

Article

Towards Robust Text-Based Person Retrieval: A Framework for Correspondence Rectification and Description Synthesis

Longlong Yu ¹, Lian Xiong ^{2,*}, Wangdong Li ² and Yuxi Feng ²¹ College of Computer Science, Chengdu University, Chengdu 610106, China; yulonglong@cdu.edu.cn² School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; lwd201723@gmail.com (W.L.); esdvian957@gmail.com (Y.F.)

* Correspondence: xionglian@cqupt.edu.cn; Tel.: +86-13983093700

Abstract

Retrieving pedestrian images using natural language descriptions remains challenging due to the prevalence of imperfect annotations in real-world training data. Most existing methods rely on the strong assumption of perfectly aligned image–text pairs, largely ignoring the detrimental impact of annotation noise, which typically manifests as coarse-grained descriptions and erroneous correspondences. These imperfections severely degrade model performance and generalization. To address these issues, we propose a novel framework centered on two key innovations. First, we develop a probabilistic noise identification mechanism that employs a dual-channel Gaussian mixture model (GMM) to assess alignment consistency at both global and local feature levels. Second, for samples identified as noisy, we implement a description synthesis pipeline that leverages a multimodal large language model (MLLM) to generate refined descriptions. A dynamic semantic consistency module then filters these synthesized texts to ensure quality. Comprehensive evaluations on three benchmark datasets—CUHK-PEDES, ICFG-PEDES, and RSTPReid—demonstrate the superior performance of our method: ICFG-PEDES Rank-1 = 68.13%, Rank-5 = 83.39%, Rank-10 = 89.02%; RSTPReid Rank-1 = 66.31%, Rank-5 = 86.87%, Rank-10 = 92.01%; CUHK-PEDES Rank-1 = 75.98%, Rank-5 = 90.34%, Rank-10 = 94.32%. These results show consistent top-k improvements over prior methods and validate the effectiveness of the proposed noise-aware pseudo-text augmentation.



Academic Editor: Arkaitz Zubiaga

Received: 27 October 2025

Revised: 19 November 2025

Accepted: 23 November 2025

Published: 25 November 2025

Citation: Yu, L.; Xiong, L.; Li, W.; Feng, Y. Towards Robust Text-Based Person Retrieval: A Framework for Correspondence Rectification and Description Synthesis. *Electronics* **2025**, *14*, 4619. <https://doi.org/10.3390/electronics14234619>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: text-based person retrieval; cross-modal alignment; annotation noise; Gaussian mixture models; large language models; fine-grained matching

1. Introduction

The rapid advancement of urban intelligent security systems and the widespread deployment of multimodal sensing networks have significantly increased the demand for reliable person re-identification (ReID) technologies [1–3]. These technologies are vital for practical applications such as video surveillance analysis, intelligent transportation management, and criminal investigation. However, traditional image-based ReID methods face limitations when query images of target individuals are unavailable. In many real-world scenarios, such as searching for missing persons, eyewitnesses can often only provide verbal descriptions rather than visual references [4,5].

This practical constraint has driven the emergence of text-based person search (TBPS) as an alternative paradigm that overcomes the limitations of conventional ReID. Essentially, TBPS is a cross-modal retrieval task that matches pedestrian images from a database based

on natural language descriptions. The core challenge lies in establishing precise semantic alignment between visual and textual content within a shared embedding space. Recent progress in vision–language pre-training (VLP) methodologies has substantially enhanced cross-modal understanding capabilities [6]. Building on these advancements, researchers have begun exploring the application of VLP pre-trained models to address the specific requirements of TBPS tasks [7,8].

Despite these improvements, a critical issue persists: the inherent imperfection of training data annotations. As shown in Figure 1, real-world datasets commonly exhibit two predominant types of annotation deficiencies. The left example shows coarse-grained textual descriptions that lack sufficient detail to capture discriminative pedestrian attributes. The right example illustrates noisy correspondences, where image–text pairs are mismatched due to annotation errors. When treated as positive training samples, these imperfect pairs introduce erroneous supervisory signals that misguide the learning process, ultimately degrading performance, promoting overfitting, and diminishing generalization ability [9,10].

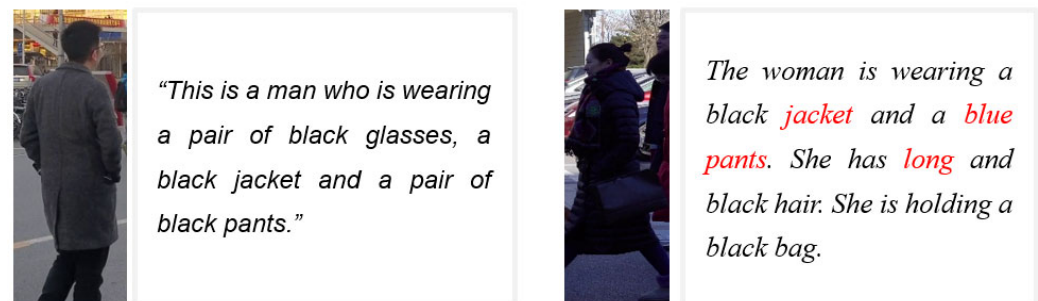


Figure 1. Illustration of annotation noise, including coarse text granularity and noisy correspondence. Both examples are from the RSTPReid dataset.

Recent studies have attempted to mitigate these data quality issues. Some approaches employ multimodal large language models (MLLMs) to generate supplemental textual descriptions [11–13]. However, these methods often overlook the crucial step of quality verification for the generated content, potentially exacerbating noise if low-quality synthetic texts are incorporated. Concurrently, while CLIP-based architectures [14] have demonstrated exceptional performance in global semantic alignment, they still struggle to capture fine-grained semantic relationships and facilitate local feature interactions—capabilities essential for high-precision cross-modal retrieval in person search.

