Consensus Labelling: Prompt-Guided Clustering Refinement for Weakly Supervised Text-based Person Re-Identification

Chengji Wang *Member, IEEE*, Weizhi Nie, Hongbo Zhang *Senior Member, IEEE*, Hao Sun *Member, IEEE*, Mang Ye *Senior Member, IEEE*

***Abstract*—Weakly supervised text-based person re-identification aims to retrieve specific pedestrians based on textual descriptions without identity labels available during training. This task remains challenging due to the inherent cross-modal heterogeneity and lack of identity annotations. There is a common issue of modality gap in vision language models, which in turn affects the performance of downstream tasks such as cross-modal retrieval and multimodal clustering. Specifically, in our research and experiments, we found that there is a problem of inter-modal misalignment between image and text modalities. However, existing methods rely on mutual enhancement strategies between image and text clustering, leading to the accumulation of clustering noise and affecting the final retrieval performance. To address this issue, we propose a Consensus Labelling: Prompt-guided Clustering refinement (CLPC) framework for weakly supervised text-based person re-identification. Specifically, we introduce a textual inversion network to learn a pseudo token that captures visual context, which is then integrated into natural language sentences as personalized textual prompt. To further improve clustering quality, we introduce a Nearest Neighbor-Guided Pseudo Label Mining (NGPM) method, which uses the clusters derived from personalized textual prompts to refine the clustering of image features. Additionally, we design a Dynamic Margin Triplet (DMT) loss, where the margin is adaptively adjusted using a sigmoid-based function to enhance the model's ability to distinguish hard negative samples. We have also introduce a Normalized Distribution Matching (NDM) loss to minimize the KL divergence between the image-text matching scores and the normalized soft matching scores. The extensive experimental results on three public datasets have demonstrated the superiority of our method. Our code is available at https://github.com/LeviWeiZhi/CLPC.**

***Index Terms*—Text-based person re-identification, weakly supervised learning, contrastive learning, prompt learning.**

I. Introduction

TEXT-based person re-identification (TPReID) [1]–[5] aims to retrieve the target person based on the given text description. TPReID has received widespread attention from both academia and industry communities in recent years. Large-scale vision-language models [6]–[8], particularly the Contrastive Language-Image Pre-Training model (CLIP) [7], have demonstrated strong capabilities in learning multimodal representations. Consequently, an increasing number of studies [2], [4], [5], [9]–[14] have incorporated CLIP into TPReID by adopting its pre-trained encoders to map visual and textual features into a shared embedding space. While significant progress has been achieved in TPReID, most existing approaches rely on supervised learning and require complete annotations, such as identity labels. However, obtaining identity annotations for large-scale pedestrian datasets is both time-consuming and costly, rendering such approaches impractical for extensive camera networks. Weakly supervised TPReID [15], [16], in which only text–image pairs are available and identity annotations are absent during training, has attracted increasing attention from both academic and industrial communities in recent years.

The absence of identity annotations renders weakly super-

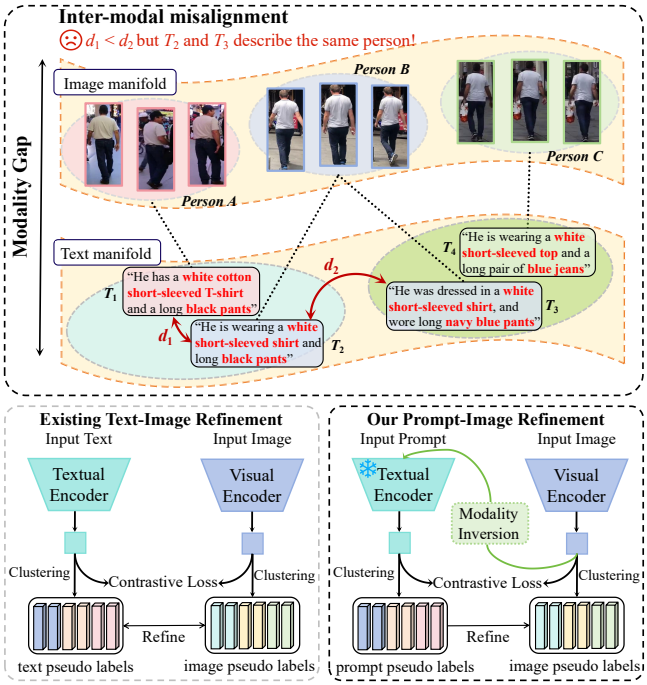

Fig. 1. Motivation and overview. **Above:** There is a modality gap between vision and text, leading to the problem of inter-modal misalignment. Specifically, image-text pairs may be assigned different pseudo labels. **Below:** The common practice of using text and image for mutual clustering refinement in existing methods is suboptimal, and refining image clustering through personalized prompts generated by modality inversion can improve performance.

This work was partially supported by the National Natural Science Foundation of China (No. 62407018), the China Postdoctoral Science Foundation (No. 2023M741305), the Natural Science Foundation of Fujian Province of China (No. 2025J01177), and in part by Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning under Grant 2025AISL004. (Corresponding author: Mang Ye.)

Chengji Wang, Weizhi Nie, Hao Sun 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; weizhi.nie@mails.ccnu.edu.cn; haosun@ccnu.edu.cn.

Hongbo Zhang is with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China (e-mail: zhanghongbo@hqu.edu.cn)

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

Manuscript received April xx, 20xx; revised August xx, 20xx.

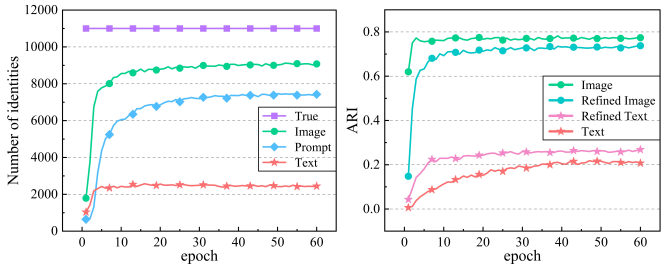

Fig. 2. **Left:** A comparison of the counts of image, prompt, and text pseudo labels and true identity counts on CUHK-PEDES. **Right:** ARI scores of original and refined image/text pseudo labels on CUHK-PEDES at each epoch.

vised TPReID extremely challenging, however, it offers a more cost-effective solution to the real world TPReID problems. Existing methods [15], [16] generate pseudo labels for both text and image modalities through unsupervised clustering within each modality. Based on these pseudo labels, they employ supervised contrastive learning to reduce intra-class variations at both intra-modality and cross-modality levels. To mitigate cross-modal matching ambiguity, these studies [15], [16] refine pseudo labels in one modality using information from the other modality to filter out unreliable pseudo labels. However, textual and visual features tend to form separate clusters within the embedding space [17]–[19], indicating a fundamental modality misalignment. As illustrated in Fig. 1, this modality gap is manifested as a clear separation between the distributions of text and visual features. Importantly, the learned representations fail to accurately preserve the semantic correspondence between paired images and texts. As a result, an image–text pair may be assigned inconsistent pseudo labels, which can significantly degrade model performance.

In this paper, we propose Consensus Labelling: Prompt-guided Clustering refinement (CLPC) for weakly supervised text-based person re-identification. We contend that textual features are inherently suboptimal, as existing benchmarks provide coarse and semantically ambiguous descriptions [12], [20], [21] that emphasize generic attributes while lacking identity-discriminative details. Moreover, the modality gap [17], [18], [22] between visual and textual representations leads to inter-modal misalignment, causing different identities to share similar textual embeddings. To support this claim, we conduct an extensive analysis of the clustering results in both the textual and visual modalities. Fig. 2 shows the Adjusted Rand Index (ARI) between text and image clustering results during training, it is impossible to achieve a consensus clustering results between two modalities. To address the inter-modal misalignment problem, we employ a lightweight modality inversion network [23] to inverse image features into their complementary modality. Instead of performing text clustering, we propose to cluster image features and inversed image embeddings. Our goal is to leverage explicit language prompts to activate cross-modal comprehension in pre-trained vision-language models. Specifically, we create personalized descriptions for invidual identities by refining the textual prompt, *i.e.* "a photo of a person", with inversed visual tokens. In addition to performing image-text matching, we also introduce supervised contrastive learning (SupCon) [24] to align the embedding of the personalized description as closely as possible with image embeddings of the same identity. We propose a Nearest Neighbor–Guided Pseudo label Mining (NGPM) method that leverages the clustering results of personalized descriptions to refine pseudo labels generated from image clustering. NGPM uses personalized descriptions to filter out unreliable labels. Consequently, CLPC can effectively avoid the inter-modal misalignment problem.

