DM-Adapter: Domain-Aware Mixture-of-Adapters for Text-Based Person Retrieval

Yating Liu[12], Zimo Liu[2], Xiangyuan Lan [24], Wenming Yang[1], Yaowei Li[23*], Qingmin Liao[1*],

[1]Shenzhen International Graduate School, Tsinghua University, China
[2] Pengcheng Laboratory, China
[3] School of ECE, Peking University, China
[4] Pazhou Laboratory (Huangpu), China
liuyatin21@mails.tsinghua.edu.cn, {liuzm, lanxy}@pcl.ac.cn,
yang.wenming@sz.tsinghua.edu.cn, ywl@stu.pku.edu.cn, liaoqm@tsinghua.edu.cn

Abstract

Text-based person retrieval (TPR) has gained significant attention as a fine-grained and challenging task that closely aligns with practical applications. Tailoring CLIP to person domain is now a emerging research topic due to the abundant knowledge of vision-language pretraining, but challenges still remain during fine-tuning: (i) Previous full-model fine-tuning in TPR is computationally expensive and prone to overfitting.(ii) Existing parameter-efficient transfer learning (PETL) for TPR lacks of fine-grained feature extraction. To address these issues, we propose **D**omain-Aware **M**ixture-of-**Adapters** (**DM-Adapter**), which unifies Mixture-of-Experts (MOE) and PETL to enhance fine-grained feature representations while maintaining efficiency. Specifically, **Sparse Mixture-of-Adapters** is designed in parallel to MLP layers in both vision and language branches, where different experts specialize in distinct aspects of person knowledge to handle features more finely. To promote the router to exploit domain information effectively and alleviate the routing imbalance, **Domain-Aware Router** is then developed by building a novel gating function and injecting learnable domain-aware prompts. Extensive experiments show that our DM-Adapter achieves state-of-the-art performance, outperforming previous methods by a significant margin.

Introduction

Text-based Person Retrieval (TPR) (Li et al. 2017) is a cross-modal task that aims to retrieve persons from a large-scale image pool based on textual descriptions instead of images, which is crucial in intelligent transportation and security scenarios, especially when witnesses can only provide textual descriptions without any target images.

Due to the remarkable generalization in cross-modal understanding, Vision-Language Pre-training (VLP) models (Zhang et al. 2024) have garnered extensive interest from both academia and industry. Among these, the most representative work is Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021), which was pretrained on a large-scale dataset of 400 million image-text pairs. FFT (Fully Fine-Tuning)-based methods, such as IRRA (Jiang and Ye 2023) and CFine (Yan et al. 2023), train the entire model and leverage CLIP's original knowledge only during initialization. These methods typically employ complex fine-grained modules to enhance performance. In contrast, the PETL-based mechanism like CSKT (Liu et al. 2024a) freezes the whole backbone of CLIP and introduces effective components such as bidirectional vision-language prompts to fine-tune a small number of parameters, which achieves comparable results while significantly reducing training and storage costs.

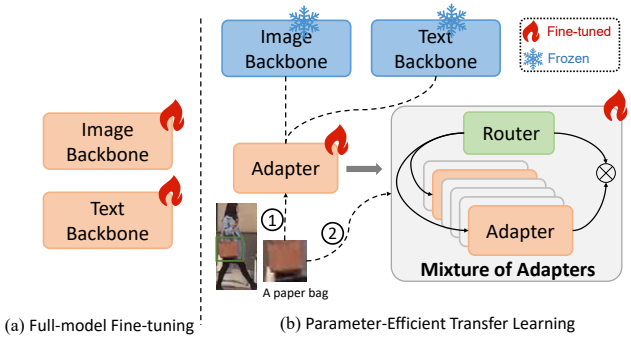

Figure 1: Evolution of CLIP-based paradigms for text-based person retrieval. (a) The FFT-based method unfreezes and trains the entire model. (b) The recent PETL-based method freezes CLIP and uses a single adapter on the input token as shown in the *left*. Our mixture-of-adapters achieves the fine-grained knowledge transferring with MOE in the *right*.

However, existing approaches still face two primary challenges: (i) FFT-based methods, despite their fine-grained transfer capabilities, require enormous time and computational consumption, and have a risk of overfitting on relatively small-scale person datasets. (ii) Recent PETL-based method CSKT, while efficient in trainable parameters, lacks of more fine-grained and specialized considerations for the specific domain of person retrieval, leading to a decline in overall performance. Meanwhile, the success of mixture-of-experts (MOE) models (Jiang et al. 2024) in foundation models that expands a feedforward block into multiple blocks, offers the potential of designing fine-grained PETL-

*Corresponding author.

