Cross-modal Generation and Alignment via Attribute-guided Prompt for Unsupervised Text-Based Person Retrieval

Zongyi Li , Jianbo Li , Yuxuan Shi * , Hefei Ling * , Jiazhong Chen , Runsheng Wang and Shijuan Huang
Huazhong University of Science and Technology, Wuhan, China
zongyili@hust.edu.cn, jianboli@hust.edu.cn, shiyx@hust.edu.cn, lhefei@hust.edu.cn, jzchen@hust.edu.cn, wrsh@hust.edu.cn, huan_shijuan@hust.edu.cn

Abstract

Text-based Person Search aims to retrieve a specified person using a given text query. Current methods predominantly rely on paired labeled image-text data to train the cross-modality retrieval model, necessitating laborious and time-consuming labeling. In response to this challenge, we present the Cross-modal Generation and Alignment via Attribute-guided Prompt framework (GAAP) for fully unsupervised text-based person search, utilizing only unlabeled images. Firstly, an Attribute-guided Prompt Caption Generation (APCG) module is proposed to generate pseudo captions by feeding the attribute prompts into a large-scale pre-trained vision-language model. These synthetic captions are meticulously selected through a sample selection for subsequent fine-tuning. To mitigate the negative effect of noise labels and mine local matching characteristics, an Attribute-guided Cross-modal Alignment (AGCA) module is introduced to align features across modalities, containing three sub-modules. The Cross-Modal Center Alignment aligns the samples with different modality centroids. Subsequently, an Attribute-guided Image-Text Contrastive Learning module is proposed to facilitate the alignment of relationships among different pairs by considering local attribute similarities. Lastly, the Attribute-guided Image-Text Matching module is introduced to mitigate noise in pseudo captions by using the image-attribute matching score to soften the hard matching labels. Empirical results showcase the effectiveness of our method across various text-based person search datasets under the fully unsupervised setting.

1 Introduction

Text-based Person Retrieval involves searching for a person of interest from a large-scale image gallery according to a provided textual query [Li *et al.*, 2017; Farooq *et al.*, 2022; Li *et al.*, 2023]. As a sub-task of person re-identification (Re-ID)[He *et al.*, 2022; Wang *et al.*, 2022a], the text-based person retrieval solely adopts the language description to locate the target person, which offers greater accessibility and closely mirrors real-world scenarios. However, this task presents challenges arising from the imperative need for feature alignment in two distinct modalities. Therefore, a large number of approaches [Zhu *et al.*, 2021; Shao *et al.*, 2022; Farooq *et al.*, 2022; Jiang and Ye, 2023; Jing *et al.*, 2024] have been developed to attain modality-invariant features by aligning fine-grained visual and textual attributes, leading to impressive results. However, these methods demand the presence of both textual descriptions

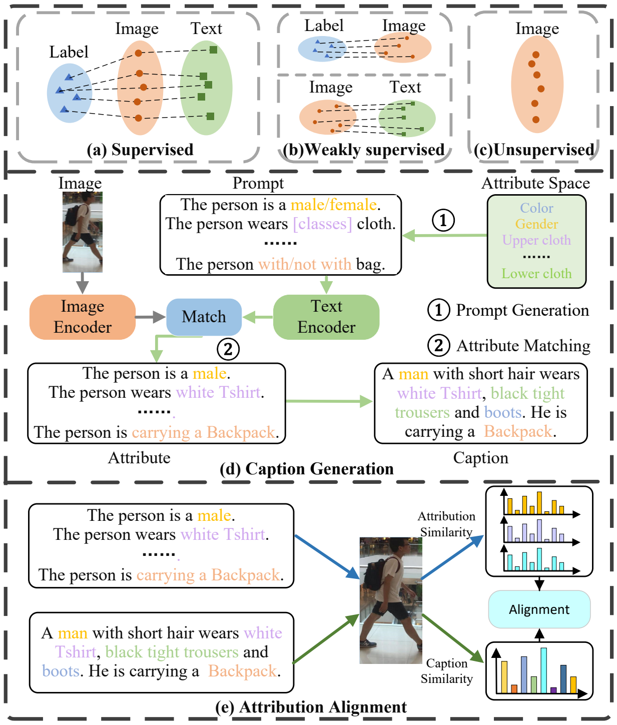

Label
Image
Text
(a) Supervised
Label
Image
Image
Text
(b)Weakly supervised
Image
(c)Unsupervised
Image
Prompt
The person is a male/female.
The person wears [classes] cloth.
......
The person with/not with bag.
Attribute Space
Color
Gender
Upper cloth
......
Lower cloth
1
Image Encoder
Match
Text Encoder
1 Prompt Generation
2 Attribute Matching
2
The person is a male.
The person wears white Tshirt.
......
The person is carrying a Backpack.
A man with short hair wears white Tshirt, black tight trousers and boots. He is carrying a Backpack.
Attribute
(d) Caption Generation
Caption
The person is a male.
The person wears white Tshirt.
......
The person is carrying a Backpack.
Attribution Similarity
Alignment
A man with short hair wears white Tshirt, black tight trousers and boots. He is carrying a Backpack.
Caption Similarity
(e) Attribution Alignment

Figure 1: Motivation for our GAAP model. (a). Supervised setting with labels and parallel image-text pairs. (b). Weakly supervised setting without texts or labels. (c). Unsupervised setting with only unlabeled images. (d) and (e) represent our framework for pseudo caption generation and cross-modality attribute alignment.

*Corresponding author

and identity labels, thus leading to significant investments in terms of time and resources.

