PLIP: Language-Image Pre-training for Person Representation Learning

Jialong Zuo [1] **Jiahao Hong** [1] **Feng Zhang** [1] **Changqian Yu** [2]
Hanyu Zhou [1] **Changxin Gao** [1*] **Nong Sang** [1] **Jingdong Wang** [3]

[1] National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, [2] Skywork AI, [3] Department of Computer Vision, Baidu Inc.
{jlongzuo, cgao}@hust.edu.cn, {wangjingdong}@baidu.com

Abstract

Language-image pre-training is an effective technique for learning powerful representations in general domains. However, when directly turning to person representation learning, these general pre-training methods suffer from unsatisfactory performance. The reason is that they neglect critical person-related characteristics, i.e., fine-grained attributes and identities. To address this issue, we propose a novel language-image pre-training framework for person representation learning, termed PLIP. Specifically, we elaborately design three pretext tasks: 1) Text-guided Image Colorization, aims to establish the correspondence between the person-related image regions and the fine-grained color-part textual phrases. 2) Image-guided Attributes Prediction, aims to mine fine-grained attribute information of the person body in the image; and 3) Identity-based Vision-Language Contrast, aims to correlate the cross-modal representations at the identity level rather than the instance level. Moreover, to implement our pre-train framework, we construct a large-scale person dataset with image-text pairs named SYNTH-PEDES by automatically generating textual annotations. We pre-train PLIP on SYNTH-PEDES and evaluate our models by spanning downstream person-centric tasks. PLIP not only significantly improves existing methods on all these tasks, but also shows great ability in the zero-shot and domain generalization settings.

1 Introduction

Person-centric tasks, such as image/text-based person re-identification, person attribute recognition, person search and human parsing, are becoming increasingly influential in widespread applications, such as security monitor, smart city, virtual reality and scene understanding. Benefiting from the advances in designing task-specific methods [78, 9, 41, 54, 50], these tasks have achieved significant progress for the past few years.

However, it has become evident through recent advancements in research that the mere development of sophisticated models based on specific tasks has already encountered a performance bottleneck. At the same time, some works [35, 15, 10, 14] on general representation learning have shown great potential to further improve model performance. Due to the above reasons, some researchers [12] have attempted to learn generic person representations by utilizing rich person images. However, their pre-training method based on sparse visual information falls short in terms of both training performance and efficiency.

Meanwhile, some works [72, 49, 85] have demonstrated that introducing the language modality helps to learn better representations in general domains, for the language naturally enjoys higher information density. However, when it comes to person representation learning, these general language-image pre-training methods like CLIP [72] are often unsatisfactory in performance. We believe the reason is that they neglect some critical person-related characteristics, i.e., fine-grained attributes and identities.

*Corresponding Author. Project Link: https://github.com/Zplusdragon/PLIP

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

On the one hand, these general language-image pre-training methods typically implement the global alignment between cross-modality representations and lack explicit consideration for fine-grained information. However, in person domain, the fine-grained information, such as the person attributes, plays a key role in distinguishing a specific person. Neglecting fine-grained attributes will easily lead to difficulty in learning discriminative person representations. On the other hand, they are based on instance level and only incorporate the concept of image-text pairs. They simply treat each image-text pair in the same person identity as different pairs, and assume that images and texts not in the same pair do not have a corresponding relationship. However, in person domain, there exists a notable concordance between the images and texts within the same person identity. If only conducting the optimization at the instance level, it will lead to instability to learn more meaningful representations.

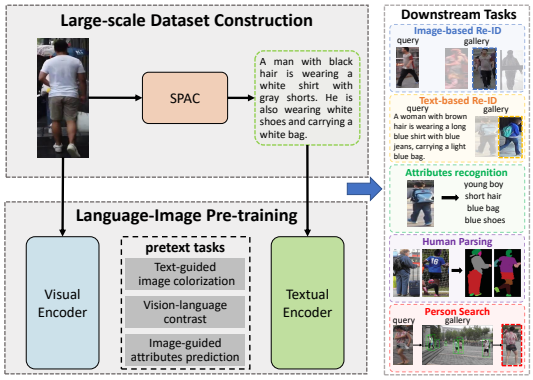

Figure 1: Illumination of our framework. Based on the constructed dataset, we pre-train a language-image model by three pretext tasks and transfer the model to some downstream person-centric tasks.

To address these limitations mentioned above, we deeply consider the characteristics of persons and attempt to introduce the language modality into person representation learning. We propose a well-motivated language-image pre-training framework for learning generic and discriminative person representations, termed PLIP, to help the downstream person-centric tasks. Also, to implement the pre-training, we construct a large-scale person dataset with rich image-text pairs. The whole framework is illustrated in Fig. 1.

Specifically, to explicitly learn fine-grained and meaningful cross-modal associations, we design three pretext tasks in PLIP:

(1)*Text-guided Image Colorization*, given a complete textual description, is designed to restore the color information of a grayscale transformed person image. This task establishes the correspondence between the person-related image regions and the fine-grained color-part textual phrases, which robustly helps the model to learn the semantic concept of person body parts.

(2) *Image-guided Attributes Prediction*, by exploiting the paired colorful images, is designed to predict the masked attribute phrases in textual descriptions. This task primarily focuses on predicting attributes, rather than predicting any random masked words as in the general domains. Through this multi-modal masked language modeling, it helps the model to understand the key areas and fine-grained attribute information of the person body in the image, which is crucial to identifying a person.

(3) *Identity-based Vision-language Contrast* is designed to associate representations between vision and language at the identity level rather than the instance level in general domains. This means it is optimized by narrowing the distance between any images and texts in the same person identity and widening the distance between those not in the same person identity. By taking identity into consideration, this task achieves more robust and meaningful association between different modalities.

As is well known, the quality and quantity of training data is essential for learning rich representations. However, there exists huge domain gap between the large-scale image-text pairs used in general domains and the specific person data. Also, the scale of the existing person datasets [52, 24] with manual textual descriptions is limited due to expensive hand-labeled annotations. Therefore, we construct a new large-scale person dataset with image-text pairs named SYNTH-PEDES based on the LUPerson-NL and LPW datasets [30, 83]. The text annotations are automatically synthetized by our proposed person image captioner named Stylish Pedestrian Attributes-union Captioning (SPAC). The dataset contains 312,321 identities, 4,791,711 images and 12,138,157 textual descriptions. At the same time, extensive experiments have been conducted to verify the competitive quality of our synthetic dataset compared to manually annotated datasets [52, 24, 122], which guarantees the superior performance of the learned person representations.

We utilize PLIP to pre-train our models on the SYNTH-PEDES dataset, and then evaluate the model performance on spanning downstream person-centric tasks. Extensive experiments show that our pre-trained models have learned generic person representations, pushing many state-of-the-art methods to a higher level on a wide range of person-centric tasks without bells and whistles. For example, for unsupervised person Re-ID, by applying our pre-trained ResNet50 model on PPLR [20], we improve the mAP metric on Market1501 [114] and MSMT17 [94] by 5.1% and 14.7%, respectively. The key contributions of this paper can be summarized as follows:

(1) We propose a novel language-image pre-training framework with three pretext tasks, termed PLIP, which deeply takes person-related characteristics into consideration. It can facilitate fine-grained cross-modal association and learning generic person representations explicitly.

(2) To implement the pre-training, we construct a large-scale person dataset with generated text annotations, called SYNTH-PEDES. The dataset provides rich image-text pre-training data for this community.

(3) We pre-train PLIP on SYNTH-PEDES and the learned representations perform remarkable ability in various downstream person-centric tasks. It is demonstrated as generic to bring significant improvements to various baseline methods.

2 PLIP: Representation Learning Framework

This section presents our proposed language-image based person representation learning framework PLIP via three pretext tasks, *i.e.,* text-guided image colorization (TIC), image-guided attributes prediction (IAP), and identity-based vision-language contrast (IVLC). As illustrated in Fig. 2, the whole architecture is a dual branch encoder structure, and generic person representations can be learned through joint training of these three pretext tasks.

2.1 Text-guided Image Colorization

The TIC task is designed to restore the original color information of grayscale transformed images by exploiting the complete textual descriptions. Such cross-modal colorization promotes the construction of image-text association. The reason is that the attribute phrases in the descriptions generally contain fine-grained person-centric information especially color information and this colorization process naturally enables the model to understand the key components in textual descriptions and achieve a relationship construction between the textual phrases and visual regions. The overall task can be converted into a pixel-wise regression problem.

As illustrated in Fig. 2, for a pair of a gray image and a complete textual description $\{\mathbf{i}_{gray}, \mathbf{t}_{complete}\}$, in the encoding stage, the input gray image $\mathbf{i}_{gray}$ is firstly fed into a hierarchical network, which is as the visual encoder. Secondly, the hierarchical features are up-sampled to the same scale and concatenated to produce the feature $\mathbf{F}_{gray}$. Then, we feed the complete textual description $\mathbf{t}_{complete}$ to the textual encoder [23] and adopts the average of the hidden-state outputs as the textual global embedding $\mathbf{T}_{global}$.

In the decoding stage, the visual feature $\mathbf{F}_{gray}$ and textual global embedding $\mathbf{T}_{global}$ should be fused for colorization. Specifically, we adopt the multi-modal SE-blocks [95] as the cross-modal feature fusing module, so that the textual global embedding can play a role in the visual feature channels. In the block, the visual feature $\mathbf{F}_{visual}$ is compressed into a feature vector $\mathbf{V}_f$ through max pooling operation. Then we concatenate the visual feature vector and the textual global embedding $\mathbf{T}_{global}$, and feed the concatenated vector into several fc layers and a softmax layer to generate an attention vector $\mathbf{A}_f$. Finally, we utilize $\mathbf{A}_f$ on the visual feature $\mathbf{F}_{visual}$ to generate a multi-modal feature $\mathbf{Z}_m$. The decode is also made up of several de-convolution layers, which are employed to restore the feature dimensions. Finally, we generate a mutilmodal feature map with same dimensions as the input gray image $\mathbf{i}_{gray}$ and it can be utilized to predict the target color image $\mathbf{i}_{color}$.

We denote θ_{tic} as the parameters of the trainable regression model mentioned above. It maps the textual global embedding $\mathbf{T}_{global}$ and the gray image extracted feature $\mathbf{F}_{gray}$ to the output recovered color image $\mathbf{i}_{color}$ as a target. TIC is supervised by:

$$\mathcal{L}_{tic} = \frac{1}{N}\sum^{N}\mathcal{D}(\mathbf{i}_{color}, \theta_{tic}(\mathbf{T}_{global}, \mathbf{F}_{gray})), \tag{1}$$

where $\mathcal{D}$ can be any differentiable distance function such as Euclidean distance we adopt and N represents the total number of samples within a batch.

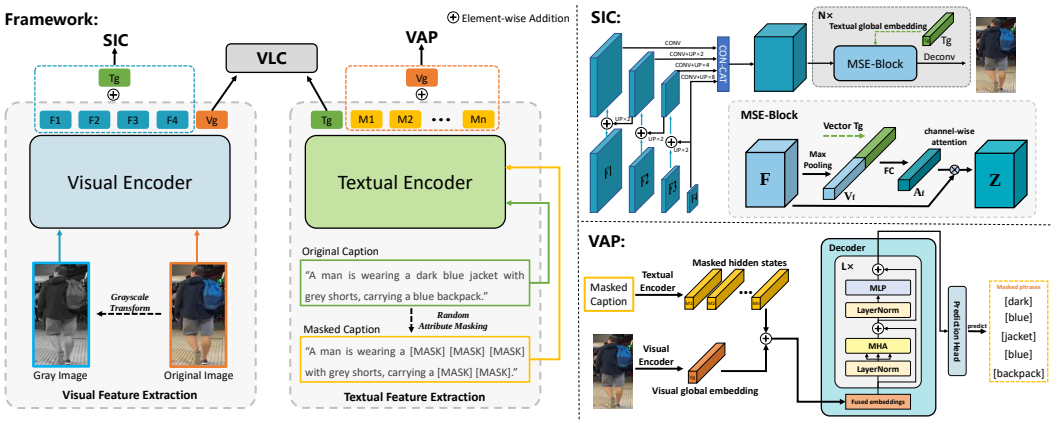

Figure 2: Overview of our proposed framework incorporating a text-guided image colorization task, an image-guided attributes prediction task and an identity-based vision-language contrast task.

As displayed in Fig. 5 of the appendix, altering the color word in textual description significantly affects the colorization of image. However, our model may not fully understand the semantics of more detailed image regions. As shown in the last row, our model fails to distinguish between the blue clothing region and the red shoulder strap region (marked with yellow boxes), instead blending the two into a unified coloration. This is due to the fact that the level of detail in manually annotated datasets is still not sufficient, resulting in the model being unable to theoretically learn representations with higher levels of detail and greater discrimination capabilities. However, it is undeniable that our model has a preliminary understanding of the meaning of attributes and colors, and can associate them with related image regions, rather than simple memorization. This ability to distinguish between different parts of the person body guarantees the superior performance on the subsequent person-centric tasks.

2.2 Image-guided Attributes Prediction

The IAP task requires utilizing original color images to predict the masked attribute phrases in textual descriptions. Unlike previous methods [42, 95] that randomly mask any words or only color words in a description, our method focuses on masking attribute phrases. For each sentence, the attribute phrases are partially masked to create a masked textual description. In this multi-modal masked language modeling (MMLM) way, the correlation between images and texts can be bridged more in depth and more discriminative representations can be learned. The reason is that the prediction process enables the model to further understand the person-centric regions in images and extract the key semantic information. Meanwhile, by exploiting the visual representations, the MMLM enhances the perception of context and strengthens the interaction between vision and language modality.

In the encoding stage, as illustrated in Fig. 2, for a pair of a color image $\mathbf{i}_{color}$ and a masked textual description $\mathbf{t}_{masked}$ with masked words $\mathbf{w}_m = \{\mathbf{w}_{m_1}, \ldots, \mathbf{w}_{m_M}\}$ (M is the number of masked words), we feed them to respective encoders to extract the visual global embedding $\mathbf{V}_{global}$ and textual hidden-state outputs $\mathbf{h}_t = \{\mathbf{h}_{t_1}, \ldots, \mathbf{h}_{t_L}\}$ (L is the length of the tokens in the masked textual description). The visual global embedding $\mathbf{V}_{global}$ is obtained from a pooling operation on the last stage feature output of the visual encoder.