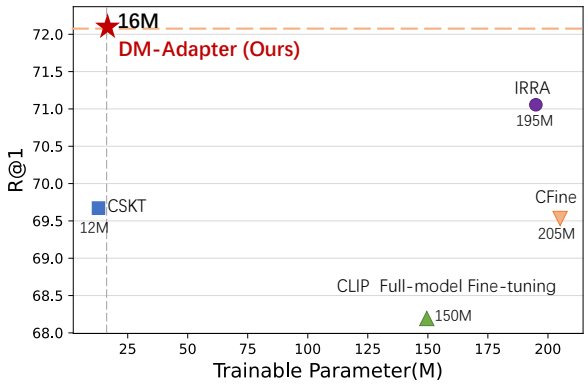

Figure 2: Comparison with CLIP-based methods. Our approach achieves the best trade-off between performance and parameter efficiency.

related components.

In this paper, we propose **DM-Adapter**, *i.e.*, **D**omain-Aware **M**ixture-of-**Adapters** for Text-Based Person Retrieval to achieve robust performance while keeping parameter efficiency. To make the vanilla adapter more fine-grained and specialized, we first design vision-language **Sparse Mixture-of-Adapters (SMA)**, which are composed of a Top-K router and multiple adapters spanning MLP layer. SMA enables different adapter experts to specialize in distinctive aspects of person characteristics, thereby achieving fine-grained transfer learning. Meanwhile, to alleviate the routing imbalance in MOE and incorporate domain information related to person retrieval, **Domain-Aware Router (DR)** is proposed to establish a coupling relationship between domain information and router by designing a domain-aware gating function and injecting learnable prompts to the gate, which helps our model to select expert adapters more effectively.

In overall, DM-Adapter transfers the person-related knowledge of CLIP effectively based on mixture-of-experts and parameter-efficient transfer learning. As illustrated in Figure 2, our approach surpasses previous state-of-the-art methods while training only 16M parameters.

Our contributions are summarized as follows:

- To our knowledge, our research is the first to explore a MOE framework based on PETL for TPR, which implicitly mines fine-grained person knowledge without requiring any additional complex interaction modules.
- We design a novel domain-aware router by incorporating domain information by several learnable prompts to alleviate the imbalanced routing issue.
- We conduct comprehensive experiments on three public benchmarks *i.e.*, CUHK-PEDES, ICFG-PEDES and RSTPReid, and the results demonstrate the superiority of the proposed DM-Adapter framework.

Code: https://github.com/Liu-Yating/DM-Adapter

Related Work

Text-based Person Retrieval

Text-based Person Retrieval (TPR) along with the benchmark dataset CUHK-PEDES, was first proposed by Li *et al.* (Li et al. 2017) to solve the problem that the target query images are not always available. Earlier research predominantly adopted separate uni-modal backbones (Shu et al. 2022; Farooq et al. 2022) including ResNet, ViT, LSTM or BERT to extract the vision and language features, and their representations are then aligned by global (Zhang and Lu 2018; Chen et al. 2022b) or local (Ding et al. 2021; Gao et al. 2021) matching methods. Ding *et al.* proposed a cross-modal implicit relation reasoning and aligning framework based on Contrastive Language–Image Pre-training (CLIP) (Radford et al. 2021). Yan *et al.* (Yan et al. 2023) also designed a multi-grained matching method based on CLIP to mine cross-modal correspondences from coarse to fine. CLIP containing abundant vision and language knowledge simultaneously has emerged as a key backbone in TPR (Song, Hu, and Zhao 2024; Zhao et al. 2024; Liu et al. 2024b; Cao et al. 2024). Liu *et al.* (Liu et al. 2024a) first developed a novel parameter-efficient transfer learning method CSKT based on CLIP for TPR, which outperforms the performance of full-tuning CLIP with only fine-tuning 12M parameters. This motivates us to further explore how to adapt CLIP to text-based person retrieval with fewer parameters for more efficient and effective performance.

Parameter-Efficient Transfer Learning

Parameter-Efficient Transfer Learning (PETL) (Han et al. 2024) provides a modular and efficient way to address the challenges of full-model fine-tuning (FFT) such as catastrophic forgetting and high computation costs. They work by keeping most of the pretrained model's weights fixed, only updating a small subset of parameters to efficiently adapt the model to a new task. Adapter (Houlsby et al. 2019) and Prompt (Liu et al. 2021) were introduced to facilitate the transfer of large language models to specific downstream tasks by inserting *additional* parameters to models. LoRA (Hu et al. 2021) as a *reparameterized* fine-tuning methods, utilizes low-rank decomposition to reconstruct the weight matrices. Subsequently, cross-modal prompt MaPLe (Khattak et al. 2023) were proposed in vision-language and further achieve cross-modal interactions. In CSKT (Liu et al. 2024a), PETL was first successfully incorporated in CLIP for text-based person retrieval, which designed bidirectional prompts and dual-branch adapters to achieve superior performance compared to FFT. However, it still relies on fundamental PETL configurations, and CLIP-based PETL has not yet reached its limits in TPR.

Mixture-of-Experts

Mixture-of-Experts (MOE) has been extensively explored in computer vision (Riquelme et al. 2021), natural language processing (Shazeer et al. 2017) and vision-language pre-training (Chen et al. 2024), which designs multiple separate experts to scale up models, and integrates a gate function to

modulate the contributions of each expert. Sparse Mixture-of-Experts (SMoE) (Jiang et al. 2024) has recently gained widespread attention in large language models, which strategically activates distinct experts for input via a router, thereby yielding noteworthy efficiency enhancements. Despite the powerful capabilities of MoE, there is still rare exploration about MOE in transfer learning for cross-modal downstream tasks.