In response to the challenges of costly labeling, researchers have directed their attention toward mitigating the expenses associated with annotating images from multiple non-overlapping cameras with captions and identity labels. Specifically, MAN [Jing *et al.*, 2020] proposes a moment alignment network to address the cross-domain text-based person search task, where pairwise text-image identity labels are absent in the target domain. Similarly, CMMT [Zhao *et al.*, 2021] introduces a Cross-Modal Mutual Training framework for weakly supervised person search, where solely image-text pairs are available without any identity annotations. GTR [Bai *et al.*, 2023b] proposes a generation-then-retrieval framework without parallel Image-Text data, liberating it from the dependence on manually labeled text descriptions. However, the assignment of one hot positive and negative label fails to accurately capture the non-strict mutual exclusion relations between image and text pairs, particularly when the generated caption is not sufficiently precise.

Therefore, to eliminate the cost associated with manual annotation, we aim at the Unsupervised Text-based Person Retrieval, as depicted in Figure 1(c), which involves solely image data without any additional information. To solve this challenging task, we leverage the potential of the vision-language pre-training model and propose a Cross-modal Generation and Alignment via Attribute-guided Prompt (GAAP) framework, encompassing two key processes: Attribute-guided Prompt Caption Generation and Attribute-guided Cross-modal Alignment. GAAP employs attribute-based prompts to generate pseudo text and subsequently aligns the coarse textual information with finely detailed attribute prompts during the finetuning process. The overall architecture of GAAP is illustrated in Figure 2.

To alleviate the issue of missing image-text pairs, Attribute-guided Prompt Caption Generation is proposed to generate pseudo captions. Specifically, prompts for various person attributes are established for fine-grained captions. These prompts are then combined with the images and input into the text and image encoder within the pre-trained vision-language model BLIP[Li *et al.*, 2022]. By calculating the similarity between the image and attribute prompt embeddings, the corresponding attributes can be determined by choosing the optimal matching. All identified attributes are then aggregated to form appropriate captions. Furthermore, to enhance sentence coherence and style diversity, we leverage Large Language Model (LLM) to reconstruct the generated captions, resulting in more comprehensive and varied sentences. Additionally, the image and its newly constructed caption are sent to the ITM head, facilitating the computation of a matching score. This matching score serves as the basis for sample selection, wherein only reliable pairs are utilized for the subsequent fine-tuning process.

Leveraging the generated image-text pairs, we present an Attribute-guided Cross-modal Alignment module to align cross-modal features, which is mainly comprised of three sub-modules: Cross-Modal Center Alignment (CMCA), Attribute-guided Image-Text Contrastive learning (AITC), and Attribute-guided Image-Text Matching (AITM). In the CMCA module, we assign the image-text pairs with the same attribute set with identical identity labels. And the image-based and text-based class memories are established by computing average features from identical identities. The pairs with notably significant disparities in class similarities are subsequently omitted. Considering the potential overlap in local attributes across different image-text pairs, we leverage the local attribute information to guide the cross-modalities interaction. Specifically, in the AITC module, the attribute-image similarity is employed to guide the text and image relations, weakening the strict cross-modal constraint. Similarly, the AITM loss is proposed to align image caption matching score with attribute-based similarity by introducing soft targets rather than original strict positive or negative labels. This approach can mitigate the adverse impact of the noise pairs and the unpaired samples exhibiting certain overlaps.

The main contribution of our proposed framework can be summarized as follows:

1. We propose a Cross-modal Generation and Alignment via Attribute-guided Prompt framework for unsupervised text-based person retrieval. By leveraging the attribute-based prompt, GAAP effectively generates text captions and aligns the cross-modal relation.

2. We propose an Attribute-guided Prompt Caption Generation module for caption generation. Incorporating with LLM and BLIP, we generate image captions and select reliable image-text pair for unsupervised finetuning.

3. We propose an Attribute-guided Cross-modal Alignment module to align the cross-modal relation by employing fine-grained attribute-image similarity as softened targets. Extensive experiments demonstrate the effectiveness of our proposed GAAP framework, which brings significant improvements on several text-based person retrieval datasets.

2 Related Work

2.1 Pre-Training Methods

Pre-trained models have demonstrated impressive performance across a wide array of tasks and can be broadly classified into three main categories: Vision pre-training Models, Vision-Language pre-training Models, and Language pre-training Models. Vision pre-training Model usually performs self-supervised [He *et al.*, 2020] on large image datasets, while Vision-Language pre-training Model effectively learns modality-invariance features from large image-text datasets, with CLIP [Radford *et al.*, 2021] as a prominent example. Subsequent to CLIP, SoftCLIP [Gao *et al.*, 2023] and PyramidCLIP [Gao *et al.*, 2022] employ a softened target to achieve a soft cross-modal alignment. Moreover, BLIP [Li *et al.*, 2022] leverages CapFilt to generate image descriptions and eliminate noisy text effectively. In the text-based person search field, PLIP [Zuo *et al.*, 2023] generates attributes and stylish textual captions based on these attributes. Similarly, APTM [Yang *et al.*, 2023] generates pedestrian images and corresponding captions based on attributes, and regularizes the model training by leveraging both the attribute recognition task and the text-based person retrieval task.

Early Language pre-training Models like BERT [Devlin *et al.*, 2018] and GPT-2 [Radford *et al.*, 2019] play a foundational role in the development of the Large Language Models. Subsequently, GPT-3 [Brown *et al.*, 2020] emerges as a revolutionary model capable of achieving remarkable performance across diverse tasks, eliminating the need for gradient updates or fine-tuning. GLM [Du *et al.*, 2022] operates on the principle of autoregressive blank infilling. On the other hand, ChatGLM2-6B showcases the dual advantages of being lightweight and high-performing.

In our paper, we explore the potential of employing the pre-training models for fully unsupervised text-based person search. Subsequently, ChatGLM2-6B was employed to generate diverse styles of texts, while BLIP was adopted for caption generation and retrieval.

