P-CLIP: Progressive Discrepancy Learning for One-Shot Text-to-Image Person Re-identification

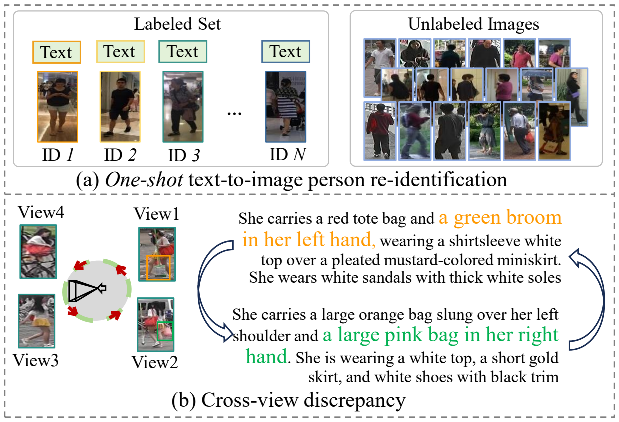
Chengji Wang, Ming Dong *Member, IEEE*, Mang Ye *Senior Member, IEEE*, Hao Sun *Member, IEEE*, Xingpeng Jiang *Member, IEEE*

***Abstract*—One-shot Text-to-Image Person Re-Identification (One-shot TIReID) aims to construct a TIReID model using only a single labeled image-text pair per identity, along with a large pool of unlabeled person images. While supervised learning in text-to-image person re-identification has demonstrated high effectiveness, the requirement for extensive annotated data, both in terms of identities and corresponding textual descriptions, makes it impractical for large-scale camera networks. One-shot TIReID presents a promising approach to reduce the annotation burden. The primary challenge in one-shot TIReID lies in establishing consistent visual-textual correspondences across diverse viewing conditions, particularly in the absence of cross-view paired data. To address this challenge, we propose a novel progressive discrepancy learning framework, termed P-CLIP, which aims to establish a shared embedding space that is robust to view-specific biases. To achieve this goal, we dynamically construct multi-view image-text pairs based on a single labeled pair and simultaneously project the multi-view data into a unified embedding space. Specifically, we propose a Progressive Multi-View Generation method (MVG) to generate multiple noisy views from a single labeled instance for training. To mitigate cross-view ambiguities, we introduce a Cross-View Discrepancy Learning module (CDL) that leverages the discrepancies among different views to guide the learning of cross-view visual-textual correspondences. This approach effectively integrates multimodal error correction into the person re-identification domain. Furthermore, to enhance the effectiveness of visual-textual correspondence learning, we propose a Compact Cross-Modal Matching Loss (CCM), which suppresses unmatched pairs while emphasizing matched ones. Extensive experiments were conducted on three benchmark datasets, and the experimental results demonstrate the effectiveness of our proposed method. The data and codes are available at https://github.com/Itachjw/P-CLIP/tree/main.**

***Index Terms*—Text-to-image person re-identification, one-shot learning, multimodal error correction, contrastive learning, cross-modal retrieval**

I. Introduction

Text-to-image person re-identification (TIReID) aims to find a person from non-overlapping views in a multi-camera system based on a textual description [1], [2]. It focuses on addressing the challenging problem of establishing correspondences between words and visual objects to distinguish different individuals. Current state-of-the-art methods [3], [4], [5], [6], [7] rely on supervised learning with large amounts of annotated data. Recent studies [8], [9], [10], [11] focus on establishing trustworthy visual-textual correspondence by addressing the inherent uncertainty in pairwise similarity measurements. Specifically, they introduce self-evaluation mechanisms to quantify uncertainty in retrieval outcomes, thereby enhancing model interpretability and result reliability. However, acquiring textual annotations and identity labels for a large set of unlabeled images is an extremely time-consuming and cumbersome task.



She carries a red tote bag and a green broom in her left hand, wearing a shirtsleeve white top over a pleated mustard-colored miniskirt. She wears white sandals with thick white soles

She carries a large orange bag slung over her left shoulder and a large pink bag in her right hand. She is wearing a white top, a short gold skirt, and white shoes with black trim

Fig. 1. (a) One-shot text-to-image person re-identification only labeled one image-text pair for each identity. (b) Cross-view discrepancy of the same identity arises due to changes in object sizes and styles. Each image-text pair contains view-specific biases, for example, "green broom" and "pink bag".

Unlike previous work, we focus on a challenging and practically important semi-supervised task called **one-shot text-to-image person re-identification** (one-shot TIReID). As shown in Fig. 1(a), each identity has only one labeled image-text pair. The objective of the one-shot TIReID is to utilize a small number of labeled image-text pairs (one labeled pair per identity) along with a larger set of unlabeled images to obtain a TIReID model[1]. Variations in illumination and occlusion conditions result in view-specific biases [12], [13]. The key challenge in one-shot TIReID is to establish consistent visual-textual correspondences across different views without the aid

Manuscript received x.x.2021; revised x. x. 2021.

This work was supported in part by the National Natural Science Foundation of China under Grant 62407018, in part by the China Postdoctoral Science Foundation under Grant 2023M741305.

Chengji Wang, Ming Dong, Hao Sun and Xingpeng Jiang are with the Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning and National Language Resources Monitoring and Research Center for Network Media, School of Computer Science, Central China Normal University, Wuhan, China (e-mail: wcj@ccnu.edu.cn; dongming@ccnu.edu.cn; haosun@ccnu.edu.cn; xpjiang@ccnu.edu.cn).

Mang Ye is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: yemang@whu.edu.cn).

[1] In our one-shot TIReID setting, the unlabeled set exclusively contains images of identities already present in the labeled set, with no introduction of unseen identities.

of cross-view annotations, regardless of variations in viewing conditions.

One-shot learning has been widely studied in image-based person re-identification [14], [15], [16], [17]. Current one-shot person re-identification works mitigate cross-view ambiguity by estimating the pseudo identity labels of unlabeled images and then performing supervised learning to obtain view-consistent image representations. In one-shot TIReID task, pseudo-labeled candidates lack corresponding text annotations. Large pre-trained Visual-Language Models (VLPs), such as BLIP [18] and Flamingo [19], have demonstrated strong image captioning capabilities. However, directly utilizing descriptions generated by these models often introduces substantial noise. These descriptions frequently omit crucial details while incorporating irrelevant or extraneous information. Alternatively, fine-tuning a VLP on annotated data requires significant computational resources and the design of effective fine-tuning strategies. Using a single text annotation to obtain inter-modal correlations is insufficient to generalize to unseen textual descriptions from another view. In fact, a textual description is generated by annotating a particular image, but it may not always align well with images of the same person from different camera views [20]. Changes in viewpoint can cause objects within a person to appear in different sizes and styles. People typically focus on primary objects and automatically ignore smaller ones, leading to cross-view discrepancies. As shown in Fig 1(b), “a large pink bag in her right hand” in view2 is absent in the description of view1. Due to the limited diversity of labeled data, one-shot TIReID models tend to learn view-specific visual-textual correspondences, which significantly hinder the utilization of cross-view identity information.

In this paper, we propose a Progressive Discrepancy Learning framework (P-CLIP) for one-shot text-to-image person re-identification. P-CLIP addresses the one-shot learning challenge through two aspects: progressive multi-view generation (MVG) and cross-view discrepancy learning (CDL). In Progressive Multi-view Generation, we construct multiple noisy views (image-text pairs) from a single labeled data, leveraging the unlabeled images in a self-supervised manner. The cross-view discrepancy learning module aims to recover information from noisy views. At each epoch, we first estimate pseudo-labels for unlabeled images using k-reciprocal distance [21], and select the most reliable and easiest ones as candidates to extend the labeled set. To assign text annotations for selected images, we construct multiple noisy text annotations by text augmentation. Two methods are adopted: 1) random replacement, we randomly replace some key words in text annotation to generate pseudo captions; 2) LLM-based text Paraphrasing, rewriting text with Large Language Models, such as ChatGPT. Cross-view discrepancy refers to the differences in visual content described by the text annotations (visual words). We use a cross-modal attention mechanism to extract these visual words and then utilize the discrepancy from between views to recover the masked information. Cross-view discrepancy learning relies on the learned fine-grained correspondence between two modalities, excavating the discrepancy of visual words to guide learning cross-view visual-textual correspondence. Our main idea is inspired by multimodal error correction [22], [23]. In this work, we make a pioneering effort attempt to incorporate multimodal error correction into person re-identification.

Pre-trained contrastive language-image models, such as CLIP [24], commonly use the InfoNCE loss. This loss function employs categorical cross-entropy to identify matched image-text pairs among a set of unrelated pairs in the batch. However, the InfoNCE loss places excessive emphasis on indistinguishable samples, rather than enhancing the correlation between matched images and texts. As a result, persons with similar appearances will have high matching scores, while cross-view image-text pairs of the same person will have low scores. To address this issue and explore a more effective cross-modal matching objective, we propose a compact cross-modal matching (CCM) loss. This loss optimizes the dot product of text-to-image and image-to-text similarity distributions. Consequently, a high text-to-image similarity score will be suppressed by a low image-to-text similarity score via the dot product, and vice versa. The proposed CCM loss enables the TIReID model to precisely control the multimodal embedding space and learn compact visual-textual correspondences. Importantly, our proposed method does not require any additional annotations or extra pre-trained models. Our main contributions can be summarized as follows:

- We introduce and formalize the one-shot TIReID task, to the best of our knowledge, this is one of the first works to formally define and tackle one-shot TIReID.
- We present a novel progressive discrepancy learning framework. We introduce cross-view discrepancy learning, which repurposes text augmentation to explicitly learn from semantic discrepancies, coupled with a compact cross-modal matching loss for more effective embedding learning.
- We provide a mutual information maximization perspective that explains why our framework works, P-CLIP can be understood as maximizing a lower bound on the mutual information (MI) between different views of an image-text pair.
- Extensive experiments demonstrate our method's effectiveness, establishing a strong baseline for one-shot TIReID while requiring no additional annotations or pre-trained models.

The remainder of the paper is organized as follows: Section II reviews the related works; Section III introduces the proposed P-CLIP in detail; Section V reports extensive experimental results and analysis; and finally the paper is summarized in Section VI.