To comprehensively address these challenges, we propose a robust framework that systematically tackles annotation noise while enhancing fine-grained semantic alignment. Our methodology begins with a dual-feature analysis using CLIP to extract both global and local representations from images and text. We then employ a dual-channel Gaussian mixture model (GMM) to dynamically identify potentially noisy image–text pairs by modeling loss distributions at different feature levels. For samples identified as noisy, we deploy Qwen-VL [15] as our MLLM to generate refined descriptions. These are then subjected to a semantic consistency matching mechanism to ensure quality before integration into training. To further strengthen cross-modal alignment, we incorporate the Implicit Relation Reasoning (IRR) module [8], which enhances fine-grained semantic interactions through a hybrid attention mechanism combined with masked language modeling.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces our proposed method, including dual-channel GMM noise identification, MLLM-based pseudo-text synthesis, the semantic consistency filter, and the IRR module. Section 4 describes datasets, evaluation metrics, and experimental settings

and presents comparison and ablation studies. Finally, Section 5 concludes with discussions and future work. The principal contributions of this work are summarized as follows:

- We directly address the pervasive yet underexplored problem of annotation noise in TBPS, explicitly targeting both coarse-grained descriptions and mismatched image–text pairs.
- We introduce an integrated approach combining GMM-based noise identification with MLLM-driven text augmentation and filtering, enabling effective noise suppression while enhancing training supervision quality.
- We demonstrate the efficacy of our method through extensive experiments on three public benchmarks, with ablation studies empirically validating each component’s contribution to overall performance.

2. Related Work

2.1. Text-Based Person Search

The primary challenge in TBPS is effectively aligning image and text features into a joint embedding space for accurate and efficient person re-identification. Early studies [16–19] often utilized unimodal pre-trained models and attention mechanisms, employing ResNet-50 [20] or ViT [21] as image encoders and LSTM [22] or BERT [23] as text encoders. However, using unimodal backbone networks to extract embeddings often overexploits intra-modal information, making cross-modal alignment and network optimization more difficult.

Recently, VLP [10,24,25] has become the primary solution for various cross-modal tasks. Han et al. [26] first introduced the CLIP model for text-to-image person retrieval, using a momentum contrastive learning framework to transfer knowledge from large-scale generic image–text pairs. CFine [27] employed CLIP’s image encoder to enrich cross-modal correspondence and used BERT to replace the text encoder, avoiding distortion of intra-modal information. TP-TPS [7] explored the potential of CLIP’s text component in TBPS by aligning images with multiple complete descriptions and attribute prompts. RaSa [28] developed two new loss functions under the ALBEF [9] backbone for cross-modal alignment.

Although VLP has revolutionized cross-modal retrieval, learning local relations remains key to mining hidden fine-grained information from image patches. For instance, IRRRA [8] introduced an implicit relation learning network that uses image features to predict caption words, aligning fine-grained cross-modal information. Despite their promising performance, these methods almost universally assume that all input training pairs are correctly aligned—an assumption difficult to satisfy in practice due to ubiquitous noise.

2.2. Learning with Noisy Data

The issue of noisy data in cross-modal scenarios has garnered increasing attention. The core challenge is learning robust representations from mismatched multimodal pairs, which can be broadly categorized into two types of solutions:

Sample Selection [29–31]: These methods leverage the memory effect of deep neural networks [32] to gradually distinguish noisy data, focusing more on clean samples and reducing attention to noisy ones. However, if the proportion of noisy samples is high, the model may struggle to learn effectively, potentially overfitting to clean data. Sample selection strategies can also introduce bias, affecting generalization.

Robust Loss Functions [32–36]: These methods aim to develop noise-tolerant loss functions, improving model robustness under noisy correspondence. However, complex loss functions often incur significant computational overhead, impacting training efficiency.

Recently, RDE [37] simultaneously employed sample selection and robust loss functions to identify and handle noise, achieving promising results. However, for

identified noisy pairs, the model discards them during parameter updates, using only clean data for training. This approach wastes data resources and limits further performance improvements.

3. Method

As illustrated in Figure 2, we propose a robust framework based on a CLIP backbone, integrating two GMMs for noisy pair identification and an MLLM for pseudo-text generation and filtering. We also introduce an IRR module to enhance fine-grained cross-modal alignment. The system employs a progressive update mechanism during training: first identifying noise, then replacing noisy text with pseudo-text, and continuously optimizing the alignment representation.