To improve learning under noisy supervision, we employ a momentum version of the model, which maintains a moving average of its parameters, to generate soft targets as additional supervision. During training, we use the normalized matching scores between images and personalized descriptions from the momentum model as soft labels. Unlike the one-hot labels used in InfoNCE [7] and SupCon [24], which primarily emphasize hard negative samples that may have weak correlation with the anchor, soft labels effectively guide the model to capture weak correlations across modalities. On one hand, they enhance the capability of unimodal encoders to better capture the semantic meanings of images and texts. On the other hand, they help reduce gradient contributions from noisy image–text pairs, thereby improving the overall robustness and performance of the model. To achieve this goal, we propose a Normalized Distribution Matching loss to minimize the KL divergence between image–text matching scores and the normalized soft matching scores. Metric learning with triplet loss is a commonly used hard example mining method. We observe that image–text matching scores dynamically increase during training, and a fixed margin may delay model convergence. Therefore, we introduce an adaptive margin generation strategy that gradually increases the margin throughout training. The main contributions can be summarized as follows:

- We reveal the inter-modal misalignment in weakly supervised TPReID and demonstrate that it obstructs the inference of consensus labels across modalities. We propose CLPC, which first incorporates prompt learning into weakly supervised TPReID to infer consensus labels between modalities.
- We propose a nearest neighbor-guided pseudo label mining method. During training, NGPM leverages personalized prompts to filtering untrustworthy pseudo labels, rather than relying on text clustering.
- We adopt a momentum distillation strategy to enhance unimodal feature learning under noisy supervision, which constructs soft matching scores to guide the model in exploring untagged weak correlations. A dynamic margin mechanism is applied in the triplet loss to help the model better distinguish hard negative samples.
- Extensive experimental evaluations and visualization analyses demonstrate that our method achieves superior performance on three benchmark datasets without requiring any additional identity annotations, outperforming existing weakly supervised approaches.

II. Related Work

A. Text-Based Person Re-Identification

Text-based person re-identification task was first introduced by Li *et al.* [1], who proposed the first challenging CUHK-PEDES dataset. Unlike typical image-based person re-identification [25], TPReID [9], [12], [26], [27] aims to find a specific person by free-form natural language. Early works [1], [12], [26]–[30] apply two unimodal encoders, such as CLIP [7], ViT [31], BERT [32], to project images and texts into a shared embedding space, following a supervised learning paradigm that requires image–text matching labels and identity annotations. These works focus on exploring fine-grained visual–textual representations and designing sophisticated loss functions to align features across the two modalities. Identity information is crucial for supervised text-based person re-identification. Zhao et al. [15] introduced a weakly supervised TPReID setting by removing identity annotations for both images and texts. The absence of identity annotations prevents weakly supervised TPReID from effectively mitigating inter- and intra-modality variations.

B. Weakly Supervised Text-Based Person Re-Identification

Identity annotations are costly, which limits the development of text-based person re-identification. To address this limitation, weakly supervised TPReID aims to build TPReID models without using identity annotations. CMMT [15] introduced weakly supervised text-based person re-identification for the first time, employing pseudo labels for self-training within each modality and using the Text-IoU score as similarity soft labels to facilitate cross-modal matching and hard sample mining. CPCL [16] introduces a prototype multimodal memory for cluster-level contrastive learning and instance-level cross-modal matching, and proposes a two-stage outlier pseudo label mining module that leverages existing image–text pairs to identify outlier samples. Both CMMT [15] and CPCL [16] refine each other through mutual clustering of images and texts, and both require maintaining memory of clustering centroid features for the image and text modalities. Recently, unsupervised TPReID task has attracted widespread attention from researchers. Li et al. [33] incorporate with Large Language Model (LLMs) and BLIP [34] to generate captions for pedestrian images, then selecting reliable image-text pair for fnetuning. How to synthesize large-scale image-text pairs for training has also attracted attention. Song et al. [35] utilize diffusion model to synthesize person images. Cao et al. [36] and Sun et al. [37] analyze the availability of synthetic data experimentally. In this work, we innovatively propose using personalized prompt clustering to refine image clustering, without requiring additional memory to store clustering centroid features.

C. Prompt Learning

Prompt learning [38]–[40], initially developed in natural language processing (NLP), enables task-specific customization of pre-trained models through textual instructions called prompts. While early implementations relied on manually designed prompts for individual tasks, recent advances automate prompt generation during fine-tuning, effectively mitigating instability and knowledge bias. With the rapid development of vision–language pre-training models (VLP) [7] and multimodal learning, prompt learning has been successfully extended to computer vision. The seminal work CoOp [41] first adapted prompt learning for vision–language models, with CoCoOp [42] subsequently enhancing generalization through conditional prompting mechanisms. Notable extensions include video understanding adaptations via optimized CLIP prompt vectors [43], and CLIP-ReID [44] designed a two-stage framework to generate coarse pedestrian descriptions for re-identification tasks. Inspired by these works, we propose using prompt learning to generate personalized pedestrian descriptions and refine image clustering through the results of personalized prompt clustering.

D. Textual Inversion

Originally developed for personalized text-to-image generation [23], textual inversion operates by identifying pseudo-tokens within embedding spaces that capture both abstract visual semantics and fine-grained details. This methodology has recently achieved success in zero-shot composed image retrieval tasks [45], [46] using networks trained on large-scale unlabeled image datasets. Building upon these advances, PromptSG [47] pioneered the adaptation of textual inversion for image-based person re-identification. Unlike previous implementations that focused on visual modality alignment, we are the first to apply this paradigm to cross-modal text-to-image person re-identification.

III. Methodology

A. Problem Statement and Notations

Let $\mathcal{X} = \{\mathcal{X}^{\mathcal{V}}, \mathcal{X}^{\mathcal{T}}\}$ be an identity-free image-text training dataset, where $\mathcal{X}^{\mathcal{V}} = \{x_1^{\mathcal{V}}, x_2^{\mathcal{V}}, ..., x_N^{\mathcal{V}}\}$ denotes the N pedestrain images, and $\mathcal{X}^{\mathcal{T}} = \{x_1^{\mathcal{T}}, x_2^{\mathcal{T}}, ..., x_N^{\mathcal{T}}\}$ denotes the N person descriptions. Let $f^{\mathcal{V}}(\cdot)$ and $f^{\mathcal{T}}(\cdot)$ denote the visual and textual encoders, respectively. Weakly supervised TPReID aims to build a cross-modal retrieval model by learning shared representations from the unlabeled image-text pairs $\mathcal{X}$, thereby enabling the accurate retrieval of the most semantically relevant person image given a textual query.

To achieve this goal, a common approach is to employ classical clustering methods, such as DBSCAN [48], to generate pseudo labels for each modality. Contrastive learning is then conducted to facilitate cross-modal matching. Specially, given the visual encoder $f^{\mathcal{V}}(\cdot)$ and textual encoder $f^{\mathcal{T}}(\cdot)$, we first extract visual features $\mathcal{F}^{\mathcal{V}} = \{f_1^{\mathcal{V}}, f_2^{\mathcal{V}}, ..., f_N^{\mathcal{V}}\}$ and textual features $\mathcal{F}^{\mathcal{T}} = \{f_1^{\mathcal{T}}, f_2^{\mathcal{T}}, ..., f_N^{\mathcal{T}}\}$. A clustering algorithm is then applied to generated pseudo labels for images and texts,

$$\begin{aligned} \mathcal{Y}^{\mathcal{V}} &= DBSCAN(\mathcal{F}^{\mathcal{V}}), \quad \mathcal{Y}^{\mathcal{V}} = \{y_i^{\mathcal{V}}\}_{i=1}^{N} \\ \mathcal{Y}^{\mathcal{T}} &= DBSCAN(\mathcal{F}^{\mathcal{T}}), \quad \mathcal{Y}^{\mathcal{T}} = \{y_i^{\mathcal{T}}\}_{i=1}^{N} \end{aligned} \tag{1}$$

where $y_i^{\mathcal{V}} \in \{-1, 0, 1, 2, \ldots\}$ and $y_i^{\mathcal{T}} \in \{-1, 0, 1, 2, \ldots\}$ denote the pseudo labels for images and texts, respectively, with -1 indicating un-clustered instances that are removed from the training set. Existing methods [15], [16] perform

