CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Align What Truly Matters: Pedestrian-Relevant Hierarchical Parsing Network for Text-Based Person Retrieval

Jiajun Su[1] | Yifan Shi[1] | Pudu Liu[1] | Jianqing Zhu[1] | Huanqiang Zeng[1,2] | Liu Liu[3,4] | Zhen Lei[5,6,7]

[1]College of Engineering, Huaqiao University, Quanzhou, China | [2]School of Optoelectronic and Communication Engineering, Xiamen University of Technology, Xiamen, China | [3]School of Artificial Intelligence, Beihang University, Beijing, China | [4]Hangzhou International Innovation Institute, Beihang University, Hangzhou, China | [5]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China | [6]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China | [7]School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China

Correspondence: Jianqing Zhu (jqzhu@hqu.edu.cn)

Received: 28 February 2026 | **Revised:** 20 May 2026 | **Accepted:** 4 June 2026

Keywords: cross-modal retrieval | mixture of experts | pedestrian feature parsing

ABSTRACT

Fine-grained alignment is crucial for text-based person retrieval, which searches for relevant pedestrian images using a text query. However, background clutter and semantically vacuous words can cause interference and misalignment, hindering retrieval performance. Existing methods employ masked language modelling to randomly mask and predict words, or leverage multimodal large language models to generate diverse and pure samples. However, these methods rely on external tools and are not sufficiently stable or controllable. In this paper, we propose a pedestrian-relevant hierarchical parsing (PHP) module to extract well-aligned fine-grained visual and textual features for alignment. First, we design a coarse relevant feature mapping (CRFM) module, which uses learnable unified tokens to project both modalities into a shared low-dimensional space, enabling coarse-level semantic filtering. Next, we design an expert-driven feature parsing (EFP) module, which integrates the representational power of a mixture of experts with a modality-aware gating mechanism to uncover deep semantic associations between text and image features. Extensive experiments on three public large-scale datasets demonstrate that our approach outperforms existing state-of-the-art methods, achieving, in particular, a 5.68% improvement in mAP and a 4.81% boost in R1 accuracy on the CUHK-PEDES dataset.

1 | Introduction

Text-based person retrieval (TBPR) aims to locate a specific individual within large-scale image collections based on natural language queries. Unlike traditional image-based [1, 2] or attribute-based [3, 4] person search methods, TBPR offers greater flexibility and expressiveness, enabling intuitive interaction in visual search engines, multimedia management systems and advanced human–computer interfaces. These advantages have sparked substantial research interest in multimodal intelligence and collaborative perception.

Recent progress in large-scale vision–language pre-training (VLP) models [5, 6] has greatly advanced cross-modal representation learning. Among them, contrastive language image pre-training (CLIP) [5] stands out for its capability to learn rich semantic associations from image-text pairs. Consequently, many TBPR approaches [7–9] use CLIP as the feature encoder.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

However, CLIP is primarily trained with global captions and therefore often fails to capture fine-grained details—such as specific colours or spatial relationships—that are crucial for TBPR. This limitation becomes particularly problematic for fine-grained alignment in TBPR, which faces two key challenges: (1) irrelevant background clutter and semantically uninformative words in raw data, and (2) the difficulty of establishing precise part-level correspondences between modalities. As illustrated in Figure 1a, both images and their text descriptions typically contain complex backgrounds and generic phrases (e.g., "windows," "roads" "and," "in this photo"), which frequently induce semantic misalignment and hinder precise attribute matching. Although CLIP exhibits limitations in capturing fine-grained features, it remains one of the most powerful and accessible open-vocabulary vision–language foundation models. Its large-scale pretraining on diverse image–text pairs enables strong cross-modal alignment, which is crucial for open-world TBPR scenarios. In addition, CLIP's relatively lightweight architecture, compared to other vision–language models (e.g., ALBEF [6]), provides significant computational efficiency and deployment flexibility, which makes CLIP a strong foundation for TBPR.

To improve CLIP for TBPR, current methods can be broadly categorised into two directions. The first direction leverages masked language model (MLM) strategies [10–12], which modify the training objective rather than the dataset itself. MLM enhances fine-grained cross-modal interaction by predicting masked tokens, improving the alignment between visual and textual features. For instance, Jiang et al. [11] employ relation reasoning in conjunction with MLM to achieve better correspondence. The second direction focuses on data-centric generative pre-training [13–15], which uses diffusion models [16] and large multimodal language models [17, 18] to generate diverse pedestrian image–text pairs, further boosting visual–language alignment. However, both approaches have limitations. While MLM-based methods improve fine-grained alignment within the high-dimensional feature space, they still suffer from noisy representations due to irrelevant backgrounds and semantically empty words. On the other hand, generative pre-training enhances data diversity but primarily focuses on global semantic correspondences, neglecting fine-grained attribute distinctions. This indicates that data-centric improvements alone are insufficient and that explicit architectural mechanisms are crucial to achieve robust fine-grained alignment.

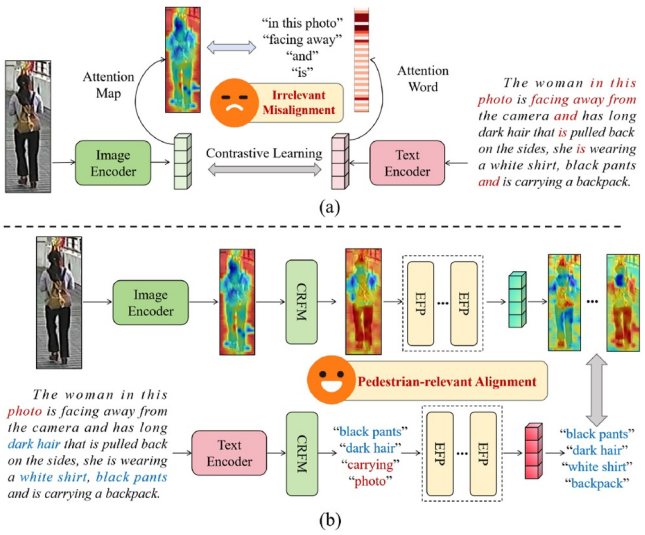


FIGURE 1 | (a) Irrelevant misalignment caused by direct feature alignment in existing methods. (b) Our proposed pedestrian-relevant hierarchical parsing (PHP) framework.

In contrast to these approaches, architectural innovations, such as the mixture of experts (MoE) paradigm [19–21], offer an alternative path for improving cross-modal alignment. MoE utilises dynamic expert selection through a gating mechanism, enabling specialisation in processing distinct semantic or attribute-level features (e.g., clothing colour, texture or spatial arrangement). This allows the model to focus on the most relevant fine-grained cues per instance, effectively decomposing the alignment task into specialised sub-tasks. Given the challenges of fine-grained alignment in TBPR, MoE's dynamic expert selection offers a promising solution to enhance precision.

Building upon these insights, in this paper, we introduce the pedestrian-relevant hierarchical parsing (PHP) framework to address two critical issues in TBPR. First, noise-induced misalignment: Irrelevant backgrounds in images and semantically vacuous words in text often create interference, leading to false attribute matches. Second, fine-grained correspondences are difficult to establish when noise persists in the high-dimensional feature space. PHP tackles these challenges with a targeted two-stage design. The coarse relevant feature mapping (CRFM) module reduces noise by projecting both modalities into a unified low-dimensional space. By using learnable semantic anchors and attention-weighted aggregation, CRFM suppresses irrelevant information before alignment, thus minimising the noise that current methods struggle to filter. The expert-driven feature parsing (EFP) module then works on the noise-reduced representations produced by CRFM, employing MoE with modality-aware gating to establish precise part-level matches. By explicitly linking each module to its corresponding problem, PHP effectively improves cross-modal alignment where existing methods fall short. Extensive experiments on CUHK-PEDES [22], ICFG-PEDES [23] and RSTPReid [24] datasets demonstrate that PHP achieves state-of-the-art accuracy and high efficiency, underscoring its potential for practical deployment.

The main contributions of this paper are as follows:

- We propose a novel pedestrian-relevant hierarchical parsing (PHP) framework. It integrates a full-parameter mixture-of-experts (MoE) architecture, in which both the backbone encoders (e.g., image and text CLIP encoders) and the expert modules are jointly trained, with all parameters updated.
- We design a coarse relevant feature mapping (CRFM) module that projects visual and textual information into a unified low-dimensional concept space, enabling coarse-level semantic filtering.
- We develop an expert-driven feature parsing (EFP) module, which integrates MoE with a modality-aware gating

strategy to enable fine-grained alignment and deeper semantic association between text and images.
- Extensive experiments on three public benchmarks validate the effectiveness and efficiency of our approach, achieving superior results compared to prior state-of-the-art methods.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 details our proposed approach. Experimental results and analysis are presented in Section 4. Section 5 concludes the paper.

2 | Related Works

In this section, we review the development of text-based person retrieval (TBPR), noise filtering methods and then discuss the mixture of experts (MoE) paradigm—a prominent advance in large model architectures for improving feature extraction and visual-textual alignment in TBPR.

2.1 | Text-Based Person Retrieval

Early TBPR methods adopt diverse unimodal backbone networks for feature extraction. For images, architectures such as ResNet and vision transformer (ViT) are typical choices, whereas LSTM and BERT are standard for encoding text. Features are then combined through global or local matching strategies. Global matching methods [25–27] project images and text into a shared embedding space using cross-modal loss functions, though these approaches may overlook fine-grained correspondences even with sophisticated losses. Li et al. [27] learn semantic-aligned feature representations by automatically aligning visual and textual features across modalities. Local matching, by contrast, explicitly models correspondences between salient visual regions (e.g., body parts) [28, 29] and specific text phrases [23, 30, 31], enhancing alignment at a detailed level. For example, Wu et al. [31] propose a local alignment framework that models image–phrase relations with unbiased masking strategies.

Later, implicit matching mechanisms have emerged, utilising masked language and vision modelling [32] for more nuanced visual-textual cues. For example, Farooq et al. [32] proposed a network that dynamically leverages multi-scale features from both modalities for shared semantic recalibration. More recently, vision-language pre-training models, such as CLIP [5] and ALBEF [6], have enabled strong cross-modal representation. Current research thus often focuses on transferring these pretrained knowledge structures to TBPR.

To further improve fine-grained alignment, two main strategies are prevalent. Masked language modelling (MLM)-based methods mask text tokens and predict them from image regions, driving implicit alignment (see, e.g. [10, 11]). Generative pre-training approaches synthesise diverse data via diffusion models [16] or prompt-based template generation [14], using large language/multimodal models [17, 33]. Although the above methods improve representation diversity, they still struggle to fully address the inherent diversity and complexity of TBPR's multimodal alignment needs.

2.2 | Noise Filter Method

To address noisy correspondence in cross-modal retrieval, existing methods can be broadly categorised into sample selection [34, 35] and robust loss functions [36, 37]. Sample selection methods typically leverage the memorisation effect of DNNs to distinguish noisy data from clean data, thereby allocating more attention to reliable samples during training. The meta similarity correction network [34] tackles noisy correspondence in cross-modal retrieval by proposing a meta similarity correction with meta-learning for reliable similarity computation and data purification to filter out mismatched pairs. In contrast, robust loss function approaches aim to develop noise-tolerant objectives that improve model robustness against label noise. The dually robust training method [37] tackles twin noise labels (noisy annotation and correspondence) by leveraging memorisation effects to compute clean confidence and introducing a dual robust loss for noise-resilient training.