In the decoding stage, to perform the IAP task, we specially design a simple but effective mutil-modal fusion module that mainly consists of self-attention blocks and a prediction head. Firstly, we adopt the element-wise summation of $\mathbf{h}_t$ and $\mathbf{V}_{global}$ as the preliminary fused embeddings. Then, the embeddings will be served as query($\mathcal{Q}$), key($\mathcal{K}$) and value($\mathcal{V}$) simultaneously. Finally, we obtain the multi-modal fused embeddings for each masked position by:

$$\{\mathbf{h}_{m_i}\}_{i=1}^{M} = Blocks(\mathcal{Q}, \mathcal{K}, \mathcal{V}), \tag{2}$$

where M is the total number of masked tokens in the masked textual description, and $Blocks$ are the self-attention blocks.

For each embedding representing the masked word, we use a prediction head to realize the corresponding probability prediction. The IAP can be optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_{iap} = -\frac{1}{MN}\sum_{k=1}^{N}\sum^{M}\log P(\mathbf{w}_{m_k}|\mathbf{h}_{m_k}), \tag{3}$$

where M is the total number of masked words in a masked textual description, N denotes the number of samples within a batch and P denotes the probability distribution mapping.

2.3 Identity-based Vision-language Contrast

In order to further strengthen the correlation between vision and language modalities, we must optimize the model to learn a unified cross-modal feature space. A preliminary approach based on contrastive learning [72] is to shorten the distance of the representations in image-text pairs and simultaneously amplify the distance of representations not in the same pair. This approach considers image-text pairs as instances and ignores the identity, where different image-text pairs of the same identity are treated as negative samples. However, in person-centric field, the identity plays a crucial role in distinguishing different people. Therefore, unlike the usual practices in general domain, we must take the identity into consideration.

For a group of a color image, a complete textual description and a corresponding identity $\{\mathbf{i}_{color}, \mathbf{t}_{complete}, Id\}$. The color image is firstly fed into the visual encoder. Then the last-stage feature is pooled to get the visual global embedding $\mathbf{V}_{global}$. The description is directly fed into the textual encoder and the average pooling of hidden-states will be served as the textual global embedding $\mathbf{T}_{global}$.

Then, given a batch of N image-text pairs, for each visual global embedding $\mathbf{V}^i_{global}$, we construct a set of visual-textual embedding pairs as $\{(\mathbf{V}^i_{global}, \mathbf{T}^j_{global}), y_{i,j}\}^N_{j=1}$, where $y_{i,j} = 1$ means that the pair is matched and from the same identity, and $y_{i,j} = 0$ indicates the unmatched pair. Let $sim(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top\mathbf{z}/\|\mathbf{x}\|\|\mathbf{z}\|$ denotes the matching probability of $\mathbf{x}$ and $\mathbf{z}$. Then, the probability of matching pairs can be calculated with the following function:

$$p_{i,j} = \frac{\exp\left(sim\left(\mathbf{V}^i_{global}, \mathbf{T}^j_{global}\right)\right)}{\sum^N_{k=1}\exp\left(sim\left(\mathbf{V}^i_{global}, \mathbf{T}^k_{global}\right)\right)}. \tag{4}$$

Then, the IVLC loss from vision to language in a batch can be computed by:

$$\mathcal{L}_{v2l} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right), \tag{5}$$

where $q_{i,j} = y_{i,j}/\sum^N_{k=1} y_{i,k}$ is the true matching probability and ϵ is a small number to avoid numerical problems.

Similarly, the IVLC loss from language to vision $\mathcal{L}_{l2v}$ can be be computed by exchange V_{global} and T_{global} in above equations. Finally, the IVLC task can be optimized by:

$$\mathcal{L}_{IVLC} = \mathcal{L}_{v2l} + \mathcal{L}_{l2v}, \tag{6}$$

Indeed, the IVLC loss can be any cross-modal contrastive loss that takes identity into consideration such as the Cross-Modal Projection Matching (CMPM) loss [112] we adopt, which promotes the representation association between multiple modalities by incorporating the cross-projection into KL divergence.

Then, to supervise the model to learn discriminative and highly generic person representations, the overall multi-task loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{ivlc} + \lambda_1\mathcal{L}_{tic} + \lambda_2\mathcal{L}_{iap}, \tag{7}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}^+$ are hyper-parameters to control the importance of each pretext task.

Figure 3: Visualization of some examples in our SYNTH-PEDES dataset.

3 SYNTH-PEDES: A Large-scale Image-text Person Dataset

We build the SYNTH-PEDES dataset to pre-train our PLIP models at a large-scale. In this section, we show the general process of constructing our SYNTH-PEDES dataset, which can be described as three steps. The complete construction details can be found in Sec. A.5 of the appendix.

Firstly, we collect and process two large-scale person datasets to form the image dataset. The first is LUPerson-NL [30]. It is a new variant of LUPerson [29] on top of raw videos from LUPerson and assign the noisy labels to each person image with automatically generated tracklet. It consists of $10M$ images with about $430K$ identities collected from $21K$ scenes. The second is LPW [83]. It consists of 2,731 different persons and 592,438 images collected from three different crowded scenes.

Secondly, we propose an image captioner named SPAC to generate satisfactory textual descriptions for the images. Given an input person image, there is no specific work targeting at generating captions that detailedly describe the person's appearance. To this end, we propose a simple but effective method for person image captioning. It can generate attribute-annotations and stylish textual-descriptions, which simulate the diverse perspectives that different annotators may have on the same person image. The specific technics for SPAC can be found in Sec. A.5.1 of the appendix.

Thirdly, we adopt some post-processing approaches to eliminate the noises and improve the dataset quality. We propose Seed Filter Strategy to filter the noises in LUPerson-NL, which includes three processes of Filter-out, Reassignment and Merger. Meanwhile, we propose Data Distribution Strategy to ensure the quality of generated attributes, the consistency of gender annotation, and identity distribution balance. The specific details for the strategies can be found in Sec. A.7 of the appendix.

Thanks to the outstanding generating ability of our proposed SPAC, the SYNTH-PEDES dataset is full of high-quality textual descriptions in a variety of styles, which can be utilized to train the representation learning model. Compared with existing person datasets, SYNTH-PEDES has the following advantages:

Diversified. Our dataset contains a wide range of variations in the textual descriptions. Unlike the previous person datasets with only one or two image-text pairs, most images of our dataset are annotated with three textual descriptions.

High-quality. As some typical qualitative examples can be seen in Fig. 3, the generated annotations achieve an accurate and detailed description of the person appearance. The further experiments conducted on the dataset quality evaluation can be found in Sec. A.9.2 of the appendix. Researchers can use this dataset with confidence to conduct relevant studies.

Large-scale. In Tab. 11 of the appendix, we have compared the properties of SYNTH-PEDES with other popular person datasets. As we can see, SYNTH-PEDES is the largest real person dataset with high-quality image-text pairs by far, which contains 312,321 identities, 4,791,711 images, and 12,138,157 textual descriptions.

4 Experiments

Implementation. During the training of PLIP, we adopt four types of backbone as the visual encoder, *i.e.,* ResNet50, ResNet101, ResNet152 and Swin Transformer Base. The pre-trained BERT [23] is utilized as the textual encoder with the last 5 layers unfrozen. We train our model on 4 × Geforce 3090 GPUs for 70 epochs. For each person-centric downstream task, we reproduce a range of state-of-the art methods as the baselines. If not specially stated, we perform the experiments by just replacing the backbone in each baseline to the our pre-trained models. We perform in-depth experiments on eleven datasets for five downstream person-centric tasks. Meanwhile, we perform thorough ablation studies and analyses in Sec. A.9 of the appendix. More details and experiments can be found in the appendix.

Table 1: Performance comparisons of our PLIP with some other SoTA pre-trained models on downstream tasks. The table is divided into two parts: the upper part covers some general-domain pre-trained models and the lower part focuses on some person-domain pre-trained models. The baseline downstream methods for the five tasks are CMPM/C [112], ABDNet [9], Rethinking [41], SeqNet [54] and SCHP [50], respectively. Due to the inability of certain non-hierarchical models to be applied to some downstream methods, their performance results are indicated as "-".

Method	Backbone	Text Re-ID		Image Re-ID		Attribute Recog.		Person Search		Human Parsing	
		CUHK	ICFG	Market	Duke	PETA	PA100K	SYSU	PRW	LIP	PASCAL
Baseline	RN50	54.81/83.22	47.61/75.48	88.3/95.6	78.6/89.0	83.96/86.35	80.21/87.40	94.8/95.7	47.6/87.6	59.36	71.46
Baseline	RN101	56.77/84.86	51.78/77.92	89.3/95.7	79.8/88.9	85.17/86.44	81.61/87.62	95.1/96.1	48.4/87.7	60.57	72.12
Baseline	RN152	58.64/85.84	53.13/79.18	89.6/95.8	80.1/89.0	85.81/86.64	82.17/87.96	95.3/96.3	48.8/87.9	61.71	72.87
MoCov3 [17]	RN50	52.17/81.94	46.51/74.86	87.7/95.0	77.3/87.7	82.87/85.19	79.12/86.56	94.1/94.5	46.2/86.1	57.63	69.16
MoCov3 [17]	ViT-S	53.21/82.63	46.77/74.62	-/-	-/-	83.51/85.98	79.64/87.12	-/-	-/-	-	-
MoCov3 [17]	ViT-B	55.46/83.82	50.28/76.22	-/-	-/-	84.82/86.83	80.96/87.28	-/-	-/-	-	-
CLIP [72]	RN50	57.92/85.12	51.30/77.64	89.5/95.7	79.7/88.9	84.49/86.65	81.07/87.82	92.8/93.4	48.0/86.7	59.81	70.93
CLIP [72]	RN101	58.04/85.88	52.80/79.31	90.6/96.0	81.4/89.9	85.67/87.04	81.82/87.95	93.3/93.8	48.5/87.4	60.74	72.39
CLIP [72]	ViT-B	58.91/86.21	53.73/79.98	-/-	-/-	86.37/87.95	82.41/88.41	-/-	-/-	-	-
BLIP [49]	ViT-B	60.37/86.63	54.64/80.23	-/-	-/-	86.97/88.42	82.81/88.46	-/-	-/-	-	-
BLIP [49]	ViT-L	63.23/88.74	56.96/82.79	-/-	-/-	87.08/88.55	83.23/88.62	-/-	-/-	-	-
LUP [29]	RN50	57.51/84.86	50.52/76.63	90.2/95.9	80.7/89.3	84.17/86.10	81.94/88.25	94.3/94.9	47.7/87.3	59.43	71.89
LUP [29]	RN101	57.84/85.36	53.08/78.93	91.4/96.5	83.1/90.9	85.66/86.57	82.31/88.21	94.8/95.5	47.9/87.4	60.41	72.35
LUP [29]	RN152	59.41/86.22	53.79/79.51	91.7/96.7	83.7/91.9	85.92/87.44	82.63/88.47	95.3/95.9	48.6/88.1	62.05	73.16
LUP-NL [30]	RN50	57.85/84.97	50.64/76.55	90.8/96.1	81.7/89.7	84.09/86.32	81.71/88.08	95.4/95.9	49.3/87.5	59.88	71.67
LUP-NL [30]	RN101	58.12/85.74	53.38/78.85	91.4/96.5	82.1/90.9	85.77/86.93	82.39/88.28	95.5/95.7	49.6/87.6	60.92	72.51
LUP-NL [30]	RN152	59.27/86.06	54.21/80.12	91.5/96.5	81.8/90.6	86.18/88.09	82.54/88.65	95.8/96.1	49.5/88.2	61.98	73.12
SOLIDER [12]	Swin-T	51.22/80.27	46.51/74.16	87.2/94.6	77.5/87.8	84.12/86.27	83.81/87.91	94.9/95.7	56.8/86.8	57.52	68.42
SOLIDER [12]	Swin-S	56.43/84.78	51.19/77.82	89.6/95.5	80.2/89.3	85.47/87.14	85.32/89.08	95.5/95.8	**59.8**/86.7	60.21	69.81
SOLIDER [12]	Swin-B	58.06/85.64	52.37/79.62	90.1/95.6	81.2/89.8	85.78/87.46	**85.91/89.66**	94.9/95.5	59.7/86.8	60.50	70.02
PLIP(ours)	RN50	71.13/92.81	64.88/87.32	91.4/96.8	81.7/91.1	85.12/86.94	82.21/88.62	96.0/96.7	53.5/88.9	60.41	72.14
PLIP(ours)	RN101	72.34/93.46	64.92/87.66	92.0/96.9	82.3/91.8	85.84/87.42	82.52/88.16	96.2/96.8	54.6/89.4	61.32	72.63
PLIP(ours)	RN152	73.68/94.22	65.52/88.40	92.6/97.1	83.1/92.1	86.28/88.17	82.91/88.83	**96.5/97.0**	55.7/89.8	62.44	73.51
PLIP(ours)	Swin-B	**75.36/94.87**	**66.17/88.94**	**93.2/97.3**	**84.4/92.2**	**87.12/88.84**	83.65/89.17	96.4/96.7	56.6/**89.9**	**63.52**	**73.93**

4.1 Comparison With Other Pre-trained Models

We have compared the performance of our PLIP pre-trained models with other SoTA pre-trained models [17, 29, 30, 12, 72, 49, 106] on five downstream tasks. We reproduce a range of popular and open-sourced methods [112, 9, 41, 54, 50] as the baselines, which are initialized by ImageNet supervised pre-trained ResNet backbones. Then, we compare the performance of different pre-trained models by simply replacing the backbone in each method. The compared models can be divided into general-domain pre-trained models and person-domain pre-trained models, with the latter typically exhibiting better performance in person-centric tasks. The performance metrics for these tasks are R@1/R@10, mAP/R@1, mAP/F1, mAP/R@1 and mIoU, respectively. As shown in Tab. 1, our PLIP models consistently demonstrate a significant performance advantage over other pre-trained models in a fair comparison with roughly equal parameters.

4.2 Evaluation on Text-based Person Re-ID

Transfer capability. To evaluate the transfer capability of our pre-trained models, we conduct three different experiments with ResNet50 as the visual encoder. Firstly, we directly evaluate the model's zero-shot performance without any extra fine-tuning. Secondly, we perform linear probing by adding a trainable linear embedding layer to each modal frozen encoder. Finally, we unfreeze all encoders and perform fine-tuning. We use the simple CMPM loss [112] as the training target. From Tab. 2 we can see, our pre-trained model is not only competitive with some fully supervised methods [93, 24] even without fine-tune training, but also greatly exceeds them by a large margin with a simple fine-tune. These results demonstrate that our pre-trained models have excellent transfer capability for this task.