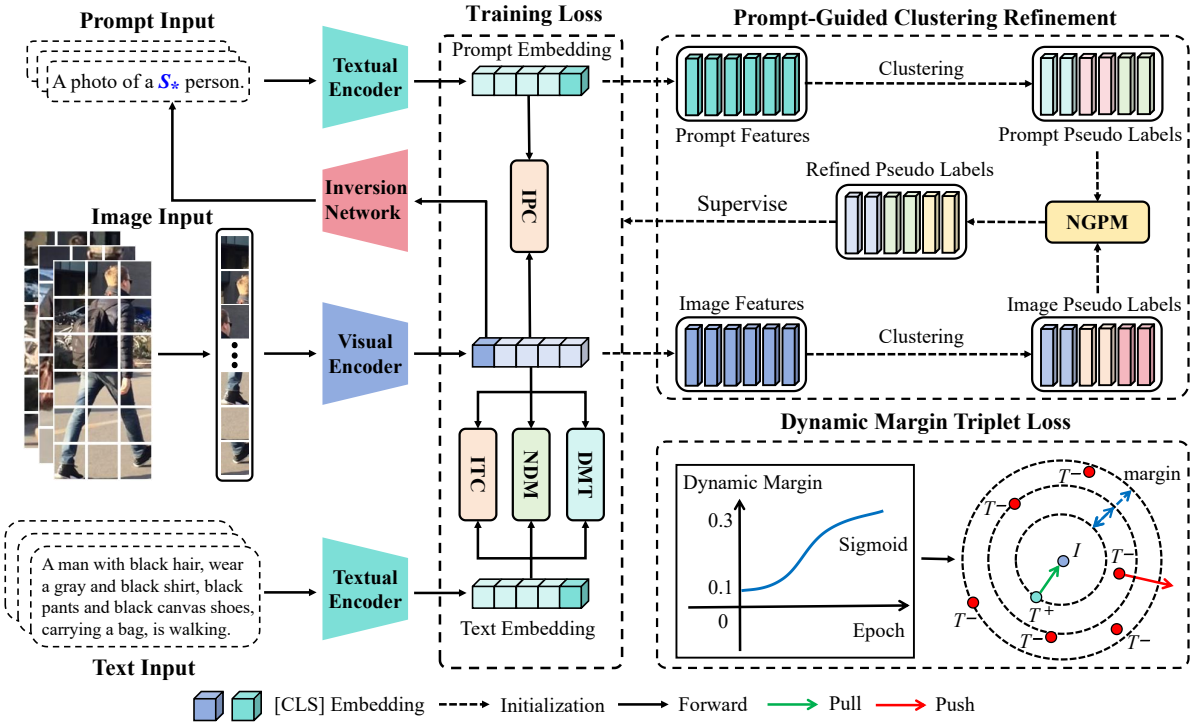

Fig. 3. Overview of our proposed Consensus Labelling: Prompt-guided Clustering refinement (CLPC) framework. The feature extraction backbone of CLPC comprises CLIP's visual and textual encoders. The Nearest Neighbor-Guided Pseudo Label Mining (NGPM) module leverages the clustering results of personalized textual prompts to refine image clustering through pairwise relationships. For instance-level matching, each image is aligned directly with its corresponding annotation text and personalized prompt via multimodal contrastive learning losses IPC and ITC. The Normalized Distribution Matching (NDM) loss is introduced to effectively bridge the gap between visual and textual modalities. Additionally, a Dynamic Margin Triplet (DMT) loss is proposed to enhance the model's ability to discriminate hard negative samples.

clustering on images and texts separately and construct pseudo labels individually. Two contrastive learning losses are applied separately to the image and text modalities for model training.

B. Prompt-Guided Consensus Labelling

Language, being inherently limited, fails to fully capture the boundlessness of visual attributes and context. This limitation in textual information hinders the creation of accurate labels. To address this, we employ the textual inversion technique [23], [47], which learns to represent visual context through a unique token.

Let $f_M : \mathcal{V} \rightarrow \mathcal{T}$ denote an inversion network that maps visual embeddings to the token embedding space. For the i-th image, we create the personalized description by enhancing the prompt, *e.g.*, "A photo of a person", with pseudo token

$$\begin{aligned} Token_i &= f_M(f_i^{\mathcal{V}}) \\ x_i^{\mathcal{P}} &= \text{“A photo of a } \underbrace{Token_i}_{\text{pseudo token}} \text{ person”} \end{aligned} \tag{2}$$

where $x_i^{\mathcal{P}}$ is the personalized description of image $x_i^{\mathcal{V}}$, $Token_i \in \mathcal{T}$ is the inversed visual embedding. The pseudo token is a projection of visual content, carrying rich fine-grained visual information to distinguish identities. We integrate the pseudo token into a language prompt to create a personalized description for each identity. $Token_i$ bears no relationship to an actual word but serves as a representation in the token embedding space. The textual inversion network is used only during training, while inference relies solely on global visual features for retrieval.

Personalized descriptions are fed into a frozen textual encoder, producing personalized textual embeddings represented as $\mathcal{F}^{\mathcal{P}} = \{f_1^{\mathcal{P}}, f_2^{\mathcal{P}}, ..., f_N^{\mathcal{P}}\}$. Instead of clustering the raw textual features, we cluster the personalized textual embeddings to generate pseudo labels,

$$\mathcal{Y}^{\mathcal{P}} = DBSCAN(\mathcal{F}^{\mathcal{P}}), \quad \mathcal{Y}^{p} = \{y_i^{\mathcal{P}}\}_{i=1}^{N} \tag{3}$$

where $y_i^{\mathcal{P}} \in \{-1, 0, 1, 2, \ldots\}$ denote cluster assignments, with -1 indicating un-clustered ones.

Nearest Neighbor-Guided Pseudo Label Mining. Intuitively, samples that are close in the original feature space are more likely to belong to the same cluster, and this neighborhood information can enhance clustering performance [49]. Therefore, we propose a Nearest Neighbor-Guided Pseudo Label Mining (NGPM) module to mine valuable un-clustered samples. As shown in Fig. 4, NGPM leverages the clustering results of personalized prompts to refine the pseudo labels in the image modality via cross-modal pairwise relationships.

Given an un-clustered images $x_i^{\mathcal{V}}$ ($y_i^{\mathcal{V}} = -1$). If the paired prompt $x_i^{\mathcal{P}}$ is un-clustered ($y_i^{\mathcal{P}} = -1$), we keep $x_i^{\mathcal{V}}$ un-clustered. In contrast, if $y_i^{\mathcal{P}} \neq -1$, we get the neighbours $\mathcal{C}_i^{\mathcal{V}}$ of images $x_i^{\mathcal{V}}$, $\mathcal{C}_i^{\mathcal{V}}$ is an image set which has same labels in the prompt clustering space, where

$$\mathcal{C}_i^{\mathcal{V}} = \{x_j^{\mathcal{V}} | y_j^{\mathcal{P}} = y_i^{\mathcal{P}} \wedge y_j^{\mathcal{V}} \neq -1 \wedge j \neq i\} \tag{4}$$

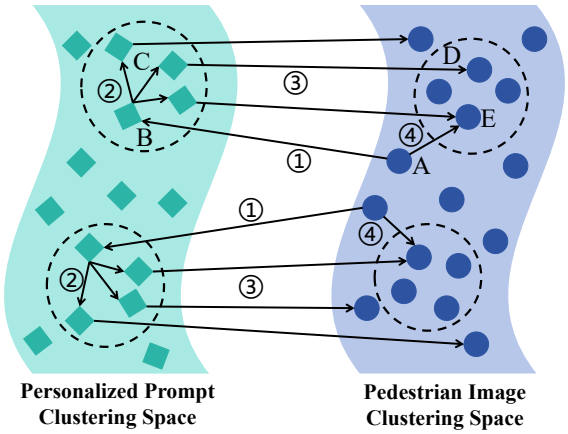

Fig. 4. Illustration of NGPM. ① For each un-clustered instance A, we first search its paired instance B in the other modality. ② If B is clustered, we then identify all other instances of the cluster to which B belongs, denoted as set C. ③ Next, we find the paired instance of C in the original modality, denoted as set D. ④ Finally, we identify the instance E in set D nearest to A and assign A to the cluster that E belongs to, if E is clustered.

In here, we remove the un-clustered images. The nearest instance can be obtained by

$$x_k^{\mathcal{V}} = \underset{x_j^{\mathcal{V}} \in \mathcal{C}_i^{\mathcal{V}}}{\arg\max} \quad cos(f_i^{\mathcal{V}}, f_j^{\mathcal{V}}) \tag{5}$$

where k denotes the index of the nearest instance, $cos(\cdot,\cdot)$ denotes the cosine similarity. The pseudo label of un-clustered image $x_i^{\mathcal{V}}$ is defined by

$$y_i^{\mathcal{V}} = \begin{cases} -1, & y_i^{\mathcal{P}} = -1 \quad \text{or} \quad \mathcal{C}_i^{\mathcal{V}} = \emptyset, \\ y_k^{\mathcal{V}}, & \text{otherwise.} \end{cases} \tag{6}$$

During NGPM, all the un-clustered instances in image modality are traversed.

Finally, the genrated pseudo labels can be represented by

$$\mathcal{Y} = NGPM(\mathcal{Y}^{\mathcal{V}}, \mathcal{F}^{\mathcal{V}}, \mathcal{Y}^{\mathcal{P}}, \mathcal{F}^{\mathcal{P}}), \quad \mathcal{Y} = \{\hat{y}_i\}_{i=1}^N \tag{7}$$

where $\mathcal{Y}$ denotes the refined pseudo labels. The consensus pseudo labels maintain cross-modal consistency according to the following assignment rule:

$$y_i^{\mathcal{P}} = y_i^{\mathcal{V}} = y_i^{\mathcal{T}} = \begin{cases} \hat{y}_i, & \text{if } \hat{y}_i \neq -1 \\ -1, & \text{otherwise} \end{cases}, \tag{8}$$

For each mini-batch consisting of N prompt-image-text triplets $\{(f_i^p, f_i^v, f_i^t)\}_{i=1}^N$, we assign corresponding pseudo label triplets $\{(y_i^p, y_i^v, y_i^t)\}_{i=1}^N$ that satisfy $y_i^p = y_i^v = y_i^t$ through cross-modal pairing relationship.

C. Multimodal Contrastive Learning

1) Image-Text Contrastive Learning: Since multiple images and text descriptions can share the same identity label, each image (or text) in a batch may have more than one positive counterpart. To bring each image closer to all text descriptions sharing the same pseudo label within the batch, we extend the original InfoNCE [50] loss as follows:

$$\mathcal{L}_{itc} = \mathcal{L}_{itc}^{i2t} + \mathcal{L}_{itc}^{t2i}. \tag{9}$$

$$\mathcal{L}_{itc}^{i2t} = -\log \frac{\sum_{i=1}^{N^p} \exp(sim(f^v, f_i^{t+})/\tau)}{\sum_{j=1}^{N} \exp(sim(f^v, f_j^t)/\tau)}, \tag{10}$$

$$\mathcal{L}_{itc}^{t2i} = -\log \frac{\sum_{i=1}^{N^p} \exp(sim(f^t, f_i^{v+})/\tau)}{\sum_{j=1}^{N} \exp(sim(f^t, f_j^v)/\tau)}. \tag{11}$$

where $sim(,)$ denotes the cosine similarity, f^v represents the image feature, f^{t+} indicates the positive text feature sharing the same pseudo identity label as f^v, N is batch size, and N^p is the number of positive samples. The temperature parameter τ controls the strength of penalties on negative samples.