Recently, several methods have been proposed to address noise in text-based person retrieval. The simple and robust correlation filtering approach [38] designs denoising filters that compute similarity between input features and learnable foreground/background filter vectors, formulating noise filtering as binary separation with a mutual-exclusion loss to suppress irrelevant backgrounds and uninformative words. The robust dual embedding method [8] focuses on noisy correspondence where mismatched image-text pairs are incorrectly labelled as matched, proposing a confident consensus division module for clean pair identification and a triplet alignment loss for stable optimisation. Furthermore, recent advances have focused on mitigating noisy correspondence for cross-modal retrieval. For example, the key phrase dynamic masking method [39] preserves adjective–noun relationships, along with a frequency-based masked language loss to facilitate more robust fine-grained alignment. The geometry-aware noisy correspondence mitigation approach [40] alleviates noisy correspondence by enforcing geometric consistency across cross-modal and intra-modal structures. The correspondence calibrating and dynamic consistency learning method [41] tackles this issue by utilising a co-teaching paradigm to calibrate confidence scores for positive samples and employing dynamic margin adaptation to accurately discriminate negative samples.

In contrast to these approaches, our proposed CRFM introduces an early semantic bottleneck with two key distinguishing properties: (i) it employs shared learnable unified tokens to project multimodal features into a unified low-dimensional space for robust alignment; and (ii) it performs noise suppression via relevance-weighted aggregation rather than explicit masking, margin adaptation or post-hoc correction. This early filtering mechanism ensures smoother optimisation and provides more stable compact representations for downstream expert routing.

2.3 | Mixture of Experts

The mixture of experts (MoE) paradigm offers an effective way to scale model capacity while preserving computational efficiency, by selectively activating specialised sub-networks for

different inputs or tasks [19–21, 42]. Recognising its rapid growth across diverse domains, Cai et al. [42] provided a comprehensive survey covering architectural principles, a novel taxonomy, key algorithmic and system design choices, open-source implementations, empirical analyses and practical applications—offering researchers a systematic roadmap for navigating this expanding field. This flexibility allows models to better handle diverse and complex input distributions. MoE has been widely studied in natural language processing, computer vision and multimodal learning: Zoph et al. [20] proposed a robust MoE for NLP and Riquelme et al. [21] scaled vision transformers via MoE designs, highlighting MoE's adaptability and efficiency across domains. Beyond centralised training, MoE has also been extended to distributed and personalised settings, such as personalised federated learning with mixture of experts by Yi et al. [19], which enables fine-grained data-level personalisation in model-heterogeneous federated learning, further broadening its applicability.

Recently, MoE has driven significant advancements in overcoming domain-specific challenges across diverse and complex fields. For example, the context-aware adaptive learning for dynamic knowledge graph embedding method [43] utilises specialised expert networks to dynamically integrate relational paths, whereas the degradation-customised mixture-of-experts approach [44] enhances multi-modal image fusion by coupling explicit degradation estimation with conditional expert routing and gated decoding. This adaptability has been particularly beneficial in person re-identification and retrieval tasks, where handling cross-domain and cross-modal variations is critical. Besides, the source-free adaptive gated expert method [45] leverages MoE to dynamically merge low-rank adapters for efficient multi-source domain adaptation. However, despite these advantages, the deployment of MoE still faces fundamental challenges, such as routing instability and expert collapse, which can hinder optimisation in unconstrained environments.

In text-based person retrieval, MoE has mainly been applied in parameter-efficient architectures [9, 46, 47]. These methods employ lightweight expert modules to achieve robust feature extraction without significant computational overhead, rendering them suitable for low-resource scenarios and helping reduce overfitting. However, the full potential of MoE for capturing complex cross-modal correspondences—such as fine-grained matching of person attributes and textual descriptions—remains largely untapped. Full-parameter MoE could in theory endow models with greater capacity for such complex modelling, but in practice it has experienced limited deployment in TBPR due to challenges like overfitting to irrelevant background or redundant text cues. Thus, how to effectively leverage the expressive power of full-parameter MoE while ensuring robust and accurate expert assignment for fine-grained retrieval remains an important open problem motivating the present work.

3 | Proposed Method

3.1 | Overview

As shown in Figure 2a, PHP comprises two modules, CRFM and EFP, forming a bottleneck-first, routing-second coarse-to-fine pipeline. In TBPR, high-capacity MoE models are often susceptible to background and lexical noise, which may lead to expert misrouting. To alleviate this issue, CRFM serves as a semantic bottleneck to suppress noise, whereas EFP leverages the purified features to enhance fine-grained expressiveness beyond CRFM alone. These two modules are not merely sequential but complementary. The aggressive coarse filtering introduced by CRFM can reduce noise but may also constrain representational capacity. EFP is therefore designed to compensate for this limitation by recovering and enriching the expressive capability. Built upon the purified features, the gating mechanism in EFP is better positioned to route features to specialised experts, facilitating more precise part-level alignment with reduced interference from noise. Overall, CRFM helps stabilise the routing process by suppressing noise, whereas EFP further enhances the fine-grained expressiveness that may be limited by the coarse bottleneck.

Given an image-text pair, we use the CLIP image and text encoders [5] to extract features. Specifically, the image encoder yields $F_{v+1} \in \mathbb{R}^{(n_1+1)\times d}$, where the first token is a learnable [CLS] token followed by n_1 visual patch embeddings. The text encoder produces $F_t \in \mathbb{R}^{n_2\times d}$, which contains n_2 textual tokens including the special [SOS] and [EOS] tokens. To obtain the partial visual features, we exclude the [CLS] token and get $F_v = F_{v+1}[1 : n_1 + 1, :] \in \mathbb{R}^{n_1\times d}$.

Next, F_v and the complete text features F_t are input to the CRFM module, resulting in a coarse unified representation F_{cur}. This initial representation is then refined by a stack of EFP modules, ultimately generating fine-grained parsing features F_p. Meanwhile, the [CLS] token output from the image encoder and the [EOS] token from the text encoder are each mapped through separate linear projection layers to produce the global image feature $g_v \in \mathbb{R}^d$ and the global text feature $g_t \in \mathbb{R}^d$. Finally, the parsed features F_p are respectively combined with the corresponding global features to form the vision-relevant feature F_{vr} and the text-relevant feature F_{tr}. The specific designs and roles of each module will be detailed in the subsequent subsections.

3.2 | Coarse Relevant Feature Mapping

CRFM acts as a semantic bottleneck to provide a cleaner foundation for fine-grained parsing. Raw multimodal inputs often contain background and lexical noise that can lead to expert misrouting. As shown in Figure 2b, learnable unified tokens aggregate pedestrian-centric cues into compact, low-dimensional representations, helping reduce the influence of irrelevant information on downstream routing.

First, we employ a unified attention mechanism, utilising a set of learnable tokens $U \in \mathbb{R}^{l\times d}$ that act as semantic anchors. Here, we take the visual feature as an example. The text feature follows the same principle. For a batch of visual features $F = \{F_v^i \in \mathbb{R}^{n_1\times d} \mid i = 1, 2, ..., b\} \in \mathbb{R}^{b\times n_1\times d}$, we compute coarse attention weights $\sigma = \{\sigma_i \in \mathbb{R}^{l\times n_1} \mid i = 1, 2, ..., b\} \in \mathbb{R}^{b\times l\times n_1}$ based on their similarity to the unified tokens, as described in Equation (1):

$$\sigma_i = \mathrm{Softmax}\left(U\left(F_v^i\right)^{\mathrm{T}}, \mathrm{dim} = -1\right) \in \mathbb{R}^{l\times n_1}, \tag{1}$$

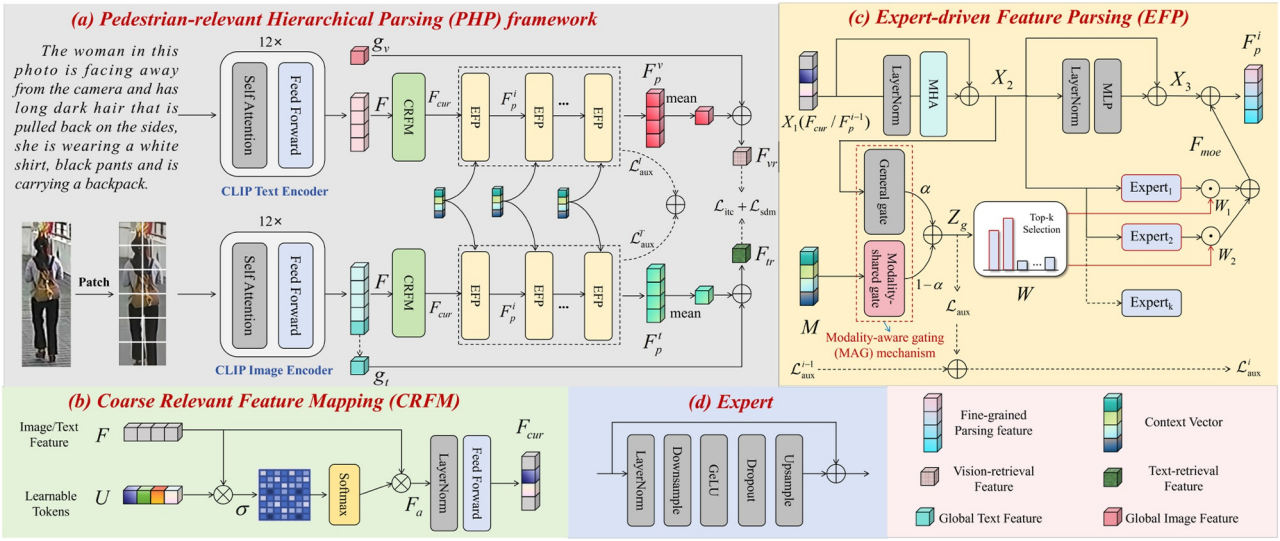


FIGURE 2 | Overview of the PHP framework, illustrating the hierarchical feature parsing via CRFM and EFP modules.

where T denotes the transpose, and softmax is applied along the patch/token dimension (i.e., the dimension regarding n_1). The attention weights σ are then used to aggregate input features in a relevance-weighted manner, resulting in F_a as calculated in Equation (2):

$$F_a = \mathrm{BMM}(\sigma, F)/\sqrt{d} \in \mathbb{R}^{b\times l\times d}, \tag{2}$$

where BMM(·) denotes batch matrix multiplication.

To further improve feature expressiveness, F_a is processed with layer normalisation (LN) and subsequently passed through a multi-layer perceptron (MLP). This produces the coarse unified features F_{cur}, as defined in Equation (3):

$$F_{cur} = \mathrm{MLP}(\mathrm{LN}(F_a)) \in \mathbb{R}^{b\times l\times d}, \tag{3}$$

where the MLP consists of two linear layers with GeLU activation and dropout (probability 0.3) applied after the first layer. The first layer, followed by a Gaussian error linear unit (GeLU) activation, expands the feature dimension to $4d$, whereas the second layer projects the feature back to dimension d.

