KPDM: Key Phrase Dynamic Masking for Robust Text-to-Image Person Retrieval

Shaofeng You[1], Tianle Miao[1], Qihang Chen[1], Xin Li[2], Zhuo Cheng[1], Dapeng Luo[1]*

[1] School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China
[2] Intelligent Technology Co., Ltd of Chinese Construction Third Engineering Bureau
{yousf93887, miaotianle, chenqh, chengzhuo, luodapeng}@cug.edu.cn
lixin@sjzn.com.cn

Abstract

Text-to-image person re-identification (TIReID) aims to retrieve the most relevant pedestrian images from an image gallery based on natural language descriptions. Recent studies have achieved significant performance improvements by leveraging Masked Language Modeling (MLM) to align fine-grained information through local matching. However, in the text feature extraction, randomly masking text tokens may disrupt the semantic relationships between these local tokens, leading to feature misalignment; on the other hand, from an image feature perspective, redundant patches in pedestrian images hinder the information interaction across modalities. Moreover, the presence of noisy image-text pairs further complicates the learning process, as the model may be misled into recognizing incorrect patterns. To address these issues, we propose a robust fine-grained local alignment framework based on Key Phrase Dynamic Mask (KPDM). First, we strengthen the semantic relationships between text tokens by implementing a "adjective + noun" phrase-level masking strategy, and design a frequency-based masked language loss (FMLM) to supervise fine-grained semantic-level local alignment. Second, we integrate cross-layer importance estimation to highlight key pedestrian image representations while removing redundant image features. Third, we propose a trusted consensus partitioning mechanism, utilizing intra-identity image-text similarity distributions to identify noisy pairs, enhancing the model robustness. Extensive experiments show that our method achieves 67.95% Rank-1 and 51.88% mAP on the RSTPReid dataset, exceeding the previous state-of-the-art by 2.6% and 1%. Furthermore, KPDM achieves Rank-1 accuracies of 75.97% on the CUHK-PEDES dataset and 67.78% on the ICFG-PEDES dataset, outperforming earlier methods.

Introduction

In the evolving landscape of smart cities and public safety, text-to-image person re-identification (TIReID) is a critical research domain bridging computer vision and natural language processing (Miech et al. 2021; Wang et al. 2015). It aims to utilize natural language descriptions to retrieve a target individual from a large-scale image gallery.

Early methods (Farooq et al. 2022; Yan et al. 2023c) typically involve extracting global features from both images and text, which are then utilized to learn a joint embedding space. In this space, the features can be directly aligned using cross-modal matching functions. However, these approaches often struggle with capturing fine-grained details and contextual nuances in complex visual and textual representations. To address these limitations, recent approaches (Chen, Xu, and Luo 2018; Suo et al. 2022) focus on local matching strategies. Most of these methods leverage Masked Language Modeling (MLM) (Fu et al. 2022) to extract fine-grained features from both images and text, enabling precise alignment of masked representations.

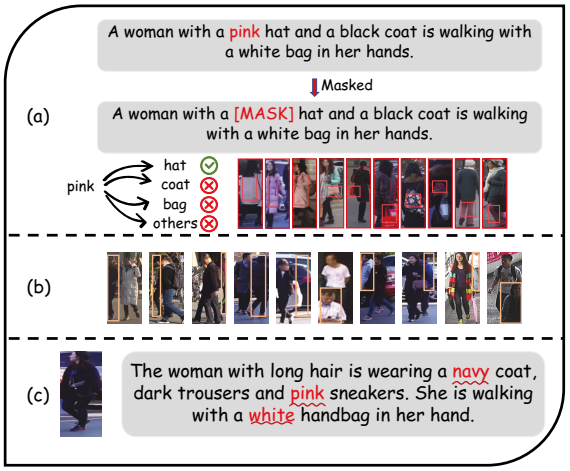

Figure 1: Existing problems in TIReID methods. (a) The omission of critical semantic cues leads to misalignment. The objects enclosed in red boxes are those associated with the masked word "pink". (b) Redundant background information with the yellow boxes. (c) Inaccuracies in text descriptions when compared to the corresponding image.

However, despite these promising developments, current methodologies still face several limitations. In the text feature representations, most MLM-based methods (Yang et al. 2023) work by randomly masking words to force the model to focus on the knowledge associated with the masked words, potentially omitting critical semantic cues and disrupting word relationships. For instance, as shown in Fig. 1(a), when "pink" is masked, the model may strug-

*Corresponding author

gle to accurately predict the relationship between the word "pink" and "hat". Instead, it generates incorrect associations with "coat" or "bag". From the perspective of image feature representation, existing methods often neglect the impact of redundant visual features. As shown in Fig. 1(b), some images contain redundant background distractions, such as unrelated pedestrians or vehicles, which can negatively impact the matching process.

Moreover, due to the subjective nature of textual descriptions, the describer may be influenced by environmental factors such as pose, viewpoint, and illumination, leading to significant discrepancies between the generated textual descriptions and the corresponding image content (Zhang et al. 2023; Lin et al. 2024). As depicted in Fig. 1(c), inaccuracies in textual descriptions can result in imperfect image-text pairs, impairing the alignment accuracy of TIReID models. A recent advance (Qin et al. 2024) addresses textual misdescriptions by utilizing dual-grained alignment quality evaluation scores, achieving superior results in distinguishing imperfect image-text pairs. However, due to the lack of semantic-level local alignment, the model struggles to capture semantic associations between words, resulting in potential feature misalignment.