2) Image-Prompt Contrastive Learning: The pseudo token can faithfully capture the semantic content of an image only when its pseudo textual features are well aligned with the corresponding visual features. Our goal is to encourage pseudo token to capture visual details specific to the same identity. Therefore, we optimize it using a bidirectional contrastive loss as follows:

$$\mathcal{L}_{ipc} = \mathcal{L}_{ipc}^{i2p} + \mathcal{L}_{ipc}^{p2i}. \tag{12}$$

$$\mathcal{L}_{ipc}^{i2p} = -\log \frac{\sum_{i=1}^{N^p} \exp(sim(f^v, f_i^{p+})/\tau)}{\sum_{j=1}^{N} \exp(sim(f^v, f_j^p)/\tau)}, \tag{13}$$

$$\mathcal{L}_{ipc}^{p2i} = -\log \frac{\sum_{i=1}^{N^p} \exp(sim(f^p, f_i^{v+})/\tau)}{\sum_{j=1}^{N} \exp(sim(f^p, f_j^v)/\tau)}. \tag{14}$$

In summary, the multimodal contrastive learning loss consists of the following two components:

$$\mathcal{L}_{mcl} = \mathcal{L}_{itc} + \lambda \mathcal{L}_{ipc}. \tag{15}$$

D. Normalized Distribution Matching

We propose a normalized distribution matching loss, where a momentum model generates soft labels to guide learning. This loss minimizes the KL divergence between image-text matching scores and normalized soft matching scores, thereby improving cross-modal alignment without identity supervision.

For a mini-batch of N image-text pairs, we first compute the normalized cross-modal similarity distribution via a softmax over cosine similarities:

$$p_{i,j} = \frac{\exp(sim(f_i^v, f_j^t)/\tau)}{\sum_{k=1}^{N} \exp(sim(f_i^v, f_k^t)/\tau)}, \tag{16}$$

where $sim(\cdot,\cdot)$ denotes the cosine similarity, τ is a temperature parameter controlling distribution sharpness, and f_i^v, f_j^t denote the visual and textual features in the batch, respectively.

Different from the one-hot pseudo label distribution obtained via clustering, we construct a normalized soft label distribution $q_{i,j}$ as follows:

$$q_{i,j} = \alpha p_{i,j}^s + (1-\alpha)\widetilde{q}_{i,j} \tag{17}$$

where $\widetilde{q}_{i,j} = y_{i,j} / \sum_{k=1}^{N} y_{i,k}$ is the matching probability based on pseudo labels. $\alpha \in [0,1]$ is a weight parameter controlling the ratio of $\widetilde{q}_{i,j}$ and $p_{i,j}^s$. Soft label $p_{i,j}^s$ is the similarity probability of the visual feature to the personalized description feature, which can be formulated as:

$$p_{i,j}^s = \frac{\exp(sim(\hat{f}_i^v, \hat{f}_j^p)/\tau_s)}{\sum_{k=1}^{N} \exp(sim(\hat{f}_i^v, \hat{f}_k^p)/\tau_s)}, \tag{18}$$

where τ_s is a temperature parameter, $\hat{f}_i^v$ and $\hat{f}_j^p$ respectively represent visual feature and personalized description feature extracted by momentum encoder.

The NDM loss from image to text is defined as the KL divergence between the predicted and target distributions:

$$\mathcal{L}_{ndm}^{i2t} = D_{KL}(p_i \| q_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right), \quad (19)$$

where ϵ is a small number to avoid numerical problems.

Symmetrically, we compute the text-to-image loss $\mathcal{L}_{ndm}^{t2i}$ by exchanging the roles of f^v and f^t in Eq. (16) and Eq. (19). The final bi-directional NDM loss is given by:

$$\mathcal{L}_{ndm} = \mathcal{L}_{ndm}^{i2t} + \mathcal{L}_{ndm}^{t2i}. \quad (20)$$

E. Dynamic Margin Triplet Loss

1) Dynamic Margin Mechanism: As shown on the left in Fig. 5, we observe that during training, the similarity of both positive image-text pairs and hardest negative image-text pairs initially increases and then stabilizes. Therefore, the traditional triplet loss [51] with a fixed margin cannot effectively distinguish hard samples. To address this issue, we introduce a dynamic margin mechanism that uses a sigmoid function to adaptively adjust the margin based on the current training epoch. The margin m is defined as:

$$\begin{aligned} m &= \beta + \gamma \cdot Sigmoid(E-\theta) \\ &= \beta + \frac{\gamma}{1+e^{-(E-\theta)}}, \end{aligned} \quad (21)$$

where E denotes the current training epoch, β is the initial margin value set to 0.1, γ controls the range of margin growth and is set to 0.2, and θ serves as a bias term to shift the midpoint of the sigmoid, set to 10. This formulation ensures a small initial margin to facilitate convergence and a progressively larger margin to enhance discriminative capability in later training.

2) Hardest Negative Triplet Loss: To enhance the model's ability to learn more discriminative features and better distinguish samples with similar but distinct pseudo identity labels, we adopt a hardest negative sample mining strategy. The triplet loss is formulated as:

$$\mathcal{L}_{dmt}^{t2i} = \sum_{i=1}^{N}\left[m - sim\left(V_a^i, T_p\right) + sim\left(V_a^i, T_n\right)\right]_+ \quad (22)$$

$$\mathcal{L}_{dmt}^{t2i} = \sum_{i=1}^{N}\left[m - sim\left(T_a^i, V_p\right) + sim\left(T_a^i, V_n\right)\right]_+ \quad (23)$$

where $[x]_+ = max(x, 0)$, $sim(\cdot,\cdot)$ denotes the cosine similarity, and m is the dynamic margin. V_a^i and T_a^i represent the visual and textual anchors, respectively; T_p and V_p are the corresponding positive samples, while T_n and V_n are the hardest negative samples with pseudo labels different from the anchors. The final dynamic margin triplet loss is defined as the sum of both directions:

$$\mathcal{L}_{dmt} = \mathcal{L}_{dmt}^{i2t} + \mathcal{L}_{dmt}^{t2i}. \quad (24)$$

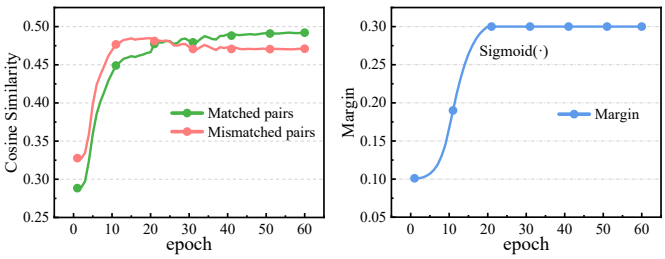

Fig. 5. The similarity variation trend of matched positive image-text pairs and mismatched hardest negative image-text pairs on CUHK-PEDES dataset.

The hardest negative mining strategy focuses on the most confusing negative samples in each batch, forcing the model to learn discriminative cross-modal representations, promoting better clustering, and improving model performance.

Optimization. In summary, the overall optimization objective function for our framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{mcl} + \mathcal{L}_{ndm} + \mathcal{L}_{dmt}. \quad (25)$$

IV. Experiment

In this section, we first introduce the three public datasets and evaluation metrics we used in the experiment. We then provided a detailed description of the implementation, including network structure, training procedure, and hyperparameters. Next, we present the comparative results and ablation studies. Finally, we conduct some additional analysis on our method, such as pseudo label analysis, hyperparameter analysis, and visualization.

A. Experiment Settings

1) Datasets: In this work, we evaluate our method on three text-based person re-identification datasets. **CUHK-PEDES** [1] contains 40,206 images and 80,412 text descriptions of 13,003 identities. Following the official protocol, the dataset is split into 34,054 images and 68,108 descriptions of 11,003 identities for training, 3,078 images with 6,158 descriptions for 1,000 identities for validation, and 3,074 images with 6,156 descriptions of another 1,000 identities for testing. **ICFG-PEDES** [29] includes more identities and textual descriptions, comprising 54,522 images of 4,102 identities. It is divided into a training set of 34,674 image-text pairs for 3,102 identities and a testing set of 19,848 image-text pairs for the remaining 1,000 identities. **RSTPReid** [30] consists of 20,505 images from 4,101 identities captured by 15 cameras. Each identity is represented by five distinct images taken from different cameras, and each image is annotated with two textual descriptions. Following the official split, the training, validation and test sets contain 3,701, 200 and 200 identities, respectively.

2) Evaluation Metrics: To comprehensively evaluate model performance, we adopt two complementary metric frameworks. The primary evaluation employs the standard Rank-k accuracy (with k=1, 5, 10), which measures top-k matching accuracy. To further ensure rigorous assessment, we incorporate two precision-oriented metrics: mean Average Precision

TABLE I
PERFORMANCE COMPARISONS WITH SOTA METHODS ON CUHK-PEDES. THE "G" AND "L" IN THE "TYPE" COLUMN REPRESENT GLOBAL/LOCAL MATCHING METHODS. "-" DENOTES THAT NO REPORTED RESULT IS AVAILABLE.