Overall, by explicitly modelling the relevance between input features and unified semantic anchors Equation (1), and by further enhancing the aggregated representation through relevance-weighting Equation (2) and nonlinear projection Equation (3), CRFM provides a transparent and interpretable mechanism for noise filtering and feature enhancement.

3.3 | Expert-Driven Feature Parsing

Whereas the coarse unified feature F_{cur} filters most cross-modal noise, its aggressive compression may limit representational capacity. As shown in Figure 2c, EFP complements this by restoring fine-grained details. Benefiting from the purified inputs, its modality-aware gating can more reliably route features to specialised experts, facilitating precise part-level alignment with reduced noise interference.

We begin by enhancing the input features using the multi-head attention (MHA) block, which establishes a solid basis for subsequent expert specialisation. This process is formally defined in Equation (4):

$$X_2 = \mathrm{MHA}(\mathrm{LN}(X_1)) + X_1 \in \mathbb{R}^{b\times l\times d}, \tag{4}$$

where X_1 denotes the original features and X_2 the enhanced features; LN indicates layer normalisation and MHA represents multi-head self-attention [48] with $h = 8$ heads and shared projections.

Building on X_2, multiple expert networks are deployed to capture different visual–textual alignment patterns. Each expert follows a residual feed-forward structure for stable optimisation, as defined by Equation (5):

$$\mathrm{Expert}(X_2) = X_2 + \mathrm{MLP}(\mathrm{LN}(X_2)) \in \mathbb{R}^{b\times l\times d}, \tag{5}$$

where $\mathrm{Expert}_i(\cdot)$ is the i-th expert. The MLP here consists of two linear layers with GeLU activation and dropout (probability 0.3) after the first layer. The first layer expands the feature dimension to $8d$ (followed by GeLU), whereas the second projects it back to d.

For adaptive expert routing, the EFP module introduces a modality-aware gating mechanism that combines two gating strategies: a general gate and a modality-shared gate, computed as in Equations (6) and (7):

$$\mathrm{E}_g(X_2) = X_2 W_g \in \mathbb{R}^{b\times l\times n}, \tag{6}$$

$$\mathrm{E}_{ms}(M) = M W_{ms} \in \mathbb{R}^{n}, \tag{7}$$

where $\mathrm{E}_g(\cdot)$ denotes the general gate with projection $W_g \in \mathbb{R}^{d\times n}$, $\mathrm{E}_{ms}(\cdot)$ is the modality-shared gate with projection $W_{ms} \in \mathbb{R}^{d\times n}$, $M \in \mathbb{R}^d$ is a modality-shared context vector and n indicates the number of experts.

The final gate logits Z_g are computed by fusing these two strategies as in Equation (8):

$$Z_g = \alpha \cdot \mathrm{E}_g(X_2) \oplus (1-\alpha) \cdot \mathrm{E}_{ms}(M) \in \mathbb{R}^{b \times l \times n}, \tag{8}$$

where $\oplus$ denotes broadcasting addition, and $\alpha \in [0,1]$ is a learnable balance parameter. This fusion allows the gating mechanism to adaptively integrate both token-level and modality-level cues.

Subsequently, by sorting Z_g along the expert dimension (i.e., the dimension regarding n), the top-k experts with the highest logits are selected to form the set $\mathrm{SExpert} = \{\mathrm{Expert}_1, \mathrm{Expert}_2, ..., \mathrm{Expert}_k\}$. The corresponding gate weights are then calculated as Equation (9):

$$W = \mathrm{Softmax}\big(\mathrm{Topk}(Z_g), \mathrm{dim} = -1\big) \in \mathbb{R}^{b \times l \times k}, \tag{9}$$

where Topk keeps the k largest logits along the expert dimension, ensuring sparse expert activation. The notation $\mathrm{dim} = -1$ specifies that the softmax function is computed over the last dimension (i.e., the dimension regarding k). The use of sparse Top-k routing is mainly motivated by two considerations. First, by activating only a subset of highly responsive experts for each token, different experts can focus on different semantic patterns or pedestrian attributes [49], which helps promote expert specialisation and facilitates fine-grained cross-modal alignment. Second, filtering out low-response expert pathways can reduce the influence of less relevant background information or noisy textual cues [21], leading to cleaner feature interactions. Meanwhile, restricting the number of active experts keeps the computational cost bounded while preserving adaptive modelling capacity.

The gate weights are then organised as $W = \{W_i \in \mathbb{R}^{b \times l \times 1} | i = 1, ..., k\} \in \mathbb{R}^{b \times l \times k}$. The mixture output of the selected experts, F_{moe}, is given by Equation (10):

$$F_{moe} = \sum_{i=1}^{k} W_i \odot \mathrm{SExpert}_i(X_2) \in \mathbb{R}^{b \times l \times d}, \tag{10}$$

where $\mathrm{SExpert}_i(\cdot)$ represents the i-th expert in the selected set SExpert, corresponding to one of the experts from the original expert pool, with its operator identical to Equation (5), and $\odot$ denotes element-wise multiplication with broadcasting.

Following this, the enhanced feature X_2 undergoes an additional feed-forward transformation with residual connection, as defined in Equation (11):

$$X_3 = X_2 + \mathrm{MLP}(\mathrm{LN}(X_2)) \in \mathbb{R}^{b \times l \times d}, \tag{11}$$

where LN denotes layer normalisation, and the MLP comprises two linear layers: the first expands to $4d$ with GeLU activation, whereas the second projects back to d.

The mixture output of the selected experts, F_{moe}, is then combined with X_3 to produce the fine-grained parsing feature F_p as in Equation (12)

$$F_p = X_3 + F_{moe} \in \mathbb{R}^{b \times l \times d}. \tag{12}$$

This sequence of operations from Equations (4–12) is conducted independently for both the image and text branches, yielding F_p^v and F_p^t, respectively. To integrate local and global information, the final retrieval features aggregate these parsed features with global representations as shown below:

$$F_{vr} = G_v + \mathrm{mean}\Big(F_p^v, \mathrm{dim} = 1\Big) \in \mathbb{R}^{b \times d}, \tag{13}$$

$$F_{tr} = G_t + \mathrm{mean}\Big(F_p^t, \mathrm{dim} = 1\Big) \in \mathbb{R}^{b \times d}, \tag{14}$$

where $G_v = \{g_v^i \in \mathbb{R}^d | i = 1, ..., b\} \in \mathbb{R}^{b \times d}$, in which g_v^i denotes the global visual representation of the i-th image sample, obtained from the [CLS] token; $G_t = \{g_t^i \in \mathbb{R}^d | i = 1, ..., b\} \in \mathbb{R}^{b \times d}$, in which g_t^i denotes the global text representation of the i-th text sample, obtained from the [EOS] token; dim = 1 denote that the mean operation is applied along the token dimension (i.e., the dimension regarding l).

As depicted in Figure 2a, multiple EFP modules are stacked following the CRFM stage, progressively refining semantic representations and enhancing cross-modal alignment. The initial EFP module receives the coarse unified feature F_{cur} as input X_1, whereas each subsequent EFP module takes the F_p from the previous stage as its X_1. This mixture-of-experts architecture enables the model to capture subtle visual–textual correspondences Equations (13) and (14) vital for accurate person retrieval while preventing experts from overfitting to irrelevant background information.

Theoretical Analysis: We adopt a principled "bottleneck first and routing second" paradigm where CRFM deliberately precedes EFP to establish stable, noise-reduced representations that facilitate reliable expert selection. To formalise this design choice, we analyse how CRFM's semantic bottleneck affects downstream expert routing in EFP.

- Noise reduction through CRFM: By modelling feature noise as additive sub-Gaussian with variance ν^2, CRFM introduces a semantic bottleneck, projecting high-dimensional sparse features into a compact unified space via relevance-weighted aggregation Equation (3) and projection Equation (2). This mechanism aggregates raw features around a small set of learnable semantic anchors U ($l \ll n_1$ for visual and $l \ll n_2$ for text), efficiently compressing information while emphasising task-relevant signals. Attention weights σ Equation (1) selectively aggregate important patches and suppress noise (e.g., background clutter).

Uncorrelated noise case (for theoretical tractability). Assuming uncorrelated noise across patches, aggregation reduces noise variance by a factor $\rho < 1$, formalised as follows:

$$\mathrm{Var}\left(\sum_{i=1}^{n} \sigma_i f_i\right) = \nu^2 \sum_{i=1}^{n} \sigma_i^2 = \rho \nu^2 \approx \frac{l}{n} \nu^2, \tag{15}$$

where $\sum_i \sigma_i = 1$ and $\rho = \sum_i \sigma_i^2 \approx l/n$ measures attention concentration. Specifically, $f_i \in \mathbb{R}^d$ denotes the feature noise variable associated with the i-th input patch, which we model as an additive sub-Gaussian random variable with zero mean and variance ν^2. This simplification is valid in the context of our model, as feature maps partially suppress background regions, leading to effective aggregation.

Correlated noise case (practical extension). In practice, background regions (e.g., walls and roads) occupy spatially contiguous patches, inducing local noise correlations. We extend the analysis by considering the pairwise covariance $\mathrm{Cov}(f_i, f_j)$ between different patches. The aggregated noise variance becomes:

$$\mathrm{Var}\left(\sum_{i=1}^{n}\sigma_i f_i\right) = \boldsymbol{\sigma}^{\top}\Sigma\boldsymbol{\sigma} = \nu^2\left(\rho + \sum_{i\neq j}\sigma_i\sigma_j\gamma_{ij}\right), \tag{16}$$

where $\gamma_{ij} = \Sigma_{ij}/\nu^2$ denotes the normalised correlation between patches i and j.

Crucially, CRFM's attention mechanism suppresses correlated background noise through the following two mechanisms: (i) learnable unified tokens U capture pedestrian-relevant semantics by assigning low weights to background patches, and (ii) the contrastive learning objective penalises attention to non-informative regions see Equation (23). Consequently, if patches i and j belong to correlated background areas, the learnt attention weights satisfy $\sigma_i \approx \sigma_j \approx 0$, which diminishes the cross-term $\sum_{i\neq j}\sigma_i\sigma_j\gamma_{ij}$. This preserves the noise reduction effect even under correlated noise conditions.