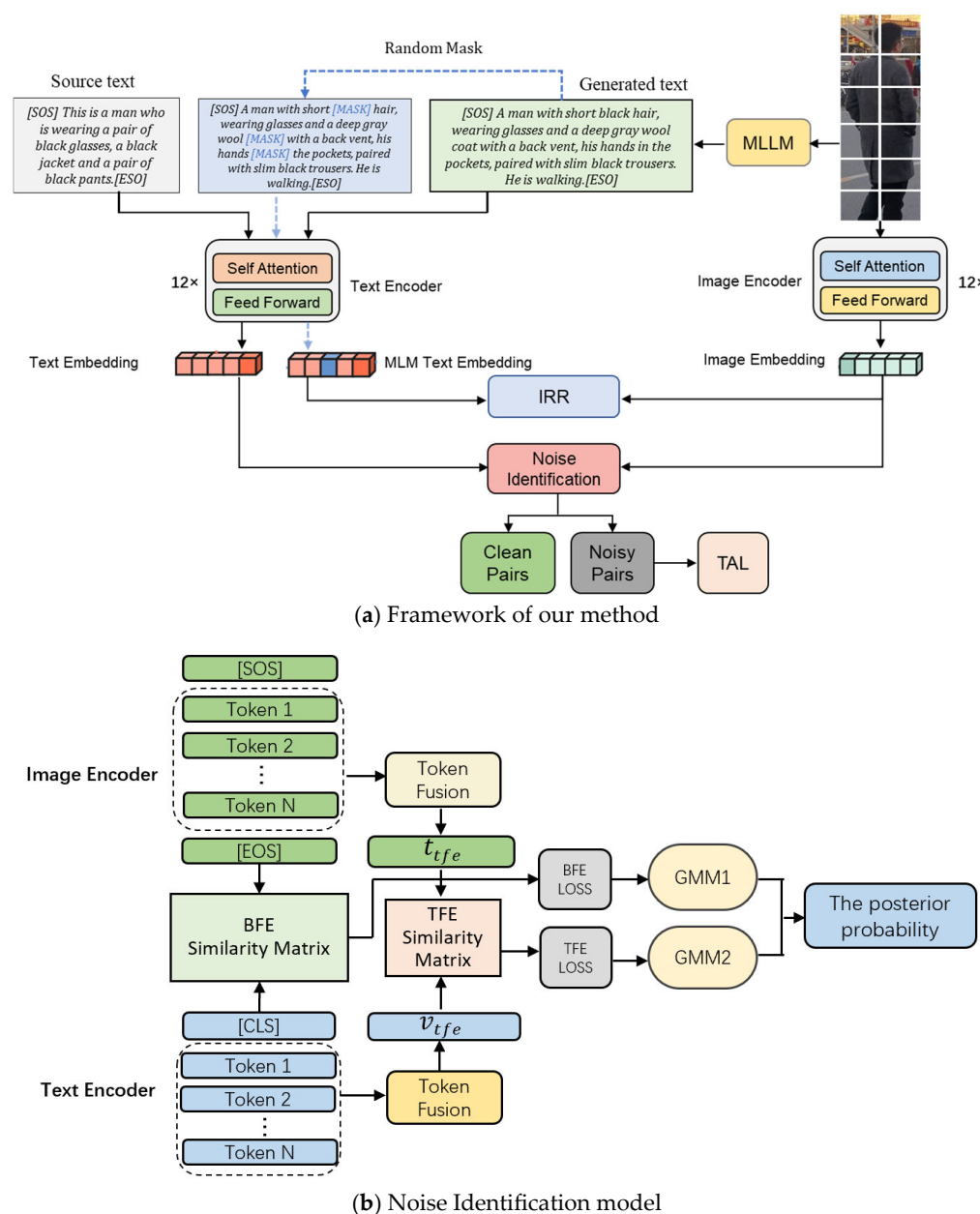


Figure 2. Framework of our method. (a) The overall architecture, which integrates image-text matching, noise identification, and MLLM-generated pseudo-text filtering to enhance cross-modal alignment. (b) The dual-channel Gaussian Mixture Model for distinguishing clean and noisy image-text pairs.

3.1. Feature Representations

We utilize the pre-trained CLIP's visual and text encoders to obtain token representations and implement cross-modal interactions through two token fusion modules.

Image Encoder: We use CLIP's pre-trained ViT model to extract image features. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, where H , W , and C denote the height, width, and number of channels of the image, respectively, we divide the image into N non-overlapping image blocks of 16×16 pixels $\{I_i | i = 1, 2, \dots, N\}$, $N = H \times W / P^2$. The N image blocks are then projected through a linear projection layer to obtain N D -dimensional image block vectors, and a learnable [CLS] embedding vector is inserted in front of the input sequence of image block vectors. To capture the relative positional relationships between image patches, positional embeddings are added to the input sequence, represented as:

$$f_i^v = [I_{cls}; F(I_1); F(I_2); \dots; F(I_p)] + P \quad (1)$$

where f_i^v denotes the embedding vector of the image input sequence and denotes the linear projection layer that maps the parcel image to a D -dimensional vector. Then the embedding vector of image features input to the ViT network can be expressed as: $\{f_{cls}^v, f_1^v, f_2^v, \dots, f_N^v\}$. After passing through the L -layer transformer blocks, the embedding vector will be mapped to the joint image–text embedding space using linear projection, where f_{cls}^v is used as the global feature representation of the image.

Text Encoder: For the input text T , we directly use the CLIP text encoder to extract the textual representation, and following IRRRA, we first tokenize the input text T using lower-cased byte pair encoding (BPE) with a 49,152 vocab size into a token sequence. The [SOS] and [EOS] embedding vectors are inserted at the beginning and the end of the text description to identify the beginning and the end of the text description sentence, respectively. The maximum length of the text description sequence is set to 77 in order to ensure the computational efficiency. In order to learn the relative positional relationship of the words in the sentence, the positional embedding $P \in \mathbb{R}^{77 \times D}$ is also added to the input sequence of word vectors. After the projection transformation, the input sequence can be represented as:

$$f_i^t = [E(T_{sos}); E(T_1); E(T_2); \dots; E(T_{eos})] + P \quad (2)$$

where f_i^t denotes the embedding vector of the text input sequence and E denotes the projection layer that maps the word vector to the D -dimensional vector. The final text-specific embedding vector input to the transformer can be expressed as:

$$\{f_{sos}^t, f_1^t, f_2^t, \dots, f_{eos}^t\} \quad (3)$$

Similarly, the projection of the output text feature embedding vectors into the joint image–text embedding space at the last layer of transformer and the vector f_{eos}^t represented by [EOS] is regarded as the global text feature representation. For computing the similarity between any image–text pair (I , T), we use the cosine similarity between their global features [CLS] and [SOS], which yields the global feature embedding (GFE) similarity:

$$S_{gfe} = \frac{f_{cls}^v f_{eos}^t}{\|f_{cls}^v\| \|f_{eos}^t\|} \quad (4)$$

Token Fusion: To enhance interaction granularity, we introduce local features for finer cross-modal alignment. Since global features ([CLS] and [EOS]) are weighted aggregates of local features, the attention weights indicate the relevance between global and local features. We select the most informative local features based on the attention weights extracted from the last transformer block. By selecting a proportion of the corresponding local token features based on correlation weights, we perform feature transformation