2.2 Text-Based Person Retrieval

Li *et al.* [Li *et al.*, 2017] first propose the text-based person retrieval task and introduce a cross-modal dataset with image-text pairs, CUHK-PEDES. Recent methods can be categorized into three groups:global-based [Zhang and Lu, 2018; Zhu *et al.*, 2021; Jiang and Ye, 2023], local-based [Shen *et al.*, 2023; Farooq *et al.*, 2022; Yan *et al.*, 2022a; Jing *et al.*, 2023], and attribute-based [Aggarwal *et al.*, 2020; Wang *et al.*, 2020; Zuo *et al.*, 2023; Yang *et al.*, 2023] approaches. In the global-based methods, IRRA [Jiang and Ye, 2023] presents a groundbreaking use of the complete CLIP [Radford *et al.*, 2021] for text-based person retrieval. In the local-based methods, CFine [Yan *et al.*, 2022a] skillfully leverages CLIP's image encoder to capture rich multimodal information. For the attribute-based methods, CMAAM [Aggarwal *et al.*, 2020] extracts attributes from the text in the training set as annotations, while ViTAA [Wang *et al.*, 2020] employs semantic segmentation to categorize pedestrians into different body parts.

Although these supervised methods can achieve satisfactory performance, they require a substantial amount of text-image labels. Therefore, researchers are devoted to exploring weakly supervised methods. MAN [Jing *et al.*, 2020] pioneers the field with a cross-domain moment alignment network for cross-domain text-based person retrieval. In contrast, CMMT [Zhao *et al.*, 2021] addresses the absence of ID labels by generating pseudo labels through clustering. However, it is important to note that all these approaches necessitate the use of text data, resulting in a substantial requirement for manual annotation efforts. In light of this, GTR [Bai *et al.*, 2023b] presents an alternative approach by generating textual descriptions corresponding to person images through VQA. Nonetheless, it falls short in considering the local similarity between different image-text pairs, which is different from our work.

3 Methodology

3.1 Overview

In the unsupervised text-based person search, a training dataset denoted as $I = \{x_i\}_{i=1}^{N}$ with only image data is given, where x_i represents the i-th image and N is the total number of images. The principal aim of this task is to learn discriminative multi-modal features with image-only data.

Our proposal Unsupervised framework, namely Cross-modal Generation and Alignment via Attribute-guided Prompt (GAAP), is composed of two principal components: Attribute-guided Prompt Caption Generation and Attribute-guided Cross-modal Alignment, as depicted in Figure 2. The former component leverages the attribute-based prompts in conjunction with a pre-trained vision-language model to generate pseudo captions. On the other hand, the latter component aims to establish alignment between image-caption relationships at a fine-grained attribute level, which is achieved through the incorporation of three distinct sub-modules: Cross-modality Center Alignment (CMCA), Attribute-guided Image-Text Contrastive Learning (AITC), and Attribute-guided Image-Text Matching (AITM).

3.2 Attribute-Guided Prompt Caption Generation

Given the image data within the realm of unsupervised text-based person search, the initial step involves generating pseudo captions for each individual's images. To facilitate this, we formulate distinct attribute text prompts denoted as $A = \{a_1, a_2, ..., a_n\}$, where n signifies the count of binary attributes. We feed the image into the visual encoder, yielding the image embedding denoted as $E_{\text{img}}(x_i)$. Concurrently, the attribute prompts P are directed to the text encoder, producing prompt embeddings $\{E_{text}(a_1), ..., E_{text}(a_n)\}$. Subsequently, we compute the cosine similarity between the image and each attribute, and the softmax is used to normalize the similarities of attributes belonging to the same category:

$$p\left(x_i, a_j\right) = \frac{\exp\left(E_{img}(x_i), E_{text}(a_j)\right)}{\sum_{a_k \in S(a_j)} \exp\left(E_{img}(x_i), E_{text}(a_k)\right)}, \qquad (1)$$

where $S(a_j)$ denotes the attribute set containing attributes of the same category with a_j, e.g., *color set: yellow, red,...blue*. Furthermore, the image and attribute prompts are sent into the cross-modal encoder to obtain the matching score, denoted as $m(a_i, x)$. Then the attribute-image similarity can be established by averaging the matching score and cosine similarity, expressed as $\hat{s}(x, a_i) = m(x, a_i) + p\left(x, a_i\right)$.

The prediction for each attribute set is realized by a systematic selection of the attribute-image score that boasts the highest value within each individual attribute set. This approach results in the acquisition of an attribute label set associated with the image x_i, denoted as $A_i = \{a_1, ..., a_{A_i}\}$. Subsequently, these attributes are fused into a template with masked sentences to construct the caption t_i for image x_i.

Subsequent to obtaining corresponding pseudo captions according to the predicted attribute, we leverage LLM to generate distinct stylistic captions by giving the prompts:" *I will give you a sentence, please return a sentence with different styles with the same semantics as if it is spoken by different people: [CAPTION]* ". Then the template-based captions and the rephrased captions combined with the images are fed into the cross-modal encoder, resulting in matching scores $m(t, x)$ between the captions and images. The text-image matching scores representing reliability are used to filter out noisy image-text pairs that could potentially impair performance. Consequently, we set the sample selection threshold $th \in [0, 1]$, whereby solely the image-text pairs with match-