II. Related Works

A. Text-to-image Person Re-identification.

Text-to-image person re-identification was initially introduced by Li et al. [1]. Researchers have proposed various solutions, including global feature learning [25], [26], [27], local feature alignment [28], [29], [30], [31], [32], [33], and models incorporating multi-modal interactions [34], [35], [36], [2], [4], [3], [37]. With the advancement of vision-language

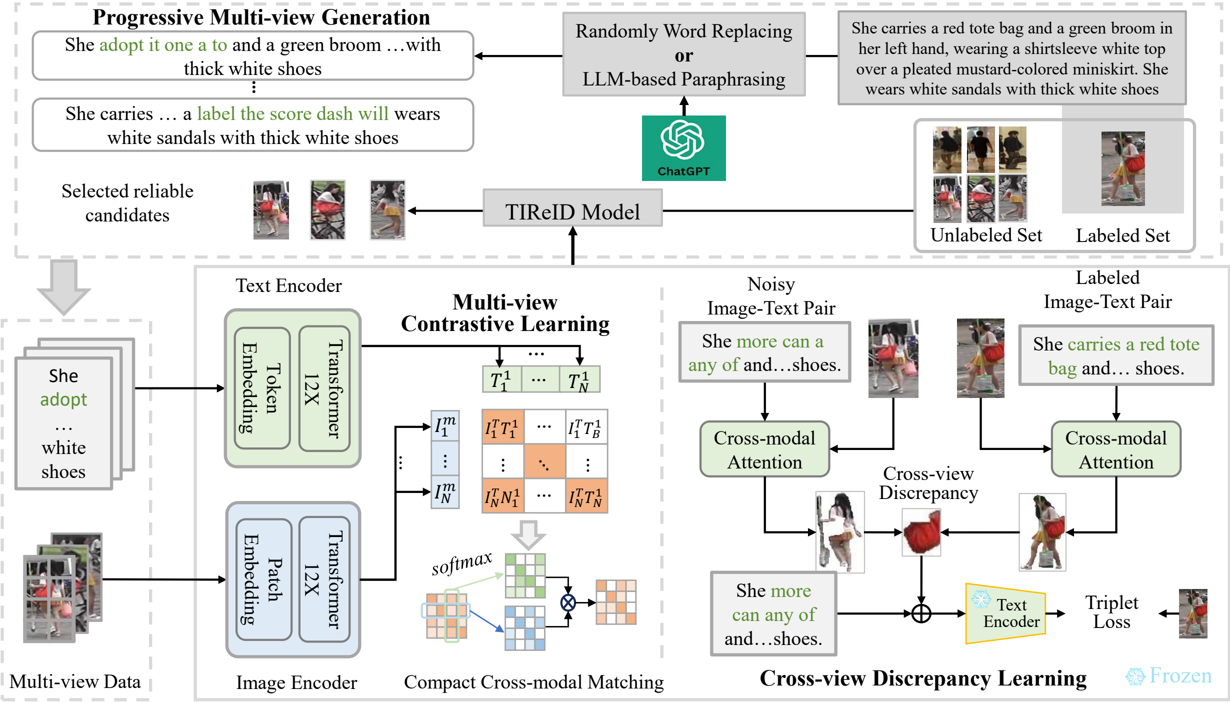

Progressive Multi-view Generation
She adopt it one a to and a green broom …with thick white shoes
She carries … a label the score dash will wears white sandals with thick white shoes
Randomly Word Replacing or LLM-based Paraphrasing
She carries a red tote bag and a green broom in her left hand, wearing a shirtsleeve white top over a pleated mustard-colored miniskirt. She wears white sandals with thick white shoes
ChatGPT
Selected reliable candidates
TIReID Model
Unlabeled Set
Labeled Set
She adopt … white shoes
Multi-view Data
Text Encoder
Token Embedding
Transformer 12X
Patch Embedding
Image Encoder
Multi-view Contrastive Learning
T_1^1 … T_N^1
I_1^m ⋮ I_N^m
$I_1^T T_1^1$ … $I_1^T T_B^1$
$I_N^T N_1^1$ … $I_N^T T_N^1$
softmax
Compact Cross-modal Matching
Noisy Image-Text Pair
She more can a any of and…shoes.
Labeled Image-Text Pair
She carries a red tote bag and… shoes.
Cross-modal Attention
Cross-modal Attention
Cross-view Discrepancy
She more can any of and…shoes.
Text Encoder
Triplet Loss
Cross-view Discrepancy Learning
Frozen

Fig. 2. Our proposed progressive discrepancy learning framework, which jointly projects multi-view data into a common space to learn view-consistent visual-textual correspondence. The progressive multi-view generation module constructs multi-view image-text pairs from a single labeled pair, we propose a cross-view discrepancy module and a compact cross-modal matching loss to learn view-consistent visual-textual correspondences. We evaluate two text augmentation methods: randomly word replacement and LLM-based text paraphrasing.

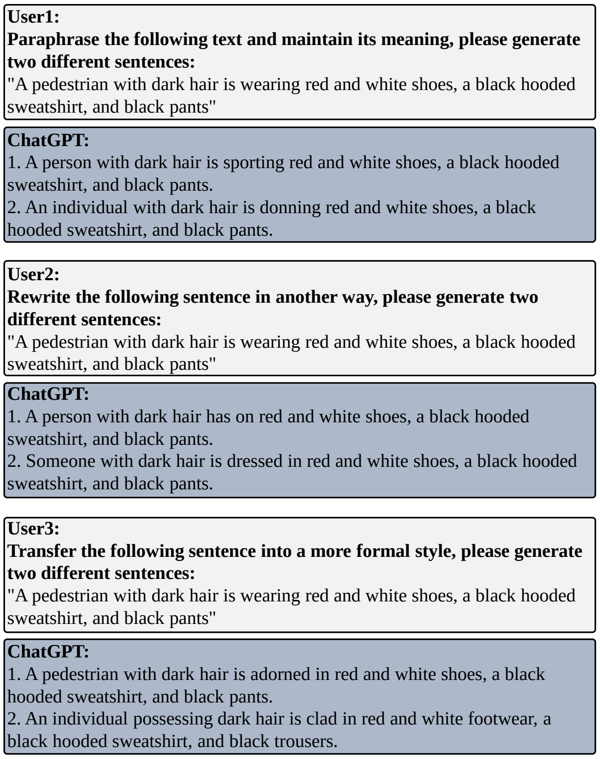
pre-training, CLIP-driven models have emerged as the state-of-the-art technology. Yan et al. [5] delve into the nuanced relationship between visual tokens and textual elements, they propose a token selection strategy to identify a set of informative tokens (discriminative patches/words). Jiang et al. [6] and Cao et al. [38] empirically investigate the effectiveness of pre-trained CLIP models. Zhang et al. [39] propose a visual perturbation network to match language descriptions with perturbed visual clues. Due to the cumbersome annotation requirements, Li et al. [40] propose generating pseudo captions by feeding attribute prompts into a large-scale pre-trained vision-language model. Gao et al. [41] design a semi-supervised setting where each person has multiple labeled data. They propose annotating the unlabeled images using a large vision-language model. Models based on synthetic data have become a recent research hotspot. Song et al. [42] propose to synthesize person images by diffusion model. Sun et al. [43] and Cao et al. [44] conduct an empirical study to explore the potential of synthetic data for TIReID.

B. One-shot Person Re-identification.

One-shot person re-identification is a semi-supervised task where only one image per identity is labeled, along with a pool of unlabeled person images. Current methods in one-shot person re-identification focus on modeling the inter-relationships among the labeled and unlabeled person images. Bak et al. [14] consider an extreme scenario where the unlabeled person image set is unavailable, and they split a metric into texture and color components. Wu et al. [16] design a progressively sampling strategy to select reliable candidates from unlabeled images step by step. Raychaudhuri et al. [45] use the temporal coherence of a person in video to estimate the labels of unlabeled images. Zhang et al. [46] utilize the nearest neighbors of labeled images to estimate the labels of unlabeled images. Wang et al. [47] design a Coarse-to-Fine hashing code search strategy to achieve both faster speed and better accuracy. Han et al. [48] define the one-shot unsupervised cross-domain person ReID task, Text-to-image person re-identification requires matching textual descriptions with images from unseen camera views, which is more challenging than other person re-identification task.

C. Multimodal Error Correction

Grammatical error correction (GEC) is the task of automatically detecting and correcting errors in text. GEC advances significantly with the development of deep learning, particularly through pre-trained large language models, such as BERT, LLM (typically ChatGPT). GEC allows language learners to better understand the nature and causes of their errors [49]. Works [50] utilizes carefully designed prompts or instructions and a training-free prompt-free framework to correct each incorrect character. Recently, multimodal error correction [23] which introduces other modality data (*e.g., audio, image*) can better correct each incorrect character. Taieb et al. [51] utilize both voice and mouse actions. Constantin et al. [23] integrate pointing gestures to enable interactive multimodal error correction. Multimodal signals is complementary which effectively enhances feature extraction and promotes



User1:
Paraphrase the following text and maintain its meaning, please generate two different sentences:
"A pedestrian with dark hair is wearing red and white shoes, a black hooded sweatshirt, and black pants"

ChatGPT:
1. A person with dark hair is sporting red and white shoes, a black hooded sweatshirt, and black pants.
2. An individual with dark hair is donning red and white shoes, a black hooded sweatshirt, and black pants.

User2:
Rewrite the following sentence in another way, please generate two different sentences:
"A pedestrian with dark hair is wearing red and white shoes, a black hooded sweatshirt, and black pants"

ChatGPT:
1. A person with dark hair has on red and white shoes, a black hooded sweatshirt, and black pants.
2. Someone with dark hair is dressed in red and white shoes, a black hooded sweatshirt, and black pants.

User3:
Transfer the following sentence into a more formal style, please generate two different sentences:
"A pedestrian with dark hair is wearing red and white shoes, a black hooded sweatshirt, and black pants"

ChatGPT:
1. A pedestrian with dark hair is adorned in red and white shoes, a black hooded sweatshirt, and black pants.
2. An individual possessing dark hair is clad in red and white footwear, a black hooded sweatshirt, and black trousers.

Fig. 3. Examples of ChatGPT 4.0-based Paraphrasing.

model performance. The complementary information between modalities can easily guide the model capturing the incorrect characters.

III. Methodology

Fig. 2 illustrates our proposed multi-view contrastive learning framework. Given a labeled image-text pair and an unlabeled image set, we first utilize the progressive multi-view generation module to construct multi-view noisy image-text pairs for each identity. We propose exploiting the cross-view discrepancy between the labeled pair and the noisy pairs to recover the labeling information from the noisy pairs. We jointly project both the labeled pair and the noisy pairs into a common space, and introduce a compact cross-modal matching loss to align the multi-view image-text pairs.

Problem statement. Consider dataset $\mathcal{D} = \{I_i\}_{i=1}^{n}$ consisting of n person images obtained from a large camera network. One-shot TIReID assumes that there exists a subset $\mathcal{D}_l \subset \mathcal{D}$, containing n_l images, $\mathcal{D}_l = \{I_i, T_i, y_i\}_{i=1}^{n_l}$, where T_i represents the textual description of image I_i, y_i denotes the identity label, and there are totally n_l distinct identities. The remaining images, $\mathcal{D}_u = \mathcal{D} - \mathcal{D}_l = \{I_j\}_{j=1}^{n_u}$, lack text and identity annotations. Our objective is to learn a discriminative TIReID model $f_\theta(\cdot)$ using both $\mathcal{D}_l$ and $\mathcal{D}_u$. During inference, given a text query χ^q and gallery person image set $\{\chi_i^g\}_{i=1}^{n_g}$, the model extracts image representations and ranks these images based on their cosine similarities to χ^q. Due to the fact that $n_l \ll n_u$ and each identity has only a single labeled image-text pair, one-shot TIReID is more challenging than other person re-identification tasks.

Dual Encoder. We use CLIP as feature extractor. Given an image $I_i \in \mathcal{D}$, we extract the image features $f_\theta^v(I_i) = \{v_g^i, v_1^i, \cdots, v_{N_v}^i\} \in \mathcal{R}^{(N_v+1)\times d}$, where N_v is the number of image patches, v_g^i is the image-level global feature, $\{v_1^i, \cdots, v_{N_v}^i\}$ are the patch-level local features. Given a text annotation $T_i \in \mathcal{D}_l$, we extract text features $f_\theta^t(T_i) = \{e_g^i, e_1^i, \cdots, e_{N_t}^i\} \in \mathcal{R}^{(N_t+1)\times d}$, where e_g^i is the sentence-level global feature, $\{e_1^i, \cdots, e_{N_t}^i\}$ are the word-level local features, with a total of N_t words. For simplicity, we use I_i and T_i to denote the image and text global features, respectively.

A. Progressive Multi-view Generation

We adopt a stepwise learning strategy. In each epoch, we first generate multi-view data based on the single labeled pair and train the TIReID model using supervised learning. For an identity y_i, we construct M noisy image-text pairs $\{I_i^m, T_i^m, y_i\}_{m=1}^{M}$ (multi-view image-text pairs) for training. In the subsequent epoch, we use the trained TIReID model to select pseudo-labeled candidates and construct multi-view data again.