Method	Reference	Type	Image Encoder	Text Encoder	Rank-1	Rank-5	Rank-10	mAP	mINP
Fully Supervised Text-Based Person Re-Identification									
CTL [52]	TCSVT23	G	ViT-Base	BERT	69.47	87.13	92.13	60.56	-
CFine [12]	TIP23	L	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
Wei *et al.* [53]	arXiv23	G	RN50	BERT	71.59	87.95	92.45	65.03	49.97
VGSG [11]	TIP23	G	CLIP-ViT	CLIP-Xfomer	71.38	86.75	91.86	67.91	-
IRRA [9]	CVPR23	G	CLIP-ViT	CLIP-Xfomer	73.38	89.93	93.71	66.13	50.24
CFAM [54]	CVPR24	G	CLIP-ViT	CLIP-Xfomer	72.87	88.61	92.87	64.92	-
TBPS-CLIP [55]	AAAI24	G	CLIP-ViT	CLIP-Xfomer	73.54	88.19	92.35	65.38	-
SAMC [56]	TIFS24	G	CLIP-ViT	CLIP-Xfomer	74.03	89.18	93.31	68.42	55.01
LSPM [57]	TMM24	G	CLIP-ViT	CLIP-Xfomer	74.38	89.51	93.42	67.74	53.09
FSRL [58]	ICMR24	L	CLIP-ViT	CLIP-Xfomer	74.65	89.77	94.03	67.49	-
PTMI [4]	TIFS25	L	CLIP-ViT	CLIP-Xfomer	76.02	89.93	94.14	70.85	57.85
GAHR [3]	TIFS25	G	CLIP-ViT	CLIP-Xfomer	76.64	90.46	94.10	66.81	-
CAMel [2]	TIFS25	G	SG-Former	BERT	77.24	91.80	95.16	68.32	-
DCFL [5]	TIFS25	L	ViT-Base	BERT	80.39	94.98	97.94	75.40	-
Weakly Supervised Text-Based Person Re-Identification									
CMMT [15]	ICCV21	G	RN50	LSTM	57.10	78.14	85.23	-	-
CAIBC [59]	MM22	L	RN50	BERT	58.64	79.02	85.93	-	-
CPCL [16]	arXiv24	G	CLIP-ViT	CLIP-Xfomer	70.03	87.28	91.78	63.19	47.54
CLPC (Ours)	-	G	CLIP-ViT	CLIP-Xfomer	**73.90**	**89.27**	**93.25**	**65.89**	**49.94**

TABLE II
PERFORMANCE COMPARISONS WITH SOTA METHODS ON ICFG-PEDES.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
Fully Supervised Text-Based Person Re-Identification						
CFine [12]	L	60.83	76.55	82.42	-	-
Wei *et al.* [53]	G	60.93	77.96	84.11	36.44	7.79
VGSG [11]	G	63.05	78.43	84.36	-	-
IRRA [9]	G	63.46	80.25	85.82	38.06	7.93
CFAM [54]	G	62.17	79.57	85.32	36.34	-
TBPS-CLIP [55]	G	65.05	80.34	85.47	39.83	-
LSPM [57]	G	64.40	79.96	85.41	42.60	11.65
FSRL [58]	L	64.01	80.42	85.86	39.64	-
PTMI [4]	L	66.54	81.50	86.50	46.05	14.60
GAHR [3]	G	68.69	82.83	87.40	42.10	-
CAMeL [2]	G	68.70	83.11	88.32	41.58	-
DCFL [5]	L	73.02	91.98	96.13	49.74	-
Weakly Supervised Text-Based Person Re-Identification						
CMMT [15]	G	54.27	71.17	77.86	33.17	5.73
CPCL [16]	G	62.60	79.07	84.46	36.16	6.31
CLPC (Ours)	G	**65.16**	**80.74**	**86.03**	**37.80**	**6.87**

TABLE III
PERFORMANCE COMPARISONS WITH SOTA METHODS ON RSTPREID.

Method	Type	Rank-1	Rank-5	Rank-10	mAP	mINP
CFine [12]	L	50.55	72.50	81.60	-	-
Wei *et al.* [53]	G	56.65	77.40	84.70	45.27	26.02
IRRA [9]	G	60.20	81.30	88.20	47.17	25.28
CFAM [54]	G	59.40	81.35	88.50	46.04	-
TBPS-CLIP [55]	G	61.95	83.55	88.75	48.26	-
FSRL [58]	L	60.20	81.40	88.60	47.38	-
PTMI [4]	L	63.35	81.95	89.10	52.46	32.36
GAHR [3]	G	68.85	86.50	91.10	53.60	-
CAMeL [2]	G	68.50	87.40	92.70	53.61	-
Weakly Supervised Text-Based Person Re-Identification						
CMMT [15]	G	52.25	76.45	84.55	41.98	22.00
CPCL [16]	G	58.35	81.05	87.65	45.81	23.87
CLPC (Ours)	G	**61.60**	**82.55**	**88.65**	**47.53**	**25.63**

(mAP), which reflects overall ranking quality, and mean Inverse Negative Penalty (mINP), which specifically quantifying hard negative sample discrimination. Higher values of Rank-k, mAP, and mINP indicate better performance.

3) Implementation Details: We adpot the pre-trained CLIP [7] model as the backbone network, which consists of the CLIP-ViT-B/16 image encoder and text encoder, *i.e.*, CLIP text Transformer. In addition, our framework incorporates a randomly initialized textual inversion network, implemented as a lightweight three-layer MLP with dropout and ReLU, which outputs a 512-dimensional pseudo text embedding and generates one pseudo token per image. To enhance input diversity during training, we apply data augmentation strategies to both modalities. For images, we employ random horizontal flipping, random cropping, and random erasing. For texts, we perform token-level augmentation by randomly masking a subset of tokens with a fixed probability. All input images are resized to 384×128, and the maximum textual sequence length is set to 77 tokens. The model is trained for 60 epochs with a batch size of 64 using the Adam optimizer, an initial learning rate of 1×10^{-6} and cosine learning rate decay. The temperature parameter τ is set to 0.02, and λ in Eq. (15) is set to 0.5 across all datasets. Following CMMT [15], we perform clustering using DBSCAN [48] before each training epoch. The entire framework is implemented using PyTorch and experiments are conducted using a single RTX4090 GPU with 24GB VRAM.

B. Comparisons with the State-of-the-art Methods

In this section, we present comparison results with state-of-the-art methods on three public benchmark datasets.

1) Performance Comparisons on CUHK-PEDES: We first evaluate our proposed method on the CUHK-PEDES dataset. As shown in Table I. Our CLPC achieves a Rank-1 accuracy of 73.90% and an mAP of 65.89%. Compared with existing weakly supervised methods, our method CLPC achieves state-of-the-art performance across all metrics. Furthermore, when compared to fully supervised methods that also use CLIP as the backbone, our method outperforms CFAM [54] and TBPS-CLIP [55] in all metrics, surpasses IRRA [9] method in Rank-1 accuracy, and achieves comparable performance to LSPM [57] and FSRL [58]. These results demonstrate the effectiveness of the proposed CLPC framework.

2) Performance Comparisons on ICFG-PEDES: The experimental results on the ICFG-PEDES dataset are presented in Table II. Our CLPC achieves 65.16%, 80.74% and 86.03% on Rank-1, Rank-5 and Rank-10, respectively. Compared with existing weakly supervised methods, our approach achieves the best performance across all metrics, surpassing the recent state-of-the-art method CPCL [16] by +2.56% in Rank-1 accuracy and +1.64% in mAP. Notably, both the CMMT [15] and CPCL [16] independently cluster the images and texts to generate pseudo labels and maintain memory for clustering centroid features in each modality. In contrast, our method does not require additional memory. Moreover, when compared to the fully supervised methods, CLPC outperforms several methods, such as IRRA [9], LSPM [57] and FSRL [58], across Rank-1, Rank-5, and Rank-10 metrics.

3) Performance Comparisons on RSTPReid: We further report the experimental results on the RSTPReid dataset in Table III. Compared with existing weakly supervised methods, our method CLPC achieves state-of-the-art performance across all metrics. Specifically, our proposed CLPC surpasses the recent cluster-based method CPCL [16] by +3.25% in Rank-1 accuracy and +1.72% in mAP, demonstrating its effectiveness in reducing cross-modal ambiguity and enhancing clustering consensus. Additionally, CLPC outperforms most fully supervised methods in terms of Rank-1 accuracy and mAP, further highlighting its strong competitiveness.

In summary, our CLPC method has achieved superior performance across all three benchmark datasets, establishing state-of-the-art results under weakly supervised settings. Moreover, it remains competitive with most fully supervised methods, further demonstrating the effectiveness of our proposed CLPC framework.

C. Comparisons on Generalization Ability

To evaluate the cross-domain generalization ability of CLPC, we conduct cross-dataset retrieval experiments, where the model is trained on one dataset and directly tested on another. Cross-domain retrieval is particularly challenging due to the distribution shift between different datasets. As summarized in Table IV, the proposed weakly supervised CLPC achieves competitive cross-domain performance under both C→I and I→C settings. Specifically, CLPC attains 45.23% Rank-1 when trained on CUHK-PEDES and tested on ICFG-PEDES, and 37.22% Rank-1 in the reverse direction, outperforming or matching most fully supervised methods, while achieving Rank-1 improvements of 3.3% and 6.2% over IRRA [9] under the C→I and I→C settings, respectively, despite using fewer model parameters than IRRA. These results demonstrate the strong cross-domain robustness of CLPC under significant distribution shifts. Specifically, the synergy of the textual inversion network, prompt-guided pseudo label refinement, and the joint optimization of DMT and NDM enhances discriminative cross-modal alignment, enabling stable generalization across datasets.