To address these challenges, we propose a robust fine-grained TIReID model based on the key phrase dynamic mask (KPDM). Rather than relying on traditional word random masking, KPDM designs a key phrase masking rule focused on "adjective+noun" pairs to enhance the model's ability to grasp nuanced semantic relationships. By identifying and masking these key phrases, KPDM encourages the model to concentrate on the context in which these phrases occur, allowing for a more accurate representation. Building upon this, we further propose a key visual feature extraction mechanism based on cross-layer token selection. This mechanism identifies relevant image patches while minimizing the influence of background features. Moreover, we propose a trusted consensus partitioning mechanism to identify noisy image-text pairs, thereby enhancing the model's robustness in learning inaccurate representations. A similarity distribution matching (SDM) (Jiang and Ye 2023) loss is utilized to assess the matching quality of image-text pairs and facilitate the refinement of noisy image-text pairs.

Extensive experiments on three public benchmarks validate our method's superiority. On RSTPReid, our method achieves 67.95% R-1 and 51.88% mAP, surpassing the state-of-the-art RDE (Qin et al. 2024) by 2.6% and 1.0%. On the CUHK-PEDES dataset, our approach reaches 75.97% R-1 and 68.14% mAP. Similarly, on the ICFG-PEDES dataset, our approach achieves 67.78% R-1 and 40.30% mAP, exceeding the performance of earlier methods. Notably, employing the larger CLIP-ViT-L/14 backbone further enhances performance. The method achieves 78.02% R-1 on CUHK-PEDES and 69.13% R-1 on ICFG-PEDES, comprehensively outperforming existing approaches.

The main contributions of this paper are as follows:

- We propose a Key Phrase Dynamic Mask (KPDM) method to perform a fine-grained local matching process. In the text domain, we apply a phrase-level masking strategy to identify and mask "adjective + noun" combinations to emphasize key textual representations.Building on this, we introduce a frequency-based masked language loss (FMLM) to supervise cross-modal local alignment.
- We propose to integrate cross-layer importance estimation in the image domain to select the most relevant key patches.
- We introduce a trusted consensus partitioning mechanism that partitions training data into clean and noisy subsets based on the level of image-text SDM loss, enhancing the model's robustness in noisy data scenarios.

Related Work

Text-to-Image Person Re-identification

Existing TIReID methods can be classified into two main groups based on the alignment level: global matching methods (Zheng et al. 2020b; Zhu et al. 2021) and local matching methods (Niu et al. 2020; Zheng et al. 2020a; Shu et al. 2022). The former initially extract image-text features using pre-trained single-modality models such as ViT (Dosovitskiy et al. 2020), BERT (Devlin et al. 2019), and DeiT (Touvron et al. 2021). They introduced innovative losses such as CMPM/C loss (Zhang and Lu 2018) and Triplet Ranking Loss (Faghri et al. 2017) to align the global representations of both modalities within a shared latent space.

To achieve finer-grained alignment, some methods designed additional local feature learning mechanisms, utilizing human segmentation (Chen et al. 2022), body parts (Wang et al. 2020), color information (Wang et al. 2022b), and other explicit features (Wu et al. 2021; Wang et al. 2022a) to achieve significant improvements. However, these methods introduce extra computational complexity due to the increased feature extraction during inference.

Masked Language Modeling in TIReID

Recent approaches (Jiang and Ye 2023; Fujii and Tarashima 2023; Yan et al. 2024) attempted to leverage masked language modeling (MLM) for implicit local alignment, leading to notable improvements. The core idea is that by hiding certain words, the model learns to predict the masked tokens, thereby enabling fine-grained alignment. However, the aforementioned methods use random masking rules, which fail to effectively capture semantic relationships between words, often leading to misaligned image-text features. In this paper, we propose a key phrase masking strategy utilizing the "adjective + noun" structure to preserve these semantic structures, allowing key textual features to be closely associated in the form of phrases.

Moreover, existing methods primarily focus on text tokens, often neglecting the redundant background patches. In terms of extracting key patches from visual tokens, some methods attempt to leverage hyperpixels (Ding et al. 2021; Li et al. 2021) or incorporate additional information like pose (Jing et al. 2020) to segment images. However, these methods necessitate pre-labeled datasets or the use of pre-trained segmentation models, which complicates their practical application. Recent studies (Shang et al. 2024; Ye et al. 2024) introduced various complex merging frameworks that

utilize the [CLS] token to filter visual tokens prior to multimodal interaction. However, such approaches also require additional fine-tuning on the TIReID datasets. In this paper, we propose a training-free method for selecting visual tokens solely on their relationship with the [CLS] token, enhancing the comprehensive preservation of visual token information in TIReID.

Method

The overall architecture of our proposed model is illustrated in Fig. 3. First, the global feature representations are produced by the CLIP model to perform global matching. Then, a token selection aggregation module is designed to generate fine-grained local fusion features. Since random masking can easily disrupt the word relationships in text descriptions, we propose a KPDM-based local matching method. The text tokens rely on a key phrase masking strategy to generate local features, while the image patches are selected based on integrated cross-layer importance estimation to identify key patches. Both are then used for subsequent cross-modal local matching. Frequency-based MLM is designed to encourage the model to prioritize the influence of low-frequency tokens with more distinctive features. Moreover, in order to address the noisy pairs in image-text matching, we utilize similarity distribution matching to partition the image-text pairs into noisy and clean sets, enhancing the model's robustness in handling noisy image-text pairs.