to obtain more expressive representations. The feature transformation utilizes the same embedding module as in the residual block [38], as follows:

$$\begin{cases} f_{tfe}^v = \text{MaxPool}(\text{MLP}(\hat{f}_S^v) + \text{FC}(f_S^v)) \\ f_{tfe}^t = \text{MaxPool}(\text{MLP}(\hat{f}_S^t) + \text{FC}(f_S^t)) \end{cases} \quad (5)$$

where $f_S^v = \{f_1^v, f_2^v \dots f_k^v\}$ denotes the top- k most informative local features selected, $\hat{f}_S^v = \text{L2Norm}(f_S^v)$, $\hat{f}_S^t = \text{L2Norm}(f_S^t)$ are the image and text features after L2-normalization. Finally, we calculate the cosine similarity S_{tfe} between f_{tfe}^v and f_{tfe}^t , and the global feature similarity embedding S_{gfe} , in order to evaluate the cross-modal matching degree during both training and inference.

3.2. Noise Identification

In early training stages, clean data typically exhibits smaller loss values than noisy data, as deep neural networks learn from clean samples faster [39]. We employ two GMMs to fit the triplet alignment loss (TAL), calculated from GFE and token fusion embedding (TFE) similarities for each batch. The image–text similarity is modeled as a mixture of clean and noisy distributions.

For the i_{th} sample pair we define its TAL as l_i (l_i is defined in Equation (10)). The per-sample loss is input into the GMM, which is optimized using the expectation–maximization (EM) algorithm, then we compute the posterior probability, i.e.,

$$p(k|l_i) = p(k)p(l_i|k)/p(l_i) \quad (6)$$

where the posterior probability $p(k|\ell_i)$, $k \in \{0, 1\}$ denotes the probability that the i_{th} sample is classified as a clean or noisy sample pair. By setting a threshold δ , we can classify the dataset containing M image–text pairs into clean dataset $\hat{S}^c$ and noisy dataset $\hat{S}^n$, i.e.,

$$\hat{S}^c = \bigcup_{i=1}^M p(k=0|\ell_i) > \delta, \hat{S}^n = \bigcup_{i=1}^M p(k=1|\ell_i) < \delta \quad (7)$$

We set the noisy dataset obtained from the two Gaussian mixture models (GMMs) as $\hat{S}_{gfe}^n$, $\hat{S}_{tfe}^n$, respectively, and the final segmentation yields a noisy dataset of S^n and a clean dataset of S^c , i.e.,

$$S^n = \hat{S}_{gfe}^n \cap \hat{S}_{tfe}^n, S^c = \hat{S}_{gfe}^c \cap \hat{S}_{tfe}^c \quad (8)$$

For the rest of the data, i.e., $S' = S - (S^c \cup S^n)$, we randomly classify it as either a noisy set or a clean set.

3.3. Pseudo-Text Augmentation

Multimodal large language models (MLLMs) [12] have shown significant progress in tasks like image captioning. Models like BLIP-2 [40], MiniGPT-4 [13], and Qwen-VL can generate semantically rich and detailed descriptions. We choose the open-source Qwen-VL model for its strong performance across various vision–language tasks, computational efficiency, and suitability for iterative training.

To avoid monotonous sentence structures and semantic redundancy from static prompts, we adopt a diversified text generation strategy inspired by Tan et al. [11]. We provide real description corpora from the CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets to GPT-3.5 [41] to extract descriptive templates. After several dialogue rounds, ChatGPT generated 35 templates. For each pseudo-text generation, we randomly select one template and insert it into a static prompt to form a dynamic instruction:

“Generate a description about the overall appearance of the person, including clothing, shoes, hairstyle, gender, and belongings, in a style similar to the template: ‘{template}’. If

some requirements in the template are not visible, you can ignore them. Do not imagine any contents that are not in the image.”

We use Qwen-VL-Chat-7B [15] for its strong multimodal understanding, efficient local deployment, and stable large-scale batch generation, which avoids cloud API dependencies, reduces costs, and mitigates privacy concerns. Figure 3 shows an example of generated pseudo-text.

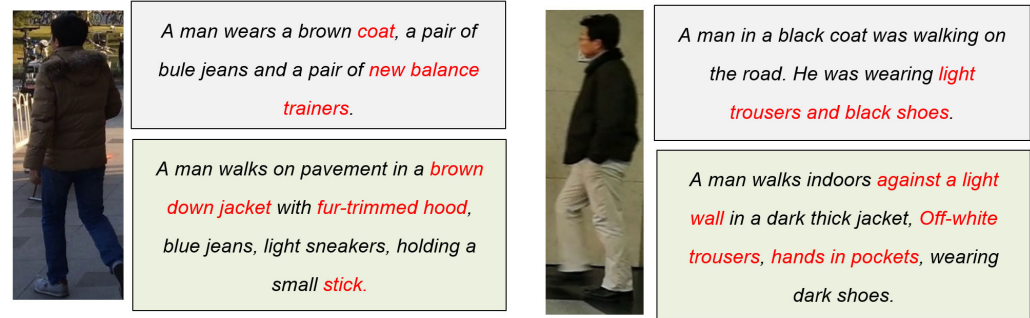


Figure 3. Pseudo-text generation example. The original dataset text (gray background) is replaced by the MLLM-generated pseudo-text (green background).

Despite MLLMs’ strong capabilities, generated text in low-quality image scenarios (e.g., severe occlusion, blur) may be inaccurate. To prevent “pseudo-text misrepresentations,” we propose a semantic consistency-based filtering mechanism. Let T_i^{noisy} be the original noisy text corresponding to an image T_i^{pseudo} and T_i^{pseudo} be the generated pseudo-text. We compute the semantic matching scores using the current model:

$$S_i^{noisy} = \cos(f_i^I, f_{T_i^{noisy}}^I), S_i^{pseudo} = \cos(f_i^I, f_{T_i^{pseudo}}^I) \quad (9)$$

If $S_i^{pseudo} > S_i^{noisy}$, the pseudo-text is considered higher quality and replaces the noisy text. Otherwise, the pair is discarded in the current training round to prevent low-quality descriptions from harming learning.