- Improved routing reliability: Reducing the noise variance to $\rho\nu^2$ helps stabilise top-k gating in EFP, which is based on the logits Z_g see Equation (8). Consider a batch of b sequences, each of length l, with gating logits $Z_g \in \mathbb{R}^{b\times l\times n}$ over n experts. For each token $t \in \{1, 2, ..., b \times l\}$, let $Z_g^{(t)} \in \mathbb{R}^n$ be the logits for that token across all experts. In top-k gating, the k experts with the largest values in $Z_g^{(t)}$ are selected for each token t. Define $Z_{(k)}^{(t)}$ as the k-th largest value of $Z_g^{(t)}$, and $Z_{(k+1)}^{(t)}$ as the $(k+1)$-th largest value of $Z_g^{(t)}$, where $k \in \{1, 2, ..., n-1\}$. The selection margin for the batch is then given by the following equation:

$$\Delta_k := \min_t\left(Z_{(k)}^{(t)} - Z_{(k+1)}^{(t)}\right), \tag{17}$$

where the minimum is taken over all tokens t in the batch. Assuming the gating logits are affected by additive sub-Gaussian noise, standard concentration inequalities for sub-Gaussian random variables (see, e.g. [50]) yield the following bound on the probability that the noisy Top-k assignment differs from the noiseless one:

$$P_{\text{mis-routing}} \leq C_k \exp\left(-\frac{\Delta_k^2}{2\rho\nu^2}\right), \tag{18}$$

where $P_{\text{mis-routing}}$ denotes the probability of mis-routing. Here, C_k arises from a union bound over mis-routing events; practical choices are $C_k \leq n-k$ (over non-selected experts) or $C_k \leq k(n-k)$ (pairwise interactions), for n experts [51]. Because $\rho < 1$, the bound improves exponentially over the no-bottleneck case ($\rho = 1$); for example, when $\rho = 0.5$, it tightens by a factor $\exp\left(\Delta_k^2/(4\nu^2)\right)$.

3.4 | Loss Functions

To guarantee the efficacy of our EFP module, and following previous works [46, 52], we incorporate a load-balancing loss to encourage balanced utilisation among the selected experts. Specifically, we first apply the softmax function to the gate logits Z_g to derive the routing probabilities P, as described in Equation (19):

$$P = \mathrm{Softmax}(Z_g, \mathrm{dim} = -1) \in \mathbb{R}^{b\times l\times n}, \tag{19}$$

where $\mathrm{dim} = -1$ indicates that the softmax function is applied along the expert dimension n. We then compute the average routing probabilities R for experts as defined in Equation (20):

$$R = \mathrm{mean}(P, \mathrm{dim} = [0, 1]) \in \mathbb{R}^n, \tag{20}$$

where $\mathrm{dim} = [0, 1]$ indicates that the mean operation is taken over the batch and sequence length dimensions, denoted by b and l, respectively.

We maintain a multi-hot tensor $O = \{O_{i,j,r} | i = 1, 2, ..., b, j = 1, 2, ..., l, r = 1, 2, ..., n\} \in \mathbb{R}^{b\times l\times n}$ that records the expert selection status, where $O_{i,j,r} = 1$ if the r-th expert is among the top-k experts selected at position (i, j), and $O_{i,j,r} = 0$ otherwise. The average selection frequency S for experts is then computed as in Equation (21):

$$S = \mathrm{mean}(O, \mathrm{dim} = [0, 1]) \in \mathbb{R}^n, \tag{21}$$

The auxiliary load-balancing loss is thus defined as in Equation (22):

$$\mathcal{L}_{\text{aux}} = R^{\top} \cdot S. \tag{22}$$

This formulation encourages balanced utilisation across all experts, which mitigates expert collapse and ensures that all experts contribute effectively to the model's overall performance.

To reinforce global alignment between image and text modalities, we adopt the image-text contrastive (ITC) loss [6], which encourages high similarity for matched pairs and low similarity for mismatched pairs. Specifically, for a batch of b image-text pairs, we define the set $\left\{\left(F_{vr}^i, F_{tr}^j\right), y_{i,j}\right\}_{j=1}^{b}$, where $y_{i,j} = 1$ if (i, j) form a matched pair (i.e., correspond to the same identity), and $y_{i,j} = 0$ otherwise. The ITC loss is formulated in Equation (23):

$$\mathcal{L}_{\text{itc}} = -\frac{1}{b}\sum_{i=1}^{b}\left[\log\frac{\exp\left(\mathrm{sim}\left(F_{vr}^i, F_{tr}^i\right)/\tau\right)}{\sum_{j=1}^{b}\exp\left(\mathrm{sim}\left(F_{vr}^i, F_{tr}^j\right)/\tau\right)} + \log\frac{\exp\left(\mathrm{sim}\left(F_{vr}^i, F_{tr}^i\right)/\tau\right)}{\sum_{j=1}^{b}\exp\left(\mathrm{sim}\left(F_{vr}^j, F_{tr}^i\right)/\tau\right)}\right], \tag{23}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter.

Additionally, we also employ the similarity distribution matching (SDM) loss [11] during training. The predicted matching probability for each pair is obtained via the softmax function, as in Equation (24):

$$p_{i,j} = \frac{\exp\left(\text{sim}\left(F_{vr}^{i}, F_{tr}^{j}\right)/\tau\right)}{\sum_{m=1}^{b} \exp\left(\text{sim}\left(F_{vr}^{i}, F_{tr}^{m}\right)/\tau\right)}, \tag{24}$$

where $p_{i,j}$ represents the predicted probability (based on the cosine similarity between F_{vr}^{i} and F_{tr}^{j}), and τ determines the sharpness of the probability distribution. The ground-truth probability is calculated by normalising the binary label: $q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{b} y_{i,k}}$. The SDM loss from image to text is formulated as the KL divergence between the predicted and ground-truth distributions, as shown in Equation (25):

$$\mathcal{L}_{\text{i2t}} = \frac{1}{b} \sum_{i=1}^{b} \sum_{j=1}^{b} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \epsilon}\right), \tag{25}$$

where ϵ is a small constant added to prevent numerical instability. For symmetric cross-modal supervision, the overall SDM loss is defined in Equation (26):

$$\mathcal{L}_{\text{sdm}} = \mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}, \tag{26}$$

where $\mathcal{L}_{\text{t2i}}$ is analogously defined to measure the distributional divergence in the reverse direction.

Finally, the entire framework is optimised in an end-to-end manner, with the total objective combining all the aforementioned components as shown in Equation (27):

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{sdm}} + \beta \cdot \left(\mathcal{L}_{\text{aux}}^{I} + \mathcal{L}_{\text{aux}}^{T}\right), \tag{27}$$

where $\mathcal{L}_{\text{aux}}^{I}$ and $\mathcal{L}_{\text{aux}}^{T}$ denote the auxiliary load-balancing losses for the image and text branches, respectively, and β controls the relative contribution of the auxiliary loss terms.

4 | Experiments and Analyses

4.1 | Datasets

We select three widely used datasets, such as CUHK-PEDES [22], ICFG-PEDES [23] and RSTPReid [24] due to their widespread adoption, large scale, and complementary characteristics: CUHK-PEDES is the most established benchmark; ICFG-PEDES offers richer, fine-grained annotations and RSTPReid reflects more realistic and diverse surveillance scenarios, making them collectively representative and comprehensive for evaluating TBPR models.

CUHK-PEDES [22] is the most widely used benchmark dataset for TBPR. It provides 40,260 images and 80,440 textual descriptions for 13,003 identities, with each image annotated by two natural language descriptions. The dataset is split into training, validation and test subsets with non-overlapping identities. Specifically, the training set includes 11,003 identities with 34,054 images and 68,108 descriptions, while the validation and test sets each contain 1000 identities with approximately 3000 images and 6100 descriptions.

ICFG-PEDES [23] is a large-scale and more fine-grained TBPR dataset constructed from MSMT17 [53], containing 54,522 images of 4102 pedestrians. Each image is paired with a single textual description averaging 37 words. Compared to CUHK-PEDES, it offers more identity-centric and detailed annotations. The dataset includes 34,674 text-image pairs for 3102 identities in the training set and 19,848 pairs for 1000 identities in the test set.

RSTPReid [24] is a recent dataset for real-world TBPR applications, also based on MSMT17 [53]. It consists of 20,505 images of 4101 identities captured across 15 cameras. Each identity is represented by five images from distinct cameras, each annotated with two textual descriptions. It is split into 3701 training, 200 validation and 200 test identities.

4.2 | Setup

Our PHP method is trained by using PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory. We employ CLIP [5] as the backbone visual encoder and text encoder, coupled with CRFM and EFP modules instantiating our hierarchical parsing framework. The image is resized to 384 $\times$ 128, and the length of textual token sequence is 77. Our PHP framework is trained using Adam optimiser for 60 epochs, with a batch size of 64, $\beta = 0.5$, 4 EFP modules for CUHK-PEDES and ICFG-PEDES, and 6 EFP modules for RSTPReid and an initial learning rate 3×10^{-6}. The number of experts is set to 4, the default number l of unified tokens is set to 9 and top-k is set to 2. Data augmentation is applied to the training images and includes random horizontal flipping, padding, random cropping and random erasing. The temperature parameter τ of SDM in Equation (24) is set to 0.02.

4.3 | Comparison to State-of-the-Art

We compare our method with state-of-the-art approaches recently published at premier conferences and journals, categorising them into three kinds: non-CLIP-based, CLIP-based and generative pre-training based methods. The comparative results are summarised in Table 1.

For non-CLIP-based approaches, RASA [12] (IJCAI23) introduces relation- and sensitivity-aware learning to improve person search under noisy conditions. However, our PHP method outperforms RASA by 6.29% in R1 and 6.02% in mAP on CUHK-PEDES. On ICFG-PEDES, PHP yields more substantial gains of 32.04% in R1 and 36.83% in mAP, and on RSTPReid, it achieves increases of 9.80% in R1 and 11.16% in mAP. For CLIP-based methods, PHP consistently delivers superior results over both parameter-efficient models such as DM-Adapter [46] and UP-Person [9], and masked language modelling techniques including IRRA [11], UM2SA [10] and RDE [8]. Compared to RDE [8] (CVPR24), PHP achieves higher R1 and mAP by 6.86% and 7.84% on CUHK-PEDES. On ICFG-PEDES and RSTPReid, PHP further delivers improvements of 29.64%/38.06% and 11.35%/12.59% in R1/mAP, respectively.

Beyond the above methods, another prominent line of work employs generative data pre-training to further boost retrieval performance. PHP also demonstrates marked superiority over such approaches, including HAM-PEDES [13], LuPerson-T [54] and SYNTH-PEDES [15], which leverage multimodal LLMs for fine-grained feature alignment. Compared to the latest HAM-PEDES [13] (CVPR25), PHP achieves increments of 4.81% in

TABLE 1 | Performance (%) comparison with state-of-the-art methods on CUHK-PEDES [22], ICFG-PEDES [23] and RSTPReid [24].