Domain generalization. To verify our models' domain generalization capability, we carry out experiments with cross-domain settings. We use the CMPM loss as the training target. As illustrated in Tab. 3, our model achieves improvements by large margins when compared with all

Table 2: The results of our transfer experiments. We show the best score in bold. *z-s*: zero-shot setting; *l-p*: linear-probing setting; *f-t*: fine-tune setting. ‡ stands for the results reproduced with public checkpoints released by the authors.

Method	CUHK-PEDES			ICFG-PEDES		
	R@1	R@5	R@10	R@1	R@5	R@10
ViTAA [93]	55.97	75.84	83.52	50.98	68.79	75.78
SSAN [24]	61.37	80.15	86.73	54.23	72.63	79.53
LapsCore [95]	63.40	-	87.80	-	-	-
TIPCB‡ [18]	63.63	82.82	89.01	54.96	74.72	81.89
LGUR [78]	64.21	81.94	87.93	57.42	74.97	81.45
PLIP+*z-s*	52.86	74.69	82.46	49.86	68.78	76.27
PLIP+*l-p*	63.63	82.85	89.36	58.51	77.83	84.24
PLIP+*f-t*	**70.11**	**86.60**	**91.89**	**64.58**	**81.30**	**86.82**

Table 3: Comparison on domain generalization. "C" and "I" denote CUHK-PEDES and ICFG-PEDES, respectively.

Method	$C \rightarrow I$			$I \rightarrow C$		
	R@1	R@5	R@10	R@1	R@5	R@10
Dual Path [118]	15.41	29.80	38.19	7.63	17.14	23.52
SCAN [43]	21.27	39.26	48.83	13.63	28.61	37.05
SSAN [24]	29.24	49.00	58.53	21.07	38.94	48.54
LGUR [78]	34.25	52.58	60.85	25.44	44.48	54.39
RaSa [1]	50.59	67.46	74.09	50.70	72.40	79.58
PLIP	**56.64**	**75.65**	**82.38**	**57.34**	**77.60**	**84.49**

other methods. Specifically, our model outperforms LGUR by 22.4% and 31.9% in terms of Rank-1 metric on the $C \rightarrow I$ and $I \rightarrow C$ settings, respectively. These results demonstrate that our pre-trained models have great capability in domain generalization for this task.

Improvement over existing methods. We reproduce three representative baseline methods [112, 24, 78] and explore the performance difference by changing the encoders with different pre-trained models. From Tab. 4, we can see that equipped with our pre-trained model, all the baseline methods achieve higher and best accuracy on each dataset. It is worth noting that, due to the fact that our PLIP models have already learned an excellent joint visual-textual feature space through large-scale pre-training, achieving outstanding performance is easily attained by fine-tuning with the simple CMPM/C [112] loss rather than complicated designing such as SSAN [24] and LGUR [78].

Comparison with state-of-the-art methods. In Tab. 5, we compare our results with some SoTA methods on each dataset. The compared methods can be classified based on whether they rely on the multi-modal pre-trained models. Generally, multi-modal pre-trained models can bring about noticeable performance improvements for this task. It is worth noting that RaSa utilizes a larger image resolution of 384×384, while APTM adopts a two-stage inference method similar to the re-rank mechanism. These approaches lead to RaSa and APTM achieving optimal performance, however, introducing additional training and inference costs. Instead, our PLIP achieves competitive performance without bells and whistles. Specifically, with ResNet50 as backbone, PLIP outperforms LGUR by 6.9% and 7.5% rank-1 on each dataset respectively.

4.3 Evaluation on Image-based Re-ID

Unsupervised methods achieve significant improvements. With simply replacing the backbone, our pre-trained models benefit unsupervised image-based person Re-ID methods significantly. We evaluate the improvement brought by different pre-trained ResNet50 models to the SoTA unsupervised methods PPLR [20] and ISE [110]. As shown in Tab. 6, PLIP outperforms all other pre-trained models by a large margin. Specifically, applied to ISE, PLIP achieves new SoTA performance, outperforming the previsous SoTA by 2.9% and 11.4% mAP on Market1501 and MSMT17, respectively.

Comparison with state-of-the-art methods. We compare our results with existing SoTA image-based person Re-ID methods on Market1501 and DukeMTMC. Any results gained from post-processing techniques like re-rank [121] are excluded for a fair comparison. As indicated in Tab. 7, by applying our PLIP

Table 4: Comparison on three methods by using different pre-trained models. All results are shown in Rank-1/Rank-10.

	Pre-train	CMPM/C [112]	SSAN [24]	LGUR [78]
CUHK-PEDES	Baseline.	54.81/83.22	61.37/86.73	64.21/87.93
	MoCov3	52.17/81.94	61.97/86.63	65.33/88.47
	CLIP	57.92/85.12	62.09/86.89	64.70/88.76
	LUP	57.51/84.86	63.91/88.36	65.42/89.36
	LUP-NL	57.85/84.97	63.71/87.46	64.68/88.69
	PLIP	**71.13/92.81**	**65.90/89.44**	**67.62/89.90**
ICFG-PEDES	Baseline.	47.61/75.48	54.23/79.53	57.42/81.45
	MoCov3	46.51/74.86	55.27/79.64	59.90/82.94
	CLIP	51.30/77.64	53.58/78.96	58.35/82.02
	LUP	50.52/76.63	56.51/80.41	60.33/83.06
	LUP-NL	50.64/76.55	55.59/79.78	60.25/82.84
	PLIP	**64.88/87.32**	**60.70/83.17**	**62.38/84.28**

Table 5: Comparison with the state-of-the-art methods on text-based person Re-ID. We show the best score in bold.

Method	Image	Text	CUHK-PEDES			ICFG-PEDES		
			R@1	R@5	R@10	R@1	R@5	R@10
GNA-RNN [52]	VGG	LSTM	19.05	-	53.64	-	-	-
CMPM/C [112]	RN50	LSTM	49.37	-	79.27	43.51	65.44	74.26
ViTAA [93]	RN50	LSTM	55.97	75.84	83.52	50.98	68.79	75.78
NAFS [31]	RN50	BERT	59.94	79.86	86.70	-	-	-
SSAN [24]	RN50	LSTM	61.37	80.15	86.73	54.23	72.63	79.53
LapsCore [95]	RN50	BERT	63.40	-	87.80	-	-	-
TIPCB‡ [18]	RN50	BERT	63.63	82.82	89.01	54.96	74.72	81.89
LGUR [78]	RN50	BERT	64.21	81.94	87.93	57.42	74.97	81.45
IVT [82]	ViT-B	BERT	65.59	83.11	89.21	56.04	73.60	80.22
CFine [103]	ViT-B	BERT	69.57	85.93	91.15	60.83	76.55	82.42
IRRA [42]	ViT-B	Xformer	73.38	89.93	93.71	63.46	80.25	85.82
APTM [106]	Swin-B	BERT	76.17	89.47	93.57	**68.22**	82.87	87.50
RaSa [1]	ViT-B	BERT	**76.51**	90.29	94.25	65.28	80.40	85.12
PLIP	RN50	BERT	71.13	87.57	92.81	64.88	81.70	87.32
PLIP	RN101	BERT	72.34	89.15	93.46	64.92	81.63	87.66
PLIP	RN152	BERT	73.68	90.35	94.22	65.52	82.64	88.40
PLIP	Swin-B	BERT	75.36	**90.86**	**94.87**	66.17	**83.37**	**88.94**

Table 6: Comparison on two baseline methods by using different pre-trained models. The best results are shown in bold.

	Pre-train	Market1501				MSMT17			
		mAP	R@1	R@5	R@10	mAP	R@1	R@5	R10
PPLR [20]	Baseline	81.5	92.8	97.1	98.1	31.4	61.1	73.4	77.8
	MoCov2 [15]	79.6	91.6	96.6	97.9	28.7	57.5	69.6	74.6
	CLIP [72]	75.0	89.0	95.6	97.1	7.6	20.3	29.9	34.9
	LUP [29]	57.5	78.2	85.9	88.9	22.5	48.9	62.0	66.9
	LUP-NL [30]	84.2	93.4	97.7	98.6	25.0	50.3	63.1	68.7
	PLIP	**86.6**	**94.5**	**98.0**	**98.7**	**46.1**	**73.4**	**83.7**	**87.0**
ISE [110]	Baseline	84.7	94.0	97.8	98.8	35.0	64.7	75.5	79.4
	MoCov2 [15]	84.9	93.5	97.6	98.7	34.1	64.5	75.0	79.0
	CLIP [72]	79.5	92.0	96.8	98.0	14.9	35.2	46.3	51.3
	LUP [29]	84.5	94.2	97.6	98.4	27.8	56.7	68.8	73.4
	LUP-NL [30]	87.4	95.0	98.2	98.9	33.9	62.5	73.2	77.3
	PLIP	**87.6**	**95.3**	**98.2**	**98.9**	**46.4**	**74.9**	**84.3**	**87.5**

Table 7: Comparison with SoTA methods on fully supervised image-based person Re-ID. The best results are shown in bold.

Method	Settings		Market1501		DukeMTMC	
	Backbone	Pretrain	mAP	R@1	mAP	R@1
MGN [89]	RN50	IMG	87.5	95.1	79.4	89.0
ABDNet [9]	RN50	IMG	88.3	95.6	78.6	89.0
GCP [69]	RN50	IMG	88.9	95.2	78.6	87.9
ISP [123]	RN50	IMG	88.6	95.3	80.0	89.6
TransReID [38]	ViT-B	IMG	88.2	95.0	80.6	89.6
UPReID [107]	RN50	LUP	91.1	97.1	-	-
LUP [29]	RN50	LUP	91.0	96.4	82.1	91.0
LUP-NL [30]	RN50	LUP-NL	91.9	96.6	84.3	92.0
PASS [124]	ViT-B	LUP	93.0	96.8	-	-
PLIP	RN50	SYNTH	91.4	96.8	81.7	91.1
PLIP	RN101	SYNTH	92.0	96.9	82.3	91.8
PLIP	RN152	SYNTH	92.6	97.1	83.1	92.1
PLIP	Swin-B	SYNTH	**93.2**	**97.3**	**84.4**	**92.2**

on ABD-Net [9], we achieve competitive performance on all datasets. Moreover, we achieve new SoTA performance with Swin-Base as the backbone. This demonstrates that the learned representations benefits this task significantly.

4.4 Evaluation on Person Search

Improvement over existing methods. Our pre-trained models bring significant performance gain to existing person search methods. To verify this, we choose two representative methods SeqNet [54] and GLCNet [70] as the baselines, and evaluate the improvements brought by different pre-trained models. As shown in Tab. 8, under the ResNet50 setting, our model brings maximum performance improvements to all methods. Specifically, when applying our pre-trained ResNet50 to GLCNet, it achieves 96.3% and 53.7% mAP on the CUHK-SYSU and PRW datasets, respectively.

Comparison with state-of-the-art methods. We compare our results with existing SoTA person search methods in Tab. 9. By applying our pre-trained models on GLCNet [70], we achieve new SoTA performance on each dataset. Specifically, under ResNet50 setting, our method surpasses the previous SoTA GLCNet by 5.9% mAP on PRW dataset. Also, with Swin-Base as the backbone, we achieve the best performance compared with all other methods. These results demonstrate that our pre-training framework shows great potential in learning discriminative person representation for this task.

4.5 Ablation studies and analyses

We perform ablation studies with ResNet-50 as the visual encoder and pre-training on the sub-dataset of SYNTH-PEDES, which has 10,000 identities, 139,564 images and 353,617 textual descriptions. To assess the **effectiveness of pre-text tasks** on the generalizability of our models, we directly evaluate the zero-shot performance of pre-trained models on downstream datasets. As we can see in Tab. 10, each task contributes to the model's zero-shot capability and combining all of the tasks leads to the best performance. Meanwhile, to validate the **effectiveness of textual diversity** in SYNTH-PEDES, we have studied four different degrees of textual diversity from weak to strong. As shown in Fig. 4, the fourth case with the highest degree of textual diversity has the best performance. More thorough ablation studies and analyses about effectiveness of each component, dataset quality evaluation, pre-training settings and so on can be found in Sec. A.9 of the appendix.

Table 8: Comparison on two baseline methods by using different pre-trained models. We show the best score in bold.

	Pre-train	CUHK-SYSU				PRW			
		mAP	R@1	R@5	R@10	mAP	R@1	R@5	R10
SeqNet [54]	Baseline	94.8	95.7	98.1	98.7	47.6	87.6	94.4	95.4
	MoCov2 [15]	94.0	94.4	98.2	98.5	48.3	87.3	94.6	95.8
	CLIP [72]	92.8	93.4	97.8	98.2	48.0	86.7	94.5	95.9
	LUP [29]	94.3	94.9	98.2	98.7	47.7	87.3	94.5	95.7
	LUP-NL [30]	95.4	95.9	98.2	98.8	49.3	87.5	94.3	95.9
	PLIP	**96.0**	**96.7**	**98.7**	**99.1**	**53.5**	**88.9**	**95.2**	**96.3**
GLCNet [70]	Baseline	95.8	96.2	98.3	98.8	47.8	87.8	94.5	95.5
	MoCov2 [15]	95.1	95.6	98.0	98.4	48.4	87.8	94.6	95.9
	CLIP [72]	93.0	93.6	97.7	98.1	48.2	87.1	94.5	96.0
	LUP [29]	95.5	95.8	98.3	98.9	48.1	87.7	94.5	95.8
	LUP-NL [30]	96.0	96.4	98.4	99.0	49.8	87.8	94.6	96.0
	PLIP	**96.3**	**97.0**	**99.0**	**99.3**	**53.7**	**89.0**	**95.4**	**96.4**

Table 9: Comparison with state-of-the-art methods on the end-to-end person search. We show the best score in bold.

Method	Backbone	CUHK-SYSU		PRW	
		mAP	R@1	mAP	R@1
OIM [98]	RN50	75.5	78.7	21.3	49.9
NPSM [59]	RN50	77.9	81.2	24.2	53.1
CTXGraph [104]	RN50	84.1	86.5	33.4	73.6
QEEPS [67]	RN50	88.9	89.1	37.1	76.7
BINet [25]	RN50	90.0	90.7	45.3	81.7
NAE+ [5]	RN50	92.1	92.9	44.0	81.1
SeqNet [54]	RN50	94.8	95.7	47.6	87.6
GLCNet [70]	RN50	95.8	96.2	47.8	87.8
SOLIDER [12]	Swin-B	94.9	95.5	**59.7**	86.8
PLIP	RN50	96.3	97.0	53.7	89.0
PLIP	RN101	96.5	97.2	55.1	89.4
PLIP	RN152	**96.7**	97.3	56.2	89.9
PLIP	Swin-B	96.6	**97.5**	57.8	**90.3**

Table 10: Ablation study on the effectivenss of each pretext task, all using default settings.