3.4. Loss Function

Common loss functions in TBPS include CLIP loss (InfoNCE) [42], triplet ranking loss (TRL) [43], and similarity distribution matching (SDM) loss [8]. However, these can be challenging due to difficulties in selecting the hardest negative samples and similarity margins, especially under annotation noise, potentially leading to local optima or model collapse.

In contrast, the triplet alignment loss (TAL) proposed by RDE [37] applies an upper bound constraint on all negative samples, mitigating overoptimization on the hardest negatives and enabling more stable training:

$$\mathcal{L}_{tal}(I_i, T_i) = m - f_v^I \hat{f}_t^I + \tau \log(\sum_{j \neq i}^K e^{(f_v^I \hat{f}_t^I / \tau)}) \quad (10)$$

where m denotes the margin value of the difference between positive and negative samples, f_v^I and f_t^I represent a batch of images and texts, and $\hat{f}_t^I$, f_t^I , and f_t^I are the hardest negative, negative, and positive samples of the anchor sample f_v^I samples in the batch.

While TAL stabilizes image–text pair matching at the global level, it does not explicitly guide fine-grained semantic alignment. TBPS queries often involve specific details (e.g., “red coat,” “black backpack”), which require implicit structural correspondences with image regions. Solely relying on global vector supervision may overlook such mappings.

To enhance fine-grained alignment, we incorporate the Implicit Relation Reasoning (IRR) module from IRRA [8]. This module randomly masks key words in the input text (e.g., colors, attire, actions, etc.) and uses a cross-attention mechanism between image regions and text context to guide the learning of implicit semantic correspondences. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{TAL} + \mathcal{L}_{IRR} \quad (11)$$

4. Experiments

4.1. Datasets

We evaluate our method on three public benchmarks using standard TBPS metrics.

ICFG-PEDES [38]: Includes 54,522 images of 4102 individuals, each with one text description. The training set has 34,674 pairs from 3102 individuals; the test set has 19,848 pairs from 1000 individuals. It features challenging backgrounds and lighting variations.

RSTPReid [44]: Consists of 20,505 images from 4101 pedestrians captured by 15 cameras. Each pedestrian has 5 images, each annotated with two texts. It contains complex indoor/outdoor scene transitions and varying backgrounds. The split is 3701 training, 200 validation, and 200 test individuals.

CUHK-PEDES [45]: Contains 40,206 images and 80,412 text descriptions for 13,003 identities. Each image has two manually annotated descriptions with rich appearance details. The training set includes 11,003 identities, 34,054 images, and 68,108 texts. A comprehensive summary of the key statistics for all three datasets is provided in Table 1.

Table 1. Dataset statistics.

Dataset	Identities	Images	Texts	Train Split	Val Split	Test Split
CUHK-PEDES	13,003	40,206	80,412	11,003 IDs, 34,054 images, 68,108 texts	-	2000 IDs/ images
ICFG-PEDES	4102	54,522	54,522	3102 IDs, 34,674 pairs	50.93	1000 IDs, 19,848 pairs
RSTPReid	4101	20,505	41,010	3701 IDs	200 IDs	200 IDs

4.2. Evaluation Protocols

We evaluate retrieval performance using the Rank-k accuracy ($k = 1, 5, 10$), where Rank-k denotes the percentage of text queries for which the correct image appears within the top-k retrieved results. In addition, we report the mean average precision (mAP) to assess the overall ranking quality and the mean inverse negative penalty (mINP), which reflects the model's ability to retrieve the most challenging samples. The computation of mAP follows the standard definition as the average of the AP values over all queries, while the formulation of mINP is adopted from reference [46]. Higher Rank-k, mAP, and mINP scores indicate better performance.

4.3. Experimental Details

All experiments use a CLIP-based visual-text backbone (CLIP-ViT-B/16). For text pre-processing we use the CLIP tokenizer with a fixed sequence length of 77 tokens, consistent with CLIP's pre-training. Short texts are right-padded with the tokenizer's pad token; texts longer than 77 tokens are truncated on the right (keeping the initial tokens). Token embedding and text encoder follow the CLIP pre-trained weights; only newly introduced modules (TFE, IRR, and MLLM adapter layers) are initialized randomly. Optimization is performed using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with weight decay = 1×10^{-4} . Learning rates are set to 1×10^{-5} for CLIP parameters and 1×10^{-3} for newly initialized modules. The primary batch size used in main experiments is 64 (we report batch-size sensitivity

separately). Models are trained for 60 epochs; noise identification is triggered at epoch 40 and pseudo-text replacement is performed at epoch 41. The GMMs used for dual-channel noise identification are two-component Gaussians (clean vs. noisy) fitted by expectation–maximization on TAL distributions. For pseudo-text generation we used a locally deployed MLLM (Qwen-VL-Chat-7B) with deterministic decoding (beam size = 1) and language set to English.

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

We compare our method with state-of-the-art approaches on three benchmarks. Based on preliminary experiments, we trigger noise identification at epoch 40 and pseudo-text generation at epoch 41. The generated pairs then replace noisy pairs for subsequent training. Results are shown in Tables 2–4. We compare our approach against several recent state-of-the-art methods for text-based person retrieval, including IRRA, RDE, RaSa, and TBPS-CLIP (see Tables 1–3). To guarantee a fair comparison, we align key experimental settings: all methods use a CLIP-ViT-B/16 backbone where applicable, are trained for 60 epochs, and adopt comparable input resolution and basic augmentation strategies. As reported in Tables 1–3, our method achieves Rank-1 = 68.13% on ICFG-PEDES, Rank-1 = 66.31% on RSTPReid, and Rank-1 = 75.98% on CUHK-PEDES, outperforming or matching the best reported baselines on these benchmarks. The superior top-k performance of our method is principally attributed to the dual-channel GMM noise identification and subsequent MLLM-based pseudo-text refinement combined with semantic consistency filtering, which together reduce the adverse impact of noisy annotations and strengthen discriminative supervision.