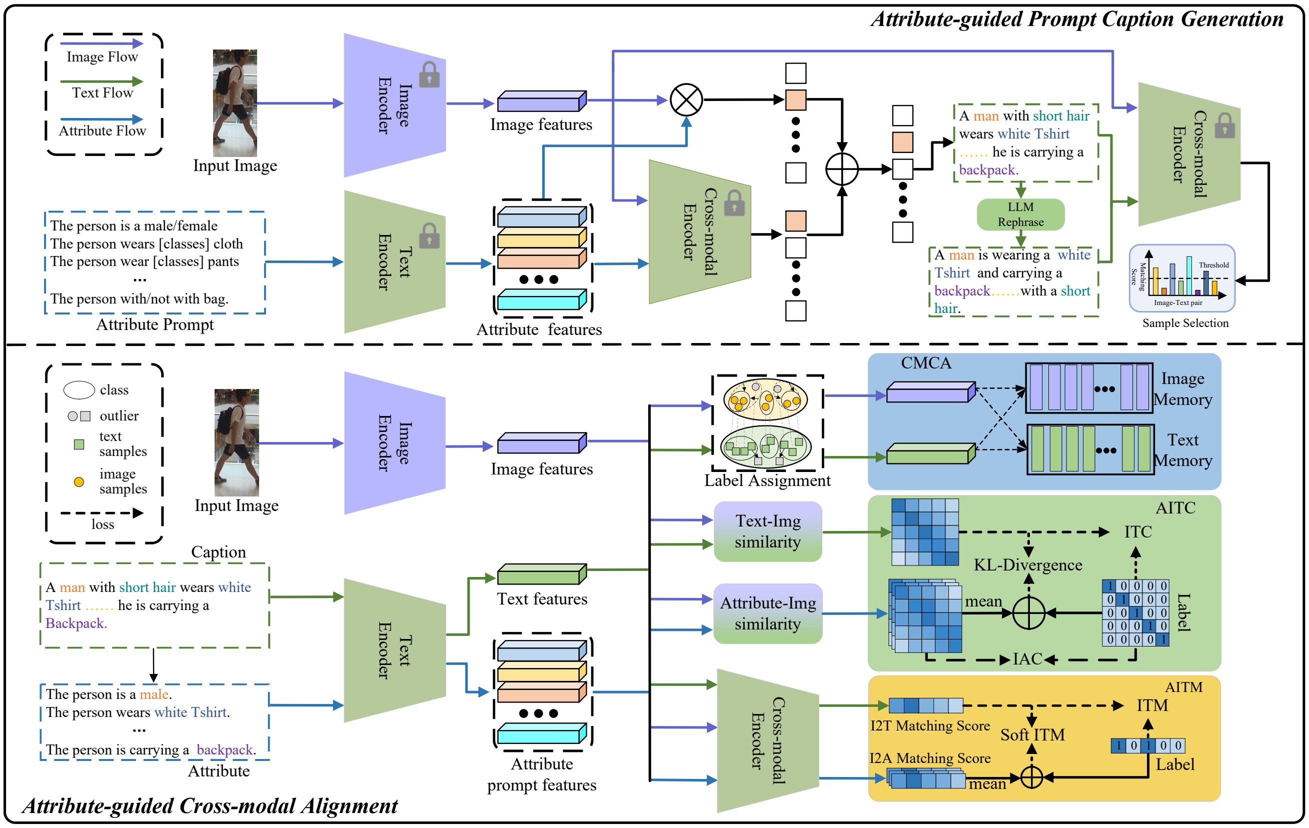


Figure 2: An overview of our proposal Unsupervised Text-based Person Search pipeline GAAP. Based on the similarity between images and attributes, the Attribute-guided Prompt Caption Generation module generates texts. The Attribute-guided Cross-modal Alignment module narrows the gap between images and Corresponding texts through CMCA, AITC, and AITM.

ing scores of $m(t, x) > th$ are preserved, while other samples are considered outliers.

3.3 Attribute-Guided Image-Text Alignment

Cross-Modal Center Alignment. After the acquisition of the pseudo image-text pairs, the absence of identity labels for these image-text pairs also presents a notable challenge. To address this issue, we adopt a strategy of assigning identical identity labels to image-text pairs that share a common attribute set. However, employing the ID labels solely based on the generated attribute set can potentially result in visual dissimilarity among images due to potential noise introduced during the generation process. To mitigate this issue, we employ the similarity between samples and class centers as a metric for assessing cluster reliability, enabling the filtration of samples with lower ID similarity. Specifically, we define the cluster reliability as the cosine similarity between the image feature and the corresponding center feature, expressed as $R(x_i) = cos(E_{img}(x_i), C_I[y_i])$, where y_i is the assigned identities of the image, and $C_I[y_i]$ represents the image class center obtained by averaging the image features belonging to the same identities. Consequently, images with lower ID similarity with $R(x_i) < th_2$ are regarded as cluster outliers. This selection operates on the class level by measuring the sample-class similarity, while sample selection is performed on the pair level.

After assigning the pseudo-identity label Y, we construct image and caption class-level memory by averaging the features belonging to the same identities, which can be represented as $C_I = \{c_i^1, ..., c_i^{N_c}\}$ and $C_T = \{c_t^1, ..., c_t^{N_c}\}$. Therefore the image class-level contrastive loss can be denoted as a nonparametric Softmax loss[Xiao *et al.*, 2017]:

$$\mathcal{L}_{\text{ic}} = -\sum_{i=0}^{N} \log \frac{\exp\left(\left\langle C_I\left[y_i\right], E_{img}\left(x_i\right)\right\rangle / \tau\right)}{\sum_{j=1}^{N_c} \exp\left(\left\langle C_I[j], E_{img}\left(x_i\right)\right\rangle / \tau\right)}, \quad (2)$$

where N represents the total number of images, y_i is the identity label of image x_i and τ is temperature hyper-parameter. During each iteration, the memory bank is updated by:

$$C_I\left[y_i\right] \leftarrow mC_I\left[y_i\right] + (1-m)E_{img}\left(x_i\right), \quad (3)$$

where m is the momentum parameter for updating the features in class memory.