TABLE IV
DOMAIN GENERALIZATION COMPARISON OF STATE-OF-THE-ART METHODS. HERE "C" AND "I" DENOTES CUHK-PEDES AND ICFG-PEDES, RESPECTIVELY.

	Method	Reference	Rank-1	Rank-5	Rank-10
C → I	Fully Supervised Text-Based Person Re-Identification				
	LGUR [20]	MM22	34.25	52.58	60.85
	ASAMN [60]	TIP23	30.22	50.51	59.59
	IRRA [9]	CVPR23	41.89	61.56	69.04
	RDE [61]	CVPR24	48.17	66.25	73.66
	GAHR [3]	TIFS25	49.00	66.34	73.59
	DCFL [5]	TIFS25	45.64	70.21	79.27
	Weakly Supervised Text-Based Person Re-Identification				
	CLPC(Ours)	-	**45.23**	**63.53**	**71.08**
I → C	Fully Supervised Text-Based Person Re-Identification				
	LGUR [20]	MM22	25.44	44.48	54.39
	ASAMN [60]	TIP23	17.99	35.30	44.75
	IRRA [9]	CVPR23	31.04	52.18	63.53
	RDE [61]	CVPR24	38.14	59.23	68.50
	GAHR [3]	TIFS25	40.60	62.46	71.62
	DCFL [5]	TIFS25	28.57	52.73	63.97
	Weakly Supervised Text-Based Person Re-Identification				
	CLPC(Ours)	-	**37.22**	**58.85**	**68.02**

D. Ablation Study

In this section, we conduct a comprehensive empirical analysis to thoroughly evaluate the contributions of different components in CLPC. The Rank-1, Rank-5, and Rank-10 accuracies (%) are reported in Table V. Specifically, our Baseline solely includes visual and textual encoders initialized by CLIP, trained with the InfoNCE loss [50]. The $\mathcal{L}_{mcl}$ represents the multimodal contrastive learrning loss, $\mathcal{L}_{dmt}$ represents the dynamic margin triplet loss, and $\mathcal{L}_{ndm}$ denotes the normalized distribution matching loss. Finally, NGPM refers to the nearest neighbor-guided pseudo label mining module.

1) Contributions of Multimodal Contrastive Learning: We first investigate the contribution of Multimodal Contrastive Learning (MCL) in strengthening the alignment between images and their associated textual descriptions, including personalized text prompts. As shown in No.0 and No.1 in Table V, incorporating $\mathcal{L}_{mcl}$ improves the Rank-1 accuracy by 0.8% over to baseline. This result demonstrates that MCL effectively enhances cross-modal alignment.

2) Contributions of Normalized Distribution Matching: To demonstrate the effectiveness of our proposed Normalized Distribution Matching (NDM) loss, simply adding the NDM to $\mathcal{L}_{mcl}$ improves the Rank-1 accuracy by 1.26% and 1.35% on the two datasets, respectively. Furthermore, the efficacy

TABLE V
ABLATION STUDIES ON THE EFFECTIVENESS OF EACH COMPONENT OF CLPC ON CUHK-PEDES AND ICFG-PEDES.

No.	Methods	Components				CUHK-PEDES			ICFG-PEDES		
		$\mathcal{L}_{mcl}$	$\mathcal{L}_{ndm}$	$\mathcal{L}_{dmt}$	NGPM	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
0	Baseline					70.45	88.21	92.84	59.23	77.18	83.56
1	$+\mathcal{L}_{mcl}$	✓				71.23	88.69	93.23	60.09	78.24	84.59
2	$+\mathcal{L}_{mcl}+\mathcal{L}_{ndm}$	✓	✓			72.49	88.85	93.12	61.44	79.30	84.95
3	$+\mathcal{L}_{mcl}+\mathcal{L}_{dmt}$	✓		✓		73.08	88.48	92.87	64.19	79.89	85.45
4	$+\mathcal{L}_{mcl}+\mathcal{L}_{ndm}+\mathcal{L}_{dmt}$	✓	✓	✓		73.21	88.82	93.13	64.68	80.45	85.88
5	$+\mathcal{L}_{mcl}+\mathcal{L}_{ndm}+$ NGPM	✓	✓		✓	72.68	88.87	93.39	61.94	79.52	85.08
6	$+\mathcal{L}_{mcl}+\mathcal{L}_{dmt}+$ NGPM	✓		✓	✓	73.65	88.51	93.02	64.72	80.46	85.53
7	CLPC	✓	✓	✓	✓	**73.90**	**89.27**	**93.25**	**65.16**	**80.74**	**86.03**

TABLE VI
COMPARISON OF CLUSTERING REFINEMENT USING DIFFERENT FEATURES IN THE NGPM MODULE ON CUHK-PEDES AND ICFG-PEDES.

Methods	Rank-1	Rank-5	Rank-10	mAP
(a) CUHK-PEDES				
CLPC w/o NGPM	73.21	88.82	93.13	65.63
CLPC w/ NGPM(text)	73.26	88.52	93.03	65.60
CLPC w/ NGPM(prompt)	**73.90**	**89.27**	**93.25**	**65.89**
(b) ICFG-PEDES				
CLPC w/o NGPM	64.68	80.45	85.88	37.69
CLPC w/ NGPM(text)	64.80	80.65	85.72	37.76
CLPC w/ NGPM(prompt)	**65.16**	**80.74**	**86.03**	**37.80**

TABLE VII
PERFORMANCE COMPARISON OF OUR METHOD USING DIFFERENT NUMBERS OF PSEUDO TOKENS IN PERSONALIZED DESCRIPTIONS.

Method	CUHK-PEDES		ICFG-PEDES	
	Rank-1	mAP	Rank-1	mAP
$M=1$	**73.90**	**65.89**	**65.16**	**37.80**
$M=2$	73.73	65.53	63.65	36.87
$M=4$	73.68	65.48	63.48	36.88
$M=8$	73.57	65.62	63.41	36.86
$M=16$	73.77	65.58	63.81	36.91

of NDM is further revealed via the experimental results of No.3 *vs*. No.4 and No.6 *vs*. No.7 in Table V. These results demonstrate that the proposed NDM loss effectively promotes feature alignment between the two modalities.

3) Contributions of Dynamic Margin Triplet Loss: The effectiveness of the proposed Dynamic Margin Triplet (DMT) loss is evidenced by the comparisons in No.1 *vs*. No.3, No.2 *vs*. No.4 and No.5 *vs*. No.7 of Table V. Specifically, compared with No.1, introducing the DMT loss yields Rank-1 accuracy improvements of 1.85% and 4.1% on the two datasets, respectively. Similarly, the comparison between No.5 and No.7 shows further gains of 1.22% and 3.22%. These results clearly demonstrate that the DMT loss enhances the model's discriminative ability, particularly in distinguishing hard negative samples.

4) Contributions of NGPM Module: According to the experimental results of No.4 and No.7 in Table V, the NGPM can promotes the Rank-1 accuracy by 0.69% and 0.48% on two datasets, respectively. In addition, the experimental results of No.2 *vs*. No.5 and No.3 *vs*. No.6 also demonstrate the effectiveness of the NGPM. These results demonstrate that NGPM effectively refines image clustering with the aid of personalized prompt features, thereby enhancing the quality of consensus pseudo labels.

5) Analysis of Clustering Refinement: To intuitively demonstrate the effectiveness of our proposed prompt-guided clustering refinement method, we compare improvements in image clustering performance achieved by utilizing clustering results based on personalized prompt features and those based on textual features on two datasets. As shown in Table VI, shows that the use of personalized textual prompt features for refining image clustering consistently outperforms the use of text features. This result confirms that prompt features not only enhance clustering performance more effectively but also mitigate clustering noise caused by inter-modal misalignment between the image and text modalities.

6) Analysis of the Number of Pseudo Tokens: To analyze the effect of pseudo token quantity, we vary the number of learned pseudo tokens M in the personalized descriptions and report the results on CUHK-PEDES and ICFG-PEDES in Table VII. The best performance occurs at M=1, where CLPC achieves 73.90% Rank-1 and 65.89% mAP on CUHK-PEDES, and 65.16% Rank-1 and 37.80% mAP on ICFG-PEDES. As M increases, the results show a slight decline, suggesting that a single pseudo token already captures the key global visual semantics for cross-modal alignment. Additional tokens introduce redundant interactions and increase optimization difficulty. Overall, these results indicate that lightweight prompt refinement is sufficient, while larger pseudo token dimensionality offers no further benefit.