Table 2. Performance comparisons on ICFG-PEDES dataset.

Methods	ICFG-PEDES				
	R-1	R-5	R-10	mAP	mINP
DSSL [44]	39.05	62.60	73.95	-	-
SSAN [38]	43.50	67.80	77.15	-	-
IVT [19]	46.70	70.00	78.80	-	-
ISANet [47]	57.73	75.42	81.72		
LGUR [48]	57.42	74.97	81.45	-	-
BILMA [49]	63.83	80.15	85.74	38.26	
IRRA [8]	63.46	80.25	85.82	38.06	7.93
TBPS-CLIP [50]	65.05	80.34	85.47	39.83	-
RaSa [28]	65.28	80.40	85.12	41.29	9.97
RDE [37]	67.54	82.31	87.12	40.11	8.05
Ours	68.13	83.39	89.02	41.25	9.08

Bold numbers indicate the best performance.

Table 3. Performance comparisons on RSTPReid dataset.

Methods	RSTPReid				
	R-1	R-5	R-10	mAP	mINP
ViTAA [51]	50.98	68.79	75.78	-	-
SSAN [38]	54.23	72.63	79.53	-	-
IVT [19]	56.04	73.60	80.22	-	-
ACSA [52]	48.40	71.85	81.45	-	-
LGUR [48]	57.42	74.97	81.45	-	-
BILMA [49]	61.20	81.50	88.80	48.51	
TP-TPS [7]	50.65	72.45	81.20	43.11	-
IRRA [8]	60.20	81.30	88.20	47.17	25.82

Table 3. Cont.

Methods	RSTPReid				
	R-1	R-5	R-10	mAP	mINP
TBPS-CLIP [50]	61.95	83.55	88.75	48.26	-
RaSa [28]	65.90	86.50	91.35	52.31	29.23
RDE [37]	65.42	84.92	90.01	50.93	29.02
Ours	66.31	86.87	92.01	52.41	29.01

Table 4. Performance comparisons on CUHK-PEDES dataset.

Methods	CUHK-PEDES				
	R-1	R-5	R-10	mAP	mINP
DSSL [44]	59.89	80.41	87.56	-	-
SSAN [38]	61.37	80.15	86.73	-	-
IVT [19]	65.59	83.11	89.21	-	-
ACSA [52]	63.56	81.40	87.70	-	-
SAF [53]	64.13	82.62	88.40	58.61	-
TP-TPS [7]	70.16	86.10	90.98	66.32	-
IRRA [8]	73.38	89.93	93.71	66.13	50.24
UUMSA [54]	74.25	89.93	93.58	66.15	-
TBPS-CLIP [50]	73.54	88.19	92.35	65.38	-
RaSa [28]	76.51	90.29	94.25	67.49	51.11
RDE [37]	75.94	90.21	94.12	67.56	51.44
Ours	75.98	90.34	94.32	67.49	51.37

4.4.1. Comparison of ICFG-PEDES

As shown in Table 2, our method achieves the highest Rank-1 (68.13%), Rank-5 (83.39%), and Rank-10 (89.02%) accuracy on the ICFG-PEDES dataset. Although mAP (41.25) and mINP (9.08) are slightly lower than those of the best competitor, the superior top-k performance demonstrates its effectiveness in high-confidence retrieval, which is crucial for practical pedestrian–image matching.

4.4.2. Comparison of RSTPReid

Table 4 shows that our method achieves the best performance on RSTPReid, with Rank-1 (66.31%), Rank-5 (86.87%), and Rank-10 (92.01%) accuracy surpassing previous state-of-the-art methods by +0.41%, +0.37%, and +0.78%, respectively. This indicates its robustness on challenging datasets.

4.4.3. Comparison of CUHK-PEDES

Results in Table 4 show that our method achieves competitive performance on CUHK-PEDES. While Rank-1 (75.98%) is slightly lower than the best, it achieves the highest Rank-5 (90.34%) and Rank-10 (94.32%) accuracy, demonstrating strong top-k retrieval capability. The slightly lower Rank-1 may be attributed to the relatively higher image–text consistency in CUHK-PEDES (with accurate manual annotations), where the noise identification module might occasionally misclassify some valid pairs as noisy, introducing unnecessary pseudo-text generation.

4.5. Ablation Study

We conduct ablation studies on the three benchmarks to evaluate each component’s contribution. We compare the baseline (CLIP + TAL), the baseline integrated with the IRR module (+IRR), the baseline with MLLM pseudo-text augmentation and filtering (+MLLM), and our full model.

Results in Tables 5–7 show that introducing the IRR module consistently improves performance across all datasets, with Rank-1 accuracy increasing by 1.64%, 1.87%, and 1.11% on ICFG-PEDES, RSTPReid, and CUHK-PEDES, respectively. This confirms the importance of fine-grained semantic alignment via masked language modeling.

Table 5. Ablation study on the ICFG-PEDES.

Methods	ICFG-PEDES				
	R-1	R-5	R-10	mAP	mINP
Baseline	66.12	81.76	88.01	40.12	7.94
+IRR	67.76	82.14	88.14	40.95	8.32
+MLLM	67.88	82.16	88.71	40.84	8.49
Our	68.13	83.39	89.02	41.25	9.08

Table 6. Ablation study on the RSTPReid.

