJCAI23	76.51	90.29	94.25	69.38
AUL (Li et al. 2024)	AAAI24	77.23	90.43	94.41	-
Unsupervised					
IRRA* (Jiang and Ye 2023)	CVPR23	32.94	54.37	64.67	30.87
BLIP* (Li et al. 2022a)	ICML22	51.41	71.41	78.76	44.73
GTR (Bai et al. 2023c)	MM23	47.53	68.23	75.91	42.91
MUMA (ours)	-	**59.52**	**77.79**	**84.65**	**52.75**

Table 1: Performance comparison on CUHK-PEDES dataset. * denotes the model is trained using the caption generated by LLaVA-1.5(Liu et al. 2024).

encoder, and cross-modal encoder are initialized using the weights of the BLIP model (Li et al. 2022a). We set the parameter K to 3, indicating that three captions are sampled per epoch based on their reliability. And M is set to 5. Additionally, the threshold th and α are configured to 0.5 and 0.4. During training, the input image size is fixed at 256×256 pixels, with a batch size of 16 and a total of 20 epochs. We adopt the AdamW optimizer with an initial learning rate of 10^{-5} and cosine learning rate decay. The temperature parameter τ is initialized to 0.07. Image data augmentation techniques employed during training include random horizontal flipping, random cropping with padding, and random erasing. The length of textual tokens is set to 77. Experimental procedures are conducted on two RTX4090 GPUs.

Comparison with the State-of-the-Art

To evaluate the effectiveness of our proposed framework, we conducted experiments using three widely utilized text-based person search datasets. Our baseline model is constructed based on BLIP (Li et al. 2022a), employing captions exclusively generated by LLaVA-1.5 with a single granularity to train the retrieval model. Additionally, we present the performance of IRRA under the unsupervised setting, leveraging captions generated by LLaVa-1.5.

CUHK-PEDES. We conducted a comparative analysis of our framework against state-of-the-art methods using the CUHK-PEDES dataset, as depicted in Table 1. Our approach yields notable results with 59.52% Rank-1 accuracy and 52.75% mAP, surpassing the performance of the GTR method by 11.99% and 9.84%, respectively. Moreover, our method exhibits substantial improvements, achieving an increase of 8.91% in Rank-1 accuracy and 8.46% in mAP compared to the baseline approach.

ICFG-PEDES. Our framework demonstrates competitive performance on the ICFG-PEDES dataset, as illustrated in Table 2. Our method achieves a new SOTA Rank-1 accuracy of 38.11% and mAP of 19.02%, surpassing the BLIP method by 6.53% in Rank-1 accuracy and 5.82% in mAP.

Methods	Ref	ICFG-PEDES			
		R-1	R-5	R-10	mAP
Fully Supervised					
Dual Path (Zheng et al. 2020)	TOMM20	38.99	59.44	68.41	-
CMPM/C (Zhang and Lu 2018)	ECCV18	43.51	65.44	74.26	-
ViTAA (Wang et al. 2020)	ECCV20	50.98	68.79	75.78	-
IVT (Shu et al. 2022)	ECCV22	56.04	73.60	80.22	-
CFine (Yan et al. 2022)	arXiv22	60.83	76.55	82.42	-
LCR^2S (Yan et al. 2023)	MM23	57.93	76.08	82.40	38.21
IRRA (Jiang and Ye 2023)	CVPR23	63.46	80.25	85.82	38.06
RaSa (Bai et al. 2023b)	IJCAI23	65.28	80.40	85.12	41.29
RDE (Qin et al. 2024a)	CVPR24	67.68	82.47	87.36	40.06
AUL (Li et al. 2024)	AAAI24	69.16	83.32	88.37	-
Unsupervised					
IRRA* (Jiang and Ye 2023)	CVPR23	21.23	37.37	46.04	11.47
BLIP* (Li et al. 2022a)	ICML22	31.58	52.03	61.73	13.20
GTR (Bai et al. 2023c)	MM23	28.25	45.21	53.51	13.82
MUMA (ours)	-	**38.11**	**56.01**	**63.96**	**19.02**

Table 2: Performance comparison on ICFG-PEDES dataset.