Method	Reference	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
non-CLIP-based													
TIPCB [25]	Neurocomputing22	64.26	83.19	89.10	N/A	54.96	74.72	81.89	N/A	46.70	70.00	78.80	N/A
AXM-Net [32]	AAAI22	64.44	80.52	86.77	58.73	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TGDA [29]	TCSVT23	64.64	83.38	89.34	58.64	57.26	75.19	81.80	32.27	48.35	73.15	80.30	37.96
IVT [30]	ECCV22	65.59	83.11	89.21	N/A	56.04	73.60	80.22	N/A	46.70	70.00	78.80	N/A
CTLP [28]	TCSVT23	69.47	87.13	92.13	60.56	57.69	75.79	82.67	36.07	N/A	N/A	N/A	N/A
RASA [12]	IJCAI23	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	66.90	86.50	91.35	52.31
CLIP-based													
UM2SA [10]	AAAI24	74.25	89.83	93.58	66.15	65.62	80.54	85.83	38.78	63.40	83.30	90.30	49.28
CFine [7]	TIP23	69.57	85.93	91.15	N/A	60.83	76.55	82.42	N/A	50.55	72.50	81.60	N/A
VGSG [54]	TIP24	71.38	86.75	91.86	67.91	63.05	78.43	84.36	N/A	—	—	—	—
RDE [8]	CVPR24	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
IRRA [11]	CVPR23	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
DM-Adapter [46]	AAAI25	72.17	88.74	92.85	64.33	62.64	79.53	85.32	36.50	60.00	82.10	87.90	47.37
UP-Person [9]	arXiv25	74.17	89.70	93.88	65.91	65.02	80.98	86.16	38.32	63.15	83.45	89.75	48.15
Generative pre-training based													
MALS + APTM [55]	MM23	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
MALS + IRRA [55]	MM23	74.05	89.48	93.64	66.57	64.37	80.75	86.12	38.85	61.90	80.60	89.30	48.08
LuPerson-T + IRRA [14]	ICCV23	74.37	89.51	93.97	66.60	64.50	80.24	85.74	38.22	62.20	83.30	89.75	48.33
SYNTH-PEDES + IRRA [15]	NIPS24	74.25	89.49	93.68	66.52	65.79	81.94	87.32	39.43	64.35	83.75	91.00	50.93
LUPeron-MLLMs + IRRA [56]	CVPR24	76.82	91.16	94.46	69.55	67.05	82.16	87.33	41.51	68.50	87.15	92.10	53.02
HAM-PEDES + IRRA [13]	CVPR25	77.99	91.34	95.03	69.72	68.25	83.30	88.15	42.30	71.69	87.85	93.30	55.47
PHP (Ours)	**Ours**	**82.80**	**93.55**	**96.20**	**75.40**	**97.32**	**99.38**	**99.74**	**78.12**	**76.70**	**92.85**	**96.25**	**63.47**

Note: The bold values indicate the best results on the corresponding metrics among three datasets.

R1 and 5.68% in mAP on CUHK-PEDES. Significant gains are also observed on ICFG-PEDES (29.07% R1, 35.82% mAP) and RSTPReid (5.01% R1, 8.00% mAP).

Whereas retrieval accuracy is essential, model efficiency and scalability are crucial for real-world applications. As shown in Figure 3, methods like RASA [12], built on the ALBEF [6] architecture, contain 420M parameters—comprising 210M fine-tuned parameters and an additional 210M from its momentum model updated via EMA. For a fair comparison, we focus on the 210M fine-tuned parameters of RASA. Despite having only 165M parameters, our proposed PHP outperforms RASA, demonstrating a better trade-off between efficiency and effectiveness. Furthermore, generative pre-training methods such as APTM [55] and HAM-PEDES [13] require similar parameter scales but come with higher storage costs due to the extensive data requirements (around 1M samples). On the other hand, parameter-efficient models such as DM-Adapter [46] (16M) and UP-Person [9] (7.4 M) reduce the fine-tuned parameters but suffer slight performance degradation. In contrast, our PHP method achieves state-of-the-art retrieval accuracy (82.80% R1 on CUHK-PEDES) while maintaining moderate complexity.

4.4 | Analysis

To comprehensively investigate the factors underlying the effectiveness of our proposed PHP framework, we conduct a series of systematic analyses and experiments. We begin by evaluating the individual contributions of each core module in PHP, revealing how their integration leads to improved retrieval performance. Subsequently, we perform sensitivity analysis on key hyperparameters to assess their influence on overall outcomes. In addition, we analyse the benefits of the parameter-sharing architecture. We further examine the effects of different loss function configurations on feature learning and retrieval accuracy. Then, we analyse expert utilisation in the EFP module and evaluate the computational efficiency of PHP model. Finally, we present qualitative visualisations to provide an intuitive analysis of our approach in real-world retrieval scenarios.

4.4.1 | *Effectiveness of Core Modules*

As illustrated in Figure 4, we conduct comprehensive ablation studies on the CUHK-PEDES [22], ICFG-PEDES [23] and RSTPReid [24] datasets to systematically analyse the individual and joint contributions of our CRFM and EFP modules. The Baseline refers to the setting where both CRFM and EFP modules are removed. The CRFM case corresponds to enabling only the CRFM module while disabling the EFP module, allowing us to evaluate the effect of coarse-grained feature refinement in isolation. The EFP case refers to enabling only the EFP module with the CRFM module removed, thus focussing solely on the impact of expert fine-grained processing. The CRFM + EFP case represents our complete model where both modules are jointly incorporated, enabling us to examine their combined effect and synergy.

As shown in Figure 4a, on the CUHK-PEDES dataset, introducing CRFM alone yields moderate improvements compared to the Baseline, with R1 increasing from 66.44% to 69.77% (+3.33%) and mAP rising from 59.17% to 62.55% (+3.38%). This demonstrates the effectiveness of CRFM in capturing coarse-grained pedestrian-relevant semantics. Adding only EFP leads to even more substantial gains, with R1 reaching 75.75% (+9.31%) and mAP achieving 68.00% (+8.83%), indicating that fine-grained expert processing can effectively extract discriminative details. Remarkably, when both CRFM and EFP are integrated, their synergy brings the best performance, with R1 reaching 82.80% (+16.36%), R5 achieving 93.55% (+8.20%), and mAP attaining 75.40% (+16.23%), significantly outperforming both individual modules and demonstrating clear complementary benefits.

This consistent trend of complementary improvement is observed across all three datasets with varying characteristics. On the ICFG-PEDES dataset, as shown in Figure 4b, CRFM alone improves R1 from 53.94% to 60.19% (+6.25%) and mAP from 30.52% to 34.74% (+4.22%). The EFP module demonstrates particularly strong individual performance, boosting R1 to 92.09% (+38.15%) and mAP to 63.34% (+32.82%), suggesting that expert-based fine-grained processing is especially effective on this dataset. The combination of both modules further enhances performance, achieving R1 of 97.32% (+43.38%), R5 of 99.38% (+25.50%) and mAP of 78.12% (+47.60%), indicating that CRFM provides essential complementary coarse-grained semantic refinement even when EFP is already highly effective. Similarly, on RSTPReid shown in Figure 4c, CRFM provides modest gains over the baseline (R1: 57.35% vs. 55.95%, +1.40%; mAP: 44.74% vs. 44.50%, +0.24%), whereas the EFP alone delivers more substantial improvements (R1: 71.30%, +15.35%; mAP: 56.78%, +12.28%). When both modules are integrated, the performance improves markedly to R1 of 76.70% (+20.75%), R5 of 92.85% (+12.55%), and mAP of 63.47% (+18.97%).

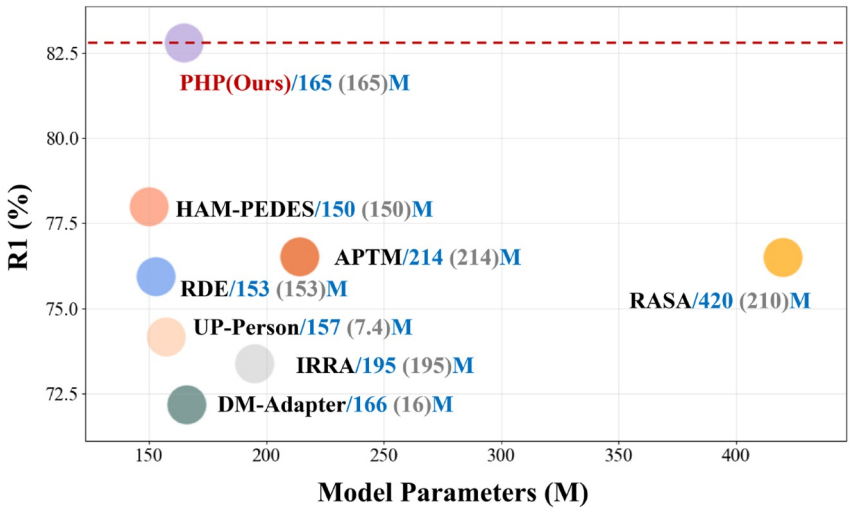


FIGURE 3 | R1 and model parameter for our PHP method and several state-of-the-art approaches on the CUHK-PEDES [22] dataset, with model parameter highlighted in blue. The (·) in grey denotes fine-tuned parameters.

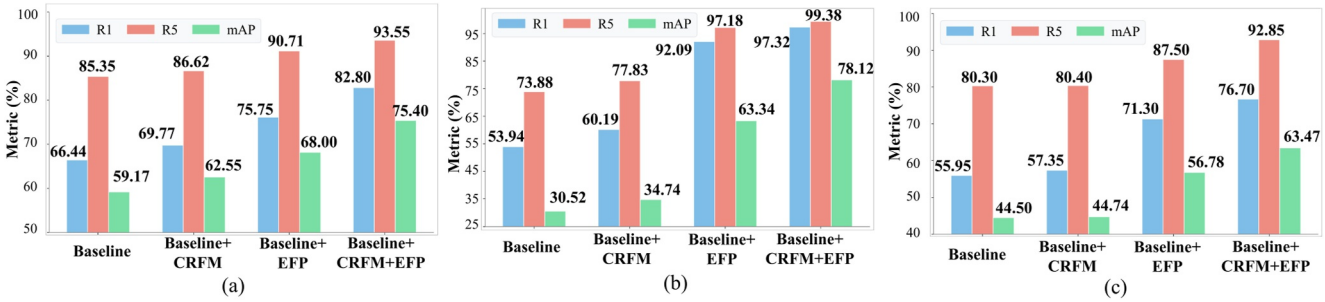


FIGURE 4 | Ablation results for the core modules of our PHP method on CUHK-PEDES [22], ICFG-PEDES [23], and RSTPReid [24].

Furthermore, the bound in Equation (18) suggests that reducing noise variance ($\rho < 1$) through the CRFM bottleneck leads to a decrease in the probability of expert mis-routing. Because routing accuracy is closely related to expert specialisation and feature alignment, reducing mis-routing is expected to improve matching performance. This theoretical insight is consistent with the empirical results shown in Figure 4. On the CUHK-PEDES dataset, applying EFP to raw features (baseline + EFP) achieves 75.75% R1, while introducing the CRFM bottleneck (baseline + CRFM + EFP) provides a denoised feature space that improves routing reliability and boosts the performance to 82.80% R1 (+7.05%). Similar trends are observed on ICFG-PEDES and RSTPReid, where the combination of CRFM and EFP consistently outperforms individual components. These results provide empirical support for the mis-routing bound in Equation (18), suggesting that reducing noise-induced mis-routing contributes to improved performance in practice.

In summary, these results underscore a key principle of our framework: CRFM and EFP address different aspects of the TBPR task and yield complementary benefits. CRFM effectively captures coarse-grained pedestrian-relevant semantics by suppressing background noise and aligning cross-modal representations, while EFP excels at extracting fine-grained discriminative details through expert specialisation. Notably, whereas each module brings standalone improvements over the baseline, their combination consistently achieves the best results across all three datasets. This improvement pattern confirms that the two modules target distinct but synergistic semantic levels of the feature space—coarse-grained semantic alignment followed by fine-grained attribute discrimination—thereby validating the rationale behind our coarse-to-fine architecture design.