We design a progressive multi-view generation module to generate multi-view image-text pairs from single labeled data. Given the labeled set $\mathcal{D}_l = \{I_i, T_i, y_i\}_{i=1}^{n_l}$, the unlabeled set $\mathcal{D}_u = \{I_j\}_{j=1}^{n_u}$, and a TIReID model, we first obtain M nearest neighbors $\{h_i^m\}_{m=1}^{M}$ from $\mathcal{D}_u$ for each I_i using k-reciprocal distance [21]. Here, $dist_{i,m}$ denotes the distance between I_i and h_i^m. The set $\{h_i^m\}_{m=1}^{M}$ represents M pseudo-labeled candidates for identity y_i,

$$I_i^m = \begin{cases} h_i^m, & \text{if } dist_{i,m} \le \sigma \\ \text{ImageAug}(I_i), & \text{if } dist_{i,m} > \sigma \end{cases} \tag{1}$$

Here, σ represents the confidence threshold for reliable pseudo-labeled candidates, and $\text{ImageAug}(I_i)$ denotes image augmentation techniques such as random cropping, random erasing, and jittering. If $dist_{i,m} > \sigma$, we use the augmented images as candidates.

To assign a text annotation to each candidate, we adopt two pipelines: randomly word replacement and LLM-based Paraphrasing:

Random word replacement. It employs random noise (S^m) to artificially create cross-view discrepancies. Specially, we sample k words from the vocabulary to form a short sentence S^m. Given the text annotation T_i, we randomly select a fragment of k words from T_i and replace it with S^m.

$$T_i^m = S^m + mask * T_i \tag{2}$$

where $mask$ is a binary code used to mask words. This process is repeated M times, generating M noisy text annotations $\{T_i^m\}_{m=1}^{M}$. In practice, the number of words replaced depends on the length of the sentence, $k = p * len(T_i)$, where p is the replacement ratio. During training, random word replacement serves as an online augmentation technique, dynamically applied by the dataloader during each training iteration.

LLM-based Paraphrasing. ChatGPT 4.0 is the most popular tools for text paraphrasing. As shown in Fig. 3, we directly apply ChatGPT4 to paraphrasing annotated pedestrian descriptions. We use three differents prompts:

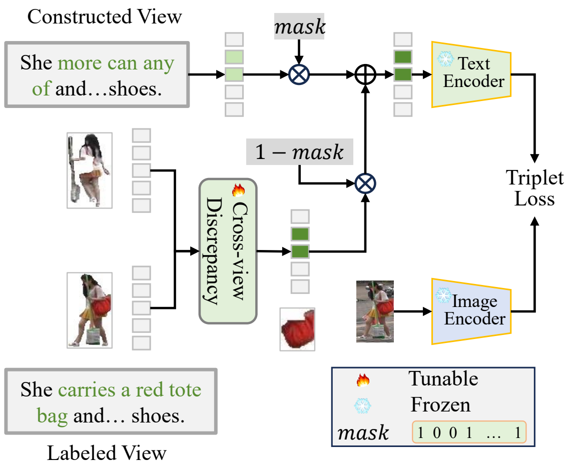


Fig. 4. Cross-view Discrepancy Learning extract conflicting visual contents (cross-view discrepancy) between constructed view and labeled view, then using this discrepancy to repair the noisy annotations.

1) Paraphrase the following text and maintain its meaning.
2) Rewrite the following sentence in another way.
3) Transfer the following sentence into a more formal style.

LLM-based Paraphrasing is implemented as an offline preprocessing stage. We generate six diverse paraphrases using three distinct prompts prior to training, creating a static caption pool containing seven variants per identity (one original and six paraphrased). During training, $M(M < 6)$ captions are randomly sampled from this pool to construct noisy text views, while the original caption is consistently retained for the labeled view.

We employ a dual-strategy approach to enhance the reliability of pseudo image-text pairs. First, the k-reciprocal neighbor requirement serves as a primary filter that effectively eliminates most unidirectional similarity matches through mutual verification. Second, we apply a confidence threshold (σ) to further refine sample selection, ensuring only high-confidence candidates are retained. A smaller σ value results in fewer candidate images being selected, consequently requiring more iteration steps and longer training time. Conversely, setting σ too aggressively may introduce unreliable samples in early stages, leading to training instability and unreliable convergence. For random word replacement, we control the noise level in pseudo captions by adjusting the replacement ratio p. As p increases, so does the noise level in the generated captions. A smaller replacement ratio introduces less noise, thereby enhancing in-context retrieval capabilities. Previous studies have demonstrated that moderate masking ratios (e.g., 0.15) benefit language understanding by promoting more distinguishable attention patterns through concentration on non-masked tokens [52], [53]. Our experimental results align with these findings, showing that a replacement ratio of 0.15 yields optimal performance across most datasets. In practical applications, these parameters represent a trade-off between efficiency and accuracy. The main limitation of our approach lies in the potential need for extensive tuning of both the confidence threshold σ and replacement ratio p across different deployment scenarios.

B. Cross-view Discrepancy Learning

Cross-view discrepancy between the labeled view and the noisy view can be utilized to recover the masked information. As shown in Fig. 4, cross-view discrepancy refers to the differences in the visual content described by texts from multiple views. We employ a cross-modal attention mechanism to extract the visual words from different views, and obtain the cross-view discrepancy by comparing the visual words from the labeled view with those from the noisy views.

Given a labeled image-text pair $\{I_i, T_i\}$ with local features $\{v_1^i, \dots, v_{N_v}^i\}$ and $\{e_1^i, \dots, e_{N_t}^i\}$. Following multi-head attention, we use the local word features as queries ($\mathcal{Q}$), and the local patch features as keys ($\mathcal{K}$) and values ($\mathcal{V}$). The cross-modal attention block is computed as:

$$Z_i = MHA(\mathcal{Q}, \mathcal{K}, \mathcal{V}) \tag{3}$$

where $Z_i = \{z_1^i, \dots, z_{N_t}^i\}$ are the visual words, and MHA$(\cdot)$ denotes multi-head attention. Given a noisy pair $\{I_i^m, T_i^m\}$, we can obtain the visual words $Z_i^m = \{z_1^{i,m}, \dots, z_{N_t}^{i,m}\}$. Visual words constitute semantically-aligned visual representations generated through cross-modal attention. These are formed by using textual token embeddings as queries to selectively aggregate relevant local image patch features (serving as keys and values). Visual words function as dynamic, multimodal representations that fuse linguistic semantics with their corresponding visual evidence in a shared embedding space, thereby enabling fine-grained cross-modal alignment.

We compute the cross-view discrepancy between $\{I_i, T_i\}$ and $\{I_i^m, T_i^m\}$ by:

$$\hat{Z}_i^m = (1 - C^{i,m}) \cdot Z_i, \tag{4}$$

where $C^{i,m} \in \mathcal{R}^{N_t \times N_t}$ is the cross-view attention map between two views,

$$c_{hj}^{i,m} = \frac{exp^{cos(z_h^{i,m}, z_j^i)}}{\sum_j exp^{cos(z_h^{i,m}, z_j^i)}} \tag{5}$$

here, $cos(\cdot, \cdot)$ is cosine distance. We reverse the $C^{i,m}$ to identify which elements are ignored by the noisy view. The cross-view attention map $C^{i,m}$ serves as a semantic difference detector. It directly measures how visual words change between clean and noisy views to obtain the meaningful visual differences between views. Mathematically, $1 - C^{i,m}$ reverses the attention to isolate visual semantics that the noisy view fails to capture.

We combine $\hat{Z}_i^m$ and T_i^m to repair the noisy annotation,

$$\hat{T}_i^m = (1 - mask) \cdot \hat{Z}_i^m + mask \cdot T_i^m \tag{6}$$

The repaired sentence $\hat{T}_i^m$ should be closer to the text annotation T_i than T_i^m. Here, we use the triplet margin loss,

$$\mathcal{L}_{cdl}^m = [\alpha - cos(\hat{T}_i^m, I_i) + cos(\hat{T}_i^m, T_i^m)]_+ \tag{7}$$

TABLE I
SUMMARY STATISTICS OF THREE DATASETS. WE COUNT THE VOCABULARIES AND AVERAGE LENGTH OF CAPTIONS IN TRAINING SET.

Dataset	Train			Val			Test			Average Lens	Words
	Labeled Pairs	Unlabeled Images	IDs	Images	Texts	IDs	Images	Texts	IDs		
CYHK-PEDES	11003	23051	11003	3078	6156	1000	3074	6148	1000	34.14	4955
ICFG-PEDES	3102	31572	3102	-	-	-	19848	19848	1000	49.75	1831
RSTPReid	3701	14804	3701	1000	2000	200	1000	2000	200	33.21	2991

C. Multi-view Contrastive Learning

We jointly project multi-view image-text pairs into a common space with the goal of maximizing the mutual information between the representations of the multimodal data of the same individual.

Compact Cross-Modal Matching. To guide the cross-modal matching of image-text pairs, we propose a compact cross-modal matching loss to align images and texts within the same view. Given a mini-batch of N image-text pairs $\{I_i^m, T_i^m\}_{i=1}^N$, we define the image-text similarity matrix $S \in \mathbb{R}^{N\times N}$ by:

$$S_{i,j}^m = \frac{exp^{cos(I_i^m, T_j^m)/\tau}}{\sum_{i=1}^N exp^{cos(I_i^m, T_j^m)/\tau}} \cdot \frac{exp^{cos(I_i^m, T_j^m)/\tau}}{\sum_{j=1}^N exp^{cos(I_i^m, T_j^m)/\tau}} \quad (8)$$

where, τ is a temperature hyperparameter that controls the peaks of the probability distributions. We utilize the softmax function to convert the image-to-text and text-to-image similarities into probability distributions. The compact cross-modal matching loss is computed by:

$$\mathcal{L}_{ccm}^m = \mathcal{L}_{i2t}^m + \mathcal{L}_{t2i}^m \quad (9)$$

$$\mathcal{L}_{i2t}^m = -\frac{1}{N}\sum_{i=1}^N y_{i,j}^m log \frac{exp^{S_{i,j}^m}}{\sum_{j=1}^N exp^{S_{i,j}^m}} \quad (10)$$

$$\mathcal{L}_{t2i}^m = -\frac{1}{N}\sum_{j=1}^N y_{i,j}^m log \frac{exp^{S_{i,j}^m}}{\sum_{i=1}^N exp^{S_{i,j}^m}} \quad (11)$$

If $\{I_i^m, T_j^m\}$ is a matched pair, $y_{i,j}^m = 1$; else, $y_{i,j}^m = 0$. The CCM loss guides the model to learn a compact and well-structured cross-modal embedding space, characterized by tight within-class clustering and clear between-class separation. This is achieved by optimizing the dot product between text-to-image and image-to-text similarity distributions, which simultaneously enforces two complementary constraints: (1) positive image-text pairs must exhibit high mutual similarity in both directions, and (2) incorrect pairings are consistently suppressed through coupled similarity reduction. This dual-constraint mechanism serves as an effective compactness regularizer, discouraging reliance on spurious one-way matches while encouraging positive pairs to occupy concentrated, discriminative regions in the shared embedding space.