#	Components			CUHK-PEDES			Market1501		
	IVLC	TIC	IAP	R@1	R@5	R@10	R@1	R@5	R@10
1	✓			30.2	53.3	64.0	62.3	79.9	85.7
2	✓	✓		31.4	55.2	65.9	62.9	80.7	86.2
3	✓		✓	30.5	54.4	64.3	62.7	80.5	86.1
4		✓	✓	-	-	-	39.3	61.4	70.4
5	✓	✓	✓	**32.5**	**56.3**	**66.6**	**63.1**	**80.8**	**86.3**

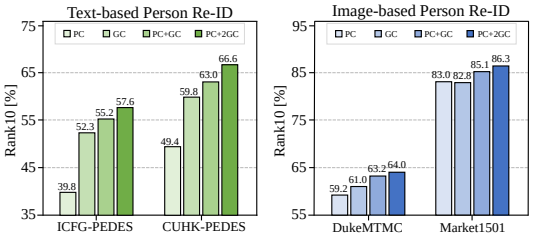

Figure 4: The diversity of textual descriptions matters. PC and GC mean prompt caption and generated caption, respectively.

5 Conclusion

In this paper, we propose a novel language-image self-supervised person representation learning framework named PLIP, which consists of three well-motivated pretext tasks. Also, we build a large-scale real-scenario image-text person dataset SYNTH-PEDES by auto-captioning procedures. We achieve good generic person representation learning by utilizing PLIP and SYNTH-PEDES. Equipped with our pre-trained models, we push many existing methods to a much higher level without bells and whistles. We hope that our simple and effective framework can inspire researchers to devote further attention to this area.

Acknowledgements. This work was supported by the National Natural Science Foundation of China No.62176097, and the Hubei Provincial Natural Science Foundation of China No.2022CFA055.

References

[1] 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.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021.

[5] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *CVPR*, 2020.

[6] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *CVPR*, pages 12615–12624, 2020.

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016.

[8] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, pages 1879–1887. IEEE, 2018.

[9] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8351–8361, 2019.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[11] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[12] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *CVPR*, 2023.

[13] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1979–1986, 2014.

[14] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[17] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[18] 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.

[19] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *TCSVT*, 32(10):6994–7004, 2022.

[20] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7298–7308, 2022.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[22] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACMMM*, pages 789–792, 2014.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

[24] 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.

[25] W. Dong, Z. Zhang, C. Song, and T. Tan. Bi-directional interaction network for person search. In *CVPR*, 2020.

[26] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. In *CVPR*, pages 2585–2594, 2020.

[27] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[28] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2009.

[29] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, pages 14750–14759, 2021.

[30] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *CVPR*, pages 2476–2486, 2022.

[31] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021.

[32] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, pages 7450–7459, 2019.

[33] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, pages 6757–6765, 2017.

[34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NIPS*, 33:21271–21284, 2020.

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[38] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, October 2021.

[39] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[40] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICLR*, pages 4904–4916, 2021.

[41] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021.

[42] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2023.

[43] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[44] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115. IEEE, 2015.

[45] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *TIP*, 28(4):1575–1590, 2018.

[46] He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *ACMMM*, 2021.

[47] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017.

[48] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[50] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 44(6):3260–3271, 2022.

[51] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. In *AAAI*, volume 33, pages 8634–8641, 2019.

[52] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017.

[53] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *ECCV*, pages 562–579. Springer, 2022.

[54] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *AAAI*, volume 35, pages 2011–2019, 2021.

[55] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *TPAMI*, 41(4):871–885, 2018.

[56] Shengcai Liao and Ling Shao. Interpretable and Generalizable Person Re-Identification with Query-Adaptive Convolution and Temporal Lifting. In *ECCV*, 2020.

[57] Shengcai Liao and Ling Shao. TransMatcher: Deep Image Matching Through Transformers for Generalizable Person Re-identification. 2021.

[58] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[59] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. In *ICCV*, pages 493–501, 2017.

[60] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. Cdgnet: Class distribution guided network for human parsing. In *CVPR*, pages 4473–4482, June 2022.

[61] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018.

[62] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, pages 1–9, 2017.

[63] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[64] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *ECCV*, pages 418–434, 2018.

[65] Zhou Luowei, Palangi Hamid, Zhang Lei, Hu Houdong, Jason J. Corso, and Gao Jianfeng. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

[66] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[67] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *CVPR*, pages 811–820, 2019.

[68] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *TIP*, 29:5542–5556, 2020.

[69] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *AAAI*, volume 34, pages 11839–11847, 2020.

[70] Jie Qin, Peng Zheng, Yichao Yan, Quan Rong, Xiaogang Cheng, and Bingbing Ni. Movienet-ps: A large-scale person search dataset in the wild. In *ICASSP*, 2023.

[71] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yi Xu, Xiang Wang, Changxin Gao, Rong Jin, and Nong Sang. Self-supervised learning from untrimmed videos via hierarchical consistency. *TPAMI*, 2023.

[72] 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 *ICML*, pages 8748–8763, 2021.

[73] 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.

[74] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015.

[76] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI*, volume 33, pages 4814–4821, 2019.

[77] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, pages 5814–5824, 2019.

[78] 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 *ACMMM*, pages 5566–5574, 2022.

[79] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, pages 486–504, 2018.

[80] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2401.06825*, 2024.

[81] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, pages 11218–11228, 2023.

[82] 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. *arXiv preprint arXiv:2208.08608*, 2022.

[83] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, volume 32, 2018.

[84] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, pages 402–419, 2018.

[85] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightning-dot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, 2021.

[86] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[87] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. In *CVPR*, pages 11952–11961, 2020.

[88] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. *TNNLS*, 32(5):2142–2156, 2020.

[89] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, pages 274–282, 2018.

[90] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, pages 531–540, 2017.

[91] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *CVPR*, pages 5703–5713, 2019.

[92] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing Real-World Source Training Data: Random 3D Characters for Generalizable Person Re-Identification. In *ACMMM*, 2020.

[93] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *ECCV*, pages 402–420, 2020.

[94] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[95] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *ICCV*, pages 1624–1633, 2021.

[96] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, pages 6769–6778, 2017.

[97] Suncheng Xiang, Zirui Zhang, Mengyuan Guan, Hao Chen, Binjie Yan, Ting Liu, and Yuzhuo Fu. Vtbr: Semantic-based pretraining for person re-identification. *arXiv*, pages arXiv–2110, 2021.

[98] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[99] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, July 2017.

[100] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022.

[101] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022.

[102] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057. PMLR, 2015.

[103] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification, 2023.

[104] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *CVPR*, pages 2153–2162, 2019.

[105] Xuezhi Liang Yanan Wang and Shengcai Liao. Cloning Outfits from Real-World Images to 3D Characters for Generalizable Person Re-Identification. *CVPR*, 2022.

[106] 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. In *ACMMM*, 2023.

[107] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In *CVPR*, pages 14278–14287, 2022.

[108] Ye Yuan, Wuyang Chen, Yang Yang, and Zhangyang Wang. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *CVPR*, pages 354–355, 2020.

[109] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *CVPR*, 2021.

[110] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *CVPR*, pages 7369–7378, 2022.

[111] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021.

[112] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701, 2018.

[113] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan. Recurrent attention model for pedestrian attribute recognition. In *AAAI*, volume 33, pages 9275–9282, 2019.

[114] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[115] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017.

[116] Ruochen Zheng, Changxin Gao, and Nong Sang. Viewpoint transform matching model for person re-identification. *Neurocomputing*, 433:19–27, 2021.

[117] Ruochen Zheng, Lerenhan Li, Chuchu Han, Changxin Gao, and Nong Sang. Camera style and identity disentangling network for person re-identification. In *BMVC*, page 66, 2019.

[118] 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.

[119] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017.

[120] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782, 2017.

[121] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017.

[122] 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 *ACMMM*, pages 209–217, 2021.

[123] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*. Springer, 2020.

[124] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. *arXiv preprint arXiv:2203.03931*, 2022.

[125] Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. *arXiv preprint arXiv:2409.18569*, 2024.

[126] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *CVPR*, pages 22010–22019, 2024.

A Appendix

A.1 Related Work

General Representation Learning. To avoid the labor-intensive manual annotation process, many general representation learning approaches have been explored to utilize unlabeled image data. Several representative works [36, 101, 10, 34, 16, 2, 35, 4, 14] have achieved performance comparable to, or even surpassing, supervised methods. For example, CAE [14] presents a novel masked image modeling approach and shows the benefit to representation learning through an encoder-regressor-decoder architecture. Besides to pre-training on unlabeled image data, there is currently a popular trend of attempting to establish the relationship between vision and language to learn more discriminative representations. For example, CLIP [72] and ALIGN [40] perform cross-modal contrastive learning on hundreds or thousands of millions of image-text pairs crawled from the web and can directly perform open-vocabulary image classification. BLIP [49] shows that the noisy web texts are suboptimal for vision-language learning, and proposes a Captioning and Filtering method to improve the quality of the text corpus.

Person Representation Learning. In person related field, it is common practice to leverage the backbones pre-trained on ImageNet [21], which ignores domain gap between general images and person-related images, and leads to limited performance. To address such a problem, LUP [29] constructs a large-scale unlabeled person dataset and makes the first attempt of performing a general unsupervised pre-training method MoCov2 [15] for learning person representations. The good experimental results on the person Re-ID task verified the effectiveness of it.

However, simply adopting the general pre-training method may result in the ignorance of person fine-grained characteristics. Therefore, considering the particularity of Re-ID tasks, the proposed UP-ReID [107] introduces an intra-identity regularization and is the first attempt toward a Re-ID specific pre-training framework by explicitly pinpointing the difference between the general pre-training and Re-ID pre-training. Before long, LUP-NL [30] develop a large-scale pre-training framework utilizing noisy labels, and demonstrate that learning directly from raw videos is a promising alternative for pre-training, which utilizes spatial and temporal correlations as weak supervision. This simple pre-training task provides a scalable way to learn good Re-ID representations from scratch without bells and whistles. Also, PASS [124] is proposed to use several learnable tokens to extract part-level features offering fine-grained information. It helps the ViTs set good performance on several person Re-ID datasets. More recently, CION [125] is proposed to learn identity-invariant representations across large-scale different videos for person Re-ID pre-training.

However, these approaches are only target at learning person representations especially for promoting the person Re-ID performance. They have poor generalization ability and perform poorly on other person-centric tasks. To learn a general human representation, SOLIDER [12] is proposed recently by introducing prior knowledge and more semantic information.

In general, these pure-vision based works are rather aimed at only promoting the person Re-ID performance, or limited to the person-centric visual tasks. Their performance in learning generic person representation is still unsatisfactory and lacks the ability of multi-modal understanding.

Person-centric Tasks. In computer vision community, there are many tasks directly or indirectly related to person, *i.e.,* image/text based person Re-ID, person attribute recognition, person pose estimation, person search, multi-object tracking and human parsing. We name such tasks as person-centric tasks and sort out five representative tasks among them for study.

1) *Text-based person Re-ID* aims to search for person images of a specific identity by textual descriptions. Existing works can be divided into attention-based and attention-free methods. The former [8, 52, 68, 24, 78] attempts to establish region-text correspondences but ignores the efficiency. To better align the multi-modal features, the latter usually focuses on designing various objective functions [112, 27, 77] and model structures [93, 118].

2) *Image-based person Re-ID* aims to search for person images by given person images. Most works are based on supervised learning. The hard triplet loss [11, 39, 108] and classification loss [79, 119] are introduced to learn a global feature. Also, some works [84, 86] focus on learning a part-based feature instead. For example, Sun *et al.* [86] proposed to represent features as horizontal stripes and learn with separate classification losses. Some works utilize camera style [117] and viewpoints [116] to train a more robust model. Some works [81, 80] focus on semi-supervised or unsupervised setting.

3) *Person attribute recognition* aims to identify the person's attributes. Many methods [44, 62, 61] treat this task as a multi-label classification problem, while some others [90, 51, 113] adopt recurrent neural networks for exploring the attribute context. Also, some works [19, 53] introduce an additional language modality to get better performance.

4) *Person search* aims to jointly localize and identify a query person from natural and uncropped images. Existing works can be generally categrized into two-step and one-step ones. In two-step works [26, 87], persons are detected and then fed into a Re-ID model for identification, while one-step approaches [54, 70, 6] makes the joint framework more effective and efficient.

5) *Human parsing* is a fine-grained semantic segmentation task in human analysis. Some early works [33, 96] combine human parsing with pose estimation. Moreover, some works [47, 32] study the human parsing task in a multi-person scenario, which needs to distinguish the human instances. Meanwhile, the work [50] proposes a self-correction mechanism and leads to better performance.

A.2 Broader Impact

This paper proposes a language-image pre-training framework for person representation learning, and contributes a large-scale synthetic person-centric dataset with rich image-text pairs. Our pre-trained models and dataset can help diverse person-related applications such as person re-identification, person attribute recognition and human parsing, thus boosting the development of smart retail, smart transportation, smart security systems and so on in the future metropolises.

Nevertheless, the application of person Re-ID and attribute recognition, such as for recognizing and tracking pedestrians in surveillance systems, might raise privacy concerns. It typically depends on the utilization of surveillance data for training without the explicit consent of the individuals being recorded. Therefore, governments and officials need to carefully establish strict regulations and laws to control the usage of these technologies. Otherwise, these technologies can potentially equip malicious actors with the ability to surveil pedestrians through multiple cameras without their consent. Furthermore, we should be cautious of the misidentification of the Re-ID systems to avoid possible disturbance. Also, note that the demographic makeup of the datasets used is not representative of the broader population.

At the same time, we have utilized a substantial amount of person-containing video data from the internet for pre-training purposes. Consequently, it is inevitable that the resulting models may inherently contain information about these persons. Researchers should adhere to relevant laws and regulations, and strive to avoid using our models for any improper invasion of privacy. We will have a gated release of our models and training data to avoid any misuse. We will require that users adhere to usage guidelines and restrictions to access our models and training data. Meanwhile, all our open-sourced assets can only be used for research purpose and are forbidden for any commercial use.

A.3 Limitations

PLIP presents a preliminary attempt to introduce language modality into generic person representation learning. Despite its effectiveness on existing public datasets, PLIP may still be difficult to learn good fine-grained person representations for it does not explicitly achieve local information correlation across different modalities. Also, its pretext tasks require multiple forward propagation, directly increasing the memory overhead and affecting computational efficiency. Meanwhile, as we have followed the conventional practice of previous works by utilizing a large amount of person-containing internet video data to implement pre-training, there is a potential for privacy and security issues to some extent. Therefore, in our subsequent work, we will focus on addressing fine-grained issues and improving the efficiency of our framework. We will take every possible measure to prevent the misuse of our models and dataset as well.