E. Further Analysis

1) Analysis of Hyper-parameter: Our method involves four key hyper-parameters, and their influence on retrieval performance is quantitatively evaluated in Fig. 6 and Fig. 7. As shown in Fig. 6, the retrieval performance of our model is largely insensitive to these two parameters, with optimal results achieved at $\lambda = 0.5$ and $\alpha = 0.9$. Fig. 7 further presents retrieval performance under different combinations of β and γ in the DMT loss, with the best results observed at $\beta = 0.1$ and $\gamma = 0.2$. It is worth noting that when $\gamma = 0$, it represents the triplet loss using the traditional fixed margin. As shown in the figure, methods using a fixed margin consistently

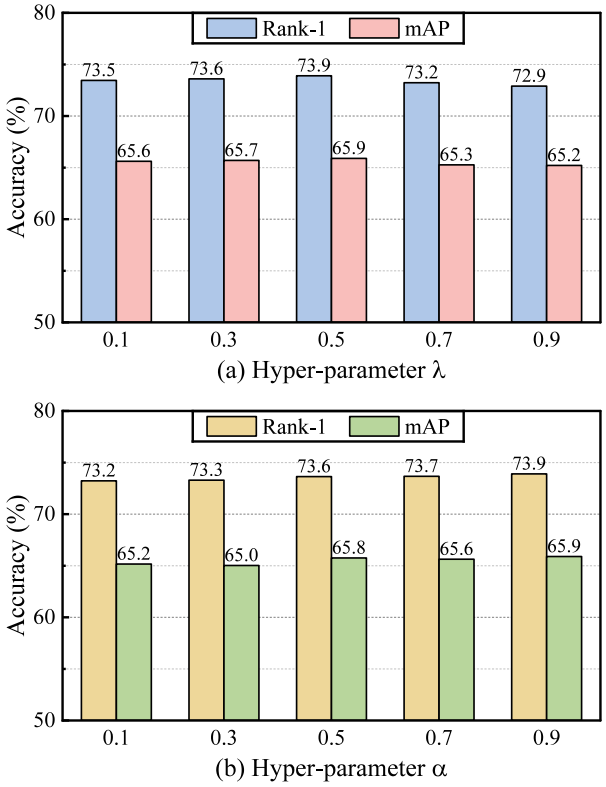

Fig. 6. The influence of two important hyper-parameters under different values on the CUHK-PEDES dataset.

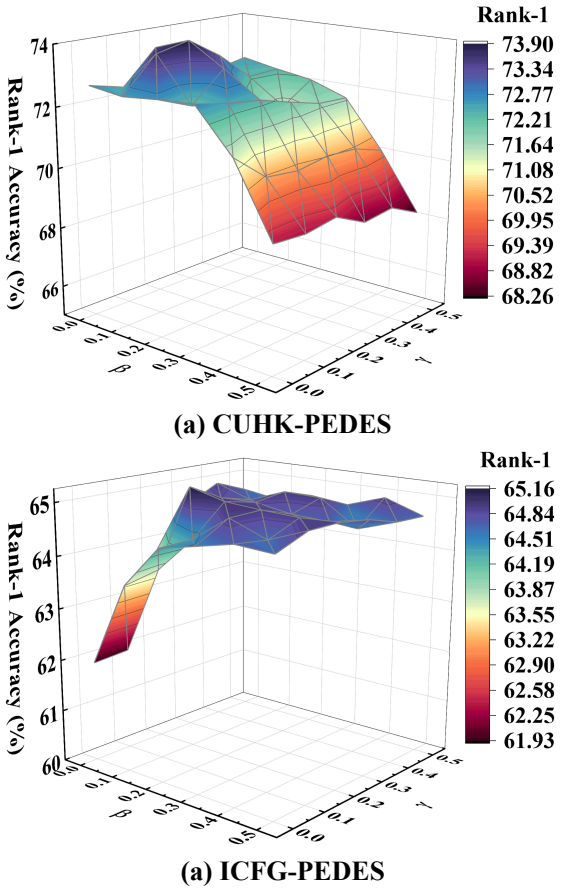

Fig. 7. Rank-1 accuracy of CLPC under different combinations of β and γ in the DMT loss on the CUHK-PEDES and ICFG-PEDES datasets.

achieve lower Rank-1 accuracy compared to our approach with the dynamic margin mechanism, which further validates the effectiveness of our proposed dynamic margin mechanism.

2) Comparisons on Computational Efficiency: We evaluate the computational efficiency of CLPC by comparing its parameter size, FLOPs, runtime, and GPU memory consumption with several representative CLIP-based methods, as summarized in Table VIII. Notably, the reported runtime measures one full training epoch or a complete inference pass over the CUHK-PEDES dataset, respectively. Since CLPC performs feature clustering at the beginning of each training epoch, the clustering time is included in the reported training runtime. For fair comparison, the same clustering procedure is also applied to all baseline methods. As shown in Table VIII, CLPC maintains a model size comparable to the CLIP [7] backbone and smaller than the other compared methods, while requiring 26.17 GFLOPs. Notably, the introduced textual inversion network is lightweight, contributing fewer than 1M additional parameters. In terms of runtime, CLPC achieves training and inference efficiency comparable to fully supervised baselines. The peak GPU memory consumption during training is slightly lower than that of IRRA [9], and the inference memory footprint remains equally lightweight. Overall, CLPC offers a favorable trade-off between effectiveness and efficiency.

3) Distance Distributions Visualization: To verify that our approach effectively mitigates modality differences, we visualize the intra-class and inter-class distance distributions on the CUHK-PEDES and ICFG-PEDES datasets, as shown in Fig. 8. By comparing (1) with (3) and (2) with (4), it is clear that our method increases the separation between intra-class and inter-class distances, as indicated by the relative positions of the vertical lines, where $\delta_1 < \delta_3$ and $\delta_2 < \delta_4$. For instance, comparing (2) with (4), we observe that the intra-class distances in (4) are notably smaller than those in (2), indicating that our method reduces intra-class variations. Conversely, the inter-class distances in (4) are larger than those in (2), demonstrating that our method increases inter-class separability. Overall, compared with the baseline, our method effectively alleviates the distribution discrepancy between images and texts, leading to more compact intra-class samples and larger inter-class separations.

4) Feature Distributions Visualization: We visualized the image and text feature distribution with t-SNE [62] in the 2-D embedding space, which contains 10 randomly selected identities. Circles and diamonds represent textual and visual

TABLE VIII
COMPARISON OF PARAMETER COUNTS (M), FLOPS (GFLOPS), RUNTIME (MINUTES) AND GPU MEMORY (G) OF SEVERAL METHODS.

Method	Params	FLOPs	Runtime		GPU Memory	
			Train	Infer	Train	Infer
CLIP	149.62	20.21	6.06	0.12	7.83	1.72
IRRA	194.54	26.32	7.04	0.12	12.49	2.07
RDE	152.78	21.07	7.76	0.21	8.52	1.75
CLPC(Ours)	150.41	26.17	7.02	0.12	11.03	2.07

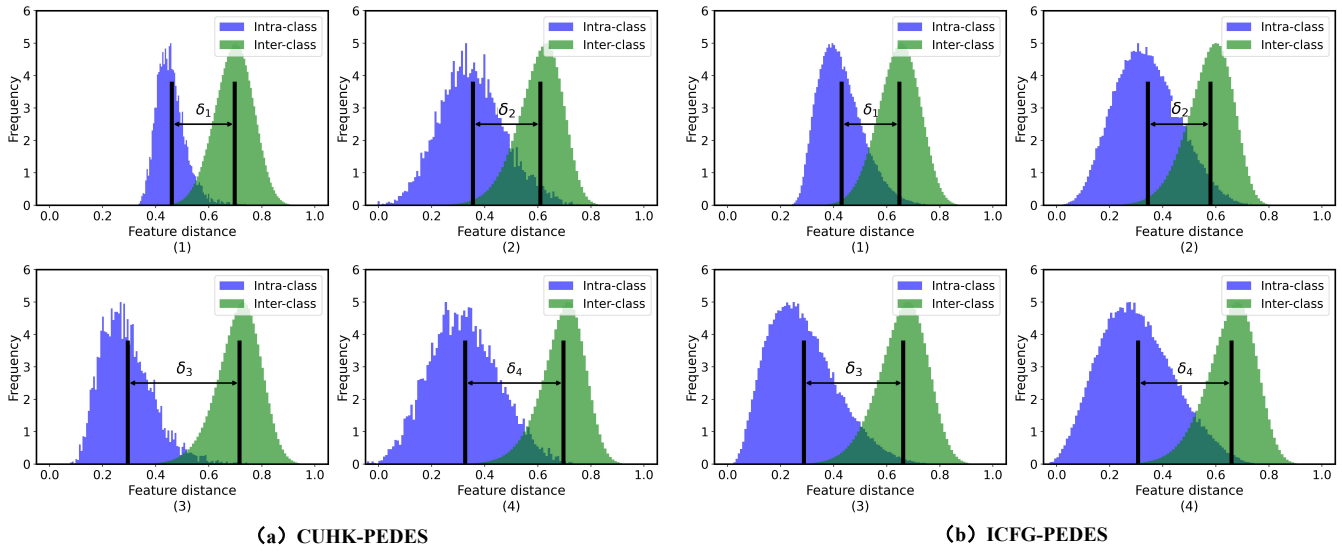

Fig. 8. (1–4) present the feature distance distributions of inter-modal (left) and image intra-modal (right) representations learned by the Baseline and our CLPC on the CUHK-PEDES and ICFG-PEDES datasets. (1) and (2) correspond to the Baseline, while (3) and (4) correspond to our approach.

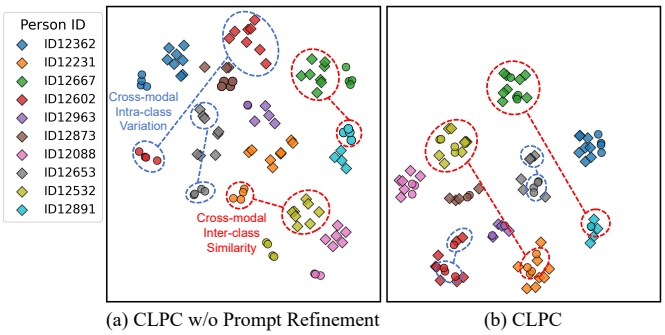

Fig. 9. The t-SNE Visualization for feature distribution by randomly sampling 10 identities of the CUHK-PEDES dataset.