Image-Text Feature Extraction

Global Feature Extraction We employ the pre-trained CLIP model (Radford et al. 2021) as our feature extraction backbone. As shown in Fig. 3(a), for a given image $I_i \in V$, the visual encoder f^v produces a set of embeddings:

$$V_i = f^v(I_i) = \left\{ v_i^g, v_i^1, v_i^2, \cdots, v_i^{N^*} \right\}^T \in \mathbb{R}^{(N^*+1)\times d}, \tag{1}$$

where v_i^g is the global image feature from the [CLS] token, $\{v_i^j\}_{j=1}^{N^*}$ denote the local patch features, and N^* denotes the number of patches. Similarly, for a text $T_i \in L$, $\{t_i^j\}_{j=1}^{N^\diamond}$ denote the word-level features of the text, and the feature of the [EOS] token t_i^e is the sentence-level global feature.

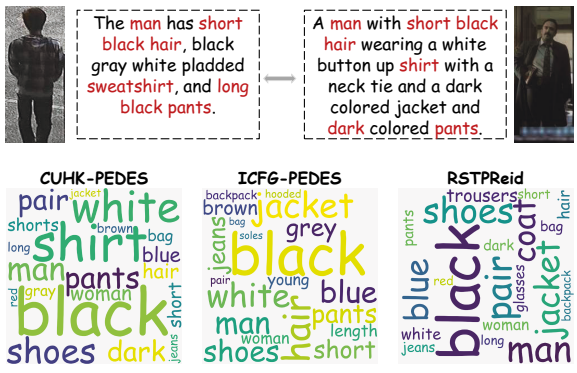

Figure 2: Retrieval examples on the three public dataset.

Local Feature Extraction To further enhance key feature extraction and minimize redundancy, we filter and aggregate the local image and text features to generate discriminative fine-grained key features. Similar to (Yan et al. 2023b; Zhu et al. 2022; Qin et al. 2024), we select the top $R\%$ of tokens with the highest scores, resulting in filtered feature sets $V_i^s \subset \{v_i^j\}$ and $T_i^s \subset \{t_i^j\}$. This filtered local feature set is further optimized through an embedding transformation, as detailed in the subsequent process:

$$\begin{aligned} v_i^{fil} &= \text{MaxPool}\left(\text{MLP}\left(\hat{V}_i^s\right) + \text{FC}\left(\hat{V}_i^s\right)\right), \\ t_i^{fil} &= \text{MaxPool}\left(\text{MLP}\left(\hat{T}_i^s\right) + \text{FC}\left(\hat{T}_i^s\right)\right), \end{aligned} \tag{2}$$

where $\hat{V}_i^s$ and $\hat{T}_i^s$ represent the results of features after layer normalization, and $MLP(\cdot)$ denotes a multilayer perception consisting of two fully connected layers with residual connections. Ultimately, we obtain the global features v_i^g and t_i^e for the image and text, respectively. Simultaneously, the fine-grained aggregated features v_i^{fil} and t_i^{fil} are computed to remove redundant information. These aggregated features form the foundation for alignment in cross-modal matching tasks.

KPDM-based Local Feature Matching

Our proposed Key Phrase Dynamic Mask (KPDM) based local matching method consists of three core components: the phrase-based masking strategy, the removal of redundant image patches, and a frequency-based masked language loss (FMLM) that supervises multimodal feature fusion.

Phrase-based Masking Strategy Instead of the traditional random masking approach, we propose a phrase-based masking strategy focusing on "adjective+noun" pairs. This approach enables the model to capture semantic relationships between words more effectively, alleviating issues related to misalignment of internal features.

Specifically, we utilize SpaCy (Vasiliev 2020) to lexically analyze text tokens and extract key phrases in the "adjective+noun" structure for masking. However, significant feature overlap may occur, even among texts referring to different identities. As shown in Fig. 2, text descriptions from different pedestrians may exhibit substantial similarity. High-frequency token features often obscure the distinction between ambiguous text pairs. To mitigate the impact of high-frequency words and emphasize the importance of rare key features, we apply nonlinear compression to the loss weights of high-frequency words. The weight scores are based on token frequencies: for tokens that belong to the 25 most frequent words, we introduce a nonlinear factor λ to compress their weights. The weight score is defined as follows:

$$w_i = (1 - f_i)^{\lambda}, \tag{3}$$

where f_i denotes the frequency of the token, and λ is a predefined factor that controls the degree of compression. Tokens that do not belong to the set of high-frequency words are assigned a weight score of 1.