A.4 Altering Color Word Affects Image Colorization: Visualization

As displayed in Fig 5, in our pretext task of text-guided image colorization, altering the color word in textual description significantly affects the colorization of image. However, our model may not fully understand the semantics of more detailed image regions. As shown in the last row, our model fails to distinguish between the blue clothing region and the red shoulder strap region (marked with yellow boxes), instead blending the two into a unified coloration. This is due to the fact that the level

of detail in manually annotated datasets is still not sufficient, resulting in the model being unable to theoretically learn representations with higher levels of detail and greater discrimination capabilities. However, it is undeniable that our model has a preliminary understanding of the meaning of attributes and colors, and can associate them with related image regions, rather than simple memorization. This ability to distinguish between different parts of the person body guarantees the superior performance on the subsequent person-centric tasks.

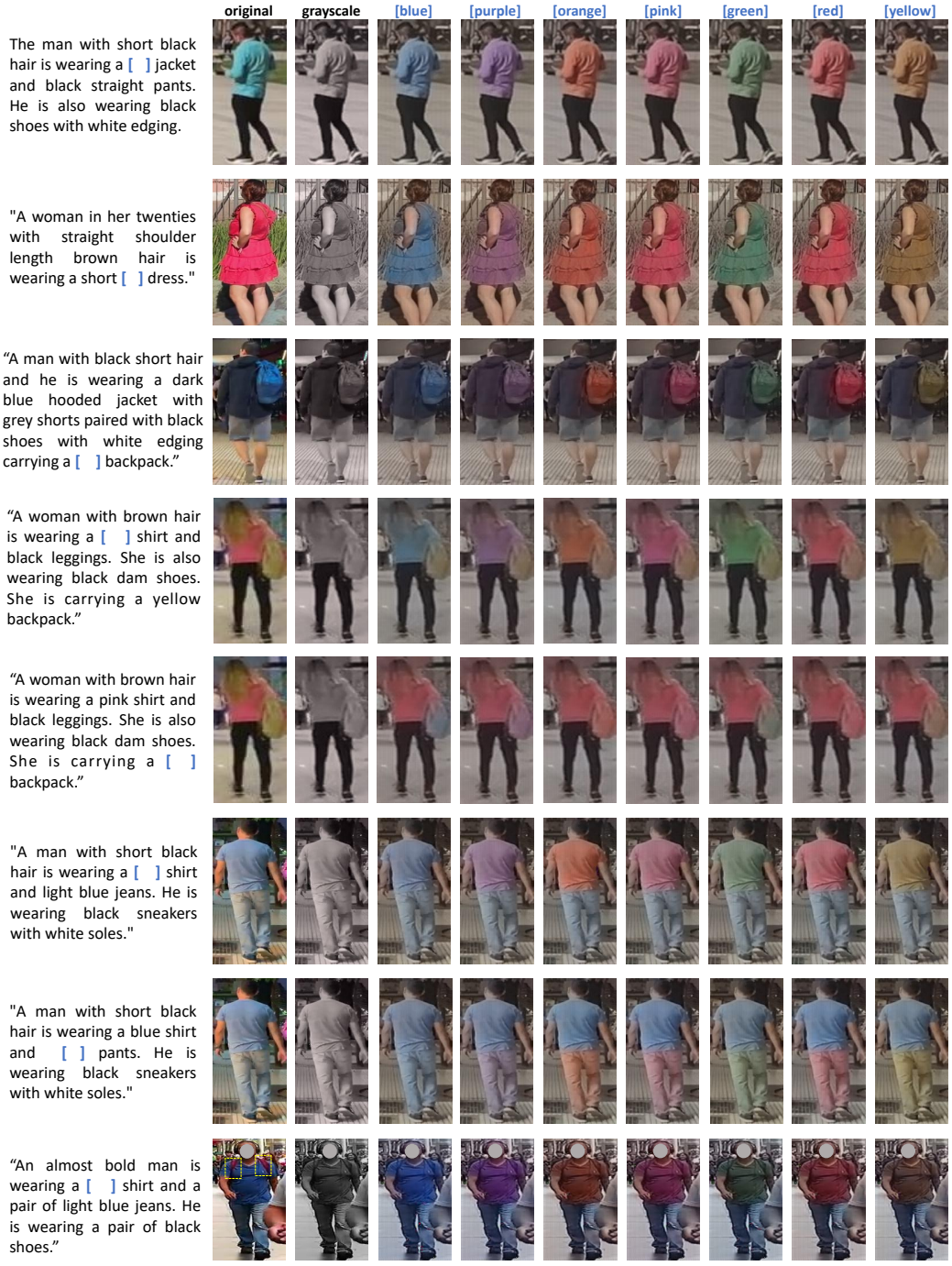

Figure 5: Visualization of gray-scale person image colorization results by changing the color words in textual descriptions.

A.5 How to Construct SYNTH-PEDES and its Properties

We build the SYNTH-PEDES dataset to pre-train the PLIP models at a large-scale. In this section, we show the details of how to construct our SYNTH-PEDES dataset and its characteristic properties.

A.5.1 Dataset Construction

The dataset construction process can be described as three steps. Firstly, we collect several person datasets [30, 83] to form the large-scale image dataset. Secondly, as a person image captioner, a simple but effective method is proposed to achieve automatic diversified text annotation with high-quality. Finally, we adopt some post-processing approaches to eliminate the noises and improve the dataset quality.

Image Collection.

We collect and process two large-scale person datasets to form the image dataset. The first is LUPerson-NL [30]. It is a new variant of LUPerson [29] on top of raw videos from LUPerson and assign the noisy labels to each person image with automatically generated tracklet. It consists of 10M images with about 430K identities collected from 21K scenes. The second is LPW [83]. It consists of 2,731 different persons and 592,438 images collected from three different crowded scenes.

Image Captioner (SPAC).

Given an input person image, there is no specific work targeting at generating captions that detailedly describe the person's appearance. To this end, we propose a simple but effective method for person image captioning. It can generate attribute-annotations and stylish textual-descriptions, which simulate the diverse perspectives that different annotators may have on the same person image.

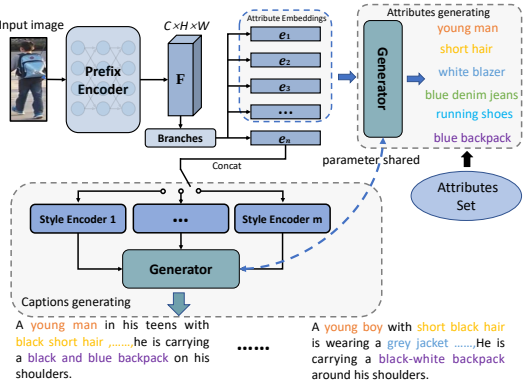

Figure 6: Architecture of Stylish Pedestrian Attributes-union Captioning. It mainly consists of a prefix encoder and a generator. We use it to generate descriptions from a person image.

As illustrated in Fig. 6, Stylish Pedestrian Attributes-union Captioning (SPAC) mainly comprises two modules, *i.e.,* a prefix encoder and a shared generator. Specifically, we use ResNet101-FPN [37, 58] as the encoder and GPT2 [73] as the generator to capture rich person image details and generate high-quality texts. In the existing image-text person datasets [52, 24], many unique workers were involved in the labeling tasks. The same images usually have inconsistent-style language descriptions. Constructing image-text pairs for an image with multiple descriptions will lead to unstable training and affect the model performance. Thus, in order to replicate the stylized variations that arise from multiple workers' labels and mitigate the issue of redundant real labels, we have incorporated style encoders into our pipeline. We let the concatenated prefixes pass through different style encoders to get the prefixes of their own style and then send them to the generator for the subsequent generation.

The entire training process can be seen as an autoregressive problem. Given a dataset of paired images, attributes and captions $\left\{\mathbf{x}^i, \mathbf{A}^i, \mathbf{y}^i\right\}_{i=1}^{N}$, where $\mathbf{A}^i$ is an attribute set of image $\mathbf{x}^i$ containing six attribute descriptions, the learning goal is to generate meaningful attribute descriptions and captions from an unseen person image. The attributes and captions can be referred as a sequence of padded tokens $\mathbf{A}^i = \left\{\mathbf{a}_1^{i,k}, \ldots, \mathbf{a}_{\ell_1}^{i,k}\right\}_{k=1}^{n-1}, \mathbf{y}^i = \mathbf{y}_1{}^i, \ldots, \mathbf{y}_{\ell_2}{}^i$, with maximal lengths ℓ_1, ℓ_2 accordingly.

Following recent works [66, 65], our key solution is to jointly train a prefix encoder and a generator. The former is to capture the semantic embeddings as the prefixes from the image, and the latter, as an autoregressive language model, is to use the prefixes to predict the next token one by one. As shown in Fig. 6, we first feed the input image $\mathbf{x}^i$ into the prefix encoder PE and different branches $\{BR_k\}_{k=1}^{n}$ to get the n attribute-and-relation prefix embeddings:

$$\mathbf{e}_1^i, \ldots, \mathbf{e}_n^i = \left\{BR_k(PE(\mathbf{x}^i))\right\}_{k=1}^n . \tag{8}$$

And then the concatenated embeddings are fed into different style encoders $\{SE_k\}_{k=1}^m$ to get the stylized caption embeddings:

$$\mathbf{c}_1^i, \ldots, \mathbf{c}_m^i = \left\{SE_k(concat([\mathbf{e}_1^i, \ldots, \mathbf{e}_n^i]))\right\}_{k=1}^m , \tag{9}$$

each embedding has the same dimension as a token embedding. We then concatenate the obtained embeddings to the atrribute and caption token embeddings, where $\mathbf{e}^i$ is selected from the stylized caption embeddings in turn:

$$\begin{aligned} \{\mathbf{Z}_k^i &= concat([\mathbf{e}_k^i, \mathbf{a}_1^{i,k}, \ldots, \mathbf{a}_{\ell_1}^{i,k}])\}_{k=1}^{n-1}, \\ \mathbf{Z}_n^i &= concat([\mathbf{e}^i, \mathbf{y}_1{}^i, \ldots, \mathbf{y}_{\ell_2}{}^i]. \end{aligned} \tag{10}$$

Finally, we feed the embeddings $\{\mathbf{Z}^i\}_{i=1}^N$ into the shared generator G to predict the attribute and caption tokens in an autoregressive fashion, using the cross-entropy loss:

$$\mathcal{L}_c = \sum_{i=1}^{N} \sum_{j=1}^{\ell_2} \log G\left(\mathbf{y}_j^i \mid \mathbf{e}^i, \mathbf{y}_1^i, \ldots, \mathbf{y}_{j-1}^i\right), \tag{11}$$

$$\mathcal{L}_a = \{\sum_{i=1}^{N} \sum_{j=1}^{\ell_1} \log G\left(\mathbf{a}_j^{i,k} \mid \mathbf{e}_k^i, \mathbf{a}_1^{i,k}, \ldots, \mathbf{a}_{j-1}^{i,k}\right)\}_{k=1}^{n-1}. \tag{12}$$

Define $\lambda \in \mathbb{R}^+$ as a balance factor, then the overall loss $\mathcal{L}_{spac}$ is computed as:

$$\mathcal{L}_{spac} = -\mathcal{L}_c - \lambda \mathcal{L}_a. \tag{13}$$

Post-Processing.

Noise Filter Strategy. To filter the noises in LUPerson-NL, we propose Seed Filter Strategy. Specifically, it mainly includes three processes. 1) *Filter-out*. For all samples with the same identity, we cycle to calculate the similarity between the current sample and the center of other samples and exclude the samples whose similarity does not meet the threshold until the similarity of all samples with the same identity meets it. 2) *Reassignment*. According to the similarity and threshold, we reassign the excluded samples in excluded dataset to the correct identity within a certain identity continual range. 3) *Merger*. We merge the samples that should belong to the same identity but are divided into different identities. Through this process, the samples with incorrect identity label annotations can be well filtered out or included in their expected identity set. More details can be found in Sec. A.7

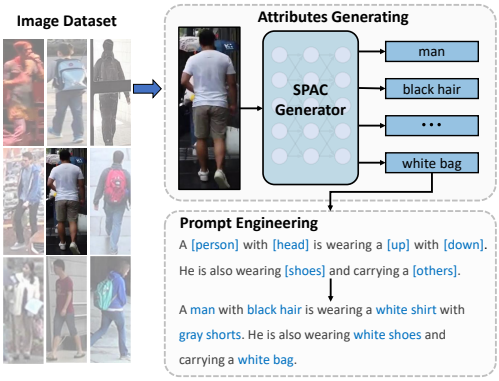

Figure 7: Visual example of Attributes Prompt Engineering. Given an image, we generate the attribute annotations based on SPAC and embed them into the masked sentence randomly chosen from the sentence library to form a complete caption.

Data Distribution Strategy. There are some samples with poor imaging conditions in the image part. Also, the number of images per identity in LUPerson-NL is very unbalanced. To this end, we have adopted some strategies to ensure the quality of generated attributes, the consistency of gender annotation, and identity distribution balance. For example, we adopt a gender voting mechanism to automatically synchronize the gender annotation of an identity in dispute. For the text part, we aim to generate three sentences for an image. Two of them are directly generated by SPAC while the another is generated by Attributes Prompt Engineering exploiting the attribute annotations generated by SPAC, as shown in Fig. 7. There are averagely 2.53 sentence per image. More details can be found in Sec. A.7

Table 11: Statistics comparison on existing popular datasets. SYNTH-PEDES is by far the largest person dataset with textual descriptions without any human annotation effort.