4.4.2 | *Hyperparameter Sensitivity*

As illustrated in Figure 5, we first fix the number of EFP modules at 4 and systematically vary the number of experts from 2 to 10. On the smaller RSTPReid [24] dataset, performance follows an inverted-U curve: Increasing the number of experts from 2 (R1: 73.45%, mAP: 62.33%) to 4 (R1: 74.66%, mAP: 63.47%) yields improvement, but further increases result in a decline (10 experts: R1: 71.65%, mAP: 60.87%). This indicates that excessive capacity can lead to overfitting, especially when data diversity is limited. In contrast, larger-scale datasets exhibit strong robustness. For CUHK-PEDES [22], performance remains stable across all settings, with R1 ranging from 82.16% to 82.80% and mAP consistently around 75%. Similarly, ICFG-PEDES [23] shows only negligible fluctuations (R1: 97.08%–97.32%, mAP: 77.93%–78.47%). This robustness can be attributed to two key aspects of our hierarchical design: (1) the CRFM module serves as an effective noise filter, providing high-quality semantic features regardless of the number of experts; and (2) parameter-sharing and modality-aware gating in EFP enable experts to dynamically focus on different semantic groups, promoting specialisation while mitigating overfitting. Next, with the number of experts fixed at 4, we vary the number of EFP modules from 2 to 8. The results are shown in Table 2. We observe that CUHK-PEDES and ICFG-PEDES again display stable near-optimal performance at 4 modules (82.80% and 97.32% R1 respectively) with only minor changes across other settings, highlighting the broad parameter insensitivity when data diversity is sufficient. However, for the small-scale RSTPReid, performance improves up to 6 modules (76.70% R1) before dropping, indicating a narrower capacity sweet spot and confirming that limited sample diversity necessitates more careful tuning to avoid overfitting.

Additionally, in our framework, l denotes the number of learnable unified tokens $U \in \mathbb{R}^{l \times d}$, which determines the representational capacity of the unified semantic space. To balance expressive power and model complexity, we evaluate a range of l values to empirically identify an optimal setting. As shown in Table 3, increasing l from 3 to 9 consistently improves performance across all datasets. For example, on CUHK-PEDES, R1 increases from 80.82% ($l = 3$) to 82.80% ($l = 9$), and mAP from 73.82% to 75.40%. Similar trends are observed for ICFG-PEDES and RSTPReid, with all metrics peaking or plateauing at $l = 9$. However, further increasing l to 12 does not yield additional gains and, in some cases, results in slight performance degradation, suggesting possible over-parameterisation or redundancy. Based on these findings, we recommend $l = 9$ as a default value, which

TABLE 2 | Effect of varying the number of EFP modules on R1 (%).

Dataset	2	4	6	8
CUHK-PEDES	81.56	**82.80**	82.03	81.60
ICFG-PEDES	97.18	**97.32**	97.10	97.03
RSTPReid	74.30	74.66	**76.70**	74.40

Note: The bold values indicate the best results among three datasets.

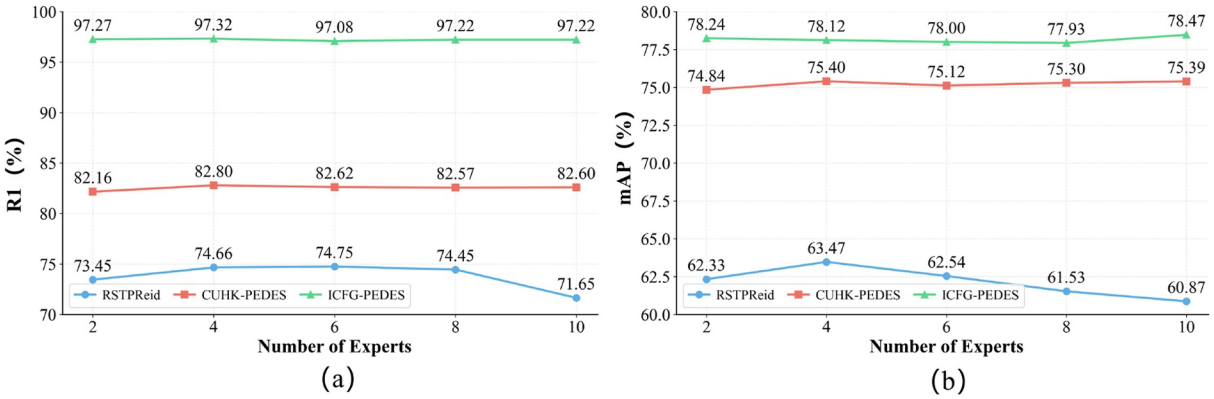


FIGURE 5 | Sensitivity analysis of expert numbers in the EFP module: (a) R1, and (b) mAP performance on three datasets.

empirically achieves the best balance between model capacity and complexity, while avoiding over-parameterisation.

As shown in Table 4, we provide an ablation study on the load-balancing weight β. When $\beta = 0.0$, performance consistently drops across all datasets (e.g., mAP decreases from 75.40% to 74.31% on CUHK-PEDES), suggesting the importance of $\mathcal{L}_{aux}$ for stabilising expert routing. We further observe that the top-performing results are concentrated within a moderate range of $\beta \in [0.5, 1.5]$, which reflects a trade-off between load-balancing regularisation and representation learning. Larger values of β (e.g., > 1.5) lead to performance degradation, suggesting that excessive regularisation may hinder the learning of informative cross-modal features. Within this range, $\beta = 0.5$ mostly achieves competitive performance across all datasets. Therefore, we use $\beta = 0.5$ as the default setting in our experiments.

4.4.3 | *Parameter-Sharing Architecture Advantage*

As shown in Table 5, we observe substantial performance improvements when PHP operates with parameter sharing (S) between text and visual branches, while the non-sharing (NS) mode yields modest or even degraded performance. Taking ICFG-PEDES as an example, the parameter sharing mode achieves 97.32% R1 and 78.12% mAP, representing improvements of 75.56% and 63.96% over the non-sharing mode (21.76% and 14.16%). This finding aligns with our original motivation and intuition. In the non-sharing setting, cross-modal alignment between textual and visual features relies on two independent branches. Although this design provides each branch with greater learning flexibility, it also makes the model more susceptible to cross-modal noise—for example, images often contain large regions of irrelevant background, whereas descriptions may lack specificity. Such noise introduces undesirable signals and hinders accurate fine-grained cross-modal alignment, ultimately degrading the alignment of the enhanced features and reducing overall performance. In contrast, in the sharing setting, enforcing parameter sharing between the two branches appropriately constrains this flexibility, encouraging the model to capture semantically consistent patterns across modalities. This leads to more stable and effective alignment and improves overall performance. This observation is also consistent with the classic Occam's razor principle in machine learning: do not increase model complexity unless necessary.

4.4.4 | *Impact of Loss Function Design*

Table 6 presents a comprehensive ablation study evaluating various loss function configurations within the PHP framework on three datasets. First, regarding the impact of auxiliary loading balance loss $\mathcal{L}_{\text{aux}}$, we note that including $\mathcal{L}_{\text{aux}}$ (Row 4, with $\mathcal{L}_{\text{aux}}$) consistently improves performance compared to the setting without it (Row 3, w/o $\mathcal{L}_{\text{aux}}$). For instance, on RSTPReid, the inclusion of $\mathcal{L}_{\text{aux}}$ increases R1 from 72.45% to 76.70%, R5 from 90.40% to 92.85%, and mAP from 61.17% to 63.47%. Similar improvements are observed across all other datasets. These results demonstrate that the auxiliary loss provides effective regularisation for expert routing, leading to improved robustness and generalisation.

Second, building on the auxiliary load-balancing loss $\mathcal{L}_{\text{aux}}$, we observe that, when comparing the second row to the first, replacing the image–text contrastive loss ($\mathcal{L}_{\text{itc}}$) with the similarity distribution loss ($\mathcal{L}_{\text{sdm}}$) yields consistent improvements across all evaluation metrics on three datasets. Specifically, on CUHK-PEDES, R1 increases from 81.53% to 82.18%; (+0.65%); on RSTPReid, from 71.85% to 75.75%; (+3.90%); and on ICFG-PEDES, from 96.77% to 97.21%; (+0.44%). This demonstrates that ($\mathcal{L}_{\text{sdm}}$) is more effective than ($\mathcal{L}_{\text{itc}}$) at enhancing fine-grained feature discrimination. In addition, the full combination of all three losses (Row 4) achieves the highest performance

TABLE 3 | Impact of varying the number l of unified tokens U in CRFM on R1 (%) and mAP (%).

	$l = 3$		$l = 6$		$l = 9$		$l = 12$	
Datasets	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
CUHK-PEDES	80.82	73.82	82.31	74.98	82.80	75.40	81.88	74.71
ICFG-PEDES	97.19	77.50	97.06	77.50	97.32	78.12	97.33	78.43
RSTPReid	74.15	61.10	74.65	61.18	76.70	63.47	75.15	61.47

TABLE 4 | Results under different values of the relative contribution factor β.

	CUHK-PEDES [22]			ICFG-PEDES [23]			RSTPReid [24]		
β	**R1**	**R5**	**mAP**	**R1**	**R5**	**mAP**	**R1**	**R5**	**mAP**
0.0	81.53	93.10	74.31	96.77	99.30	76.46	71.85	90.30	60.92
0.5	**82.80**	93.55	**75.40**	97.32	**99.38**	**78.12**	**76.70**	**92.85**	**63.47**
1.0	82.29	93.52	74.89	97.33	99.33	78.09	75.25	91.55	62.89
1.5	82.23	**93.56**	75.18	**97.36**	99.34	78.10	74.70	91.50	63.16
2.0	81.95	93.50	74.83	97.31	99.30	77.99	76.50	91.95	62.95
2.5	82.03	93.27	74.60	97.29	99.32	77.95	74.50	91.65	62.45
3.0	82.06	93.30	74.88	97.27	99.37	77.95	75.20	92.05	62.85

Note: The bold values indicate the best results on the corresponding metrics among three datasets.

across all datasets: CUHK-PEDES (82.80% R1, 93.55% R5, 75.40% mAP), RSTPReid (76.70% R1, 92.85% R5, 63.47% mAP) and ICFG-PEDES (97.32% R1, 99.38% R5, 78.12% mAP).

Overall, these results indicate that the auxiliary load-balancing loss, the contrastive loss and similarity distribution loss offer complementary benefits. The presence of $\mathcal{L}_{\text{aux}}$ facilitates stable expert routing, enabling $\mathcal{L}_{\text{itc}}$ and $\mathcal{L}_{\text{sdm}}$ to fully leverage their cross-modal alignment capabilities. This coordinated optimisation ultimately leads to more discriminative pedestrian features and enhanced retrieval performance.

4.4.5 | *Analysis of Expert Utilisation in EFP*

We conduct a quantitative evaluation of expert utilisation using Shannon entropy, with larger entropy values indicating a more uniform distribution across experts [57]. The auxiliary load-balancing loss $\mathcal{L}_{aux}$ regulates routing probabilities to encourage balanced expert utilisation, thereby playing a critical role in shaping expert assignment distributions within EFP modules. As shown in Figure 6a, when the auxiliary loss is applied (i.e., $\beta = 0.5$), the entropy is consistently higher than that of the variant trained without $\mathcal{L}_{aux}$ (i.e., $\beta = 0$), indicating a more uniform expert assignment. A similar trend is observed in the text branch of the EFP modules, as illustrated in Figure 6b. Furthermore, Table 4 shows that incorporating $\mathcal{L}_{aux}$ leads to improved performance compared to the case without it. These results suggest that a well-balanced expert assignment within the EFP module more effectively exploits the representational capacity of the mixture-of-experts architecture, ultimately yielding performance gains.