Multi-view Alignment. There are multiple positive pairs from the same identity appearing in the mini-batch. To perform multi-view alignment, we consider all possible image-text pairs within the batch. We concatenate multi-view image-text pairs, denoted as $\{f_i^v\}_{i=1}^{(M+1)N} = [I_i, I_i^1, \cdots, I_i^M]$ and $\{f_j^t\}_{j=1}^{(M+1)N} = [T_j, T_j^1, \cdots, T_j^M]$. Then, we use triplet alignment loss (TAL) [11] to align the multi-view data.

$$\mathcal{L}_{tal} = [\beta - S_{i2t}^+(f_i^v) + \tau log(\sum_{j=1}^{N_{neg}} (1 - y_{i,j}) exp^{cos(f_i^v, f_j^t)/\tau})]_+$$
$$+ [\beta - S_{t2i}^+(f_i^t) + \tau log(\sum_{j=1}^{N_{neg}} (1 - y_{j,i}) exp^{cos(f_j^v, f_i^t)/\tau})]_+ \quad (12)$$

$$S_{i2t}^+(I_i) = \sum_{i=1}^{M+1} \frac{y_{i,j} exp^{cos(f_i^v, f_j^t)/\tau}}{\sum_j^{M+1} y_{i,j} exp^{cos(f_i^v, f_j^t)/\tau}} cos(f_i^v, f_j^t) \quad (13)$$

$$S_{t2i}^+(T_i) = \sum_{i=1}^{M+1} \frac{y_{i,j} exp^{cos(f_i^v, f_j^t)/\tau}}{\sum_j^{M+1} y_{i,j} exp^{cos(f_i^v, f_j^t)/\tau}} cos(f_i^v, f_j^t) \quad (14)$$

where β is the margin between positive pairs and negative pairs, and τ is a temperature hyperparameter that controls the probability distribution peaks, $N_{neg} = (M+1)N - (M+1)$. If f_i^v and f_j^t form a matched pair, then $y_{i,j} = 1$; otherwise, $y_{i,j} = 0$. $S_{i2t}^+(I_i)$ and $S_{t2i}^+(T_i)$ represent the weighted average of positive pairs. In these calculations, noise pairs with lower similarity have smaller weights.

In order to relieve the impact of unreliable pairs, the TAL loss incorporates a similarity-weighted mechanism that automatically diminishes the influence of potential false positives during training. Specifically, it employs a weighted aggregation scheme to combine representations (Eq.13-14) from all positive pairs, where unreliable or mismatched pairs are assigned lower weights. This design ensures robust representation learning while effectively mitigating the impact of semantic inconsistencies.

D. Optimization

We jointly optimize the cross-view discrepancy loss, the compact cross-modal matching loss, and the triplet alignment loss in an end-to-end manner. The overall optimization objective is defined as:

$$\mathcal{L} = \sum_{m=1}^M (\mathcal{L}_{ccm}^m + \mathcal{L}_{cdl}^m) + \mathcal{L}_{ccm} + \mathcal{L}_{tal} \quad (15)$$

where $\mathcal{L}_{ccm}$ denotes the compact cross-modal matching loss for the labeled pair $\{I_i, T_i\}$.

IV. A MUTUAL INFORMATION MAXIMIZATION PERSPECTIVE

In this section, we provide an alternative information-theoretic interpretation of the P-CLIP framework. Specifically, we show that both Cross-view Discrepancy Learning (CDL) and Compact Cross-modal Matching (CCM) can be formulated as objectives that maximize a lower bound on the mutual information between different views of an image-text pair.

Mutual information measures the statistical dependence between random variables. In self-supervised learning [54], [55], given two variables v_1 and v_2 representing different augmentations of the same input, the goal is to learn representations invariant to such transformations by maximizing the mutual information $MI(v_1; v_2)$. In vision-language representation learning, we treat v_1 and v_2 as different views of an image-text pair that preserve its underlying semantics. In practice, we maximize a lower bound on $MI(v_1; v_2)$ via the InfoNCE loss:

$$\mathcal{L}_{NCE} = -\mathbf{E}_{p(v_1,v_2)}[log\frac{exp(s(v_1,v_2))}{\sum_{\hat{v}_2 \in B} exp(s(v_1,\hat{v}_2))}] \tag{16}$$

where $s(v_1, v_2)$ is a scoring function (e.g., cosine similarity), and B includes one positive sample v_2 and $|B| - 1$ negative samples negative samples.

As shown in prior works [56], [57], the visual and textual modalities can be treated as two views of an image-text pair. Training unimodal encoders then corresponds to maximizing the MI between image and text views for positive pairs. The CCM loss (Eq.8-Eq.10) can be rewritten as:

$$\begin{aligned}\mathcal{L}_{ccm} &= -\mathbf{E}_{p(I,T)}[log\frac{exp(s(I,T))}{\sum_{i\in B} exp(s(I,T_i))} * \frac{exp(s(T,I))}{\sum_{i\in B} exp(s(T,I_i))}] \\ &= -\mathbf{E}_{p(I,T)}[log\frac{exp(s(I,T))}{\sum_{i\in B} exp(s(I,T_i))} + log\frac{exp(s(T,I))}{\sum_{i\in B} exp(s(T,I_i))}]\end{aligned} \tag{17}$$

Minimizing $\mathcal{L}_{ccm}$ is equivalent to maximizing a symmetric variant of the InfoNCE loss. Therefore, CCM effectively maximizes the MI between image and text views.

Cross-view discrepancy learning treats labeled and noisy views as dual augmentations of an image-text pair, aiming to maximize the conditional mutual information $MI(\hat{T}^m; I|T^m)$ to ensure the repaired text $\hat{T}^m$ preserves maximal semantic information about image I given noisy text T^m. This objective is achieved by minimizing the KL divergence between noisy and labeled view distributions:

$$\mathcal{L}_{cdl} = -\sum_m \mathbb{E}_{p(I,T^m)} \log\frac{p(I,T^m)}{q(I,T)} \tag{18}$$

Random word replacement serves as an information bottleneck [58] that destroys superficial lexical correlations $MI(I; T|S)$ while forcing reliance on core semantic content $MI(I; S)$, enabling learning of vocabulary-invariant cross-modal representations robust to linguistic variations.

V. Experiments

To rigorously validate the performance, we conduct a series of experiments comparing our P-CLIP method with existing approaches. Additionally, we evaluate the contributions of each component through ablation studies and parameter analyses.

A. Experiment Setup

We present the experimental results and corresponding analyses for one-shot TIReID on three widely-used benchmark datasets.

Datasets. We validate our method on three challenging datasets: CUHK-PEDES, ICFG-PEDES and RSTPReid. We follow the official data partitions to split these datasets into training, validation, and test sets. For the training sets, we reconstruct them such that each identity retains only a single image-text pair and remove the identity labels and text annotations from the remaining data. CUHK-PEDES [1] contains a total of 40,206 images for 13,003 identities. Following the official data split, training set has 11,003 identities, validation set and test set have 1000 identities, respectively. We modify the training set by selecting one image-text pair for each identity and removing the textual annotations and identity labels of the remaining 23,051 images. The validation set and test set keep unchanged. ICFG-PEDES [59] contains 54,522 images for 4,102 identities. The dataset is divided into a training set and a test set. As same as CUHK-PEDES, we select 3,102 labeled image-text pairs from 3,102 unique identities and 31,572 unlabeled images for training. The test set comprises 1,000 unique identities with 19,848 image-text pairs. RSTPReid [60] contains 20,505 images of 4,101 identities. We select 3,701 labeled image-text pairs from 3,701 unique identities and 14,804 unlabeled images for training. The test set and validation set have 1,000 images with 2,000 text annotations from 200 unique identities, respectively. The statistics of datasets can be seen in Table I.

Evaluation metrics. Rank-k metrics are commonly used evaluation indicators in the research field. Specifically, Rank-k represents the proportion of correctly retrieved images within the top-k positions in the gallery when a textual query is provided. To compare retrieval performance, we present Rank-k for $k = 1, 5, 10$. Additionally, to ensure a comprehensive evaluation, we also report the mean Average Precision (mAP).

Implementation details. For input images, we uniformly resize them to 384x128 pixels. Image augmentations such as random cropping with padding, random horizontal flipping, and random erasing are applied. For input text, the token sequence length is unified to 77 tokens. Local and global feature dimensions are standardized to 512. We implement our model based on the popular pre-trained CLIP using the PyTorch library. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory, using a batch size of 32. Each model is trained for 60 epochs using the Adam optimizer. The multi-head attention block has 8 heads. The initial learning rate is set to $1 \times 1e^{-5}$. We adopt a linear warm-up with a cosine annealing strategy, gradually increasing the learning rate from $1 \times 1e^{-6}$ to $1 \times 1e^{-5}$ over 5 warm-up epochs. The temperature parameter τ is set to 0.02. We set $\alpha = 0.2$ and $\beta = 0.5$. In the final models, we construct 3, 2, and 1 noisy views for CUHK-PEDES, ICFG-PEDES and RSTPReID, respectively. The replacement ratio p is set to 0.15, 0.15 and 0.1 for the three datasets, while the confidence threshold σ is set to 0.15, 0.1 and 0.05, respectively.

B. Comparison with State-of-the-Art Methods

We evaluate the performance of our P-CLIP on three popular benchmarks. For a comprehensive comparison, we compare our method with several state-of-the-art methods, including

TABLE II
Performance comparisons on CUHK-PEDES, ICFG-PEDES and RSTPReid. The best and second-best results are in **bold** and underline, respectively. Baseline is the CLIP model fine-tuned with original InfoNCE loss. 'P-CLIP(LLM)' is the model with ChatGPT4-based text augmentation. 'P-CLIP(Random)' is the model with randomly word replacing. 'P-CLIP(LLM)' operates without the CDL module; the CCM and TAL components remain consistent with the 'P-CLIP(Random)' framework.

Model	Image-Text Encoder	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
MANet(2021)	RN50-LSTM	40.77	64.38	73.29	32.39	19.37	38.76	48.94	7.75	22.40	46.90	60.80	13.56
SSAN(2024)	RN50-LSTM	33.09	56.25	67.06	27.33	16.02	33.19	42.75	7.05	17.40	42.30	53.30	11.46
SRCF(2022)	RN50-BERT	42.66	65.64	74.59	34.96	22.69	42.05	51.59	10.01	20.55	44.75	53.90	12.94
LGUR(2022)	DeiT-BERT	43.29	64.96	73.85	35.24	18.05	36.73	46.51	7.75	15.40	33.40	47.70	9.93
SAF(2022)	ViT-BERT	43.67	67.50	76.53	34.66	18.85	38.44	48.75	8.40	21.40	43.80	56.75	13.05
CFine(2023)	CLIP-BERT	45.99	68.92	78.18	38.50	25.89	47.22	57.25	12.62	29.80	58.45	70.40	19.93
IRRA(2023)	CLIP-CLIP	57.62	77.65	85.01	51.45	40.59	61.40	69.14	21.55	47.95	71.65	81.00	36.49
MGCC(2024)	CLIP-CLIP	58.70	78.12	85.18	52.84	24.61	47.15	58.00	13.70	24.20	51.90	66.10	19.14
Baseline	CLIP-CLIP	53.59	75.68	83.51	48.26	34.93	55.35	64.37	19.57	41.80	67.05	77.35	32.18
P-CLIP(LLM)	CLIP-CLIP	59.13	78.92	85.33	53.39	43.20	62.63	70.57	24.19	48.14	72.46	81.85	37.87
P-CLIP(Random)	CLIP-CLIP	**60.87**	**79.87**	**86.89**	**53.85**	**45.01**	**63.94**	**71.87**	**25.03**	**48.75**	**73.25**	**82.45**	**39.03**

TABLE III
Ablation study on each component of P-CLIP on CUHK-PEDES, ICFG-PEDES and RSTPReid. 'MVG' is progressive multi-view generation; 'CDL' is cross-view discrepancy learning; 'CCM' is compact cross-modal matching; 'TAL' is the triplet alignment loss.