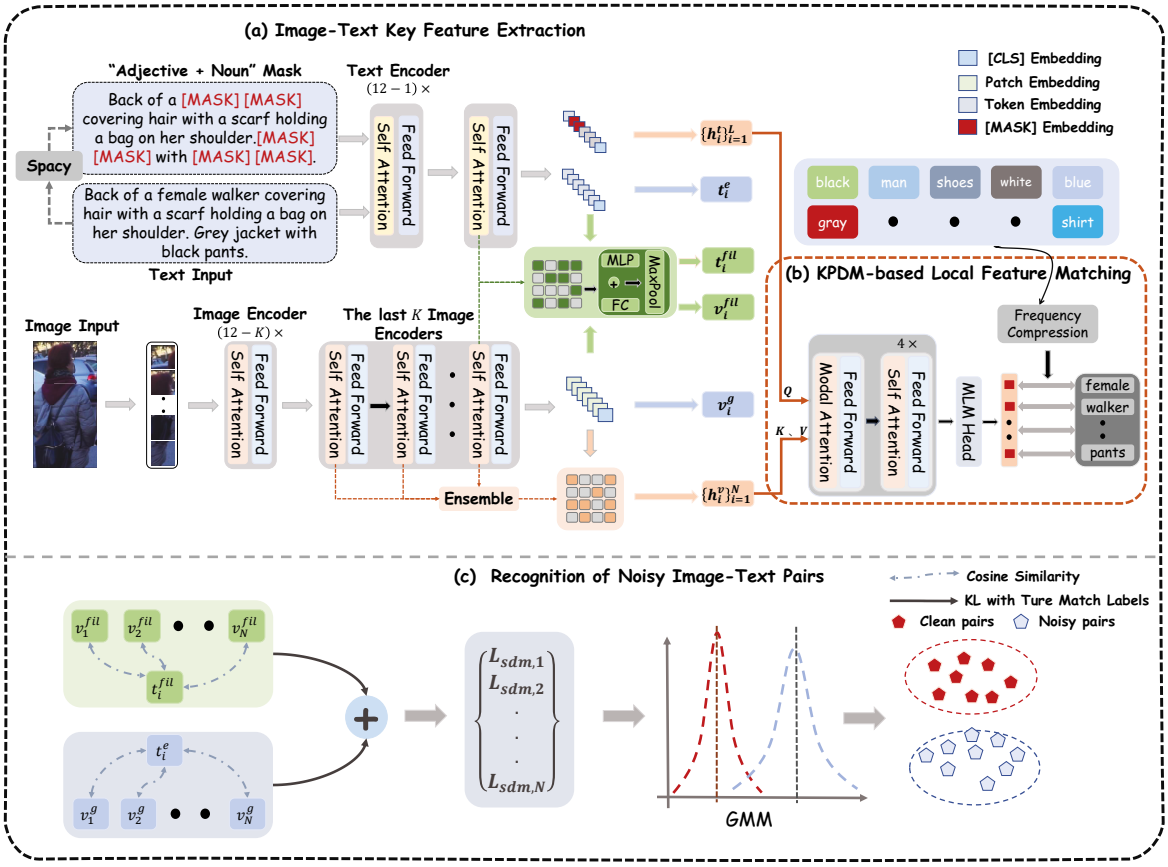

Figure 3: **Overview of the proposed KPDM framework. (a)** illustrates the CLIP-based image-text key feature extraction framework. **(b)** presents the detailed framework for cross-modal local matching between images and text. It utilizes the selected key patches to predict masked "adjective + noun" phrases, enabling the model to implicitly perform fine-grained image-text alignment. Furthermore, we take into account the impact of high-frequency tokens, encouraging the model to focus more on the influence of low-frequency tokens with more distinctive features. **(c)** illustrates the recognition mechanism for identifying noisy image-text pairs.

Image Redundancy Patch Removal To ensure the extraction of the most critical pedestrian features for efficient fine-grained local matching with masked tokens, we filter and suppress redundant local image patches. Specifically, we extract the attention scores from the last K transformer blocks, focusing on the [CLS] tokens of the visual encoder and the remaining image patches. These attention scores are denoted as $\left\{s_v^{L_v-K+1}, s_v^{L_v-K+2}, \cdots, S_v^{L_v}\right\}$, where L_v represents the number of transformer layers in the CLIP visual encoder. By default, we apply an averaging operation to aggregate the attention scores, resulting in the final importance score for the visual tokens, s_v^K:

$$s_v^K = E\left(s_v^{L_v-K+1}, s_v^{L_v-K+2}, \cdots, S_v^{L_v}\right). \tag{4}$$

We then filter the patches using Top-K strategy, selecting the top $R\%$ of patches based on their similarity scores. The selected indices are given by: $X_i^v = \left\{j | rank\left(s_{v,j}^K\right) \leq \lfloor R \cdot N^* \rfloor\right\}$. Subsequently, redundant patch blocks are eliminated. The set of selected patches is denoted as $\{h_i^v\}_{i=1}^N$.

Multimodal Feature Fusing After extracting the key pedestrian patches from the image and the key phrase masked tokens from the text description, we input them into the Transformer encoder for multimodal feature fusion. The masked text features $\{h_i^t\}_{i=1}^L$ are used as the query, while the filtered image features $\{h_i^v\}_{i=1}^N$ serve as the key and value in the cross-modal interactive encoder. These features are simultaneously input into the encoder as follows:

$$\{h_i^m\}_{i=1}^L = Transformer\left(\{h_i^t\}_{i=1}^L, \{h_i^v\}_{i=1}^N, \{h_i^v\}_{i=1}^N\right), \tag{5}$$

where $\{h_i^m\}_{i=1}^L$ represents the fused multimodal features, combining both masked text and image information. For each masked token $\{h_i^m : i \in M\}_{i=1}^L$, we use a multilayer perceptron (MLP) to predict the probability distribution of its original token:

$$\left\{m_j^i\right\}_{j=1}^{|V|} = MLP\left(h_i^m\right), \tag{6}$$

where $|V|$ is the size of the vocabulary.