Methods	Ref	RSTPReid			
		R-1	R-5	R-10	mAP
Fully Supervised					
LBUL (Wang et al. 2022)	MM22	45.55	68.20	77.85	-
IVT (Shu et al. 2022)	ECCV22	46.70	70.00	78.80	-
CFine (Yan et al. 2022)	arXiv22	50.55	72.50	81.60	-
C2A2 (Niu et al. 2022)	MM22	51.55	76.75	85.15	-
Beat (Ma et al. 2023)	MM23	48.10	73.10	81.30	-
LCR^2S (Yan et al. 2023)	MM23	54.95	76.65	84.70	40.92
IRRA (Jiang and Ye 2023)	CVPR23	60.20	81.30	88.20	47.17
RDE (Qin et al. 2024a)	CVPR24	65.35	83.95	89.90	50.88
RaSa (Bai et al. 2023b)	IJCAI23	66.90	86.50	91.35	52.31
AUL (Li et al. 2024)	AAAI24	71.65	87.55	92.05	-
Unsupervised					
IRRA* (Jiang and Ye 2023)	CVPR23	37.60	60.65	72.30	27.42
BLIP* (Li et al. 2022a)	ICML22	44.45	67.70	77.25	33.73
GTR (Bai et al. 2023c)	MM23	45.60	70.35	79.95	33.30
MUMA (ours)	-	**54.35**	**76.05**	**83.65**	**40.50**

Table 3: Performance comparison on RSTPReid dataset.

RSTPReid. As shown in Table 3, our method achieves 50.35% Rank-1 accuracy and 40.50% mAP, outperforming the second-best method by significant margins of 8.75% and 7.2%, respectively. Notably, our approach even surpasses several supervised methods, including CFine (Yan et al. 2022) and C2A2 (Niu et al. 2022), while eliminating the need for manual annotations.

Ablation Study

To analysis the effectiveness of our proposed method, A series of ablation experiments are performed on the CUHK-PEDES dataset, as shown in Table 4.

Ablations on Different Components. We establish the baseline using BLIP trained with single captions from

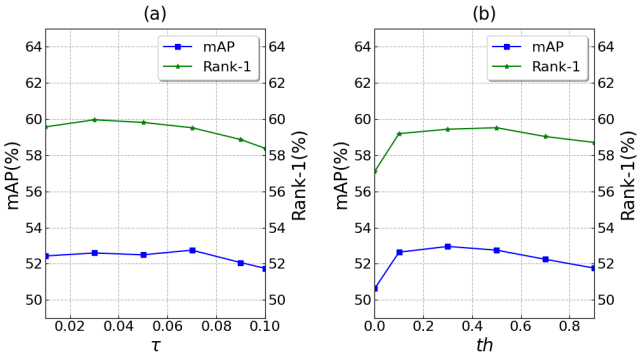

Figure 3: Performance comparison with different τ and th .

Methods	Components					CUHK-PEDS			
	MM	MG	USF	MUM	MCA	R-1	R-5	R-10	mAP
Baseline	-	-	-	-	-	51.41	71.41	78.76	44.73
Model 1	✓	-	-	-	-	52.65	72.63	80.33	47.02
Model 1	-	✓	-	-	-	53.67	74.01	81.53	48.02
Model 2	✓	✓	-	-	-	55.51	74.90	82.05	49.45
Model 3	✓	✓	✓	-	-	57.08	77.26	84.16	51.34
Model 4	✓	✓	✓	✓	-	58.62	76.64	83.79	51.68
Ours	✓	✓	✓	✓	✓	**59.52**	**77.79**	**84.65**	**52.75**

Table 4: Ablation studies on CUHK-PEDES. MM: Multi-model caption generation. MG: Multi-grained caption generation. USF: Uncertainty-guided Sample Filtration. MUM: Multi-grained Uncertainty Modeling. MCA: Multi-grained Consistency Alignment.