Moreover, to learn the class semantic cross-modal relations, the text class-level memory is used to guide the image features by a text center alignment loss:

$$\mathcal{L}_{\text{tca}} = -\sum_{i=0}^{N} \log \frac{\exp\left(\left\langle C_T\left[y_i\right], E_{img}\left(x_i\right)\right\rangle / \tau\right)}{\sum_{j=1}^{N_c} \exp\left(\left\langle C_T[j], E_{img}\left(x_i\right)\right\rangle / \tau\right)}, \quad (4)$$

where C_T is the text class memory. Similar to the image modality, the text is also supervised by both image and text memory to learn the cross-modal class relations. Therefore, the cross-modal center alignment loss can be formulated as:

$$\mathcal{L}_{\text{cmca}} = \mathcal{L}_{\text{ic}} + \mathcal{L}_{\text{ica}} + \mathcal{L}_{\text{tc}} + \mathcal{L}_{\text{tca}}. \tag{5}$$

Attribute-guided Image-Text Contrastive Learning. As a cross-modal retrieval task, our approach involves the Image-Text contrastive loss (ITC) to impose the cross-modal alignment. The ITC loss facilitates the learning of the relationship between the positive and negative pairs by pulling the matched pair and pushing the unpaired apart. Given a batch of N_B image-text pairs, we first calculate the normalized similarity between the images and text in a batch. The image-to-text similarity $\boldsymbol{p}_i^{i2t}(I) = \{\boldsymbol{p}_{i,j}^{i2t}(I)\}_{j=1}^{N}$ can be denoted as:

$$\boldsymbol{p}_{i,j}^{i2t}(I) = \frac{\exp\left(sim\left(E_{img}(x_i), E_{text}(t_j)\right)/\tau\right)}{\sum_{k=1}^{N_B}\exp\left(sim\left(E_{img}(x_i), E_{text}(t_k)\right)/\tau\right)}, \tag{6}$$

where τ is a learnable temperature parameter, $sim(\cdot)$ represents the cosine similarity, and t_i denotes the pseudo captions for i-th sample. Similarly, the text-to-image similarity $\boldsymbol{p}_i^{t2i}(T)$ can also be obtained. And the one-hot label $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{N_B}$ with the image-text paris with the same pseudo identity label as 1 and others as 0. Therefore the ITC loss can be denoted as:

$$\mathcal{L}_{itc} = \frac{1}{N_B}\sum_{i=1}^{N_B} H\left(\boldsymbol{y}_i^{i2t}, \boldsymbol{p}_i^{i2t}(I)\right) + \frac{1}{N_B}\sum_{i=1}^{N_B} H\left(\boldsymbol{y}_i^{t2i}, \boldsymbol{p}_i^{t2i}(T)\right), \tag{7}$$

where N_B is the batch size and $H(\cdot,\cdot)$ represents the cross entropy loss.

As aforementioned, such ITC loss may neglect some local attribute similarity between the image-text pairs in a mini-batch. To explore more fine-grained relations between images and text, we use the image attribute similarity to provide the implicit local relation knowledge. Formally, the text T_i consists of several corresponding attributes, $A_i = \{a_0, ...a_{N_{A_i}}\}$. Here, A_i corresponds to the predicted attribute for image x_i. Therefore, the image-attribute contrastive loss can be denoted as:

$$\mathcal{L}_{\text{iac}} = -\frac{1}{N_{A_i}}\sum_{a_j \in A_i}\log\frac{\exp\left(E_{img}(x_i), E_{text}(a_j)\right)}{\sum_{a_k \in S(a_j)}\exp\left(E_{img}(x_i), E_{text}(a_k)\right)}, \tag{8}$$

where $S(a_j)$ is the attribute set with the same category with attribute a_j. The image and the attribute prompts are fed into the image and text encoder, obtaining the attribute and image embeddings to calculate the image-attribute similarity:

$$s(x_i, A_j) = \frac{1}{N_{A_j}}\sum_{a_k \in A_j} cos\left(E_{img}(x_i), E_{text}(a_k)\right), \tag{9}$$

where x_i, A_j is the image and attribute set of i-th and j-th samples in a batch. And N_{A_j} is the number of matched attributes for j-th sample.

Therefore the normalized image-attribute similarity, denoted as $\boldsymbol{p}_i^{i2a}(I) = \{\boldsymbol{p}_{i,j}^{i2a}(I)\}_{j=1}^{N}$ can be calculated by:

$$\boldsymbol{p}_{i,j}^{i2a}(I) = \frac{\exp\left(s\left(x_i, A_j\right)/\tau\right)}{\sum_{k=1}^{N_B}\exp\left(s\left(x_i, A_k\right)/\tau\right)}. \tag{10}$$

The image-attribute similarity is used to smooth the hard target label for solving the non-strict relation between image and text, the attribute-guided image-text label can be formulated as:

$$\widetilde{\boldsymbol{y}}_i^{i2t} = (1-\alpha_1)\boldsymbol{y}_i^{i2t} + \alpha_1\boldsymbol{p}_i^{i2a}(I). \tag{11}$$

Similarly, the attribute-image similarity $\boldsymbol{p}_i^{a2i}(A) = \{\boldsymbol{p}_{i,j}^{a2i}(A)\}_{j=1}^{N}$ and attribute-guided text-image label $\widetilde{\boldsymbol{y}}_i^{t2i}$ can also be obtained through Eq 10 and 11. The attribute-guided image-text alignment loss can be calculated by incorporating local similarity with attribute-guided labels, thereby mitigating the effects of the strict regularization:

$$\mathcal{L}_{\text{aita}} = \frac{1}{N_B}\left(\sum_{i=1}^{N_B}\text{KL}\left(\widetilde{\boldsymbol{y}}_i^{i2t}\|\boldsymbol{p}_i^{i2t}(I)\right) + \sum_{i=1}^{N_B}\text{KL}\left(\widetilde{\boldsymbol{y}}_i^{t2i}\|\boldsymbol{p}_i^{t2i}(T)\right)\right). \tag{12}$$

The attribute-guided image-text contrastive loss can be computed by aggregating the image-attribute contrastive loss and the attribute-guided image-text alignment loss:

$$\mathcal{L}_{\text{aitc}} = \mathcal{L}_{\text{iac}} + \mathcal{L}_{\text{aita}} \tag{13}$$

Attribute-guided Image-Text Matching. The primary objective of the ITM loss is to predict the positivity or negativity of the given image-text pair. The image embedding and text are sent into the cross-modal encoder M, following a fully-connected layer to predict the matching score p^{itm}. Therefore the ITM can be formulated as:

$$\mathcal{L}_{itm} = \frac{1}{N}\sum_{i=1}^{N} H\left(\boldsymbol{y}_i^{\text{itm}}, \boldsymbol{p}_i^{itm}(I,T)\right), \tag{14}$$

where y^{itm} represents the ground-truth label for ITM, 1 when image text matched, 0 otherwise.