Components				CUHK-PEDES				ICFG-PEDES				RSTPReid			
MVG	CDL	CCM	TAL	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
				53.59	75.68	83.51	48.26	34.93	55.35	64.37	19.57	41.80	67.05	77.35	32.18
✓				54.33	76.00	83.60	48.43	35.98	56.97	65.96	19.66	40.80	65.70	77.95	32.32
✓	✓			55.36	76.98	84.11	48.93	36.98	57.97	66.96	19.96	42.55	66.50	77.25	32.91
✓		✓		56.14	77.70	85.12	50.24	38.41	59.13	68.11	21.89	45.65	71.20	80.90	35.98
✓	✓	✓		57.65	78.51	85.48	50.99	40.19	60.92	69.12	22.50	47.50	71.50	81.35	36.88
✓	✓		✓	57.88	78.62	85.79	51.41	41.54	61.14	68.85	22.61	45.85	71.75	80.90	36.35
✓		✓	✓	60.22	79.47	86.00	53.06	43.98	62.99	71.07	24.49	48.30	72.30	82.00	37.79
✓	✓	✓	✓	60.87	79.87	86.89	53.85	45.01	63.94	71.87	25.03	48.75	73.25	82.45	39.03

SSAN [59], MANet [61], SRCF [62], LGUR [31], SAF [63], CFine [5], IRRA [6], and MGCC [64]. The detailed results are shown in Tab. II.

From Tab. II, we can see that the proposed 'P-CLIP(LLM)' and 'P-CLIP(Random)' outperforms all state-of-the-art methods. 'P-CLIP(Random)' achieves 60.87%, 45.01% and 48.75% Rank-1 accuracy on the three benchmarks, which exceed the Baseline model by a large margin, *i,e.*, +7.28%, +10.08% and 6.95%, respectively. The mAP is increased by +5.59%, +5.46% and +6.85%, respectively. Notably, our directly fine-tuned CLIP Baseline has already achieved comparable results to recent state-of-the-art methods, with Rank-1 accuracy reaching 53.89%. 34.93% and 41.80%, respectively. 'P-CLIP(LLM)' achieves 59.13%, 43.20% and 48.14% Rank-1 accuracy on three benchmarks. The Rank-1 accuracy improves 5.44%, 8.27%, 6.34%, respectively. The performance of 'P-CLIP(LLM)' is slightly inferior to that of 'P-CLIP(Random)', this is because the ChatGPT4-based Paraphrasing can not construct cross-view discrepancy. As we can see in Fig. 3, the text rewritten by ChatGPT4 has a high similarity to the original text. The 'P-CLIP(Random)' directly replace some key words to artificially construct cross-view discrepancy. Randomly word replacing without requiring any extra models or APIs, it is a simple and economical method for text augmentation.

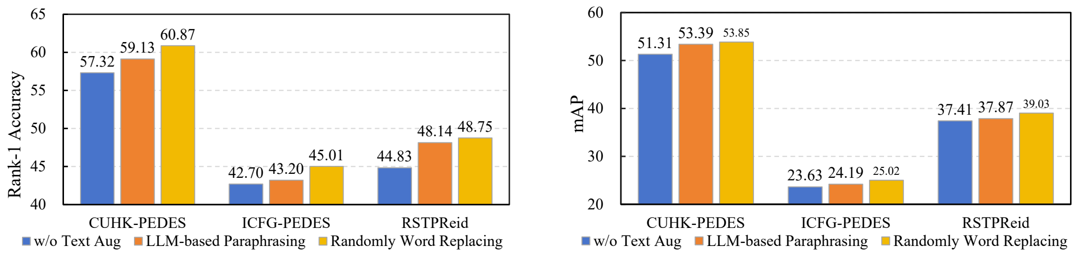
Compared with the MGCC [64], 'P-CLIP(Random)' outperforms it by **2.17%/1.01%, 20.4%/11.33% and 24.55%/19.89%** Rank-1/mAP on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets. Unlike MGCC, which relies on modeling fine-grained correspondences between image patches and words, our 'P-CLIP(Random)' primarily utilizes global contrastive learning to align images and texts. Specially, in the test set of CUHK-PEDES, each identity averages 3.07 images, while ICFG-PEDES has 19.8 images and RSTPReid has 5 images. As observed in Tab. II, the performance of MGCC drops significantly on the ICFG-PEDES and RSTPReid datasets. This is because MGCC employs a finely crafted token selection mechanism that tends to select view-specific tokens. We also observe the same phenomenon in CFine, which uses a similar token selection mechanism. Compared with the IRRA [6], 'P-CLIP(Random)' outperforms it by **3.25%/2.4%, 4.42%/3.48% and 0.8%/2.54%** Rank-1/mAP on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets. As seen in Tab. II, 'P-CLIP(LLM)' also has similar performance. Both 'P-CLIP(LLM)' and 'P-CLIP(Random)' demonstrate stable generalization ability.

These results demonstrate that fine-grained correspondences tend to learn view-specific information, while cross-view identity information is more easily encoded in global features.

C. Ablation Studies

In this section, we analyze the effectiveness of each component in our P-CLIP. We conduct a comprehensive empirical analysis on three benchmarks. The ablation results are shown in Tab. III. 'P-CLIP(LLM)' and 'P-CLIP(Random)' have similar performance, 'P-CLIP(LLM)' without the 'CDL' module. In order to save experimental costs, all experiments are conducted by 'P-CLIP(Random)' model.

Progressive Multi-view Generation. 'MVG' constructs multi-view data from single labeled image-text pair for train-



(a) (b)

Fig. 5. The Rank-1 accuracies and mAPs of different text augmentation methods on three datasets. 'w/o Text Aug' is to reuse the original text multiple times.

ing. As shown in Tab. III, 'MVG' contributes to the performance improvement on the CUHK-PEDES and ICFG-PEDES datasets. In RSTPReid, MVG vs. baseline drops at Rank-1 and Rank-5, and 'MVG+CDL' vs. 'MVG' drops at Rank-10, but the mAPs are increasing. MVG's regularization reduces reliance on superficial features, slightly impacting top-rank precision on easy samples (Rank-1/5) but improving overall ranking consistency (mAP). Subsequently, 'CDL' amplifies this by concentrating model capacity on distinguishing the hardest negatives, which is crucial for Rank-1 but may relax suppression on easier negatives affecting Rank-10. The consistent mAP increase confirms that this leads to more robust and generalizable representations. 'MVG' introduces noise by generating new pairs, which the InfoNCE loss cannot effectively handle. In Fig. 5, we report the Rank-1 accuracies and mAPs of three methods, including 'w/o Text Aug', 'LLM-based Paraphrasing' and 'Randomly Word Replacing'. Without text augmentation, 'MVG' assigns the same text annotation to multi views, leading to a lack of diversity in the constructed multi-view image-text pairs. Consequently, removing 'Text Aug' results in a significant decline in performance across all three benchmarks. 'LLM-based Paraphrasing' use ChatGPT4 to rewrite text annotations. As we can observe that 'LLM-based Paraphrasing' is better than 'w/o Text Aug'. The above experimental results show that enhancing text diversity can improve cross-modal search results. 'Randomly Word Replacing' construct cross-view discrepancy by injecting man-made noise. 'LLM-based Paraphrasing' does not construct cross-view discrepancy. As we can see, the Rank-1 accuracy and mAP of 'Randomly Word Replacing' are higher than 'LLM-based Paraphrasing'. This proves that learning cross-view discrepancy is beneficial for cross-view TIReID. Notably, the major advantage of 'Randomly Word Replacing' is the computational efficiency and low-budget (free). 'LLM-based Paraphrasing' requires an extra large language model or online APIs, 'Randomly Word Replacing' can run on CPUs.

Cross-view Discrepancy Learning. From Tab III, we observe that the Cross-Discrepancy Learning module consistently improves performance when used with 'CCM' and 'TAL'. This demonstrates that mining fine-grained representation differences between views can effectively enhance the model's cross-view retrieval capability. In Tab. IV, we replace the reconstruction loss with the Masked Language Generation module ('MLM') to predict replaced words. As shown in Tab. IV, using 'MLM' decreases the Rank-1 accuracy by 1.74%, 0.37%, and 0.25% on the three datasets, respectively. The Cross-Discrepancy Learning module generates cross-view attention maps to capture cross-view differences. In Tab. IV, we remove cross-view attention and directly use features from the labeled view to reconstruct label information. We find that 'CDL w/ CA' performs better.

TABLE IV

ANALYSIS OF THE CROSS-VIEW DISCREPANCY LEARNING MODULE. 'MLM' IS THE MASKED LANGUAGE MODULE. 'CDL W/O CA' IS CDL WITHOUT CROSS-VIEW ATTENTION. 'CDL W CA' IS CDL WITH CROSS-VIEW ATTENTION.

Method	CUHK-PEDES		ICFG-PEDES	
	Rank-1	mAP	Rank-1	mAP
MLM	59.13	52.02	44.64	24.95
CDL w/o CA	60.12	53.01	44.76	24.30
CDL w CA	60.87	53.85	45.01	25.02

TABLE V

ANALYSIS OF THE TRIPLET LOSS IN CROSS-VIEW DISCREPANCY LEARNING MODULE. T_i DENOTES THE ALIGNMENT BETWEEN THE REPAIRED SENTENCE $\hat{T}_i^m$ AND THE LABELED TEXT ANNOTATION, WHILE I_i REPRESENTS THE ALIGNMENT OF THE REPAIRED SENTENCE $\hat{T}_i^m$ WITH THE IMAGE I_i.

Method	CUHK-PEDES		ICFG-PEDES	
	Rank-1	mAP	Rank-1	mAP
T_i	60.22	53.13	44.57	24.31
I_i	60.87	53.85	45.01	25.02

Compact Cross-modal Matching. In Tab. III, replacing the InfoNCE loss with the 'CCM' loss leads to a significant performance improvement, with Rank-1 accuracy increasing by 1.81%, 2.43%, and 4.85% across the three datasets. When the Compact Cross-Modal Matching ('CMM') loss is removed from P-CLIP, the Rank-1/mAP decrease by 2.99%/2.44%, 3.47%/2.42% and 2.9%/2.68%, respectively. These results demonstrate the effectiveness of the compact cross-modal matching loss. The 'CMM' loss optimizes the dot product between image-to-text and text-to-image similarities, effectively suppressing the high scores of hard negative cases. Consequently, the model learns more accurate cross-modal correspondences. Similar performance improvements are observed in the remaining metrics.

Triplet Alignment Loss. We utilize the triplet alignment loss ('TAL') to align multi-view image-text pairs. Specifically, 'TAL' serves as a cross-view hard pair mining function, where the gradients are primarily influenced by the hard negative pairs. As shown in Tab. III, 'TAL' significantly improves

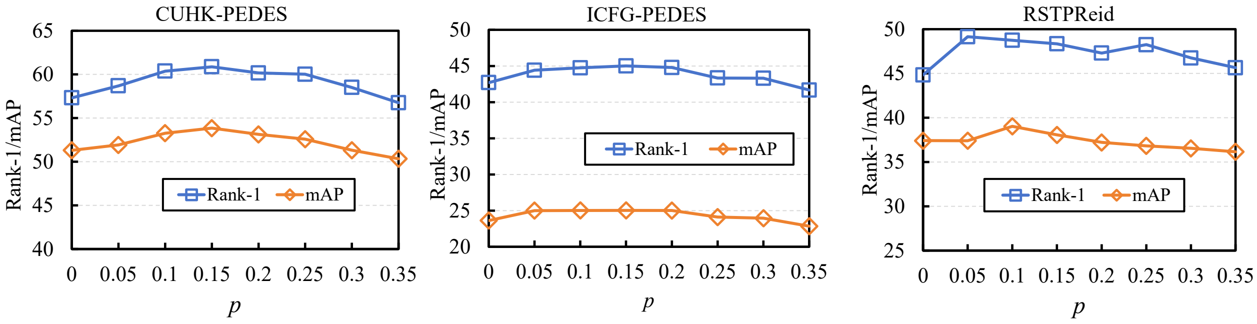


Fig. 6. The results of different values of replacement ratio p on three datasets.