Datasets	year	#images	#identities	#descriptions	view	label type	label method	crop size
Market1501 [114]	2015	32,668	1,501	-	fix camara	identity	hand+DPM [28]	128×64
DukeMTMC [120]	2017	36,411	1,852	-	fix camera	identity	hand	vary
CUHK-PEDES [52]	2017	40,206	13,003	80,412	fix camera	identity+description	hand	vary
LPW [83]	2018	592,438	2,731	-	fix camera	identity	hand+Detector+NN	vary
MSMT17 [94]	2018	126,441	4,101	-	fix camera	identity	FasterRCNN [75]	vary
SYSU30K [88]	2020	29,606,918	30,508	-	dymamic	identity	YOLOv2 [74]	vary
RSTPReid [122]	2021	20,505	4,101	41,010	fix camera	identity+description	hand	vary
ICFG-PEDES [24]	2021	54,522	4,102	54,522	fix camera	identity+description	hand	vary
LUPerson [29]	2021	4,180,243	$> 200k$	-	dymamic	no	YOLOv5	vary
LUPerson-NL [30]	2022	10,683,716	433,997	-	dymamic	identity	FairMOT [111]	vary
MALS [106]	2023	1,510,330	-	1,510,330	dynamic	description	ImaginAIry	vary
UFine6926 [126]	2024	26,206	6,926	52,412	dynamic	identity+description	hand	vary
SYNTH-PEDES	2024	4,791,771	312,321	12,138,157	dymamic	identity+description	SPAC	vary

A.5.2 Dataset Properties

Thanks to the outstanding generating ability of our proposed SPAC, the SYNTH-PEDES dataset is full of high-quality textual descriptions in a variety of styles, which can be utilized to train the representation learning model. Compared with existing person datasets in Tab. 11, SYNTH-PEDES has the following advantages:

Diversified. Our dataset contains a wide range of variations in the textual descriptions. Unlike the previous person datasets with only one or two image-text pairs, most images of our dataset are annotated with three textual descriptions.

High-quality. As some typical qualitative examples can be seen in Fig. 3, the generated annotations achieve an accurate and detailed description of the person appearance. The further experiments conducted on the quality research can be found in the following sections. Researchers can use this dataset with confidence to conduct relevant studies.

Large-scale. In Tab. 11, we compare the properties of SYNTH-PEDES with other popular person datasets. As we can see, SYNTH-PEDES is the largest real person dataset with high-quality image-text pairs by far, which contains 312,321 identities, 4,791,711 images, and 12,138,157 textual descriptions.

A.6 Details of Network Structures

In this section, we will provide a detailed explanation of the structure and some parameter settings of our representation learning network, mainly including some modifications made to the ResNet networks [37] and the decoder designs for the two pre-text tasks. The dimension of global embeddings is set to 768.

A.6.1 Modifications to ResNets

We made some modifications to the ResNet networks like CLIP [72]. Firstly, there are now 3 "stem" convolutions as opposed to 1, with an average pool instead of a max pool. For each "stem" convolutions, it is consisted of a convolution layer, a batch normalization layer and a ReLU activation funtion. Secondly, we perform anti-aliasing strided convolutions, where an avgpool is prepended to convolutions with stride > 1. Thirdly, the final pooling layer is a max pool instead of an average pool and we add a trainable linear layer to compress feature dimensions to 768.

A.6.2 Specifications of TIC Decoder

For an image with $3 \times H \times W$, we firstly obtain the multi-scale visual feature map with $1024 \times \frac{H}{4} \times \frac{W}{4}$ like FPN. Then, we send the feature map and textual global embedding to 2 Multimodal SE-Block and deconvolution layers to acquire the final feature map with $3 \times H \times W$, which is utilized to restore the color information. Next, we specify the detailed designs of the Multimodal SE-Block.

In each MSE-Block, for a visual feature map with $C \times H \times W$, we firstly perform an average pooling operation on it to obtain a visual feature embedding with $1 \times C$. Then we concat the visual feature embedding and the input texutal global embedding with 1×768 to obtain a fused feature embedding with $1 \times (C + 768)$. Then, we feed it to a FC layer, which contains a linear layer$((C + 768), C)$, a

ReLU activation, a linear layer(C, C) and a sigmoid funtion, to obtain the fused attention embedding with $1 \times C$. Finally, we apply it to the channel of the original feature map to obtain the multi-modal fused feature with $C \times H \times W$.

After each MSE-Block, there is a deconvolution layer operated on the output feature map. For the first block, the deconv is with input channel=1024, output channel=256, kernel size=3, stride=2, padding=1 and output padding=1. For the second block, the deconv is with input channel=256 output channel=3, and others the same.

After the approaches above, we finally obtain an output feature map with the same shape like the input original color image, and it can be used to predict the ground truth.

A.6.3 Specifications of IAP Decoder

For each masked token, we obtain its hidden state output as the masked embedding. We then simply concat the masked embedding and the visual global embedding to obtain the fused embedding. Then the fused embedding will be regarded as the QKV and set to a normal transformer block to improve the multimodal fusion. Finally, the output fused embedding of transformer block will be sent to a prediction head to predict the masked word. The prediction head is a simple linear layer with $(768, nvocab)$, where the $nvocab$ means the size of the dictionary.

A.7 Details of Dataset Construction

In this section, we specify the details of our gathering SYNTH-PEDES dataset. We mainly provide a detailed explanation of the designs of our texutal description generated model SPAC. Also, we demonstrate the noise filter strategy and data distribution strategy adopted in the process of dataset constrution.

A.7.1 Designs of SPAC Model

The prefix encoder is a ResNet101-FPN [58] pre-trained on ImageNet [21]. There are two types of prefix branches, one is the attribute branch and the other is the sentence branch. The detailed structures of each branch are shown in Tab. 12. Prefix dimension is set to 768. Attribute prefix length and sentence prefix length are 3 and 5, respectively.

There are totally 6 types of attribute prefix. For each type of attribute prefix, we send it to the generator GPT2 [73] and generate the attribute annotation accordingly. Also, we concat all the attribute prefixes and sentence prefix, and send them to 2 types of style encoders for obtaining the stylized prefixes. Then the stylized prefixes will be used to generate complete textual descriptions of 2 different styles accordingly. The two style encoders are actually two linear transformation matrices with different weights.

A.7.2 Details of Noise Filter Strategy

To filter various noises in the LUPerson-NL [30] dataset, we propose Seed Filter Strategy, which is illustrated in Algorithm 1. Specifically, it consists of three processes, each with its corresponding similarity threshold for filtering, reassigning or merging. 1) *Filter-out*. For all samples with same identity, we cycle to calculate the similarity between the current sample and the center of other samples and exclude the samples whose similarity does not meet the filtering threshold σ_s until the similarity of all samples with same identity meets it. 2) *Reassignment*. According to the similarity and reassigning threshold σ_r, we reassign the excluded samples in the excluded dataset to the correct identity within a certain identity continual range. 3) *Merger*. We merge the samples that should belong to the same identity but are divided into different identities according to the merging threshold σ_m. Through these processes, the samples with incorrect identity label annotations can be well filtered out or included in their expected identity set.

The following is a detailed explanation of the specific symbols in the Algorithm 1. S_{in} is the input dataset to be denoised while S_{out} is the output denoised dataset. $S_{exclude}$ is the set of image samples to be excluded. S_{merge} is the set of identities to be merged and it contains a lot of merged-identity pairs. ID_k is the $k.th$ identity of the input dataset, with n image samples x_i^k. FID_k is the output final identity. Sim is a similarity calculate function. If the variable it contained is a single identity, it will calculate the similarity between each sample and other samples' center. If it contains a image

Table 12: Structure of the attribute and sentence branch. The bias of all linear layers is set to true. n represents batch size. c represents the channel size. h and w represent the height and width of the output feature map.

Layers (CNN-MLP)	Paramters (kernel,stride,pad)	Output Size (n,c,h,w)
Conv layer	(3,2,1)	(n,c,h/2,w/2)
Avgpool	-	(n,c)
Linear	-	(n,768)
Dropout	rate=0.25	(n,768)
Linear	-	(n,2304)
Dropout	rate=0.25	(n,2304)
LeakyReLU	rate=1/5.5	(n,2304)
Rearange	-	(n,3,768)
Conv layers $\times$ 2	(3,2,1)	(n,c,h/4,w/4)
Avgpool	-	(n,c)
Linear	-	(n,1536)
Dropout	rate=0.25	(n,1536)
Linear	-	(n,3840)
Dropout	rate=0.25	(n,3840)
LeakyReLU	rate=1/5.5	(n,3840)
Rearange	-	(n,5,768)

sample and a identity, it will calculate the similarity between the image sample and the center of all samples in the identity. If it contains two identities, it will calculate the similarity between the two centers of the two identities. $Merge$ is a function that merges the identities. We use cosine similarity to calculate the similarity between samples. For the hyper-parameters, σ_s is set to 0.6. σ_r is set to 0.65. σ_m is set to 0.62. r_a and r_b are both set to 2.

Algorithm 1 Seed Filter Strategy

Input $S_{in} = \{ID_k\}_{k=1}^{N}$, where $ID_k = \{x_i^k, \ldots, x_n^k\}$
for $ID_k \in S_{in}$ **do**
  repeat
    if $\min Sim(ID_k) < \sigma_s$ **then**
      $x_j^k = \arg\min Sim(ID_k)$;
      $ID_k = ID_k - x_j^k, x_j^k \Rightarrow S_{exclude}$;
    end if
  until $\min Sim(ID_k) \geq \sigma_s$
end for
for $x_j^k \in S_{exclude}$ **do**
  if $\max\{Sim(x_j^k, ID_i)\}_{i=k-r_a}^{k+r_a} \geq \sigma_r$ **then**
    $ID_r = \arg\max\{Sim(x_j^k, ID_i)\}_{i=k-r_a}^{k+r_a}$;
    $ID_r = ID_r + x_j^k$;
  end if
end for
for $ID_k \in \{ID_i\}_{i=1}^{N}$ **do**
  if $\{Sim(ID_k, ID_{k+i})\}_{i=1}^{r_b} > \sigma_m$ **then**
    $\{ID_k, ID_{k+i}\} \Rightarrow S_{merge}$
  end if
end for
for $FID_i \in S_{merge}$ **do**
  for $FID_j \in \{FID_j\}_{j=i+1}^{i+r_b}$ **do**
    if $FID_i \cap FID_j \neq \varnothing$ **then**
      $FID_i = Merge(FID_i, FID_j)$
    end if
  end for
  $FID_i \Rightarrow S_{out}$
end for
Output S_{out};

A.7.3 Details of Data Distribution Strategy

For the image part, there are some bad pictures with blocking, blurring and multiple people. Due to the poor quality, these pictures are not a good choice for training. Also, the number of images per identity in LUPerson-NL [30] is very unbalanced. Some identities have thousands of images, while others have only few images. To further improve the quality of our dataset, we have adopted the following strategies: 1) *Ensure the qualities of generated attributes.* In our setting, the attributes generated by poor quality pictures will contain a large number of unknowns. To filter these poor quality pictures, the pictures with three or more unknown attributes will be directly filtered. Particularly, if the upbody and downbody attributes of a picture are unknown simultaneously, it will also be filtered. 2) *Ensure the consistency of gender attribute.* Ensuring the gender consistency of an id plays a role in promoting the stability of the model training. However, for those pictures with poor quality, it is often very challenging to correctly predict their gender attributes. So we adopt a gender voting mechanism. That is, for all pictures in an identity, if the gender attribute with the maximum frequency of occurrence is greater than 0.7, the gender of all pictures in the identity will be automatically changed to this gender. Otherwise, the pictures in the identity will be considered as bad pictures, and will be filtered out. 3) *Balance the distribution of identities.* If the number of images of an identity is less than the minimum number of 5, this identity will be filtered. If the number of images of an identity is greater than the maximum number of 20, randomly select 20 images from all images of the identity and filter out the others.

For the text part, we aim to generate three sentences for an image. Two of them are directly generated by SPAC while the another is generated by Attributes Prompt Engineering exploiting the attribute descriptions generated by SPAC. In many processes of our approach, we must extract the attribute phrases from a textual description. For example, in order to guild the SPAC to generate attribute descriptions, we have to extract the attribute labels from the paired textual descriptions. The extracting process is as the following: 1) *Construct the keywords set for each attribute.* Firstly, we extract all the nouns in the datasets' [52, 24] captions by nltk tools and sort them by frequency. Then, for each attribute, we manually select from the nouns and construct the keywords set. 2) *Assign the noun phrases to relevant attributes.* For a textual description, we firstly extract the noun phrases by nltk tool. Then, according to the keywords set, we assign each noun phrase to the correct attribute. On this way, we finally finish extracting the attribute phrases from a a textual description. For Attributes Prompt Engineering, as shown in Fig. 8, we have created a mask standard sentence library with many diversities. To construct the library, by exploiting the nltk tool and the above attributes extracting method, we first mask all attributes phrases of the captions in CUHK-PEDES and ICFG-PEDES. Then We manually select sentences with good sentence structure from them to form the mask standard sentence library. The library totally consists of 1664 sentences for different generated attribute missing situations.

Female:	1. The [person] has [head]. She is wearing a [up], [down], and [shoes]. She is carrying a [others]. 2. The [person] with [head] is wearing a [up] and [down]. She is also wearing [shoes] and carrying a [others]. 3. A [person] with [head] and she is wearing a [up] with [down] paired with [shoes] carrying a [others].
Male:	1. The [person] is wearing a [head] with a [up] and he is also wearing [down] paired with [shoes], carrying a [others]. 2. A [person] with [head] is wearing a [up] and [down]. He is also wearing [shoes]. He is carrying a [others]. 3. A [person] with [head] is wearing a [up] and a pair of [down]. He is also wearing [shoes] and carrying a [others].
Person:	1. An individual with [head] and is wearing a [up] with [down] paired with [shoes] carrying a [others]. 2. A [person] with [head] and is wearing a [up] with [down] paired with [shoes] carrying a [others]. 3. A [person] has [head], and wears a [up] and a pair of [down] and a pair of [shoes], is holding a [others].

Figure 8: Some examples in the mask standard sentence library for the attributes-all-known situation.

A.8 Experimental Setup

A.8.1 Implementation Details

During the training of SPAC, we adopt ResNet-101 [37] together with a Feature Pyramid Network (FPN) [58] as the prefix encoder and GPT2 [73] as the generator. The ResNet-101 and GPT2 are both pre-trained on their respective pre-training tasks. All images are resized to 384×128 and normalized with mean and std of [0.485, 0.456, 0.406], [0.229, 0.224, 0.225], which are calculated from all images in ImageNet. We use the combination of CUHK-PEDES [52] and ICFG-PEDES [24] as the training dataset. We adopt horizontally flipping to augment data, where each image has 50% probability to flip randomly. We aim to generate the textual descriptions with two style type. The learning rate is fixed at 0.0001. The balance factor λ is set as 0.15. There are six types of attributes

including gender, head, upper body, lower body, shoes and belongings. The prefix lengths of attributes and relation are 3 and 5 respectively. We train it on 4 × Geforce 3090 GPUs for 30 epochs with a batch size of 64 totally, which takes approximately 1.6 days. The optimizer is AdamW [63] with the default setting.