The image-text positive pairs may not always be matched since the generated caption may contain noise, while the negative pairs may also contain some local similarity information. However, the one-hot GT label assumes there are strictly matched positive pairs, neglecting the local attribute relations. Therefore, attribute-guided Image-Text Matching aims to soften the hard GT labels with the image-attribute matching score, which can be denoted as $\boldsymbol{m}_i(I,A) = \frac{1}{N_{A_j}}\sum_{a_k \in A_j} M\left(x_i, a_k\right)$, where M is the cross-modal encoder. This image-attribute matching score is used to soften the one hot image-text matching label, which can be denoted as:

$$\widetilde{\boldsymbol{y}}_i^{itm}(I,T) = (1-\alpha_2)\boldsymbol{y}_i^{itm} + \alpha_2\boldsymbol{m}_i(I,A), \tag{15}$$

where α_2 is the weighing parameter. The attribute-guided image-text matching loss can be denoted as:

$$\mathcal{L}_{aitm} = \frac{1}{N}\sum_{i=1}^{N} H\left(\widetilde{\boldsymbol{y}}_i^{itm}(I,T), \boldsymbol{p}_i^{itm}(I,T)\right). \tag{16}$$

Given the above objectiveness, the final optimization function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{itc} + \mathcal{L}_{itm} + \lambda_1\mathcal{L}_{aitm} + \lambda_2\mathcal{L}_{aitc} + \lambda_3\mathcal{L}_{cmca}. \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3$ are loss weight.

Methods	Ref	CUHK-PEDES			
		R-1	R-5	R-10	mAP
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
DSSL [Zhu *et al.*, 2021]	MM21	59.98	80.41	87.56	-
SSAN [Ding *et al.*, 2021]	arXiv21	61.37	80.15	86.73	-
LBUL [Wang *et al.*, 2022c]	MM22	64.04	82.66	87.22	-
TIPCB [Chen *et al.*, 2022]	Neuro22	64.26	83.19	89.10	-
CAIBC [Wang *et al.*, 2022b]	MM22	64.43	82.87	88.37	-
AXM-Net [Farooq *et al.*, 2022]	AAAI22	64.44	80.52	86.77	58.73
LGUR [Shao *et al.*, 2022]	MM22	65.25	83.12	89.00	-
BLIP [Li *et al.*, 2022]	ICML22	65.61	82.84	88.65	58.02
CFine [Yan *et al.*, 2022a]	arXiv22	69.57	85.93	91.15	-
IRRA [Jiang and Ye, 2023]	CVPR23	73.38	89.93	93.71	66.13
RaSa [Bai *et al.*, 2023a]	IJCAI23	76.51	90.29	94.25	69.38
Weakly-Supervised					
CMMT [Zhao *et al.*, 2021]	ICCV21	57.10	78.14	85.23	-
Unsupervised					
IRRA* [Jiang and Ye, 2023]	CVPR23	28.77	50.20	60.84	26.55
GTR * [Bai *et al.*, 2023b]	MM23	40.87	60.60	69.39	35.72
Baseline (ours)	-	39.36	60.96	69.98	34.35
GAAP (ours)	-	**47.64**	**67.79**	**76.08**	**41.28**

Table 1: Performance comparison on CUHK-PEDES dataset. * This is trained using our generated captions. * represents that we report the performance based on our re-implementation with GTR's VQA-based caption method under the unsupervised setting.

4 Experiments

4.1 Datasets and Protocol

Our approach is meticulously evaluated across three widely used datasets: CUHK-PEDES [Li *et al.*, 2017], ICFG-PEDES [Ding *et al.*, 2021], and RSTPReid [Zhu *et al.*, 2021]. To ensure adherence to an unsupervised approach, only image data is employed during the training process to generate captions. In the testing phase, we utilize the captions from the dataset for retrieval.

The CUHK-PEDES dataset, pioneering in text-based person retrieval, comprises 34,054/68,108 images/sentences of 11,003 identities in the training set. The validation/test set contains 3,078/3,074 images. The ICFG-PEDES dataset consists of 54,522 images of 4,102 individuals, with one caption per image. The training/testing set comprises 34,674/19,848 image-text pairs of 3,102/1,000 identities. The RSTPReid dataset encompasses 20,505 images of 4,101 identities, with each image accompanied by two textual descriptions. The training/validation/testing sets include 3,701/200/200 identities, respectively.

The widely adopted Rank-K metric serves as one of the fundamental evaluation measures. Specifically, it involves identifying the most relevant one/five/ten image(s) based on the similarities between text and images. Furthermore, we also incorporate the mean average precision (mAP) as an additional comprehensive assessment.

4.2 Implementation Details

The Image Encoder, Text Encoder, and Image-grounded Text Encoder in BLIP [Li *et al.*, 2022] have been incorporated into our methodology. In the Attribute-guided Prompt Caption Generation module, we commence the process by generating the initial caption through 8 distinct prompts, encompassing 47 attributes, and another stylistic caption is created using ChatGLM2-6B. During this process, the dimension of images is set to 256×256, while the th stands at 0.9. In the Attribute-guided Cross-modal Alignment phase, the batch size is set to 32. The optimization process employs the AdamW [Loshchilov and Hutter, 2017] optimizer, with a decay rate of 0.05 and an initial learning rate of 1e-5. α_1 and α_2 have been assigned values of 0.5 and 0.2, respectively, while the parameter m in Equation 3 has been set to 0.2. The threshold th_2 for label assignment has been established as 0.84. Image augmentation contains random horizontal flipping and RandAugment [Cubuk *et al.*, 2020] techniques.