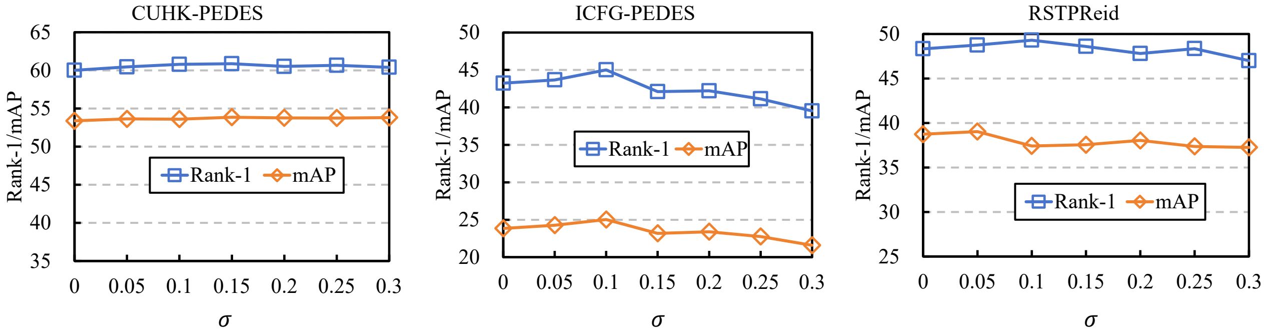


Fig. 7. The results of different values of confidence threshold σ on three datasets.

TABLE VI
PERFORMANCE ON THREE DATASETS WITH DIFFERENT SETTINGS OF VIEW NUMBER M.

M	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
1	58.58	77.65	84.74	51.30	43.96	63.17	70.97	24.81	**48.75**	**73.25**	**82.45**	**39.03**
2	60.04	79.11	85.97	52.85	**45.01**	**63.94**	**71.87**	**25.03**	48.10	70.50	80.05	38.89
3	**60.87**	**79.87**	**86.89**	**53.85**	44.49	63.83	71.56	24.93	46.15	71.30	82.00	37.13
4	60.28	79.03	85.92	52.91	44.54	64.00	72.09	24.89	47.25	70.95	80.20	37.28
5	60.15	79.22	86.10	52.59	44.05	63.39	71.36	24.42	44.85	67.75	77.90	35.07

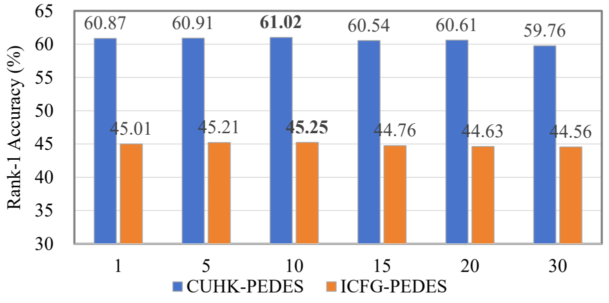


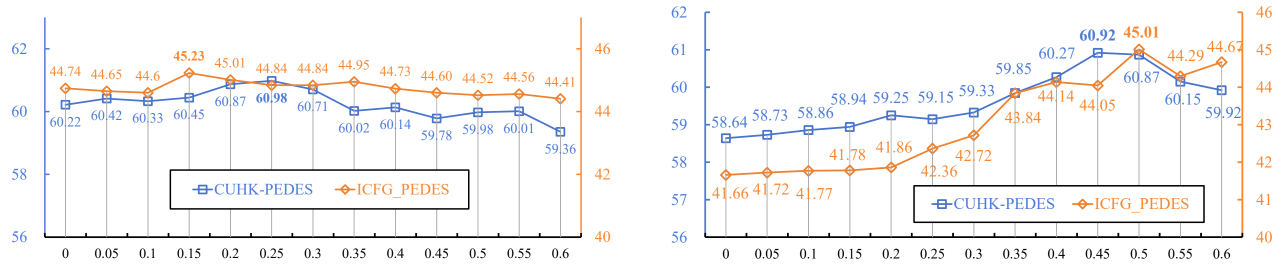
Fig. 8. Comparison of label update frequencies.

performance, demonstrating the importance of cross-view hard pair mining. As quantitatively demonstrated in Tab. V, explicitly aligning the repaired sentence $\hat{T}_i^m$ with the image I_i yields consistently superior retrieval performance compared to alignment with the original text T_i. This performance advantage stems from the role of the image I_i as a stable semantic anchor in the visual domain. By minimizing the distance between $\hat{T}_i^m$ and I_i, we ensure that the repaired text representation is semantically grounded in direct visual evidence, the ultimate source of truth for cross-modal retrieval, rather than being constrained to replicate the specific lexical formulation of T_i. This approach directly reinforces the core objective of robust image-text alignment.

D. Analysis of Progressive Multi-view Generation

In this subsection, we conduct ablation studies on a series of detailed experiments designed to evaluate the influence of various configuration parameters on pseudo-caption and identity generation. The results are presented in Fig. 6, Fig. 7, Fig. 8 and Tab. VI.

Number of View M. Tab. VI presents the results for different values of M. We report the Rank-1 accuracy and mAP. As shown in Tab. VI, for the CUHK-PEDES dataset, $M = 3$ is the optimal configuration. The test set of CUHK-PEDES consists of 3074 images from 1000 identities, with an average of 3 images per identity. For the ICFG-PEDES dataset, $M = 2$ yields the best performance. On the RSTPReid dataset, the model achieves its best performance when $M = 1$. Additionally, we evaluated the model with more views. However, when $M > 3$, a performance decline was observed across all



(a) The parameter analysis of α in CDL loss.

(b) The parameter analysis of β in TAL loss.

Fig. 9. The parameter analysis of various settings of α and β. Rank-1 accuracies on CUHK-PEDES and ICFG-PEDES are reported. The best results are in **bold**.

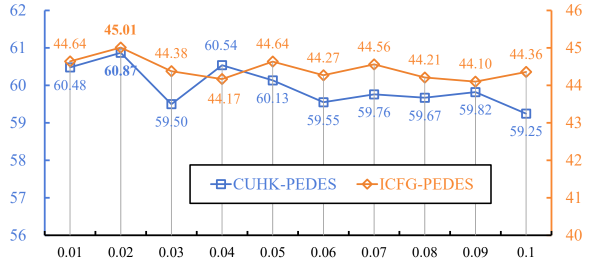


Fig. 10. The parameter analysis of scaling factor τ, Rank-1 accuracies on CUHK-PEDES and ICFG-PEDES are reported. The best results are in **bold**.

three datasets. A larger value of M introduces more noisy image-text pairs, which degrades the quality of the training data. Consequently, it becomes difficult to train a robust model using low-quality data.

Replacement ratio p**.** As shown in Fig. 6, model performance improves with moderate noise levels ($p < 0.2$) but declines beyond a critical point ($p > 0.2$). A low-to-moderate replacement ratio ($p < 0.2$) serves as an effective regularizer, where CDL refines visual-textual features while TAL leverages them to construct a noise-invariant yet semantically coherent embedding space. Consequently, the model enhances its robustness by learning to handle various noise levels. However, beyond the threshold ($p > 0.2$), the noise makes pseudo captions often contain incorrect semantics, leading to direct gradient conflict between learning objectives. This confuses the model and impedes convergence, ultimately resulting in the observed performance drop. The optimal threshold is dataset-dependent: $\alpha = 0.15$ achieves peak performance on CUHK-PEDES and ICFG-PEDES, and $\alpha = 0.05$ yields optimal results on RSTPReid.

Confidence threshold σ**.** Fig. 7 illustrates how different confidence thresholds σ affect Rank-1 accuracy and mAP across three benchmark datasets, with σ values ranging from 0 to 0.3. This parameter directly controls the number of candidate images selected from the unlabeled pool. When $\sigma = 0$, no candidates are selected, while excessively high values of σ introduce irrelevant samples that degrade performance. The optimal threshold is dataset-dependent: $\sigma = 0.15$ achieves peak performance on CUHK-PEDES, $\sigma = 0.1$ works best on ICFG-PEDES, and $\sigma = 0.05$ yields optimal results on RSTPReid. These highlight the importance of dataset-specific calibration for the confidence threshold.

Label update frequencies. Since the optimal update frequency for pseudo-identity labels is not predetermined, we evaluate multiple update intervals, as shown in Fig. 8. Experimental results indicate that per-epoch updates achieve better performance compared to slower schedules (e.g., updating every 20 or 30 epochs), although excessively frequent updates may introduce training instability. On both CUHK-PEDES and ICFG-PEDES datasets, an update interval of 10 epochs yields the optimal performance, effectively balancing the advantages of timely label refinement against the risks associated with overly frequent or infrequent updates. Unless otherwise specified, we adopt the per-epoch update strategy in our experiments.

E. Parameter Analysis

We conduct systematic analysis of key hyperparameters, with experimental results summarized in Fig. 9 and Fig. 10.

Margin α **in Triplet Loss.** The margin parameter α establishes the minimum separation between positive and negative pairs in the embedding space. Our evaluation across multiple values (Fig. 9(a)) reveals that model performance shows relative insensitivity to α variations, with no statistically significant improvements observed across different settings.

Margin β **in TAL Loss.** The Textual Alignment Loss (TAL) incorporates all positive samples during training, automatically weighting noisy pairs by their similarity scores. Parameter β governs the margin separation between positive and negative pairs. As shown in Fig. 9(b), β demonstrates substantially greater influence than α on final performance. Insufficient values ($\beta < 0.2$) fail to adequately separate labeled and noisy views, while excessively large values ($\beta > 0.5$) promote overfitting to challenging samples.

Temperature τ**.** This parameter controls the concentration of the instance probability distribution. Lower values ($\tau \rightarrow 0.01$) sharpen the distribution to emphasize hard negatives, while higher values ($\tau \rightarrow 0.1$) yield smoother distributions. Performance remains stable across the evaluated range $[0.01, 0.1]$ (Fig. 10), indicating robustness to this hyperparameter.