LLaVa1.5. As shown in Table 4, using multiple LVLMs improves Rank-1 accuracy and mAP by 1.24% and 2.29%, respectively. The Multi-Grained prompts further boost performance with 2.26% Rank-1 and 3.29% mAP improvements over the baseline. These results demonstrate that our multi-grained caption generation effectively captures diverse image descriptions, leading to more robust retrieval representations.

The proposed Multi-Grained Uncertainty Modeling and Alignment framework demonstrates strong performance with generated multi-grained captions. The Uncertainty-guided Sample Filtration effectively selects reliable image-text pairs, improving Rank-1 and mAP by 1.57% and 1.89%, respectively. The Multi-Grained Uncertainty Modeling module further enhances Rank-1 accuracy by 1.54%, while the Multi-Grained Consistency Alignment module improves mAP by 1.08%. These results confirm that knowledge from multi-grained representations can be effectively transferred and mutually reinforced, validating the effectiveness of our multi-grained caption generation and uncertainty modeling approach.

Parameter Analysis. In this section, we analyze the impact of various parameters through the implementation of three hyper-parameter ablation studies conducted on the CUHK-PEDS dataset. As illustrated in Figure 2 (a), the parameter τ is introduced in Equation (1) to control the discriminative capability between positive and negative pairs. A large value of τ has the potential to diminish the discriminative capacity between positive and negative pairs. Moreover, as depicted in Figure2(b), it is evident that the utilization of a threshold th for selecting clean pairs during training yields a notable enhancement in our model's performance. Additionally, experiments were conducted to explore the influence of the number of sampled embeddings from the multi-grained uncertainty modeling module. As depicted in Table 5, performance demonstrates enhancement with increasing m, reaching its peak at $m = 5$, beyond which performance gains become marginal.

Methods	CUHK-PEDS			
	R-1	R-5	R-10	mAP
Base	58.38	78.00	85.02	52.46
m=1	58.90	77.58	84.60	52.27
m=5	**59.52**	**77.79**	**84.65**	52.75
m=10	59.16	77.32	84.54	52.19
m=20	59.17	77.42	84.34	**52.84**

Table 5: Ablation for the number of sampled embeddings.

Methods	CUHK-PEDS			
	R-1	R-5	R-10	mAP
Attribute-based (Shao et al. 2023)	41.90	64.05	74.349	30.52
Minigpt-4(Zhu et al. 2023)	48.81	68.96	76.33	43.61
LLaVA-1.5(Liu et al. 2024)	51.41	71.41	78.76	44.73
Instruct-BLIP(Dai et al. 2024)	50.29	70.69	78.3	44.25
Qwen-VL (Bai et al. 2023a)	50.44	70.00	78.13	43.69
MUMA (ours)	**59.52**	**77.79**	**84.65**	**52.75**

Table 6: The effectiveness of different Large Vision-language Models.

Effectiveness of Different LVLMs. In this section, we provide a comparative analysis of various LVLMs. The Attribute-based method, utilizing the VLM (Bai et al. 2023c) to extract person attributes for caption generation, yields inferior performance compared to LVLM-based approaches. Table 6 reveals that distinct LVLMs achieve comparable performance, with LLaVA demonstrating the highest efficacy. Furthermore, our proposed MUMA method surpasses the single-grained approach by 8.11% and 8.02% respectively, thereby substantiating the effectiveness of our multi-grained methodology.

Conclusion

This paper proposes a multi-grained uncertainty modeling and alignment framework for unsupervised text-based pedestrian retrieval using LVLMs. Multiple LVLMs are employed to generate diverse pedestrian descriptions with varying styles and granularities. To address noise and multi-grained challenges, we propose an uncertainty-guided sample filtration module for reliable pair selection and represent captions as multivariate Gaussian distributions. A multi-level consistency distillation loss is further introduced to align multi-grained captions. Extensive experiments on three benchmark datasets demonstrate the effectiveness of the proposed approach.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62372203 and 62302186, in part by the Major Scientific and Technological Project of Shenzhen (202316021), in part by the National key research and development program of China(2022YFB2601802), in part by the Major Scientific and Technological Project of Hubei Province (2022BAA046, 2022BAA042).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023b. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv preprint arXiv:2305.13653*.