We utilize the frequency of masked tokens and the probability information of predicted original tokens to supervise the feature fusion process. The FMLM loss for a batch of size N is summarized as:

$$\mathcal{L}_{fmlm} = -\frac{\sum_{k=1}^{N}\sum_{i\in M}\sum_{j\in|V|} w_i^k y_j^i \log\left(\frac{\exp(m_j^i)}{\sum_{a=1}^{|V|}\exp(m_a^i)}\right)}{\sum_{k=1}^{N}\sum_{i\in M} w_i^k}, \tag{7}$$

where $|M|$ represents the set of masked tokens, m_j^i denotes scores of predicting the j-th word in the vocabulary at the masked location i, and y_j^i indicates whether the j-th token in the vocabulary corresponds to the original token at the masked location. Additionally, w_i^k represents the weights based on the frequency, as described in Eq.(3).

Recognition of Noisy Image-Text Pairs

To address the challenge of noisy pairs in image-text matching, we design a trusted partitioning mechanism that partitions the training data into clean and noisy subsets. Different from RDE (Qin et al. 2024), which primarily relies on learned distance metrics, we evaluate the image-text matching degree based on the alignment of similarity distribution for pairs within the same identity. Specifically, we quantify this matching degree using the similarity distribution matching (SDM) (Jiang and Ye 2023) score, which is calculated as the Kullback-Leibler (KL) divergence between image-text similarity distributions and the normalized label matching distributions. A higher SDM loss suggests a greater likelihood that the corresponding image-text pair contains noise.

For each image-text pair in a batch, we compute a bi-directional SDM loss for both global features (v^g, t^e) and aggregated local features $(v^{\text{fil}}, t^{\text{fil}})$. First, the matching probability between an image i and all texts j in the batch is calculated via softmax over their cosine similarities:

$$p_{i,j} = \frac{\exp(\text{sim}(v_i, t_j)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(v_i, t_k)/\tau)}, \tag{8}$$

where τ is a temperature hyperparameter. The SDM loss is the KL divergence between this predicted probability $p_{i,j}$ and the true matching probability $q_{i,j} = y_{i,j}/\sum_k y_{i,k}$:

$$\mathcal{L}_{\text{i2t}} = \text{KL}(p_{i,j}||q_{i,j}) = \sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\varepsilon}\right). \tag{9}$$

The final bi-directional SDM loss is then given by:

$$\mathcal{L}_{sdm} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{sdm,i} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}), \tag{10}$$

The SDM loss for each pair is computed as follows:

$$\mathcal{L}_{sdm}^{total} = \{\mathcal{L}_{sdm,i}\}_{i=1}^{N'} = \left\{(\mathcal{L}_{sdm,i}^{cls} + \mathcal{L}_{sdm,i}^{fil})/2\right\}_{i=1}^{N'}, \tag{11}$$

Following the approach in (Qin et al. 2024), we fit a two-component Gaussian Mixture Model (GMM) to the distribution of all per-sample SDM losses $\{\mathcal{L}_{\text{sdm},i}\}_{i=1}^{N'}$ across the entire training set. The GMM naturally separates the losses into two clusters: one with a lower mean, corresponding to the clean set, and one with a higher mean, corresponding to the noisy set. Based on the posterior probability of each sample belonging to the clean component, we partition the dataset U into clean (D_c) and noisy (D_n) subsets:

$$\begin{aligned} D_c &= \{\mathcal{G}(P(k=0|\mathcal{L}_{\text{sdm},i})) > 0.5, (I_i, T_i) \in U\}, \\ D_n &= \{\mathcal{G}(P(k=0|\mathcal{L}_{\text{sdm},i})) \leq 0.5, (I_i, T_i) \in U\}. \end{aligned} \tag{12}$$

This partitioning provides a pairwise label $\tilde{l}_i$ for each pair, indicating whether it is considered clean or noisy, which is then used to guide the subsequent training process.

Model Optimization

The main objective of the TIReID task is to accurately retrieve the identity of pedestrians according to text descriptions. To achieve this, the commonly utilized ID loss (Zheng et al. 2020b) is also adopted along with SDM loss and FMLM loss to optimize KPDM.

In Section , we introduce a trusted consensus mechanism to partition image-text pairs into clean and noisy subsets. To train the model robustly, we focus exclusively on the clean image-text pairs with the corrected label $\tilde{l}_i = 1$. The overall optimization objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{fmlm} + \mathcal{L}_{sdm}^{cls} + \mathcal{L}_{sdm}^{fil} + \mathcal{L}_{id}^{cls} + \mathcal{L}_{id}^{fil}, \tag{13}$$

where $\mathcal{L}_{id}^{cls}$ utilizes the global features v_i^g and t_i^e for images and text, respectively, while $\mathcal{L}_{id}^{fil}$ leverages the local aggregated features v_i^{fil} and t_i^{fil} to focus on local information.

Experiments

Datasets and Evaluation Metrics

Consistent with previous studies, we conduct extensive experiments on three challenging TIReID datasets: **CUHK-PEDES** (Li et al. 2017), **ICFG-PEDES** (Ding et al. 2021) and **RSTPReid** (Zhu et al. 2021). In line with (Qin et al. 2024; Wang et al. 2021; Wang, Yang, and Cao 2024), we evaluate model performance using the popular Rank-k metric(k = 1, 5, 10), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP).