Fig. 10. The ARI metric of image pseudo labels on CUHK-PEDES dataset at each epoch during the training process.

samples, respectively, with samples of the same color sharing the same identity. As shown in Fig. 9, compared with not using prompt refinement, our method has better intra-class compactness and inter-class separability, and the feature distributions of the same identities from different modalities are also closer. This indicates that our method CLPC reduces cross-modal differences and enhances cross-modal clustering consistency.

5) Analysis of Pseudo Label: We employ the Adjusted Rand Index (ARI) metric to evaluate the accuracy of the generated image pseudo labels on the CUHK-PEDES dataset at each training epoch. ARI measures the consistency between the predicted clustering assignments and the ground-truth labels, where higher values indicate more accurate and reliable pseudo labels. As illustrated in Fig. 10, the introduction of the personalized prompt-guided clustering refinement module improves the quality of image pseudo labels. This improvement demonstrates that the proposed module not only enhances clustering accuracy but also provides more trustworthy supervisory signals, thereby facilitating more effective model training.

6) Qualitative Analysis: To demonstrate the effectiveness of our proposed CLPC, we conducted a qualitative analysis in Fig. 11. For each query text, we compared the top-10 retrieval results between baseline and our method with and without prompt refinement. As shown in the figure, compared with the baseline, our method exhibits better retrieval results. Further combined with the prompt refinement module, our CLPC has achieved more accurate retrieval results. The above results indicate that our method can enhance cross-modal matching performance in weakly supervised scenarios, and our proposed prompt-guided clustering refinement module can further improve the reliability of pseudo labels and enhance clustering performance. However, our model also returned incorrect retrieval results, which may be attributed to our method only performs coarse-grained global matching and ignores some critical fine-grained information. In the future work, we plan to investigate this issue.

V. Conclusion

In this paper, we present the CLPC, a prompt-guided clustering refinement method for generating consensus labels

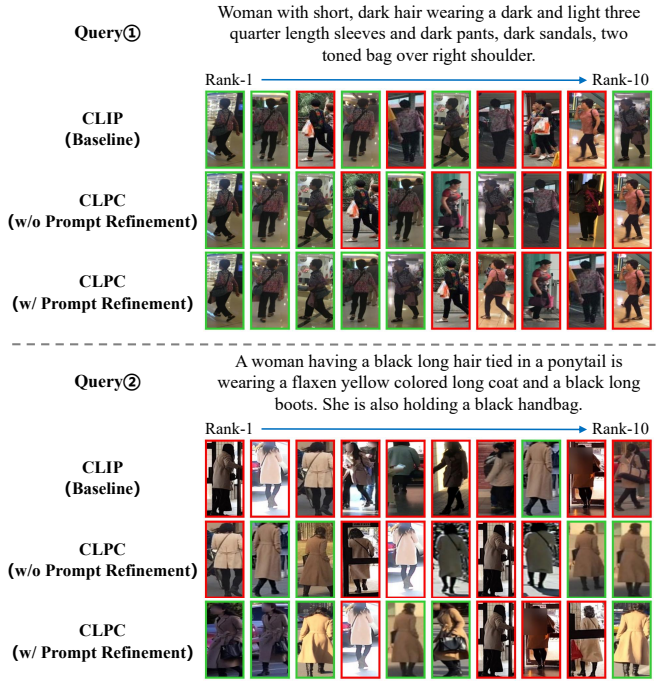

Fig. 11. Comparison of Rank-10 retrieved results on CUHK-PEDES (top) and ICFG-PEDES (bottom) between Baseline (the first row), CLPC without prompt refinement (the second row) and CLPC with prompt refinement(the third row) for each text query. The matched and mismatched person images are marked with green and red rectangles, respectively.

between images and texts. We reveal the inter-modal misalignment problem, which leads to the generation of conflicting pseudo labels. Observing that image clustering results contain less noise, we introduce a modality inversion network to construct personalized descriptions based on visual embeddings. Our method pioneers a new paradigm in weakly supervised text-based person re-identification. CLPC clusters the generated personalized descriptions, rather than the uncertain textual descriptions. We further propose a nearest neighbor-guided pseudo label mining method to refine image clustering results using the personalized descriptions. In addition, we introduce a dynamic margin triplet loss that adaptively adjusts the margin between positive and negative pairs to better discriminate hard negative samples.Compared to methods relying on clustering results from both visual and textual modalities, the labels generated by CLPC are more stable. Extensive experiments on three benchmark datasets demonstrate the superiority of our method, surpassing existing approaches across all metrics. The proposed approach highlights the effectiveness of combining prompt engineering with clustering refinement for representation learning under limited supervision, providing new insights for future research on multimodal alignment in weakly supervised settings. In futher work, we will further explore the interpretability of generated personalized descriptions and utilize them to reduce modality gap.

REFERENCES

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5187–5196.

[2] H. Yu, J. Wen, and Z. Zheng, "Camel: Cross-modality adaptive meta-learning for text-based person retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4651–4663, 2025.

[3] C. Qi, X. Yang, N. Wang, and X. Gao, "Granularity-aware hyperbolic representation for text-based person search," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 5745–5757, 2025.

[4] Z. Lu, R. Lin, Y.-P. Tan, and H. Hu, "Prompt-guided transformer and mllm interactive learning for text-based pedestrian search," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 7181–7196, 2025.

[5] S. You, C. Chen, Y. Feng, H. Liu, Y. Ji, and M. Ye, "Diverse co-saliency feature learning for text-based person retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 5465–5477, 2025.

[6] 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.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 2021, pp. 8748–8763.

[8] H. Zhao, L. Qi, and X. Geng, "Cilp-fgdi: Exploiting vision-language model for generalizable person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 2132–2142, 2025.

[9] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2787–2797.

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

[11] 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, 2024.

[12] 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.

[13] 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.

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

[15] S. Zhao, C. Gao, Y. Shao, W.-S. Zheng, and N. Sang, "Weakly supervised text-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 395–11 404.

[16] Y. Zheng, X. Zhao, C. Lan, X. Zhang, B. Huang, J. Yang, and D. Yu, "Cpcl: Cross-modal prototypical contrastive learning for weakly supervised text-based person re-identification," *arXiv preprint arXiv:2401.10011*, 2024.

[17] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.

[18] S. Yamaguchi, D. Feng, S. Kanai, K. Adachi, and D. Chijiwa, "Post-pre-training for modality alignment in vision-language foundation models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4256–4266.

[19] M. Mistretta, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bagdanov, "Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion," *arXiv preprint arXiv:2502.04263*, 2025.

[20] 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.

[21] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 22 010–22 019.

[22] W. Nie, C. Wang, H. Sun, and W. Xie, "Image-centered pseudo label generation for weakly supervised text-based person re-identification," in *Pattern Recognition and Computer Vision*, Springer. Singapore: Springer Nature Singapore, 2024, pp. 477–491.

[23] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[25] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[26] 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.

[27] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning." in *IJCAI*, 2021, pp. 1068–1074.

[28] K. You, W. Chen, C. Wang, H. Sun, and W. Xie, "Cross-modal feature fusion-based knowledge transfer for text-based person search," *IEEE Signal Processing Letters*, vol. 31, pp. 2230–2234, 2024.

[29] 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.

[30] 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 29th ACM International Conference on Multimedia*, 2021, pp. 209–217.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] 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. International Joint Conferences on Artificial Intelligence Organization*, 2024, pp. 1047–1055.

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

[35] 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.

[36] 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.

[37] 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.

[38] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[39] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[40] T. Shin, Y. Razeghi, R. L. L. IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[41] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[42] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[43] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 105–124.

[44] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 1, 2023, pp. 1405–1413.

[45] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347.

[46] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 305–19 314.

[47] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang, "A pedestrian is worth one prompt: Towards language guidance person re- identification," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 343–17 353.

[48] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 96, 1996, pp. 226–231.

[49] J. Wen, Z. Zhang, Z. Zhang, Z. Wu, L. Fei, Y. Xu, and B. Zhang, "Dimc-net: Deep incomplete multi-view clustering network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3753–3761.

[50] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[51] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[52] H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive transformer learning with proximity data generation for text-based person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7005–7016, 2024.

[53] W. Li, L. Tan, P. Dai, and Y. Zhang, "Prompt decoupling for text-to-image person re-identification," *arXiv preprint arXiv:2401.02173*, 2024.

[54] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22 010–22 019.

[55] 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.

[56] Z. Lu, R. Lin, and H. Hu, "Mind the inconsistent semantics in positive pairs: Semantic aligning and multimodal contrastive learning for text-based pedestrian search," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6409–6424, 2024.

[57] J. Li, M. Jiang, J. Kong, X. Tao, and X. Luo, "Learning semantic polymorphic mapping for text-based person retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 10 678–10 691, 2024.

[58] D. Wang, F. Yan, Y. Wang, L. Zhao, X. Liang, H. Zhong, and R. Zhang, "Fine-grained semantics-aware representation learning for text-based person retrieval," in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ser. ICMR '24, 2024, p. 92–100.

[59] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, and Y. Li, "Caibc: Capturing all-round information beyond color for text-based person retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5314–5322.

[60] 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.

[61] 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 197–27 206.

[62] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.