Bai, Y.; Wang, J.; Cao, M.; Chen, C.; Cao, Z.; Nie, L.; and Zhang, M. 2023c. Text-based Person Search without Parallel Image-Text Data. *arXiv preprint arXiv:2305.12964*.

Chen, Y.; Zheng, Z.; Ji, W.; Qu, L.; and Chua, T.-S. 2022. Composed image retrieval with text feedback via multi-grained uncertainty regularization. *arXiv preprint arXiv:2211.07394*.

Cui, K.; Liu, S.; Feng, W.; Deng, X.; Gao, L.; Cheng, M.; Lu, H.; and Yang, L. T. 2024. Correlation-aware Cross-modal Attention Network for Fashion Compatibility Modeling in UGC Systems. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ji, Y.; Wang, J.; Gong, Y.; Zhang, L.; Zhu, Y.; Wang, H.; Zhang, J.; Sakai, T.; and Yang, Y. 2023. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23262–23271.

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

Jing, P.; Cui, K.; Guan, W.; Nie, L.; and Su, Y. 2023. Category-aware multimodal attention network for fashion compatibility modeling. *IEEE Transactions on Multimedia*, 25: 9120–9131.

Jing, Y.; Wang, W.; Wang, L.; and Tan, T. 2020. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10678–10686.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, S.; He, C.; Xu, X.; Shen, F.; Yang, Y.; and Shen, H. T. 2024. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3172–3180.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and Duan, L.-Y. 2022b. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ma, Y.; Sun, X.; Ji, J.; Jiang, G.; Zhuang, W.; and Ji, R. 2023. Beat: Bi-directional One-to-Many Embedding Alignment for Text-based Person Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4157–4168.

Niu, K.; Huang, L.; Huang, Y.; Wang, P.; Wang, L.; and Zhang, Y. 2022. Cross-modal co-occurrence attributes alignments for person search by language. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4426–4434.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2023. Noisy-Correspondence Learning for Text-to-Image Person Re-identification. *arXiv preprint arXiv:2308.09911*.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024a. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2024b. Cross-modal Active Complementary Learning with Self-refining Correspondence. *Advances in Neural Information Processing Systems*, 36.

Shao, Z.; Zhang, X.; Ding, C.; Wang, J.; and Wang, J. 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*, 11174–11184.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5566–5574.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; Shu, X.; Du, X.; and Tang, J. 2023a. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8922–8931.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2024b. Boosting consistency in story visualization with rich-contextual conditional diffusion models. *arXiv preprint arXiv:2407.02482*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023b. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*.

Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Ren, B.; and Wang, X. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, 624–641. Springer.

Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; and Tao, D. 2024. Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17127–17137.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 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*, 402–420. Springer.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1984–1992.

Yan, S.; Dong, N.; Liu, J.; Zhang, L.; and Tang, J. 2023. Learning Comprehensive Representations with Richer Self for Text-to-Image Person Re-Identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6202–6211.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023a. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19883–19892.

Yang, S.; Zhou, Y.; Wang, Y.; Wu, Y.; Zhu, L.; and Zheng, Z. 2023b. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. *arXiv preprint arXiv:2306.02898*.

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

Zhao, S.; Gao, C.; Shao, Y.; Zheng, W.-S.; and Sang, N. 2021. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11395–11404.

Zhao, Z.; Liu, B.; Lu, Y.; Chu, Q.; and Yu, N. 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, 7534–7542.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y.-D. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2): 1–23.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zuo, J.; Zhou, H.; Nie, Y.; Zhang, F.; Guo, T.; Sang, N.; Wang, Y.; and Gao, C. 2024. UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22010–22019.