Implementation Details

Our experiments are conducted using two backbone configurations: CLIP-ViTB/16 and CLIP-ViTL/14. For each configuration, the text and image encoders are initialized with the pre-trained weights, while other network modules are randomly initialized. During training, all images are resized to 384×128 and augmented with random horizontal flipping, random crop with padding, and random erasing. For training text, alongside the KPM strategy for the FMLM task, we apply standard random masking to the tokens fed into the SDM and ID tasks. We utilize the Adam optimizer to train the model for 60 epochs with a cosine learning rate decay strategy. The initial learning rates are set to 1×10^{-5} for the CLIP encoders and $1e-3$ for other network modules. For hyperparameter settings, patch selection for redundancy removal utilizes scores from the last $K = 3$ transformer

blocks and the selection ratio R is 0.5; the compression factor λ for high-frequency tokens is set to 2; and the SDM temperature τ is 0.02. All experiments are trained with a batch size of 64 on a single RTX 4090 24GB GPU.

Method	Ref	R@1	R@5	R@10	mAP	mINP
Comparison with standard backbones						
DSSL	MM'21	39.05	62.60	73.95	-	-
LBUL	MM'22	45.55	68.20	77.85	-	-
CAIBC	MM'22	47.35	69.55	79.00	-	-
CFine	arXiv'22	50.55	72.50	81.60	-	-
UniPT	ICCV'23	51.85	74.85	82.85	-	-
IRRA	CVPR'23	60.20	81.30	88.20	47.17	25.28
IRLT	AAAI'24	61.49	82.26	89.23	-	-
Propot	MM'24	61.87	83.63	89.70	-	-
TBPS-CLIP	AAAI'24	62.10	83.55	88.75	48.26	-
CFAM(B/16)	CVPR'24	59.40	81.35	88.50	46.04	-
SAP-SAM	MM'24	62.85	82.65	89.85	-	-
RDE	CVPR'24	65.35	83.95	89.90	50.88	28.08
KPDM (B/16)	-	**67.95**	**84.65**	**91.10**	**51.88**	**28.20**
Comparison using ViT-L/14 backbone						
CFAM (L/14)	CVPR'24	62.45	83.55	91.10	49.50	-
KPDM (L/14)	-	**67.95**	**86.35**	**91.65**	**52.97**	**30.00**

Table 1: Performance comparisons with state-of-the-art methods on the RSTPReid dataset.

Method	Ref	Rank-1	Rank-5	Rank-10	mAP	mINP
Comparison with standard backbones						
LGUR	MM'22	65.25	83.12	89.00	-	-
LCR^{2}S	MM'23	67.36	84.19	89.62	59.24	-
CLIP (B/16)	CVPR'23	68.19	86.47	91.47	61.12	44.86
CFine	arXiv'22	69.57	85.93	91.15	-	-
IRRA	CVPR'23	73.38	89.93	93.71	66.13	50.24
TBPS-CLIP	AAAI'24	73.54	88.19	92.35	65.38	-
CFAM (B/16)	CVPR'24	72.87	88.61	92.87	64.92	-
IRLT	AAAI'24	74.46	90.19	94.01	-	-
Propot	MM'24	74.89	89.90	94.17	-	-
SAP-SAM (B/16)	MM'24	75.05	89.93	93.73	-	-
RDE	CVPR'24	75.94	90.14	94.12	67.56	51.44
KPDM (B/16)	-	**75.97**	**90.77**	**94.46**	**68.14**	**52.17**
Comparison using ViT-L/14 backbone						
CFAM (L/14)	CVPR'24	75.60	90.53	94.36	67.27	-
SAP-SAM (L/14)	MM'24	76.28	90.87	94.75	-	-
KPDM (L/14)	-	**78.02**	**91.49**	**95.00**	**69.71**	**53.68**

Table 2: Performance comparisons with state-of-the-art methods on CUHK-PEDES dataset.

Experimental Results

In this section, we present the comparative performance of the proposed KPDM against existing state-of-the-art methods (DSSL (Zhu et al. 2021), LGUR (Shao et al. 2022), LBUL (Wang et al. 2022c), LCR^{2}S (Yan et al. 2023a), CAIBC (Wang et al. 2022b), CFine (Yan et al. 2023b), UniPT (Shao et al. 2023), IRRA (Jiang and Ye 2023), IRLT (Liu et al. 2024), Propot (Yan et al. 2024), TBPS-CLIP (Cao et al. 2024), CFAM (Zuo et al. 2024), SAP-SAM (Wang, Yang, and Cao 2024), RDE (Qin et al. 2024)) across three publicly available datasets.

Method	Ref	Rank-1	Rank-5	Rank-10	mAP	mINP
Comparison with standard backbones						
CLIP(B/16)	CVPR'23	56.74	75.72	82.26	31.84	5.03
LCR^{2}S	MM'23	57.93	76.08	82.40	38.21	-
LGUR	MM'22	59.02	75.32	81.56	-	-
CFine	arXiv'22	60.83	76.55	82.42	-	-
IRRA	CVPR'23	63.46	80.25	85.82	38.06	7.93
SAP-SAM	MM'24	63.97	80.84	86.17	-	-
IRLT	AAAI'24	64.72	81.35	86.31		
TBPS-CLIP	AAAI'24	65.05	80.34	85.47	39.83	-
Propot	MM'24	65.12	81.57	86.97	-	-
CFAM(B/16)	CVPR'24	62.17	79.57	85.32	36.34	-
RDE	CVPR'24	67.68	**82.47**	**87.36**	40.06	7.87
KPDM(B/16)	-	**67.78**	82.43	87.26	**40.30**	**8.02**
Comparison using ViT-L/14 backbone						
CFAM(L/14)	CVPR'24	65.38	81.17	86.35	39.42	-
KPDM(L/14)	-	**69.13**	**83.46**	**88.10**	**42.78**	**9.77**

Table 3: Performance comparisons with state-of-the-art methods on ICFG-PEDES dataset.