Methods	Ref	ICFG-PEDES			
		R-1	R-5	R-10	mAP
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	-
SSAN [Ding *et al.*, 2021]	arXiv21	54.23	72.63	79.53	-
IVT [Shu *et al.*, 2022]	ECCV22	56.04	73.60	80.22	-
ISANet [Yan *et al.*, 2022b]	arXiv22	57.73	75.42	81.72	-
CFine [Yan *et al.*, 2022a]	arXiv22	60.83	76.55	82.42	-
IRRA [Jiang and Ye, 2023]	CVPR23	63.46	80.25	85.82	38.06
RaSa [Bai *et al.*, 2023a]	IJCAI23	65.28	80.40	85.12	41.29
Unsupervised					
IRRA* [Jiang and Ye, 2023]	CVPR23	14.52	28.91	37.54	7.00
GTR * [Bai *et al.*, 2023b]	MM23	21.75	37.1	45.18	10.26
Baseline (ours)	-	22.51	39.12	47.29	10.85
GAAP (ours)	-	**27.12**	**44.91**	**53.56**	**11.43**

Table 2: Performance comparison on ICFG-PEDES dataset.

Methods	Ref	RSTPReid			
		R-1	R-5	R-10	mAP
DSSL [Zhu *et al.*, 2021]	MM21	39.05	62.60	73.95	-
SSAN [Ding *et al.*, 2021]	arXiv21	43.50	67.80	77.15	-
LBUL [Wang *et al.*, 2022c]	MM22	45.55	68.20	77.85	-
IVT [Shu *et al.*, 2022]	ECCV22	46.70	70.00	78.80	-
CFine [Yan *et al.*, 2022a]	arXiv22	50.55	72.50	81.60	-
C2A2 [Niu *et al.*, 2022]	MM22	51.55	76.75	85.15	-
IRRA [Jiang and Ye, 2023]	CVPR23	60.20	81.30	88.20	47.17
RaSa [Bai *et al.*, 2023a]	IJCAI23	66.90	86.50	91.35	52.31
Unsupervised					
IRRA* [Jiang and Ye, 2023]	CVPR23	27.70	50.40	62.65	22.21
GTR * [Bai *et al.*, 2023b]	MM23	39.85	64.1	72.2	29.77
Baseline (ours)	-	39.60	61.95	71.25	27.61
GAAP (ours)	-	**44.45**	**65.15**	**75.30**	**31.21**

Table 3: Performance comparison on RSTPReid dataset.

4.3 Comparison with the State-of-the-Art

To verify the superiority of our proposed framework, We conducted experiments in three widely used text-based person search datasets. Our baseline is built on the basis of BLIP [Li *et al.*, 2022], using our generated captions to train the model.

CUHK-PEDES. We compare our framework with SOTA methods on CUHK-PEDES, as shown in Table 1. Our method

Methods	Components					Rank-1	mAP
	SS	LLM	CMCA	AITC	AITM		
Baseline	-	-	-	-	-	39.36	34.35
Model 1	✓	-	-	-	-	44.28	37.51
Model 2	✓	✓	-	-	-	45.37	38.50
Model 3	✓	✓	✓	-	-	46.71	40.13
Model 4	✓	✓	✓	✓	-	47.33	40.82
GAAP	✓	✓	✓	✓	✓	47.64	41.28

Table 4: Ablation experiments on CUHK-PEDES.

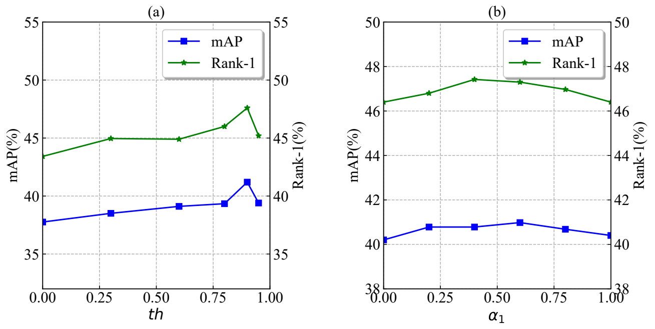


Figure 3: The influence of sample selection value th and weighting parameter α_1.

achieves 47.64% Rank-1 and 41.28% mAP, which is 6.77% and 5.56% higher than the previous method. Besides, the performance has a remarkable improvement of 8.28% at Rank-1 and 6.93% at mAP compared with baseline. Moreover, our method even outperforms some supervised methods.

ICFG-PEDES. Our framework achieves competitive performance on the ICFG-PEDES dataset, as shown in Table 2. ICFG-PEDES is the most challenging dataset with a large query in the test set. GAAP set the new SOTA Rank-1 of 27.12% and mAP of 11.43%, surpassing the baseline by 4.61% Rank-1 and 0.58% mAP.

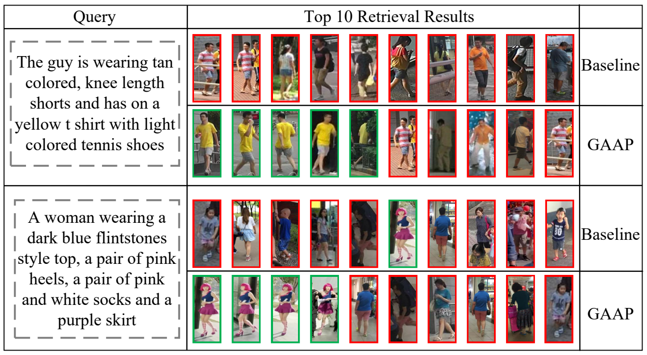
RSTPReid. As shown in Table 3, the baseline achieves 39.60% Rank-1 and 27.61% mAP accuracy on RSTPReid while GAAP reaches 44.45% Rank-1 and 31.21% mAP.