4.4.6 | *Inference Time Comparison*

As shown in Table 7, CLIP-based methods such as IRRA [11], RDE [8] and our PHP achieve much higher computational efficiency compared to the ALBEF-based RASA [12]. With a batch size of 128, the CLIP-based methods require only 1.24–1.37 ms per image and 11.04 to 11.14 G FLOPs, whereas RASA needs 7.33 ms and 49.39 G FLOPs per image. This means that RASA is approximately five times slower and more than four times more computationally expensive, clearly highlighting the efficiency advantage of CLIP-like architectures including PHP.

Among these three CLIP-based approaches, PHP maintains inference speed (1.37 ms) and FLOPs (11.14 G) that are very close to IRRA (1.24 ms, 11.04 G) and RDE (1.27 ms, 11.11 G) while delivering notable improvements in retrieval accuracy. According to Table 1, PHP improves R1 accuracy on ICFG-PEDES [23] by 33.86% (from 63.46% to 97.32%) and mAP by 40.06% (from 38.06% to 78.12%) compared to IRRA. These results demonstrate that PHP successfully balances high computational efficiency with significant gains in performance, making it an appealing choice for real-world text-based person retrieval tasks.

4.4.7 | *Robustness Capability Analysis*

To comprehensively evaluate PHP's robustness under realistic retrieval scenarios, we conduct experiments examining model performance across diverse query variations that reflect practical challenges in text-based person retrieval. We compare PHP against vanilla CLIP [5] and IRRA [11] under three complementary evaluation settings.

As illustrated in Figure 7, we assess robustness to controlled query perturbations. Starting from the origin query ("a man wearing a black coat and blue jeans walking alone"), we design three variation types: (i) *Synonym replacement*, where vocabulary substitutions are applied (e.g., "dark tee" for "black shirt" and "strolling" for "walking"); (ii) *colour variation*, where approximate colour descriptions are used (e.g., "navy blue" instead of "black");and (iii) *misleading distractor*, where semantically irrelevant words are inserted. Under synonym replacement with spatial constraints ("a guy in a dark tee and denim pants strolling near the building"), PHP and IRRA maintain reasonable performance, whereas vanilla CLIP fails to incorporate spatial cues. For colour variation ("navy blue coat"), IRRA incorrectly retrieves grey coats and vanilla CLIP defaults to black matches, whereas PHP demonstrates superior colour-aware discrimination. Under the misleading distractor ("walking elephant green jacket"), both baselines are significantly distracted, whereas PHP effectively filters the irrelevant term and maintains focus on the visually groundable attribute.

TABLE 5 | Performance (%) comparison of PHP's parameter non-shared (NS) and shared (S) modes.

	ICFG-PEDES [23]		CUHK-PEDES [22]		RSTPReid [24]	
Metric	**NS**	**S**	**NS**	**S**	**NS**	**S**
R1	21.76	**97.32**	50.29	**82.80**	38.20	**76.70**
R5	39.12	**99.38**	74.58	**93.55**	63.85	**92.85**
mAP	14.16	**78.12**	47.79	**75.40**	33.16	**63.47**

Note: The bold values indicate the best results on the corresponding metrics among three datasets.

TABLE 6 | Comparison of performance (%) for different loss function designs in PHP on three datasets, where ✓ and ✗ denote the use and absence of the respective loss function, respectively.

Loss Functions			CUHK-PEDES			RSTPReid			ICFG-PEDES		
$\mathcal{L}_{\text{aux}}$	$\mathcal{L}_{\text{itc}}$	$\mathcal{L}_{\text{sdm}}$	**R1**	**R5**	**mAP**	**R1**	**R5**	**mAP**	**R1**	**R5**	**mAP**
✓	✓	✗	81.53	93.10	74.31	71.85	90.30	60.92	96.77	99.30	76.46
✓	✗	✓	82.18	93.23	74.73	75.75	92.10	62.85	97.21	99.32	78.19
✗	✓	✓	81.74	93.23	74.72	72.45	90.40	61.17	97.01	99.32	77.86
✓	✓	✓	**82.80**	**93.55**	**75.40**	**76.70**	**92.85**	**63.47**	**97.32**	**99.38**	**78.12**

Note: The bold values indicate the best results on the corresponding metrics among three datasets.

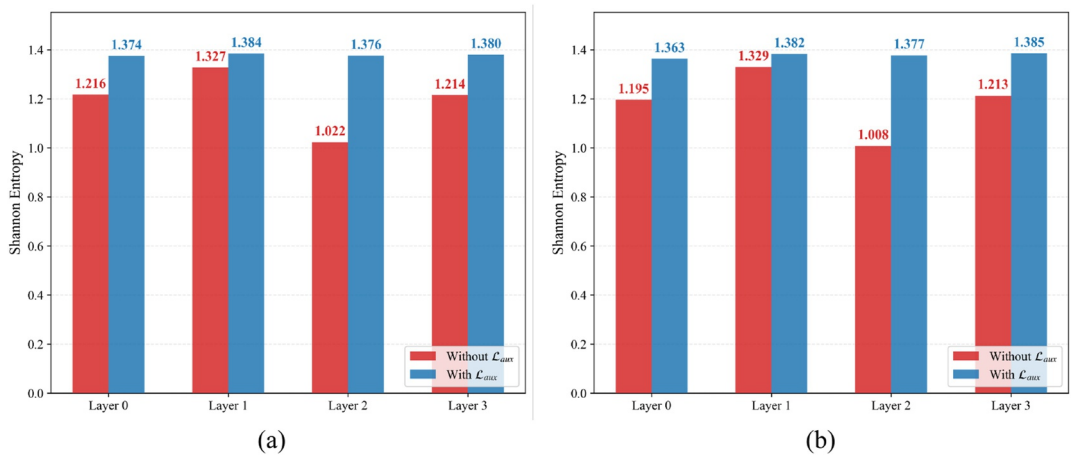


FIGURE 6 | Shannon entropy of expert utilisation in (a) image branches and (b) text branches of EFP modules on the CUHK-PEDES [22] dataset.

TABLE 7 | Inference time and floating point operations (FLOPs) at different batch sizes (BS) on ICFG-PEDES [23], measured on a single RTX 3090.

		Image encoder				Text encoder			
		Time per image (ms)			FLOPs	Time per text query (ms)			FLOPs
Methods	Arch.	BS = 64	BS = 128	BS = 256	(G)	BS = 64	BS = 128	BS = 256	(G)
RASA [12]	ALBEF	7.36	7.33	7.60	49.39	0.41	0.43	0.43	2.13
IRRA [11]	CLIP	0.26	1.24	1.30	11.04	0.36	0.26	0.25	1.94
RDE [8]		1.34	1.27	1.27	11.11	0.31	0.30	0.30	1.97
PHP (ours)		1.45	1.37	1.30	11.14	0.42	0.36	0.31	2.04



FIGURE 7 | Robustness analysis of the top-10 retrieval results. We evaluate PHP, IRRA [11], and vanilla CLIP [5] under several controlled query perturbations. Green boxes indicate correct matches; red boxes indicate mismatches.

5 | Conclusion

In this paper, we propose a pedestrian-relevant hierarchical parsing (PHP) framework, representing the first exploration of a mixture-of-experts architecture for fine-grained cross-modal alignment in text-based person retrieval. Our approach bridges the gap from coarse to fine feature matching through a hierarchical processing paradigm that directly addresses the critical challenge of cross-modal misalignment. The core innovation of PHP lies in its two-stage design: the coarse relevant feature mapping (CRFM) module projects both visual and textual information into a unified, low-dimensional concept space, thus enabling effective self-attention over coarse-grained features and establishing robust initial semantic alignment. Furthermore, the expert-driven feature parsing (EFP) module integrates a mixture-of-experts mechanism with a modality-aware gating strategy to achieve enhanced part-level alignment and deeper semantic associations between text and images. This hierarchical architecture not only maintains computational efficiency but also is highly effective at capturing nuanced cross-modal correspondences without relying on external dependencies. Extensive experiments on three public benchmarks validate both the effectiveness and efficiency of our approach, and it achieves superior performance over prior state-of-the-art methods across diverse dataset scales and retrieval scenarios.

Acknowledgements

This work was supported in part by the Key Project of Natural Science Foundation of Fujian Province under Grant No. 2026J002031, in part by

the National Key Research and Development Program of China under Grant No. 2026YFE0202200, in part by the Key Science and Technology Project of Xiamen City under Grant No. 3502Z20251011, in part by the National Natural Science Foundation of China under Grant No. 62306122, in part by the Fujian Province Science and Technology Empowering Police Research Project under Grant No. 2024Y0064, in part by the Project of Natural Science Foundation of Fujian Province under Grant No. 2026J001734, in part by the Open List and Leader Appointment Project of Fujian Province under Grant No. 2024HZ022007, in part by the Key Program of Natural Science Foundation of Fujian Province under Grant No. 2023J02022 and in part by the Quanzhou City High-Level Talent Innovation and Entrepreneurship Project under Grant No. 20241ZC010.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are publicly available. The CUHK-PEDES dataset is available at https://github.com/ShuangLI59/Person-Search-with-Natural-Language-Description. The ICFG-PEDES dataset is available at https://github.com/zifyloo/SSAN. The RSTPReid dataset is available at https://github.com/NjtechCVLab/RSTPReid-Dataset. Our code is available at https://github.com/Tedysu0916/php_V1.

References

1. Y. Wang, H. Zhang, D. Miao, and W. Pedrycz, "Multi-Granularity Re-Ranking For Visible-Infrared Person Re-Identification," *CAAI Transactions on Intelligence Technology* 8, no. 3 (2023): 770–779, https://doi.org/10.1049/cit2.12182.

2. D. Dai, X. Zhang, Z. Wu, H. Meng, and Z. Zhang, "A Keypoint-Guided Feature Partition Network for Occluded Person Re-Identification," *CAAI Transactions on Intelligence Technology* 10, no. 5 (2025): 1535–1547, https://doi.org/10.1049/cit2.70057.

3. C. Eom, G. Lee, K. Cho, H. Jung, M. Jin, and B. Ham, "Cerberus: Attribute-Based Person Re-Identification Using Semantic IDs," *Expert Systems with Applications* 259 (2025): 125320, https://doi.org/10.1016/j.eswa.2024.125320.

4. H. Zhang, M. Liu, Y. Li, et al., "Attribute-Guided Collaborative Learning for Partial Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 12 (2023): 14144–14160, https://doi.org/10.1109/tpami.2023.3312302.

5. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning* (2021): 8748–8763.

6. J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align Before Fuse: Vision and Language Representation Learning With Momentum Distillation," *Advances in Neural Information Processing Systems* 34 (2021): 9694–9705.