RSTPReid As shown in Table 1, the KPDM model significantly surpasses the other models and the state-of-the-art RDE (Qin et al. 2024) in various metrics. Specifically, our KPDM reaches on R@1 (67.95%), R@5 (84.65%), R@10 (91.10%), mAP (51.88%), and mINP (28.20%), achieving significant improvements of +2.60%, +0.70%, +1.20%, +1.00%, and 0.12%, compared to the RDE. With the CLIP-ViT-L/14 setting, KPDM demonstrates further performance improvement, also achieving a R@-1 metric of 67.95%. Attributable to the smaller segmentation patches that enable focus on fine-grained image features, the model shows significant enhancements across other metrics as well.

CUHK-PEDES As shown in Table 2, KPDM outperforms the baseline IRRA (Jiang and Ye 2023) in all key metrics. R@1 accuracy improves by 2.59% and mAP increases by 2.01%. Compared to the state-of-the-art model RDE, KPDM achieves a modest improvement of +0.58% in mAP and +0.73% in mINP. In the larger backbone CLIP-ViT-L/14, our method shows further performance improvement through richer image representations. The R@1 increases by 2.05%, reaching 78.02%.

ICFG-PEDES As shown in Table 3, KPDM outperforms the state-of-the-art model RDE, achieving a R@1 accuracy of 67.78%, and also shows stable improvements in mAP and mINP. Furthermore, utilizing the CLIP-ViTL/14 setting further enhances KPDM's performance, reaching 69.13% in R@1 accuracy and 42.78% in mAP. While KPDM outperforms RDE on R@1, it exhibits slightly lower performance on the R@5 and R@10. We attribute this primarily to the nature of the ICFG-PEDES dataset, which consists of single image-text pairs. Compared to the datasets featuring

No.	Components	TCP	KPM	IRPR	R@1	R@5	R@10
0	Baseline				61.35	81.97	88.45
1	+IRPR			✓	62.80	82.45	88.97
2	+TCP	✓			64.80	83.50	89.00
3	+KPM		✓		64.15	84.50	90.35
4	+TCP+IRPR	✓		✓	65.55	84.60	90.20
5	+KPM+IRPR		✓	✓	65.40	**84.95**	90.35
6	+TCP+KPM	✓	✓		66.35	84.55	90.75
7	KPDM	✓	✓	✓	**67.95**	84.65	**91.10**

Table 4: Ablation study on different components of our framework on RSTPReid dataset.

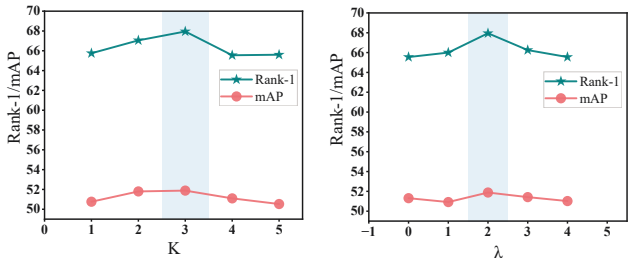

Figure 4: Variation of performance on RSTPReid dataset with different K and λ.

multiple text descriptions per image, the influence of high-frequency tokens may be less significant. This might lead the model to neglect certain textual features, resulting in a slight performance reduction on these specific metrics. We plan to optimize this issue in the future.

Ablation Study

In this section, we analyze the effectiveness of each proposed component. We use the local matching method IRRA (Jiang and Ye 2023), based on a random masking strategy, as our baseline to conduct ablation experiments. To ensure a fair comparison, we also incorporate local fusion features ($\mathcal{L}_{sdm}^{fil}$ and $\mathcal{L}_{id}^{fil}$) into the IRRA model.

We mainly analyze the effectiveness of the three key components: the trusted consensus partitioning (TCP) mechanism to distinguish noisy image-text pairs, the key phrase masking (KPM) strategy, and image redundancy patch removal (IRPR). The ablation study results are shown in Table 4. First, by comparing the experimental results of No.0 and No.1, and No.6 and No.7, it can be observed that the performance is significantly improved by the image redundancy patch removal. Compared to the baseline, R@1 has an increase of 1.45%. Furthermore, the TCP to distinguish noise in image-text pairs significantly enhances the robustness of the model. A comparison between No.0 and No.2, and No.5 and No.7 demonstrates that TCP leads to an improvement in R@1 by 3.45% compared to the baseline.

The proposed local matching method based on the KPM strategy also has a significant impact on model performance. Comparing No.0 and No.3, and No.4 and No.7, it can be seen that the inclusion of KPM leads to a notable improvement, with R@1 increasing by 2.8% compared to the baseline. Additionally, a comparison between No.3 and No.6, No.5 and No.7, demonstrates that KPM further enhances model performance through robust learning. By leveraging this approach, KPM enables the model to learn more accurate image-text representations.