Methods	RSTPReid				
	R-1	R-5	R-10	mAP	mINP
Baseline	64.05	84.68	88.89	50.05	27.84
+IRR	65.92	85.19	88.56	51.43	28.61
+MLLM	65.97	85.34	88.62	50.92	28.46
Our	66.31	86.87	92.01	52.41	29.01

Table 7. Ablation study on the CUHK-PEDES.

Methods	CUHK-PEDES				
	R-1	R-5	R-10	mAP	mINP
Baseline	73.56	88.19	92.08	65.67	50.26
+IRR	74.67	89.54	93.65	67.02	50.83
+MLLM	75.19	90.11	94.36	67.21	50.92
Our	75.98	90.34	94.32	67.49	51.37

Similarly, the pseudo-text enhancement strategy (i.e., replacing noisy descriptions with pseudo-text generated by MLLM after semantic consistency filtering) further improves retrieval performance, especially on datasets with higher noise levels like RSTPReid, where annotation quality is less reliable. When both modules are combined, the proposed method achieves the highest retrieval scores on most metrics, confirming that the two components complement each other. Although the improvements in mAP and mINP are moderate, the significant gains in Rank-1 and Rank-5 demonstrate the system’s stronger robustness and higher retrieval confidence under noisy supervision. These results validate the effectiveness of our approach in mitigating the negative effects of noisy annotations while improving fine-grained cross-modal alignment.

We additionally examine how the batch size and loss-related hyperparameters affect model behavior. On the RSTPReid dataset, we systematically varied the margin m and temperature τ on the RSTPReid dataset under identical conditions. The results (Supplementary Figure S1) show that performance peaks at $m = 0.1$; both overly small (<0.05) and large (>0.15) margins reduce retrieval accuracy. Similarly, the temperature parameter is critical for stability: values below $\tau = 0.01$ cause instability, while $\tau > 0.02$ degrades discrimination. Based on this, we selected $m = 0.1$ and $\tau = 0.015$ as optimal. These observations confirm the robustness and generalizability of the proposed pseudo-text enhancement mechanism across different noise levels.

We denote the selection ratio (threshold) as δ —the fraction of training samples flagged as noisy by the dual-channel GMM procedure and subject to pseudo-text replacement. The final δ used in all main experiments is 0.30 (30%). This value was selected by grid search on a held-out validation split using candidate values {0.10, 0.20, 0.30, 0.40, 0.50}. Selection was based on two criteria: (1) noise identification precision (the fraction of flagged samples that indeed had incorrect descriptions upon manual inspection), and (2) downstream retrieval performance (validation Rank-1 and mAP). $\delta = 0.30$ offered the best trade-off between identifying sufficient noisy samples and avoiding excessive replacement of borderline clean samples. Supplementary Table S2 presents the sensitivity analysis; an illustrative example is included below.

5. Conclusions

In this paper, we addressed the challenge of noisy annotations in TBPS by integrating a noise-aware pseudo-text generation strategy with a fine-grained semantic reasoning module. Our approach improves retrieval performance across multiple benchmarks. Empirical results show that the proposed method consistently enhances top-k retrieval accuracy, particularly in scenarios with mismatched or coarse textual descriptions. While gains in overall mAP and mNP are modest in some cases, our method exhibits clear advantages in high-confidence retrieval. Ablation studies confirm that both the IRR module and the pseudo-text replacement mechanism contribute to the model's robustness. Rather than proposing a completely new paradigm, our work provides a practical enhancement to existing CLIP-based TBPS pipelines by explicitly modeling and mitigating annotation noise. Future research may build upon this framework to explore advanced noise modeling or dynamic curriculum learning in multimodal retrieval.

Despite the strong top-k improvements achieved on three public benchmarks, the proposed framework still presents several limitations. (1) When input images are of extremely poor quality, the pseudo-text generated by the MLLM may miss key discriminative attributes, which can reduce the effectiveness of text replacement. (2) The discriminative capability of the dual-channel GMM relies on the model's ability to learn clean samples during the early training stage; under scenarios with extremely high noise ratios, additional mechanisms such as dynamic curriculum learning or more sophisticated noise modeling may be required. (3) Pseudo-text generation introduces additional computational and resource overhead. Although we deploy Qwen-VL-Chat-7B locally to mitigate latency and privacy concerns, resource-constrained environments may still face practical limitations. (4) Our experiments focus on English datasets and pedestrian description tasks; thus, the generalization of the proposed method to multilingual settings or more complex long-text scenarios remains to be validated. Future work will explore adaptive thresholding, joint noise modeling, and more lightweight locally deployed MLLMs to alleviate these issues.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/electronics14234619/s1>, Figure S1. Variation of performance with different m and τ on RSTPReid. Table S1. Effect of batch size on ICFG-PEDES dataset. Table S2. Effect of δ on ICFG-PEDES.

Author Contributions: Conceptualization, L.X. and L.Y.; Methodology, L.X. and L.Y.; Software, W.L. and Y.F.; Validation, W.L. and Y.F.; Formal analysis, L.X., W.L. and L.Y.; Investigation, W.L.; Resources, L.X. and L.Y.; Data curation, W.L. and Y.F.; Writing—original draft preparation, W.L. and Y.F.; Writing—review and editing, L.X. and L.Y.; Visualization, W.L.; Supervision, L.X.; Project administration, L.X.; Funding acquisition, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 42101348), the Natural Science Foundation of Sichuan Province (Grant No. 2025NSFSC0328), and the Natural Science Foundation of Chongqing under Grants CSTB2022NSCQ-MSX1132.