7. S. Yan, N. Dong, L. Zhang, and J. Tang, "Clip-Driven Fine-Grained Text-Image Person Re-Identification," *IEEE Transactions on Image Processing* 32 (2023): 6032–6046, https://doi.org/10.1109/tip.2023.3327924.

8. Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, *Noisy-Correspondence Learning for text-to-image Person re-identification* (IEEE Conference on Computer Vision and Pattern Recognition, 2024), 27197–27206.

9. Y. Liu, Y. Li, X. Lan, W. Yang, Z. Liu, and Q. Liao, "Up-Person: Unified Parameter-Efficient Transfer Learning for Text-Based Person Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology* 35, no. 12 (2025): 12874–12889, https://doi.org/10.1109/tcsvt.2025.3588406.

10. Z. Zhao, B. Liu, Y. Lu, Q. Chu, and N. Yu, "Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-To-Image Person Re-Identification," *AAAI Conference on Artificial Intelligence* 38, no. 7 (2024): 7534–7542, https://doi.org/10.1609/aaai.v38i7.28585.

11. D. Jiang and M. Ye, "Cross-Modal Implicit Relation Reasoning and Aligning for Text-To-Image Person Retrieval," *IEEE Conference on Computer Vision and Pattern Recognition* (2023): 2787–2797.

12. Y. Bai, M. Cao, D. Gao, et al., "Rasa: Relation and Sensitivity Aware Representation Learning for Text-Based Person Search," *International Joint Conference on Artificial Intelligence* (2023): 555–563.

13. J. Jiang, C. Ding, W. Tan, J. Wang, J. Tao, and X. Xu, "Modeling Thousands of Human Annotators for Generalizable Text-To-Image Person Re-Identification," *IEEE Conference on Computer Vision and Pattern Recognition* (2025): 9220–9230.

14. Z. Shao, X. Zhang, C. Ding, J. Wang, and J. Wang, "Unified Pre-Training With Pseudo Texts for Text-To-Image Person Re-Identification," *IEEE International Conference on Computer Vision* (2023): 11174–11184.

15. J. Zuo, J. Hong, F. Zhang, et al., "Plip: Language-Image Pre-Training for Person Representation Learning," *Advances in Neural Information Processing Systems* 37 (2024): 45666–45702.

16. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," *IEEE Conference on Computer Vision and Pattern Recognition* (2022): 10684–10695.

17. A. Hurst, A. Lerer, A. P. Goucher, et al., "Gpt-4o System Card," arXiv preprint arXiv:2410.21276 (2024).

18. Wang P., Bai S., Tan S., et al., "Qwen2-vl: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv preprint arXiv:2409.12191 (2024).

19. L. Yi, H. Yu, G. Wang, X. Liu, and Q. Hu, "pFedMoE: Data-Level Personalization With Mixture Of Experts in Model-Heterogeneous Personalized Federated Learning," *IEEE Transactions on Knowledge and Data Engineering* 38, no. 3 (2026): 1905–1918, https://doi.org/10.1109/tkde.2026.3656194.

20. B. Zoph, I. Bello, S. Kumar, et al., "St-moe: Designing Stable and Transferable Sparse Expert Models," arXiv preprint arXiv:2202.08906, 2022.

21. C. Riquelme, J. Puigcerver, B. Mustafa, et al., "Scaling Vision With Sparse Mixture of Experts," *Advances in Neural Information Processing Systems* 34 (2021): 8583–8595.

22. S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person Search With Natural Language Description," *IEEE Conference on Computer Vision and Pattern Recognition* (2017): 1970–1979.

23. Z. Ding, C. Ding, Z. Shao, D. Tao, "Semantically Self-Aligned Network for Text-To-Image Part-Aware Person Re-Identification," arXiv preprint arXiv:2107.12666 (2021).

24. A. Zhu, Z. Wang, Y. Li, et al., "Dssl: Deep Surroundings Person Separation Learning for Text-Based Person Retrieval," *ACM International Conference on Multimedia* (2021): 209–217.

25. Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "Tipcb: A Simple But Effective Part-Based Convolutional Baseline for Text-Based Person Search," *Neurocomputing* 494 (2022): 171–181, https://doi.org/10.1016/j.neucom.2022.04.081.

26. Y. Zhang and H. Lu, "Deep Cross-Modal Projection Learning for Image-Text Matching," *European Conference on Computer Vision* (2018): 686–701.

27. S. Li, M. Cao, and M. Zhang, "Learning Semantic-Aligned Feature Representation for Text-Based Person Search," *IEEE International Conference on Acoustics, Speech and Signal Processing* (2022): 2724–2728.

28. H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive Transformer Learning With Proximity Data Generation for Text-Based Person Search," *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 8 (2023): 7005–7016, https://doi.org/10.1109/tcsvt.2023.3329220.

29. L. Gao, K. Niu, B. Jiao, P. Wang, and Y. Zhang, "Addressing Information Inequality for Text-Based Person Search via Pedestrian-Centric Visual Denoising and Bias-Aware Alignments," *IEEE Transactions on Circuits and Systems for Video Technology* 33, no. 12 (2023): 7884–7899, https://doi.org/10.1109/tcsvt.2023.3273719.

30. X. Shu, W. Wen, H. Wu, et al., "See Finer, See More: Implicit Modality Alignment for Text-Based Person Retrieval," *European Conference on Computer Vision Workshops* (2023): 624–641.

31. Y. Wu, H. Wang, M. Wu, M. Cao, and M. Zhang, "Laip: Learning Local Alignment From Image-Phrase Modeling for Text-Based Person Search," *IEEE International Conference on Multimedia and Expo* (2024): 1–10.

32. A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "Axm-Net: Implicit Cross-Modal Feature Alignment for Person Re-Identification," *AAAI Conference on Artificial Intelligence* 36, no. 4 (2022): 4477–4485, https://doi.org/10.1609/aaai.v36i4.20370.

33. J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders and Large Language Models," *International Conference on Machine Learning* (2023): 19730–19742.

34. H. Han, K. Miao, Q. Zheng, and M. Luo, "Noisy Correspondence Learning With Meta Similarity Correction," *IEEE Conference on Computer Vision and Pattern Recognition* (2023): 7517–7526.

35. Z. Huang, G. Niu, X. Liu, et al., "Learning With Noisy Correspondence for Cross-Modal Matching," *Advances in Neural Information Processing Systems* 34 (2021): 29406–29419.

36. P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-Modal Retrieval With Partially Mismatched Pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 8 (2023): 9595–9610, https://doi.org/10.1109/tpami.2023.3247939.

37. M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning With Twin Noisy Labels for Visible-Infrared Person Re-Identification," *IEEE Conference on Computer Vision and Pattern Recognition* (2022): 14308–14317.

38. W. Suo, M. Sun, K. Niu, et al., "A Simple and Robust Correlation Filtering Method for Text-Based Person Search," *European Conference on Computer Vision* (2022): 726–742.

39. S. You, T. Miao, Q. Chen, X. Li, Z. Cheng, and D. Luo, "Kpdm: Key Phrase Dynamic Masking for Robust Text-To-Image Person Retrieval," *AAAI Conference on Artificial Intelligence* 40, no. 14 (2026): 12099–12107, https://doi.org/10.1609/aaai.v40i14.38199.

40. X. Yuan, S. Xie, L. Hua, C. Zhang, G. Zhao, and L. Y. Wu, "Geometry-Aware Noisy Correspondence Mitigation for Cross-Modal Text-Based Person Retrieval," *AAAI Conference on Artificial Intelligence* 40, no. 15 (2026): 12268–12276, https://doi.org/10.1609/aaai.v40i15.38218.

41. T. Yao, Y. Wu, L. Zhang, G. Sheng, Y. Li, and Q. Tian, "Correspondence Calibrating and Dynamic Consistency Learning for Noisy Cross-Modal Retrieval," *IEEE Transactions on Multimedia* (2026).

42. W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A Survey on Mixture of Experts in Large Language Models," *IEEE Transactions on Knowledge and Data Engineering* 37, no. 7 (2025): 3896–3915, https://doi.org/10.1109/TKDE.2025.3554028.

43. Z. Cao, Q. Xu, Z. Yang, X. Cao, and Q. Huang, "Cakge: Context-Aware Adaptive Learning for Dynamic Knowledge Graph Embeddings," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2026).

44. T. Chen, C. Wang, Y. Zhang, K. Xia, and P. Qian, "Dmfusion: Degradation-Customized Mixture-Of-Experts With Adaptive Discrimination for Multi-Modal Image Fusion," *IEEE Transactions on Circuits and Systems for Video Technology* (2026).

45. T. M. Nehdi, N. Mrabah, A. Belal, M. Pedersoli, and E. Granger, "Low-Rank Expert Merging for Multi-Source Domain Adaptation in Person Re-Identification," *Winter Conference on Applications of Computer Vision* (2026): 1809–1819.

46. Y. Liu, Z. Liu, X. Lan, et al., "Domain-Aware Mixture-Of-Adapters for Text-Based Person Retrieval," *AAAI Conference on Artificial Intelligence* (2025): 5703–5711, https://doi.org/10.1609/aaai.v39i6.32608.

47. Y. Wang, Y. Zhang, X. Ren, and Y. Deng, "Moda: Mixture of Domain Adapters for Parameter-Efficient Generalizable Person Re-Identification," *ACM Transactions on Multimedia Computing, Communications and Applications* 21, no. 5 (2025): 1–19, https://doi.org/10.1145/3712595.

48. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929 (2021).

49. Y. Dai, X. Li, J. Liu, Z. Tong, and L. Duan, "Generalizable Person Re-Identification With Relevance-Aware Mixture of Experts," *IEEE Conference on Computer Vision and Pattern Recognition* (2021): 16145–16154.

50. R. Vershynin, "High-Dimensional Probability: An Introduction With Applications in Data Science," *Cambridge Series in Statistical and Probabilistic Mathematics* (2018).

51. W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* 58, no. 301 (1963): 13–30, https://doi.org/10.1080/01621459.1963.10500830.

52. J. Chen, L. Guo, J. Sun, et al., "Eve: Efficient Vision-Language Pre-Training With Masked Prediction and Modality-Aware Moe," *AAAI Conference on Artificial Intelligence* 38, no. 2 (2024): 1110–1119, https://doi.org/10.1609/aaai.v38i2.27872.

53. L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," *IEEE Conference on Computer Vision and Pattern Recognition* (2018): 79–88.

54. S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "Vgsg: Vision-Guided Semantic-Group Network for Text-Based Person Search," *IEEE Transactions on Image Processing* 33 (2024): 163–176, https://doi.org/10.1109/tip.2023.3337653.

55. S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards Unified Text-Based Person Retrieval: A Large-Scale Multi-Attribute and Language Search Benchmark," *ACM International Conference on Multimedia* (2023): 4492–4501.

56. W. Tan, C. Ding, J. Jiang, F. Wang, Y. Zhan, and D. Tao, "Harnessing the Power of Mllms for Transferable Text-To-Image Person reID," *IEEE Conference on Computer Vision and Pattern Recognition* (2024): 17127–17137.

57. N. Shazeer, A. Mirhoseini, K. Maziarz, et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-Of-Experts Layer," *International Conference on Learning Representations* (2017).