Parametric Analysis

To study the impact of different hyperparameter settings on performance, we perform sensitivity analyses for two key hyperparameters: the number of self-attention layers K for selecting key patches and the compression factor λ for high-frequency tokens. From Fig. 4, we can see that:(1) Both too few and too many self-attention layers hinder the model's ability to effectively capture the information of key image patches. We choose $K = 3$ in all our experiments. (2) Excessive or insufficient compression of high-frequency tokens also impacts the model's performance. If it's too small, the model fails to emphasize rare tokens; if too large, it neglects the features of high-frequency tokens. We choose $\lambda = 2$ in all our experiments.

Qualitative Results

Figure 5: Visualization of examples on three datasets.

Fig. 5 compares the top-10 retrieval results of the baseline and our proposed KPDM. As the figure shows, KPDM retrieves more accurate results compared to the baseline. Additionally, even in cases of retrieval failure, KPDM still effectively captures fine-grained features of the pedestrian. As highlighted by the red boxes, under the "adjective + noun" phrase masking strategy, the fine-grained features of the pedestrian are tightly associated in the form of phrases.

Conclusion

In this paper, we propose a robust fine-grained local alignment framework based on key phrase dynamic mask (KPDM). First, we utilize image-text matching information to design a trusted consensus partitioning mechanism to distinguish noisy image-text pairs. Then, for clean pairs, we adopt an "adjective + noun" key phrase masking strategy to enable the TIReID model to grasp nuanced semantic relationships. Furthermore, we propose cross-layer importance estimation to select key image patches, emphasizing pedestrian regions. We validate and achieve optimal results on three existing benchmarks, demonstrating the effectiveness of our method.

Acknowledgments

We would like to thank anonymous reviewers for their suggestions and comments sincerely. This work is supported by the National Natural Science Foundation of China (62573393, 62373338).

References

Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 465–473.

Chen, T.; Xu, C.; and Luo, J. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1879–1887. IEEE.

Chen, Y.; Zhang, G.; Lu, Y.; Wang, Z.; and Zheng, Y. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

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

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

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 4477–4485.

Fu, Z.; Zhou, W.; Xu, J.; Zhou, H.; and Li, L. 2022. Contextual representation learning beyond masked language modeling. *arXiv preprint arXiv:2204.04163*.

Fujii, T.; and Tarashima, S. 2023. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2786–2790.

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

Jing, Y.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11189–11196.

Li, H.; Xiao, J.; Sun, M.; Lim, E. G.; and Zhao, Y. 2021. Transformer-based language-person search with multiple region slicing. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1624–1633.

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

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Wen, Z.; and Peng, X. 2024. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

Miech, A.; Alayrac, J.-B.; Laptev, I.; Sivic, J.; and Zisserman, A. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9826–9836.

Niu, K.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29: 5542–5556.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

Shao, Z.; Zhang, X.; Ding, C.; Wang, J.; and Wang, J. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11174–11184.

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

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

Suo, W.; Sun, M.; Niu, K.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A simple and robust correlation filtering method for text-based person search. In *European conference on computer vision*, 726–742. Springer.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Vasiliev, Y. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Wang, C.; Luo, Z.; Lin, Y.; and Li, S. 2021. Text-based person search via multi-granularity embedding learning. In *IJCAI*, 1068–1074.

Wang, P.; Ding, C.; Tan, W.; Gong, M.; Jia, K.; and Tao, D. 2022a. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. *IEEE Transactions on Multimedia*, 25: 2624–2635.

Wang, Y.; Yang, M.; and Cao, R. 2024. Fine-grained Semantic Alignment with Transferred Person-SAM for Text-based Person Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5432–5441.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XII 16*, 402–420. Springer.

Wang, Z.; Hu, R.; Yu, Y.; Liang, C.; and Huang, W. 2015. Multi-level fusion for person re-identification with incomplete marks. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1267–1270.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, 5314–5322.

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

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. LapsCore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1624–1633.

Yan, S.; Dong, N.; Liu, J.; Zhang, L.; and Tang, J. 2023a. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, 6202–6211.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2023b. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32: 6032–6046.

Yan, S.; Liu, J.; Dong, N.; Zhang, L.; and Tang, J. 2024. Prototypical Prompting for Text-to-image Person Re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2331–2340.

Yan, S.; Tang, H.; Zhang, L.; and Tang, J. 2023c. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE transactions on neural networks and learning systems*.

Yang, S.; Zhou, Y.; Zheng, Z.; Wang, Y.; Zhu, L.; and Wu, Y. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4492–4501.

Ye, X.; Gan, Y.; Huang, X.; Ge, Y.; Shan, Y.; and Tang, Y. 2024. Voco-llama: Towards vision compression with large language models. *arXiv preprint arXiv:2406.12275*.

Zhang, H.; Yang, Y.; Qi, F.; Qian, S.; and Xu, C. 2023. Robust video-text retrieval via noisy pair calibration. *IEEE Transactions on Multimedia*, 25: 8632–8645.

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

Zheng, K.; Liu, W.; Liu, J.; Zha, Z.-J.; and Mei, T. 2020a. Hierarchical gumbel attention network for text-based person search. In *Proceedings of the 28th ACM international conference on multimedia*, 3441–3449.

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

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

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4692–4702.

Zuo, J.; Zhou, H.; Nie, Y.; Zhang, F.; Guo, T.; Sang, N.; Wang, Y.; and Gao, C. 2024. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22010–22019.