4.4 Ablation Study

A series of ablation experiments are performed to analyze the effectiveness of components within the GAAP framework on the CUHK-PEDES dataset, as shown in Table 4.

Ablations on the components. We conduct experiments to investigate the efficacy of the Sample Selection and LLM rephrase module in the caption generation module. The sample selection (SS) is used to select clean image-text pairs for training. As shown in Table 4, SS contributes significantly, bringing an improvement of 4.9% and 3.2% on rank-1 and mAP, This result validate the effectiveness of removing noisy pairs, thereby enhancing overall performance. When adding the LLM Rephrase module, our method gains a commendable 1.1% for Rank-1 and 1.0% for mAP. It validates that integrating more stylish captions can bolster performance.

In the Attribute-guided Cross-modal Alignment module, CMCA is proposed to narrow the gap between text and image samples with cross-modal center features, which improves the performance by 1.3% and 1.6% on Rank-1 and mAP. Furthermore, the discernible efficacy of AITM becomes evident, culminating in performance gains of approximately 0.6% and 0.7%, as shown in Model 3. Besides, AITM gains a 0.5% improvement on mAP. As mentioned before, AITM and AITC are different from vanilla ITM and ITC, which are designed for attribute-based work.



The guy is wearing tan colored, knee length shorts and has on a yellow t shirt with light colored tennis shoes

A woman wearing a dark blue flintstones style top, a pair of pink heels, a pair of pink and white socks and a purple skirt

Figure 4: Retrieval results comparison between the baseline and GAAP methods for unsupervised text-based person search.

Parameters analysis. The Sample Selection threshold, denoted as th, serves the purpose of choosing reliable pairs whose matching score surpasses this threshold. The observations in Figure 3 elucidate that an increase in the value of th corresponds to an improvement in performance. This phenomenon arises from the utilization of cleaner samples during training. The weighting parameter α_1 aids in harmonizing the ground truth label and the image-attribute score. Consequently, as α_1 ranges from 0 to 0.4, a discernible enhancement in performance is observed. This can be attributed to the contribution of the attribute-guided score, effectively relieving the strict regulation. However, as α_1 continues to increase, the performance starts to decline, stemming from the diminishing contribution of the GT labels. This observation validates of the efficacy of our proposed approaches.

Retrieval results comparison. We additionally present a comparative analysis of the top-10 retrieval results between the baseline method and the proposed GAAP method. The integration of attribute guidance in our approach enhances its capability to discern finer details and information during the retrieval process.

5 Conclusion

In this paper, we introduce a Cross-modal Generation and Alignment via Attribute-guided Prompt framework (GAAP) for unsupervised text-based person search. Our approach leverages a pre-trained vision-language model to generate person attributes using a diverse set of attribute prompts with a sample selection module to identify reliable training samples. Subsequently, we propose an Attribute-based Cross-modal Alignment module to effectively align image and text features with attribute-guided assistance. Through extensive experimental validation, we demonstrate the efficacy of our method for unsupervised text-based person retrieval.

Contribution Statement

Zongyi Li and Jianbo Li made equal contributions. All the authors participated in designing research, performing research, analyzing data, and writing the paper.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 61972169,62302186 and 62372203, in part by China Postdoctoral Science Foundation 2022M711251, 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), in part by the Knowledge Innovation Program of Wuhan-Basic Research.

References

[Aggarwal *et al.*, 2020] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2617–2625, 2020.

[Bai *et al.*, 2023a] 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. *arXiv preprint arXiv:2305.13653*, 2023.

[Bai *et al.*, 2023b] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. *arXiv preprint arXiv:2305.12964*, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[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.

[Cubuk *et al.*, 2020] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[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. *arXiv preprint arXiv:2107.12666*, 2021.

[Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[Farooq *et al.*, 2022] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4477–4485, 2022.

[Gao *et al.*, 2022] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

[Gao *et al.*, 2023] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Wei Liu, Jie Yang, Ke Li, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. *arXiv preprint arXiv:2303.17561*, 2023.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[He *et al.*, 2022] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 879–887, 2022.

[Jiang and Ye, 2023] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[Jing *et al.*, 2020] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10678–10686, 2020.

[Jing *et al.*, 2023] Peiguang Jing, Kai Cui, Weili Guan, Liqiang Nie, and Yuting Su. Category-aware multimodal attention network for fashion compatibility modeling. *IEEE Transactions on Multimedia*, 25:9120–9131, 2023.

[Jing *et al.*, 2024] Peiguang Jing, Kai Cui, Jing Zhang, Yun Li, and Yuting Su. Multimodal high-order relationship inference network for fashion compatibility modeling in internet of multimedia things. *IEEE Internet of Things Journal*, 11(1):353–365, 2024.

[Li *et al.*, 2017] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image reidentification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Niu *et al.*, 2022] Kai Niu, Linjiang Huang, Yan Huang, Peng Wang, Liang Wang, and Yanning Zhang. Cross-modal co-occurrence attributes alignments for person search by language. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4426–4434, 2022.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[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, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Shao *et al.*, 2022] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5566–5574, 2022.

[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 *Proceedings of the 31th ACM International Conference on Multimedia*, 2023.

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

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

[Wang *et al.*, 2022a] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2540–2549, 2022.

[Wang *et al.*, 2022b] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022.

[Wang *et al.*, 2022c] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 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*, pages 1984–1992, 2022.

[Xiao *et al.*, 2017] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017.

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

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

[Yang *et al.*, 2023] Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. *arXiv preprint arXiv:2306.02898*, 2023.

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

[Zhao *et al.*, 2021] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11395–11404, 2021.

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

[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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021.

[Zuo *et al.*, 2023] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*, 2023.