During the training of PLIP, we adopt four types of backbone as the visual encoder, *i.e.,* ResNet50, ResNet101, ResNet152 and Swin Transformer Base. The pre-trained BERT [23] is utilized as the textual encoder and we only unfreeze the last 5 layers, keeping other parameters frozen. All images are resized to 256 × 128 and normalized with mean and std of [0.357, 0.323, 0.328], [0.252, 0.242, 0.239], which are calculated from our proposed SYNTH-PEDES. We adopt horizontally flipping to augment data, where each image has 50% probability to flip randomly. For ResNet50, we train our model on 4 × Geforce 3090 GPUs for 70 epochs with a batch size of 512 totally, which takes approximately 15.2 days. The base learning rate is set to 0.002 and decreased by 0.1 at the epoch of 30 and 50. Besides, the learning rate warm-up strategy is adopted in the first 10 epochs. The learning rate of BERT has a 0.1 decay. For other types of visual encoder, there are some differences on the learning rate and batch size setting. The hyper-parameters in the objective function are set to $\lambda_1 = 0.02$ and $\lambda_2 = 0.1$. The optimizer is Adan [100] with the default setting. We adopt the mixed precision training mode by Apex.

For each person-centric downstream task, we reproduce a range of state-of-the art methods as the baselines. If no special instructions are given, we perform the experiments by just replacing the backbone in each baseline to the models pre-trained by our proposed method. Meanwhile, for text-based person Re-ID, we adopt the standard metrics Rank-*k* (*k*=1,5,10) to evaluate the model performance. For image-based person Re-ID and person search, we follow the popular evaluation metrics: the mean Average Precision (mAP) and the Rank-*k* (*k*=1,5,10). For person attribute recognition, we adopt four evaluation metrics including accuracy (Acc), mean accuracy (mA), recall (Rec) and F1 value (F1). For human parsing, we use mean accuracy (mA) and mean intersection over union (mIoU) as the evaluation metrics.

A.8.2 Benchmarks

To evaluate our proposed approach, we perform in-depth experiments on eleven datasets for five downstream person-centric tasks. For text-based person Re-ID, we conduct extensive experiments on two popular pubic datasets, *i.e.,* CUHK-PEDES [52] and ICFG-PEDES [24]. For image-based person Re-ID, we use three popular public datasets, *i.e.,* Market1501 [114], MSMT17 [94] and DukeMTMC [120]. For person attribute recognition, three large-scale datasets PETA [22], PA-100K [62] and RAP [45] are used. For person search, we adopt PRW [115] and CUHK-SYSU [99] in our experiments. For human parsing, the LIP dataset [33] and PASCAL-Person-Part dataset [13] are used. Next is an introduction to representative datasets in each task.

CUHK-PEDES dataset [52] is the first and most commonly used benchmark for text-based person re-identification. It contains 40,206 images and 80,412 textual descriptions for 13,003 identities. Each person image has two corresponding textual descriptions on average. We dopt the same data split as [52]. The training set has 34,054 images of 11,003 identities. The validation and test set have 3,078 and 3,074 images of 1,000 identities, respectively.

Market1501 dataset [114] includes 32,668 images of 1,501 persons captured from 6 different cameras. The DPM detector is employed to crop the bounding boxes for these pedestrians. The dataset is divided into a training set comprising 12,936 images of 751 persons and a test set with 3,368 query images and 19,732 gallery images of 750 persons.

PA-100K dataset [62] contains 100,000 pedestrian images with resolutions ranging from 50×100 to 758×454 captured from 598 different cameras. Each pedestrian image is annotated with 26 commonly used attributes. As the official protocol, there are 80,000 images for training, 10,000 images for validation, and 10,000 images for testing.

PRW dataset [115] is a widely used dataset for person search which contains 11,816 video frames captured by 6 different cameras, and 34,304 manually annotated bounding boxes. All people are divided into labeled identities and other unknown ones. The training set comprises 5,704 images and 482 identities. The test set contains 6,112 images and 2057 query people.

LIP dataset [33] is a large human parsing dataset containing 50,462 images with manual pixel-wise annotations for 19 semantic human parts. The dataset is collected from the real-world scenarios. It contains human images with various appearances, heavily occlusions and low-resolutions, which

introduces many challenges. LIP is divided into 30,462 images for train set, 10,000 images for validation set and 10,000 images for test set.

A.9 Thorough Ablation Studies and Analyses

In this section, we perform in-depth ablation studies to analyze the effectiveness of each part of our work. If not specially stated, we use ResNet-50 as the visual encoder and pre-train on the sub-dataset of SYNTH-PEDES, which has 10,000 identities, 139,564 images and 353,617 textual descriptions. The evaluation is mainly performed on the benchmarks of text-based person Re-ID and image-based person Re-ID using the zero-shot evaluation protocol.

A.9.1 Effectiveness of Each Component

Impact of each pretext task. We have designed three pretext tasks to implement pre-training. To assess the impact of pretext tasks on the generalizability of learned representations, we directly evaluate the performance of pre-trained models on the downstream datasets without fine-tuning. As indicated in Tab. 13, each task contributes to the model's zero-shot capability on CUHK-PEDES and Market1501. Combining all tasks leads to the optimal performance, highlighting the significant role each task plays in facilitating the learning of generic person representations.

Table 13: Ablation study on the impact of each pretext task, all using default settings.

#	Components			CUHK-PEDES			Market1501		
	IVLC	TIC	IAP	R@1	R@5	R@10	R@1	R@5	R@10
1	✓			30.2	53.3	64.0	62.3	79.9	85.7
2	✓	✓		31.4	55.2	65.9	62.9	80.7	86.2
3	✓		✓	30.5	54.4	64.3	62.7	80.5	86.1
4		✓	✓	-	-	-	39.3	61.4	70.4
5	✓	✓	✓	**32.5**	**56.3**	**66.6**	**63.1**	**80.8**	**86.3**

TIC is essential for transfer capability. After pre-training with different settings, we evaluate the model's performance on four datasets under the zero-shot setting. The metric on DukeMTMC [120] is rank-5, while others [52, 24, 114] are all rank-1. As shown in Fig. 9, without TIC, the learned representation exhibits noticeably weaker performance on downstream datasets, averaging a 2.45% decrease compared to the default method on CUHK-PEDES and Market1501. However, the introduction of TIC significantly improves the transfer performance. We believe this is because TIC encourages the model to comprehend textual semantics and accomplish body part localization and coloring, which holds significant implications for learning generic representations.

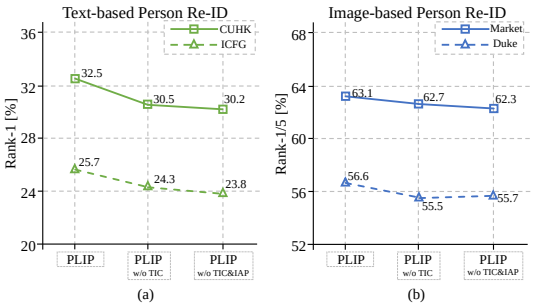

Figure 9: The effectiveness of TIC on transfer capability. TIC brings significant improvements.

IVLC contributes to stable representation learning. We utilize the IVLC loss to optimize the model at identity-level. We conduct an experiment to verify the effectiveness of IVLC in stable and meaningful representation learning. The training samples in this experiment are the first one million images from SYNTH-PEDES. We test the zero-shot performance of the trained models with different settings on the CUHK-PEDES and Market1501 datasets. As the results shown in Fig. 10, the model trained with instance-level InfoNCE loss significantly underperforms compared to the model trained with identity-level IVLC loss. Additionally, as the training process extends, the model trained with InfoNCE loss experiences a notable decline in performance, whereas the performance of the model trained with IVLC loss tends to stabilize. These results demonstrate that IVLC contributes to stable and better representation learning.

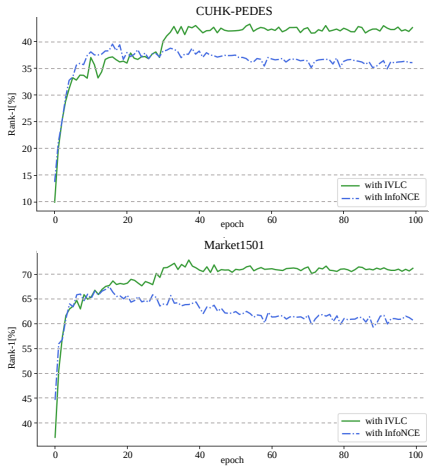

Figure 10: The effectiveness of IVLC in stable representation learning, which brings much better performance.

A.9.2 Dataset Quality Evaluation

The quality of dataset is critical for large-scale representation learning. We have conducted quantitative experiments to evaluate the quality of our dataset compared to three manual annotated datasets [52, 24, 122] and an existing synthetic dataset [97].

Manual Evaluation

We randomly select 500 image-text pairs from our SYNTH-PEDES dataset and other manually annotated or synthetic datasets, forming 2500 image-text pairs to be evaluated. The score of the evaluation is divided into five levels. The first is that the most part of the description of the image is incorrect. The second is that a small part of the description of the image is incorrect. The third is that the description is correct but rough. The fourth is that the description is correct and generally detailed. The fifth is that the description is completely correct and very detailed. As shown in Tab. 14, our SYNTH-PEDES dataset is competitive in quality compared to these manually annotated dataset at about 200 times the amount of image-text pairs.

Table 14: Manual evaluation results. Our SYNTH-PEDES dataset is competitive with manual datasets and surpasses an existing synthetic dataset by a large margin.

Dataset	Score					Average
	1	2	3	4	5	
CUHK-PEDES [52]	14	24	146	196	120	3.77
ICFG-PEDES [24]	12	44	124	150	170	3.84
RSTPReid [122]	11	26	149	146	168	3.87
FineGPR-C [97]	24	41	435	0	0	2.82
SYNTH-PEDES	18	30	162	132	158	3.76

Automatic Evaluation

We utilize some multi-modal pre-trained models like CLIP [72], ALIGN [40], BLIP [49] and BLIP2 [48] to calculate the cosine similarity of image-text pairs in each dataset. Moreover, considering the varied sensitivity of each model to image-text similarity, we employ the following normalization to ensure that the results are approximately within the same range. For each model's evaluation result, we initially identify its maximum similarity across all datasets and assign it a score of 100 points. Subsequently, the score for each dataset is determined as the average similarity relative to this highest similarity. As shown in Fig. 11, under the evaluation of five models, our SYNTH-PEDES dataset not only gains competitive scores compared to manually annotated datasets, but also surpasses the existing synthetic dataset [97] by a large margin.

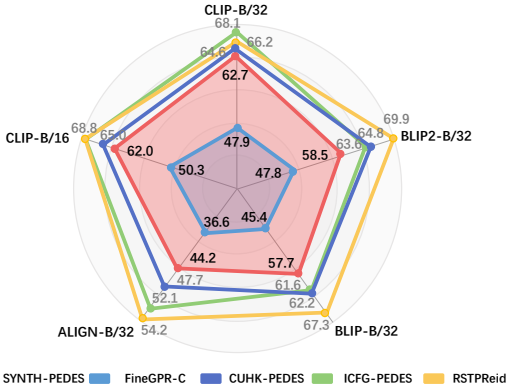

Figure 11: Automatic evaluation results by 5 pre-trained models. Our dataset are about 200 times larger than manually annotated datasets, yet still mantains competitive data quality.

A.9.3 Different Pre-training Settings

Downstream	SAT		ClipCap		SPAC	
	R@1	R@10	R@1	R@10	R@1	R@10
CUHK-PEDE	23.6	55.8	24.3	56.2	**26.6**	**59.8**
ICFG-PEDES	17.9	48.8	18.4	49.2	**20.9**	**52.3**
RSTPReid	16.2	51.0	16.5	51.4	**18.5**	**54.4**
Market1501	57.3	81.8	57.4	81.8	**58.4**	**82.8**
DukeMTMC	37.8	59.4	38.2	59.6	**39.3**	**61.0**
MSMT17	16.9	36.2	17.1	36.3	**18.3**	**37.9**

(a) Different caption methods.

Downstream	Low-quality			High-quality		
	R@1	R@5	R@10	R@1	R@5	R@10
CUHK-PEDE	28.3	51.7	62.4	**29.2**	**51.9**	**62.9**
ICFG-PEDES	22.2	43.6	53.4	**23.8**	**44.9**	**55.5**
RSTPReid	19.3	41.7	54.7	**20.0**	**42.5**	**55.2**
Market1501	57.9	77.2	83.3	**58.6**	**77.4**	**83.4**
DukeMTMC	39.6	56.0	62.8	**40.9**	**57.2**	**64.4**
MSMT17	19.3	32.6	39.1	**19.6**	**33.0**	**39.3**

(b) Different dataset quality.

Table 15: Ablation studies on caption methods in (a) and dataset quality in (b).

Pre-training with other caption methods. We have compared our proposed SPAC with the previous representative caption methods SAT [102] and ClipCap [66] in generating captions for person images. The compared methods generate only one caption per image, lacking diversity. To ensure a fair comparison, we exclusively utilize SPAC to generate a single caption for the sub-dataset derived

IVLC	TIC	CUHK R@1	CUHK R@10	Market R@1	Market R@10
$\mathcal{L}_{sd.}$	Manh.	18.7	49.0	46.4	73.8
$\mathcal{L}_{al.}$	Manh.	17.6	45.4	44.0	71.9
$\mathcal{L}_{cm.}$	Manh.	19.1	48.5	44.4	75.4
$\mathcal{L}_{sd.}$	Eucl.	30.0	63.5	63.1	86.1
$\mathcal{L}_{al.}$	Eucl.	31.8	64.1	63.0	85.7
$\mathcal{L}_{cm.}$	Eucl.	**32.5**	**66.6**	**63.1**	**86.3**

(a) Different training objectives.

$G(x)$	CUHK R@1	CUHK R@10	Market R@1	Market R@10
0.2	32.1	65.9	63.2	86.6
0.5	30.9	65.3	62.9	85.2
0.8	31.6	66.1	62.7	86.9
x^2	32.3	66.0	63.4	85.9
$\sqrt{x}$	**32.5**	**66.6**	63.1	86.3
x	31.7	65.5	**63.7**	**86.5**

(b) Different prediction difficulty.

Visual	Textual	CUHK R@1	CUHK R@10	Market R@1	Market R@10
Aver.	Pooler	26.6	60.8	55.5	81.0
Max.	Pooler.	26.2	59.5	56.2	83.0
Attn.	Pooler.	24.3	57.1	57.3	81.9
Aver.	Aver.	30.2	65.6	60.1	84.1
Max.	Aver.	**32.5**	**66.6**	**63.1**	**86.3**
Attn.	Aver.	27.9	62.4	61.3	84.9

(c) Different pooling methods.