F. Comparisons on Generalization Ability

Tab. VII presents a comparative analysis of cross-domain retrieval performance, demonstrating our method's significant

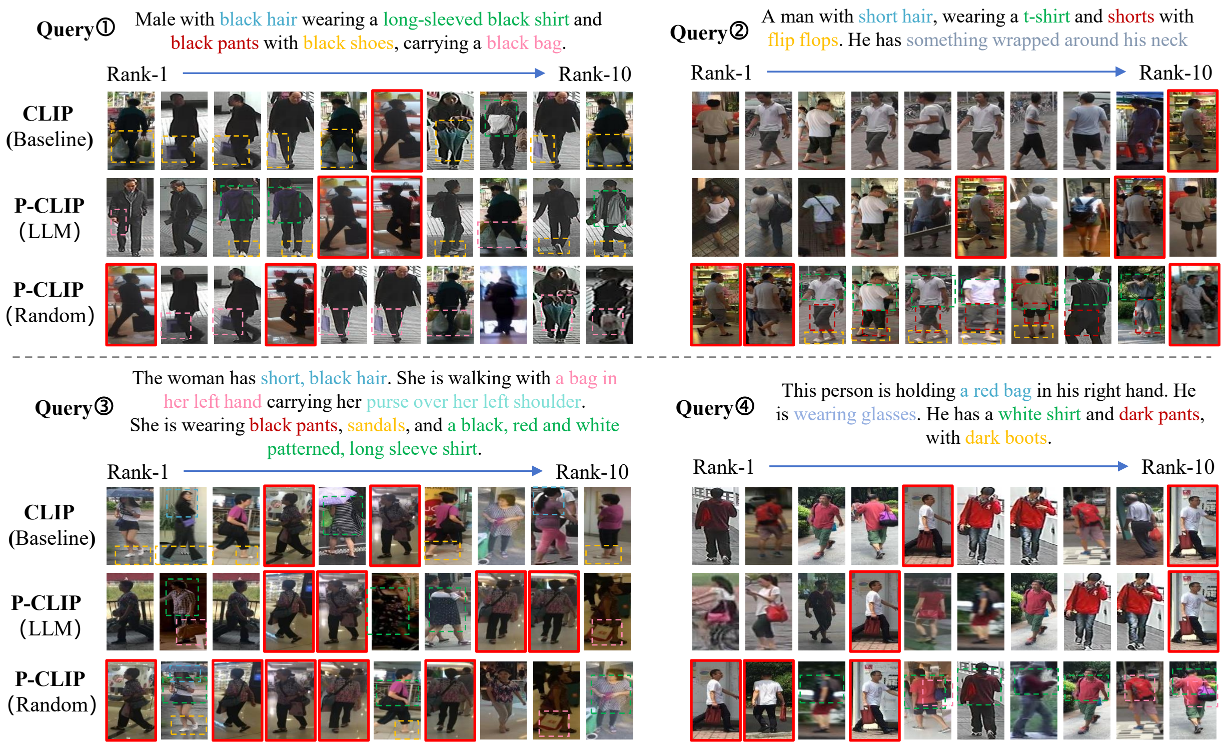

Query①
Male with black hair wearing a long-sleeved black shirt and black pants with black shoes, carrying a black bag.
Query②
A man with short hair, wearing a t-shirt and shorts with flip flops. He has something wrapped around his neck
Rank-1
Rank-10
Rank-1
Rank-10
CLIP (Baseline)
P-CLIP (LLM)
P-CLIP (Random)
Query③
The woman has short, black hair. She is walking with a bag in her left hand carrying her purse over her left shoulder. She is wearing black pants, sandals, and a black, red and white patterned, long sleeve shirt.
Query④
This person is holding a red bag in his right hand. He is wearing glasses. He has a white shirt and dark pants, with dark boots.
Rank-1
Rank-10
Rank-1
Rank-10
CLIP (Baseline)
P-CLIP (LLM)
P-CLIP (Random)

Fig. 11. Comparison of Rank-10 retrieved results on CUHK-PEDES between Baseline (the first row), P-CLIP(LLM) (the second row) and P-CLIP(Random) (the third row) for each text query. The image corresponding to query text are marked with red rectangle.

TABLE VII
DOMAIN GENERALIZATION OF CLIP-BASED METHODS. HERE, 'C' DENOTES CUHK-PEDES, WHILE 'I' REPRESENTS ICFG-PEDES.

Methods		Rank-1	Rank-5	Rank-10
C ⟶ I	CLIP	27.75	46.77	56.12
	IRRA	30.05	49.37	58.72
	P-CLIP(LLM)	30.27	50.16	59.83
	P-CLIP(Random)	33.42	51.72	60.15
I ⟶ C	CLIP	29.92	51.82	62.12
	IRRA	30.13	52.26	62.56
	P-CLIP(LLM)	30.36	52.55	62.73
	P-CLIP(Random)	31.94	53.46	63.24

TABLE VIII
RUNTIME (MINUTES) PERFORMANCE, PARAMETER COUNTS (M) AND GPU MEMORY (G) OF CLIP-BASED METHODS.

Methods	Param(M)	Runtime(m)		GPU Memory(G)	
		Train	Infer	Train	Infer
CLIP	151	56	2.03	5.73	3.68
IRRA	192	77	2.03	8.22	3.68
P-CLIP(LLM)	151	123	2.03	16.57	3.68
P-CLIP(Random)	151+3	125	2.03	16.57	3.68

improvements in generalization capability. Specifically, 'P-CLIP(Random)' achieves a 5.6% Rank-1 improvement over CLIP and outperforms IRRA by 3.4% in $C \longrightarrow I$. For $I \longrightarrow C$, it maintains a consistent 2% advantage. While 'P-CLIP(LLM)' shows marginally lower performance than the random augmentation variant, it consistently surpasses both CLIP and IRRA. These results collectively confirm the enhanced cross-distribution generalization capacity of our P-CLIP framework.

G. Comparisons on Computational Efficiency

As summarized in Tab. VIII, we report parameter counts, training time, and GPU memory usage for 'P-CLIP(LLM)' and 'P-CLIP(Random)', alongside baselines CLIP [24] and IRRA [6]. Our P-CLIP models require fewer learnable parameters than IRRA, with the CDL module adding only 3M parameters compared to the 41M parameters introduced by IRRA's MLM module. Crucially, all additional parameters in 'P-CLIP(Random)' and IRRA are used exclusively during training and introduce no computational overhead during inference. Similarly, the MVG component is only active during training. Consequently, all models maintain identical inference efficiency. While P-CLIP achieves superior performance over both CLIP and IRRA, this comes with increased training costs: pseudo-label generation extends training time, and multi-view construction increases GPU memory usage under identical batch sizes. These trade-offs between performance gains and training overhead should be considered in practical deployments.

H. Qualitative Results

Fig. 11 compares the Rank-10 retrieval results from the Baseline (CLIP) and our proposed P-CLIP. As the figure shows, both of 'P-CLIP(LLM)' and 'P-CLIP(Random)' achieve much more accurate results when CLIP fails to retrieve them. This is mainly due to our proposed multi-view contrastive learning framework, which fully exploit cross-view discriminative clues to match cross-view individuals. Firstly, our P-CLIP establishes more precise cross-modal correspondences. In Query①, 'P-CLIP(Random)' effectively

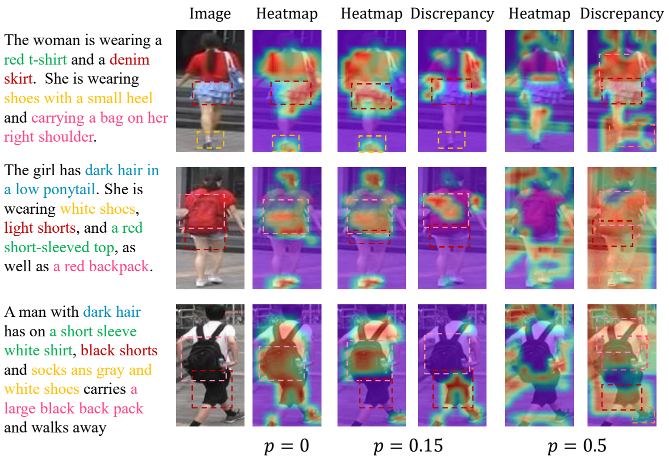
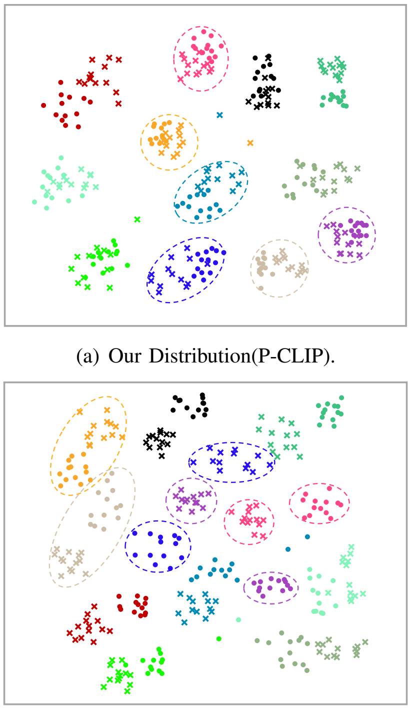


Image
Heatmap
Heatmap
Discrepancy
Heatmap
Discrepancy
The woman is wearing a red t-shirt and a denim skirt. She is wearing shoes with a small heel and carrying a bag on her right shoulder.
The girl has dark hair in a low ponytail. She is wearing white shoes, light shorts, and a red short-sleeved top, as well as a red backpack.
A man with dark hair has on a short sleeve white shirt, black shorts and socks ans gray and white shoes carries a large black back pack and walks away
$p = 0$
$p = 0.15$
$p = 0.5$

Fig. 12. Heatmap visualization.



(a) Our Distribution(P-CLIP).

(b) Baseline Distribution.

Fig. 13. Visualization for feature distribution by randomly sampling 12 identities of the CUHK-PEDES dataset.

differentiates between “long sleeved black shirt” and “black coat”, while Baseline is prone to confusion due to color information. Both of ‘P-CLIP(LLM)’ and ‘P-CLIP(Random)’ can effectively overcome the cross-view discrepancy, ‘right rear view’ and ‘left rear view’. Similar instances occur in other examples, especially for Query③, targeted pedestrian images covering front, rear and side views. In Query③, where ‘CLIP’ misidentifies “umbrella and white t-shirt” as “white patterned, long sleeve shirt”. We also notice that the Ranking of target individuals is elevated in ‘P-CLIP(Random)’ and ‘P-CLIP(LLM)’, indicating that our multi-view contrastive learning effectively discovers accurate correspondences between images and texts. Second, ‘P-CLIP(Random)’ demonstrates the cross-view matching ability, it is able to find the pedestrian with different views. In Query② and Query④, ‘P-CLIP(Random)’ accurately finds the images from different camera views, the cross-view matching capability of our ‘P-CLIP’ surpasses the ‘CLIP’. These results explicitly demonstrate the effectiveness of our proposed method, as our method mitigates the influence of view-specific bias.

To gain deeper insights into the model’s internal mechanisms, we visualize cross-modal attention distributions under varying replacement ratios ($p = 0, 0.15, 0.5$), as shown in Fig. 12. Under zero ratio ($p = 0$), the attention heatmap shows that P-CLIP successfully focuses on semantically critical regions (e.g., backpack, shorts, shoes), confirming effective cross-modal alignment and the model’s fundamental capability for cross-modal matching. At moderate ratio ($p = 0.15$), where partial textual information is omitted, the Cross-view Discrepancy Learning (CDL) module retains its ability to accurately attend to the corresponding masked visual content. However, when the replacement ratio increases to $p = 0.5$, substantial information loss challenges the model’s ability to recover precise semantic correspondences. Although severely deviated pseudo-captions reduce the precision of the cross-view attention map $C^{i,m}$, they do not completely disrupt the CDL system, as the resulting discrepancy maps still capture discriminative regions.

Fig. 13 presents t-SNE visualizations comparing feature distributions of image-text pairs between the baseline and our P-CLIP framework. In these visualizations, circles represent visual features while crosses denote textual embeddings, with identically colored samples corresponding to the same identity. Our analysis reveals that P-CLIP achieves feature distributions with superior structural properties. Specifically, compared to the baseline, P-CLIP produces more compact and well-separated clusters, demonstrating substantially reduced intra-class dispersion and enhanced inter-class separation. Furthermore, the significant reduction in erroneous cross-modal pairs indicates that our method learns more discriminative joint representations with improved alignment precision.