Data Availability Statement: The datasets analyzed during this study are available in the following public domain resources. CUHK-PEDES dataset: <https://github.com/layumi/Image-Text-Embedding/tree/master/dataset/CUHK-PEDES-prepare>, accessed on 24 November 2017. RSTPReid dataset: <https://github.com/NjtechCVLab/RSTPReid-Dataset>, accessed on 12 September 2021. Requests for additional information should be directed to xionglian@cqupt.edu.cn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ding, C.; Tao, D. Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1002–1014. [CrossRef]
2. Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; He, K. Scaling language-image pre-training via masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23390–23400.
3. Miech, A.; Alayrac, J.-B.; Laptev, I.; Sivic, J.; Zisserman, A. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9826–9836.
4. Niu, K.; Huang, Y.; Ouyang, W.; Wang, L. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Trans. Image Process.* **2020**, *29*, 5542–5556. [CrossRef]
5. Chen, T.; Xu, C.; Luo, J. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1879–1887.
6. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13–23.
7. Wang, G.; Yu, F.; Li, J.; Jia, Q.; Ding, S. Exploiting the textual potential from vision-language pre-training for text-based person search. *arXiv* **2023**, arXiv:2303.04497.
8. Jiang, D.; Ye, M. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 2787–2797.
9. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
10. Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; Hengel, A.v.D. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.* **2017**, *163*, 21–40. [CrossRef]
11. Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; Tao, D. Harnessing the power of mllms for transferable text-to-image person reid. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17127–17137.
12. Wu, J.; Gan, W.; Chen, Z.; Wan, S.; Yu, P.S. Multimodal large language models: A survey. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 2247–2256.
13. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
14. Radford, A.; Kim, J.W.; Hallacy, C. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Vienna, Austria, 18–24 July 2021; pp. 8748–8763.
15. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv* **2023**, arXiv:2308.12966.
16. Zhang, Y.; Lu, H. Deep cross-modal projection learning for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 686–701.
17. Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; Shen, Y.-D. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimedia Comput. Commun. Appl.* **2020**, *16*, 1–23. [CrossRef]
18. Gao, C.; Cai, G.; Jiang, X.; Zheng, F.; Zhang, J.; Gong, Y.; Sun, X. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv* **2021**, arXiv:2101.03036.

19. Shu, X.; Wen, W.; Wu, H.; Chen, K.; Song, Y.; Qiao, R.; Wang, X. See finer, see more: Implicit modality alignment for text-based person retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 624–641.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Graves, A.; Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer Nature: Durham, NC, USA, 2012; pp. 37–45.
23. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
24. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CSUR)* **2019**, *51*, 1–36. [[CrossRef](#)]
25. Cao, M.; Li, S.; Li, J.; Nie, L.; Zhang, M. Image-text retrieval: A survey on recent research and development. *arXiv* **2022**, arXiv:2203.14713. [[CrossRef](#)]
26. Han, X.; He, S.; Zhang, L.; Xiang, T. Text-based person search with limited data. *arXiv* **2021**, arXiv:2110.10807. [[CrossRef](#)]
27. Yan, S.; Dong, N.; Zhang, L.; Tang, J. Clip-driven fine-grained text-image person re-identification. *IEEE Trans. Image Process.* **2023**, *32*, 6032–6046. [[CrossRef](#)]
28. Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Zhang, M. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv* **2023**, arXiv:2305.13653.
29. Han, H.; Miao, K.; Zheng, Q.; Luo, M. Noisy correspondence learning with meta similarity correction. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7517–7526.
30. Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; Peng, X. Learning with noisy correspondence for cross-modal matching. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29406–29419.
31. Shao, Z.; Zhang, X.; Ding, C.; Wang, J.; Wang, J. Unified pre-training with pseudo texts for text-to-image person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 11174–11184.
32. Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M.S.; Lacoste-Julien, S. A closer look at memorization in deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 11 August 2017; pp. 233–242.
33. Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; Peng, X. Learning with twin noisy labels for visible-infrared person re-identification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14308–14317.
34. Hu, P.; Huang, Z.; Peng, D.; Wang, X.; Peng, X. Cross-modal retrieval with partially mismatched pairs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9595–9610. [[CrossRef](#)] [[PubMed](#)]
35. Yan, S.; Dong, N.; Liu, J.; Zhang, L.; Tang, J. Learning comprehensive representations with richer self for text-to-image person re-identification. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 6202–6211.
36. Qin, Y.; Peng, D.; Peng, X.; Wang, X.; Hu, P. Deep evidential learning with noisy correspondence for cross-modal retrieval. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 4948–4956.
37. Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J.T.; Hu, P. Noisy-correspondence learning for text-to-image person re-identification. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 27197–27206.
38. Ding, Z.; Ding, C.; Shao, Z.; Tao, D. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv* **2021**, arXiv:2107.12666.
39. Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8527–8537.
40. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
41. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Amodei, D. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
42. van den Oord, A.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.

43. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv* **2017**, arXiv:1707.05612.
44. Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hua, G. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 209–217.
45. Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; Wang, X. Person search with natural language description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1970–1979.
46. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [[CrossRef](#)]
47. Yan, S.; Tang, H.; Zhang, L.; Tang, J. Image-specific information suppression and implicit local alignment for text-based person search. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 17973–17986. [[CrossRef](#)]
48. Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; Ding, C. Learning granularity-unified representations for text-to-image person re-identification. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 5566–5574.
49. Fujii, T.; Tarashima, S. Bilma: Bidirectional local-matching for text-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 2786–2790.
50. Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; Zhang, M. An empirical study of clip for text-based person search. In Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 465–473.
51. Wang, Z.; Fang, Z.; Wang, J.; Yang, Y. Vitaa: Visual-textual attributes alignment in person search by natural language. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 402–420.
52. Ji, Z.; Hu, J.; Liu, D.; Wu, L.Y.; Zhao, Y. Asymmetric cross-scale alignment for text-based person search. *IEEE Trans. Multimed.* **2022**, *25*, 7699–7709. [[CrossRef](#)]
53. Li, S.; Cao, M.; Zhang, M. Learning semantic-aligned feature representation for text-based person search. In Proceedings of the ASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 2724–2728.
54. Zhao, Z.; Liu, B.; Lu, Y.; Chu, Q.; Yu, N. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 7534–7542.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.