Table 16: Ablation studies on training objectives in (a), prediction difficulty functions in (b) and pooling methods in (c).

from SYNTH-PEDES. Totally, for each method, there are 139,564 image-text pairs for pre-training. As shown in Tab. 15 (a), the performance of SPAC on six datasets is much better than the others, showing the superiority of our SPAC.

High-quality dataset leads to good pre-training. In this experiment, we consider the original LUPerson-NL dataset [30] and our dataset as low-quality and high-quality dataset, respectively. We conduct a performance comparison on downstream tasks following pre-training on each dataset. To ensure parity, an equal number of identities and image samples are randomly selected for each dataset, and captions are generated using SPAC. The pre-training process involved a total of 214,053 image-text pairs for each dataset. As shown in Tab. 15 (b), the downstream performance of the low-quality dataset is markedly inferior to ours. These results collectively validate the efficacy of the strategies employed in our dataset gathering process.

The diversity of textual descriptions matters. Our dataset has textual descriptions with different styles for each image. To validate the effectiveness of this diversity, we assess it through pre-training with datasets exhibiting varying degrees of textual diversity, followed by a comprehensive generalizability study on downstream tasks. Specifically, we have studied four different degrees of textual diversity from weak to strong, while keeping the number of person images and identities consistent. As shown in Fig. 12, the fourth case, with the highest degree of textual diversity, has the best performance. Also, compared to the first case, the superior performance observed in the second case further substantiates the significance of diversity, as the prompt caption exhibits less diversity than the generated caption.

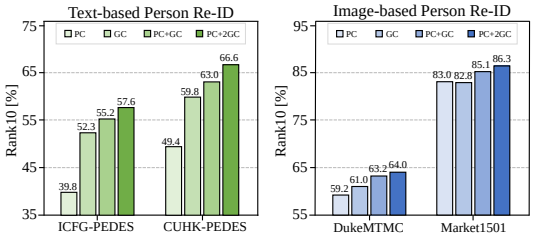

Figure 12: The diversity of textual descriptions matters. PC and GC mean prompt caption and generate caption, respectively.

A.9.4 Optional Practices

Training Objectives of IVLC and TIC. The different combinations and implementations of the loss functions for IVLC and TIC are shown in Tab. 16 (a). For the IVLC task, we have implemented three identity-based loss functions for vision-language contrastive learning. Specifically, $\mathcal{L}_{al.}$ represents the alignment loss proposed in [93], $\mathcal{L}_{sd.}$ represents the SDM loss proposed in [42], and $\mathcal{L}_{cm.}$ represents the CMPM loss proposed in [112]. For the TIC task, to measure the error between the color-restored image and the original colorful image, we tried using Euclidean distance and Manhattan distance. In fact, any function meeting the requirements can be used in our learning framework. However, as shown in Tab. 16 (a), the combination of CMPM loss and Euclidean distance leads to the best performance, and gains rank-1 32.5% and 63.1% on each dataset accordingly.

Different prediction difficulty. Curriculum learning [3, 71] demonstrates that it could be useful to organize the training samples from easy to hard. It is inspired by the human learning process and has achieved success in a wide range of tasks. To explore the effectiveness of this, we consider the masked textual descriptions with different masking probabilities as samples with different difficulty levels. Specifically, we define $G(x)$ as the difficulty function and explore a variety of optional functions, including constant, x, x^2, and $\sqrt{x}$. By using the gradually increasing functions, we start the $G(x)$ grows from 0.2 to 0.8, which means the training samples will change from easy to hard at different

rates. As shown in Tab. 16 (b), the results suggest that satisfactory performance on both datasets can be achieved with $\sqrt{x}$ in pre-training. This means the gradual difficulty sample strategy truly plays a role in encouraging the model to learn generic person representations better.

Pooling methods of global embeddings. We have studied the impact of different pooling methods of global embeddings. In the textual branch, we investigated two methods for obtaining the global embedding, namely pooler-out and average pooling. Additionally, in the visual branch, we explored three pooling operations on the last stage feature: average pooling, maximum pooling, and attention-based pooling. As shown in Tab. 16 (c), the combination of maximum pooling and average pooling consistently yields the best performance on both datasets, establishing it as the default setting.

Effectiveness of different hyper-parameters. Since our objective function consists of three losses and each loss is critical to the overall performance, we further search for the optimal hyper-parameters to balance the loss weights. As shown in Tab. 17, PLIP achieves the best performance when $\lambda_1 = 0.02$ and $\lambda_2 = 0.1$, surpassing the version of $\lambda_{1,2} = 1$ by 3.8% and 5.5% rank-1 on CUHK-PEDES and Market1501, respectively. This is mainly attributed to the loss values of three pretext tasks. The IVLC loss is the primary component of the objective function, aiming to achieve the visual and textual modality association. The TIC and IAP losses are employed to better assist in representation learning, and their loss values are relatively large. Therefore, adjusting the weights reasonably is also crucial for performance.

Table 17: Comparable results of different hyper-parameters in overall objective function. We show the best score in bold.

Settings	CUHK-PEDES			Market1501		
	R@1	R@5	R@10	R@1	R@5	R@10
$\lambda_1 = 1, \lambda_2 = 1$	28.7	51.9	62.8	57.6	77.1	83.4
$\lambda_1 = 0.2, \lambda_2 = 1$	30.3	53.6	64.1	61.7	79.2	85.5
$\lambda_1 = 0.05, \lambda_2 = 1$	31.8	55.1	65.4	62.2	79.9	85.6
$\lambda_1 = 0.02, \lambda_2 = 1$	32.3	55.8	65.9	62.6	80.3	85.9
$\lambda_1 = 0.01, \lambda_2 = 1$	31.7	55.3	65.1	62.4	80.1	85.7
$\lambda_1 = 0.02, \lambda_2 = 0.5$	32.5	55.9	66.1	62.9	80.9	85.9
$\lambda_1 = 0.02, \lambda_2 = 0.2$	32.3	56.1	66.4	62.7	**81.1**	86.1
$\lambda_1 = 0.02, \lambda_2 = 0.1$	**32.5**	**56.3**	**66.6**	**63.1**	80.8	**86.3**

A.9.5 Learning with Other Pre-training Methods

PLIP exhibits clear superiority in person representation learning compared to other general pre-training methods. To validate this, we have pre-trained a series of models on the entire SYNTH-PEDES dataset with different pre-training methods, *i.e.,* CLIP [72], BLIP [49], BLIP2 [48] and PLIP. For BLIP2, we exclusively utilize its first-stage pre-training method, aligning with the default setting in the original paper designed for image-text retrieval tasks. Subsequently, we compare their direct transfer performance on downstream datasets. The results, as reported in Tab. 18, demonstrate that under the comparable setting of models with the roughly same parameters (ViT-Small vs ResNet50 and ViT-Base vs Swin-Base), PLIP outperforms all other pre-training methods significantly on downstream datasets. Specifically, PLIP with ReNet50 achieves 52.9% and 81.1% rank-1 on CUHK-PEDES and Market1501, greatly surpassing CLIP with ResNet50 by 14.8% and 11.0%. Moreover, with Swin-Base as the backbone, PLIP achieves the best performance, with rank-1 reaching 56.3% and 83.7%, respectively. These experimental results demonstrate the superiority of PLIP in person representation learning.

Table 18: Comparable results of PLIP with other representative pre-training methods. We show the best score in bold.

Method	Backbone	CUHK-PEDES			Market1501		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	RN50	38.1	62.6	73.0	70.1	84.7	89.7
CLIP	ViT-B	41.6	65.2	74.8	72.9	85.6	90.9
BLIP	ViT-S	44.9	68.6	77.4	73.7	86.2	91.4
BLIP	ViT-B	48.2	72.7	80.6	76.3	88.6	92.5
BLIP2	ViT-S	47.3	71.6	80.1	75.9	88.1	92.4
BLIP2	ViT-B	48.8	72.6	81.2	77.3	89.4	93.3
PLIP	RN50	52.9	74.7	82.5	81.1	91.4	94.6
PLIP	Swin-B	**56.3**	**77.6**	**84.8**	**83.7**	**93.6**	**95.9**

A.10 More Evaluation on Downstream Tasks

Domain Generalization for Image-based Person Re-ID. A comparison to the SoTA in generalizable image-based person re-identification is shown in Tab. 19. The models trained on the source datasets are directly generalized to the target datasets without tuning. Several methods and datasets recently are compared, with methods including QAConv [56], TransMatcher [57], WePerson [46], and APTM [106],

Table 19: Comparisons of domain generalization performance on image-based person Re-ID.

Method	Train Set	Market1501		MSMT17	
		mAP	R@1	mAP	R@1
QAConv	RandPerson	46.9	74.5	14.0	40.6
QAConv	ClonedPerson	**59.9**	**84.5**	18.5	49.1
TransMatcher	RandPerson	49.6	77.6	16.4	45.3
TransMatcher	UnrealPerson	59.4	81.6	21.6	52.0
WePerson	WePerson	55.0	81.5	18.9	46.4
APTM	MALS	3.8	11.9	1.8	7.4
PLIP	SYNTH-PEDES	55.8	81.1	**22.1**	**52.3**

and datasets including RandPerson [92], ClonedPerson [105], UnrealPerson [109], WePerson [46] and MALS [106]. As the results indicate, even without specific design tailored for this generalizable task, our PLIP trained on SYNTH-PEDES shows competitive performance when directly generalized to downstream real-world datasets. These results verify the outstanding generalization ability of our pre-trained models for this task.

Improvement over Person Attribute Recognition Methods. Our model can also bring considerable improvement to person attribute recognition methods. We conduct experiments using two representative baseline methods [44, 41] by comparing the performance gain of different pre-trained ResNet50 models. We report the results in Tab. 20, where the two methods are based on traditional CNN structure and multi-label classification loss. As we can see, our model improves their performance significantly on the two popular datasets, and the improvements leave all other pre-trained models far behind. Particularly, in terms of *mA*, by using our pre-trained ResNet50, the average improvements of these methods are 2.1% and 1.1% on PA100k and PETA respectively. These results demonstrate that our framework helps to learn better person representations for fine-grained recognition.

Table 20: Improving two person attribute recognition baseline methods. The results of DeepMAR‡ are from a re-implementation by replacing the backbone with ResNet50, which are much better than the original.

	Pre-train	PA100k				PETA			
		mA	Acc	Rec	F1	mA	Acc	Rec	F1
DeepMar [44]	Baseline	78.3	76.8	84.0	84.3	82.7	77.1	84.6	85.2
	MoCov2 [15]	78.1	76.6	84.1	84.2	82.9	77.1	84.8	85.4
	CLIP [72]	79.4	77.5	84.9	85.2	83.3	77.6	84.8	85.4
	LUP [29]	78.5	77.3	84.5	84.6	82.8	77.3	84.5	85.3
	LUP-NL [30]	79.5	77.4	84.7	84.5	83.5	77.9	85.0	85.6
	PLIP	**80.5**	**78.8**	**86.0**	**86.5**	**83.8**	**78.2**	**85.3**	**85.8**
Rethink [41]	Baseline	80.2	79.2	87.0	87.4	84.0	78.7	85.6	86.4
	MoCov2 [15]	80.2	79.0	87.0	87.3	83.4	77.7	84.9	85.6
	CLIP [72]	81.1	79.8	87.7	87.8	84.5	79.2	86.0	86.7
	LUP [29]	80.2	79.5	87.3	87.6	84.2	78.3	85.2	86.1
	LUP-NL [30]	81.3	79.1	87.7	87.4	84.1	78.6	85.2	86.3
	PLIP	**82.2**	**81.1**	**88.6**	**88.6**	**85.1**	**79.5**	**86.3**	**86.9**

Comparison with state-of-the-art methods on human parsing. Human parsing task requires excellent perception of spatial information. By training the TIC task, as shown in the Figure 5, our models truly learn the correlation between attribute phrases and spatial regions, which guarantees the good performance of this fine-grained semantic segmentation task. In Tab. 21, we compare our results with existing SoTA human parsing methods on LIP [33] and PASCAL [13]. Under the ResNet101 setting, our method achieves 61.32% and 72.63% mIoU on LIP and PSCAL, respectively, significantly surpassing previous SoTA methods. Also, by applying our pre-trained Swin-Base model on SCHP [54], we achieve the best performance on each datasets. These results demonstrate that our pre-training framework shows good potential in learning discriminative person representation for this task.

Table 21: Comparison with SoTA methods on human parsing. We show the best score in bold.

Methods	Backbone	LIP	PASCAL
		mIoU	mIoU
Attention [7]	VGG16	42.92	56.39
MMAN [64]	RN101	46.81	59.91
JPPNet [55]	RN101	51.37	59.36
CE2P [76]	RN101	53.10	-
CNIF [91]	RN101	57.74	70.76
SCHP [50]	RN101	59.36	71.46
CDGNet [60]	RN101	60.30	-
SOLIDER [12]	Swin-B	60.50	-
PLIP	RN50	60.41	72.14
PLIP	RN101	61.32	72.63
PLIP	RN152	62.44	73.51
PLIP	Swin-B	**63.52**	**73.93**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction of our paper accurately reflect our paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We conduct an in-depth discussion of the limitations of our work, which can be found in Sec. A.3 of the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work focuses on computer vision applications, not including theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have fully provided all the details of our proposed methods in the main text and appendix, to guarantee the reproducibility of our paper's main experimental results. Meanwhile, we will open-source all of our code, data, and models after our paper is accepted. These details and open-source initiatives will ensure the reproducibility of our work and make a significant contribution to the entire community.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: To address privacy concerns associated with our dataset containing persons and the application of person ReID technology, we will implement a controlled release of our code, data, and models. Therefore, to ensure private information security, we would not provide open access to our data and code during the paper submission process. Notably, we will publicly share the application link for all our sources after our paper is accepted. Applicants will be required to adhere to relevant guidelines and regulations to access our sources, and we will conduct a strict review of their qualifications to prevent any privacy violations.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper specifies all the training and test details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Due to pre-training technology is comparably computational expensive, the error bars are not reported in the paper like nearly all related work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: our paper provides sufficient information on the computer resources needed to reproduce the experiments, which can be found in Sec. A.8.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper discusses both potential positive societal impacts and negative societal impacts of the work performed in Sec. A.2 of the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our paper describes safeguards that have been put in place for responsible release of data and models in Sec. **??** of the appendix.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators and original owners of assets used in the paper are properly credited, and the license and terms of use are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: To address privacy concerns associated with our assets including code, data, and models, we will implement a controlled release of our assets. Therefore, to ensure private information security, we would not release our assets at submission time. We will publicly share the application link for our assets after our paper is accepted. Applicants will be required to adhere to relevant guidelines and regulations to access our assets, and we will conduct a strict review of their qualifications to prevent any privacy violations.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.