I. Failure cases analysis

Despite the success of P-CLIP, our analysis reveals specific failure modes in cross-modal alignment. First, the model exhibits limitations in fine-grained attribute disambiguation. As shown in Fig. 11, when processing visually similar categories, P-CLIP occasionally fails to focus on discriminative attributes such as color or shape. In Query①, for instance, while the query describes a “black bag,” the model retrieves persons holding purple bags with higher ranking, indicating confusion in capturing subtle texture differences between bags and umbrellas. Similarly, Query④ shows confusion between a “red top” and a “red bag,” further demonstrating the model’s difficulty in resolving local semantic ambiguities. Furthermore, we observe significant performance degradation under conditions of severe information loss. As illustrated in Fig. 12, when the masking ratio reaches $p = 0.5$, the model’s semantic reconstruction capability shows marked deterioration. This reveals a clear operational boundary of the proposed CDL module: its effectiveness diminishes substantially when critical visual or textual cues are extensively absent. In such cases, while the model may still recognize individual concepts, it systematically fails to integrate them into coherent cross-modal representations.

J. Limitations

The primary limitations of our P-CLIP framework can be summarized as follows: First, the multi-view generation and cross-modal attention mechanisms introduce additional computational overhead compared to the vanilla CLIP architecture. Although these components are crucial for achieving performance gains, they present practical challenges for deployment in real-time applications or resource-constrained environments. Future work will explore more efficient architectural designs, such as model distillation techniques and sparse attention mechanisms, to reduce computational costs while preserving performance advantages. Second, while our evaluation covers multiple standard benchmarks, these datasets may not fully capture the challenges of real-world deployment, including significant domain shifts, adversarial conditions, and open-vocabulary scenarios beyond the training distribution. Future research should therefore validate the framework's robustness across more diverse application domains and under more challenging generalization conditions.

VI. Conclusion

In this paper, we introduce and investigate a novel semi-supervised task of one-shot text-to-image person re-identification (one-shot TIReID). This task utilizes a single labeled image-text pair for each identity, along with a large set of unlabeled images, to develop a TIReID model. We propose a novel progressive discrepancy learning framework(P-CLIP) to learn discriminative and view-consistent person representations. We construct multi-view image-text pairs based on a single labeled data point and perform cross-view image-text matching in a shared embedding space. Additionally, we introduce a cross-view discrepancy module to explore discrepancies between different views, thereby obtaining view-consistent visual-textual correspondences. To guide image-text matching, we propose a compact cross-modal matching loss to optimize the dot product of text-to-image and image-to-text similarities. Experimental results on three datasets demonstrate the superiority and robustness of our proposed P-CLIP framework. This study has some limitations. How to enhance the diversity of text generated by LLMs, and how to use LLMs to inject controllable noise into texts are meaningful directions for future work. We will explore these topics in our subsequent research.

References

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

[2] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, and B. Ma, "Cross-modal knowledge adaptation for language-based person search," *IEEE Transactions on Image Processing*, vol. 30, pp. 4057–4069, 2021.

[3] S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "Vgsg: Vision-guided semantic-group network for text-based person search," *IEEE Transactions on Image Processing*, vol. 33, pp. 163–176, 2023.

[4] K. Niu, T. Huang, L. Huang, L. Wang, and Y. Zhang, "Improving inconspicuous attributes modeling for person search by language," *IEEE Transactions on Image Processing*, vol. 32, pp. 3429–3441, 2023.

[5] S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-driven fine-grained text-image person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 6032–6046, 2023.

[6] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023, pp. 2787–2797.

[7] J. Zuo, H. Zhou, Y. Nie, F. Zhang, T. Guo, N. Sang, Y. Wang, and C. Gao, "Ufinebench: Towards text-based person retrieval with ultra-fine granularity," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2024, pp. 22 010–22 019.

[8] Y. Qin, L. Huang, D. Peng, B. Jiang, J. T. Zhou, X. Peng, and P. Hu, "Trustworthy visual-textual retrieval," *IEEE Transactions on Image Processing*, 2025.

[9] Q. Zha, X. Liu, Y.-m. Cheung, S.-J. Peng, X. Xu, and N. Wang, "Ucpm: Uncertainty-guided cross-modal retrieval with partially mismatched pairs," *IEEE Transactions on Image Processing*, 2025.

[10] Y. Qin, C. Chen, Z. Fu, D. Peng, X. Peng, and P. Hu, "Human-centered interactive learning via mllms for text-to-image person re-identification," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 390–14 399.

[11] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 197–27 206.

[12] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2017, pp. 994–1002.

[13] S. Wang, C. Li, Z. Yan, W. Liang, Y. Yuan, and G. Wang, "Hada: Hyper-adaptive parameter-efficient learning for multi-view convnets," *IEEE Transactions on Image Processing*, vol. 34, pp. 85–99, 2025.

[14] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2990–2999.

[15] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 1233–1245, 2019.

[16] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, 2019.

[17] M. Liu, L. Qu, L. Nie, M. Liu, L. Duan, and B. Chen, "Iterative local-global collaboration learning towards one-shot video person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 9360–9372, 2020.

[18] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International conference on machine learning*, 2022, pp. 12 888–12 900.

[19] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[20] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, and M. Zhang, "Rasa: Relation and sensitivity aware representation learning for text-based person search," in *Proceedings of the International Joint Conferences on Artificial Intelligence*, E. Elkind, Ed., 2023, pp. 555–563.

[21] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[22] T. Fang, J. Hu, D. F. Wong, X. Wan, L. S. Chao, and T.-H. Chang, "Improving grammatical error correction with multimodal feature integration," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9328–9344.

[23] S. Constantin, F. I. Eyiokur, D. Yaman, L. Bärmann, and A. Waibel, "Multimodal error correction with natural language and pointing gestures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1976–1986.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[25] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European conference on computer vision*, 2018, pp. 686–701.

[26] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 2, pp. 1–23, 2020.

[27] W. Nie, C. Wang, H. Sun, and W. Xie, "Image-centered pseudo label generation forweakly supervised text-based person re-identification," in *Pattern Recognition and Computer Vision*, 2025, pp. 477–491.

[28] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, and X. Sun, "Contextual non-local alignment over full-scale representation for text-based person search," *arXiv preprint arXiv:2101.03036*, 2021.

[29] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning." in *Proceedings of the International Joint Conferences on Artificial Intelligence*, vol. 2, 2021, pp. 1068–1074.

[30] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language," in *Proceedings of the European conference on computer vision*, 2020, pp. 402–420.

[31] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, "Learning granularity-unified representations for text-to-image person re-identification," in *Proceedings of the 30th acm international conference on multimedia*, 2022, pp. 5566–5574.

[32] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, F. Lin, X. Sun, and X. Bai, "Conditional feature learning based transformer for text-based person search," *IEEE Transactions on Image Processing*, vol. 31, pp. 6097–6108, 2022.

[33] C. Wang, Z. Luo, Z. Zhong, and S. Li, "Divide-and-merge the embedding space for cross-modality person search," *Neurocomputing*, vol. 463, pp. 388–399, 2021.

[34] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 189–11 196.

[35] Z.-J. Zha, J. Liu, D. Chen, and F. Wu, "Adversarial attribute-text embedding for person search with natural language query," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1836–1846, 2020.

[36] C. Wang, Z. Luo, Y. Lin, and S. Li, "Improving embedding learning by virtual attribute decoupling for text-based person search," *Neural Comput. Appl.*, vol. 34, no. 7, pp. 5625–5647, 2022.

[37] C. Wang, Z. Luo, and S. Li, "Omni-granularity embedding network for text-to-image person retrieval," in *IEEE International Conference on Multimedia and Expo, ICME 2024*, 2024, pp. 1–6.

[38] M. Cao, Y. Bai, Z. Zeng, M. Ye, and M. Zhang, "An empirical study of clip for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 465–473.

[39] P. Zhang, X. Yu, X. Bai, and J. Zheng, "Visual perturbation for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 058–10 066.

[40] Z. Li, J. Li, Y. Shi, H. Ling, J. Chen, R. Wang, and S. Huang, "Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024, pp. 1047–1055.

[41] D. Gao, Y. Bai, M. Cao, H. Dou, M. Ye, and M. Zhang, "Semi-supervised text-based person search," *arXiv preprint arXiv:2404.18106*, 2024.

[42] Z. Song, G. Hu, and C. Zhao, "Diverse person: Customize your own dataset for text-based person search," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4943–4951.

[43] J. Sun, H. Fei, G. Ding, and Z. Zheng, "From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2341–2351.

[44] M. Cao, Z. Zeng, Y. Lu, M. Ye, D. Yi, and J. Wang, "An empirical study of validating synthetic data for text-based person retrieval," *arXiv preprint arXiv:2503.22171*, 2025.

[45] D. S. Raychaudhuri and A. K. Roy-Chowdhury, "Exploiting temporal coherence for self-supervised one-shot video re-identification," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 258–274.

[46] Y. Zhang, B. Ma, M. Li, Y. Liu, F. Chen, and J. Hou, "Pseudo-label estimation via unsupervised identity link prediction for one-shot person re-identification," *Pattern Recognition*, vol. 146, p. 110060, 2024.

[47] G. Wang, X. Huang, S. Gong, J. Zhang, and W. Gao, "Faster person re-identification: One-shot-filter and coarse-to-fine search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3013–3030, 2024.

[48] G. Han, X. Zhang, and C. Li, "One-shot unsupervised cross-domain person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1339–1351, 2024.

[49] J. Ye, S. Qin, Y. Li, H.-T. Zheng, S. Wang, and Q. Wen, "Corrections meet explanations: A unified framework for explainable grammatical error correction," *arXiv preprint arXiv:2502.15261*, 2025.

[50] C. Davis, A. Caines, Ø. E. Andersen, S. Taslimipoor, H. Yannakoudakis, Z. Yuan, C. Bryant, M. Rei, and P. Buttery, "Prompting open-source and commercial language models for grammatical error correction of English learner text," in *Findings of the Association for Computational Linguistics: ACL 2024*, Aug. 2024.

[51] M. Taieb-Maimon and L. Romanovskii-Chernik, "Improving error correction and text editing using voice and mouse multimodal interface," *International Journal of Human–Computer Interaction*, vol. 41, no. 8, pp. 4718–4741, 2025.

[52] X. Zhuang, Z. Jia, J. Li, Z. Zhang, L. Shen, Z. Cao, and S. Liu, "Mask-enhanced autoregressive prediction: Pay less attention to learn more," in *Forty-second International Conference on Machine Learning*, 2025.

[53] K. Zheng, J. Yang, S. Liang, B. Feng, Z. Liu, W. Ju, Z. Xiao, and M. Zhang, "ExLM: Rethinking the impact of $\texttt{[MASK]}$ tokens in masked language models," in *Forty-second International Conference on Machine Learning*, 2025.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[55] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

[56] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[57] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[58] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*, 2015, pp. 1–5.

[59] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," *arXiv preprint arXiv:2107.12666*, 2021.

[60] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua, "Dssl: Deep surroundings-person separation learning for text-based person retrieval," in *Proceedings of the Acm international conference on multimedia*, 2021, pp. 209–217.

[61] S. Yan, H. Tang, L. Zhang, and J. Tang, "Image-specific information suppression and implicit local alignment for text-based person search," *IEEE Transactions on neural networks and learning systems*, vol. 35, no. 12, pp. 17 973–17 986, 2024.

[62] W. Suo, M. Sun, K. Niu, Y. Gao, P. Wang, Y. Zhang, and Q. Wu, "A simple and robust correlation filtering method for text-based person search," in *Proceedings of the European conference on computer vision*, 2022, pp. 726–742.

[63] S. Li, M. Cao, and M. Zhang, "Learning semantic-aligned feature representation for text-based person search," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 2724–2728.

[64] X. Wu, W. Ma, D. Guo, T. Zhou, S. Zhao, and Z. Cai, "Text-based occluded person re-identification via multi-granularity contrastive consistency learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6162–6170.