Methodology

Framework

As show in Figure 3, we adopt CLIP (ViT-B/16) as our backbone network and design Domain-Aware Mixture-of-Adapters (DM-Adapter) spanning MLP layers in both vision and language branches, which is capable of transferring knowledge within CLIP for TPR with only fine-tuning a small amount of parameters. Each DM-Adapter consists of a mixture-of-adapters and a domain-aware router to generate more fine-grained and specialized representations, where an auxiliary loss is incorporated to balance the load of the router.

For image encoder, the input image I is first partitioned to a sequence of N non-overlapping patches. The patches are then mapped to embeddings with a linear projection and added with positional embeddings to enhance spatial information. Subsequently, a [CLS] token is introduced at the beginning of the embeddings to denote the overall global representation of the image. The sequence of N^2+1 tokens is then fed into a series of transformer blocks, where a transformer block typically consists of a Multi-Head Attention (MHA) and a MLP. Layer normalization is omitted for simplicity in the framework. We finally obtain visual representations $\{v_{cls}, v_1, \cdots, v_N\}$ with v_{cls} being the global visual representation.

For text encoder, the input description T is tokenized to embeddings f by a simple tokenizer with a 49152 vocab size. f then adds [BOS] as the start of the sequence and [EOS] as the end flag. Thus, the overall sequence can be denoted as $\{f_{bos}, f_1, \ldots, f_{eos}\}$ and then fed into the transformer as above image encoder, where the output of f_{eos} is the global representation in language branch.

The visual representation v_{cls} and textual representation f_{eos} are finally interacted and calculated by Similarity Distribution Matching (SDM) (Jiang and Ye 2023) which is an effective matching loss function across different modalities.

DM-Adapter

Motivation & Intuition. Adapter (Chen et al. 2022a) as the most popular PETL approach, has demonstrated both its effectiveness and efficiency in fine-tuning various vision and language large models, which forms the foundation of our method. It inserts small modules into transformer layers, which employs a down-projection $\boldsymbol{W}_{\text{down}} \in \mathbf{R}^{d\times m}$ to map the input x to a lower-dimensional space defined by the bottleneck dimension m, followed by a nonlinear activation function f like ReLU and an up-projection with $\boldsymbol{W}_{\text{up}} \in \mathbf{R}^{m\times d}$. Adapter is then incorporated with a residual connection, formulated as:

$$h' \leftarrow h + f\left(x\boldsymbol{W}_{\text{down}}\right)\boldsymbol{W}_{\text{up}}, \tag{1}$$

where h is the output of the original x, and h' represents the final output with a adapter.

For the fine-grained TPR task, merely relying on a single adapter to extract comprehensive features remains challenging. The recent works like sparsely-activated mixture-of-experts (MoE) models (Jiang et al. 2024) have been shown effective in LLMs by merging and activating only a subset of FFN layers for each input. This approach is based on the assumption that a larger model capacity allows for the accommodation of more information. Thus, it offers the potential for fine-grained feature representation by leveraging MOE, where each expert can handle different aspects of the features, enabling more detailed representations.

Overview of DM-Adapter. To this end, in Figure 4, we first design **S**parse **M**ixture of **A**dapters (**SMA**) in parallel with the MLP layers of each transformer. This architecture leverages the non-linear feature transformations of MLP layers to enhance the model's adaptability by integrating sparse mixture-of-adapters. The output of SMA for a given input x is determined by the weighted sum of the selected adapters, where the weights are given by the gating network G of a router as shown in Equation (2). Gating in Equation (3) enhances the information capacity of SMA while decreasing computation costs. However, the general router process input information in a fully data-driven manner, lacking of any prior knowledge specific to person retrieval. Therefore, we further design a novel domain-aware router that embeds specific domain-aware information into the original router by injecting learnable prompts, the final output is refined in Equation (4.) In addition, to ensure the effectiveness of our designed domain-aware router, a load-balancing loss is further developed to balance load of experts.

Sparse Mixture-of-Adapters. To enable fine-grained representations, given n expert adapters, the forward process to MLP layer with SMA can be expressed as:

$$\sum_{i=0}^{n-1} G(x)_i \cdot Adapter_i(x), \tag{2}$$

where $G(x)_i$ is n-dimensional output representing the gating weight for the ith adapter as equation (1). We adopt a simple and effective gating mechanism (Jiang et al. 2024) by taking the softmax over the Top-K logits of a linear layer:

$$G(x) = \text{Softmax}\left(\text{TopK}\left(x \cdot W\right)\right), \tag{3}$$

where Top-K refers to selecting the highest K weights, W denotes the gating function weight of input tokens, and softmax performs normalization of the selected logits. This sparse mechanism ensures that if the K is fixed, the model's capacity is enhanced with the increase of n while its computation costs remain consistently stable. Considering the above, the output y after a MLP layer and n expert adapters

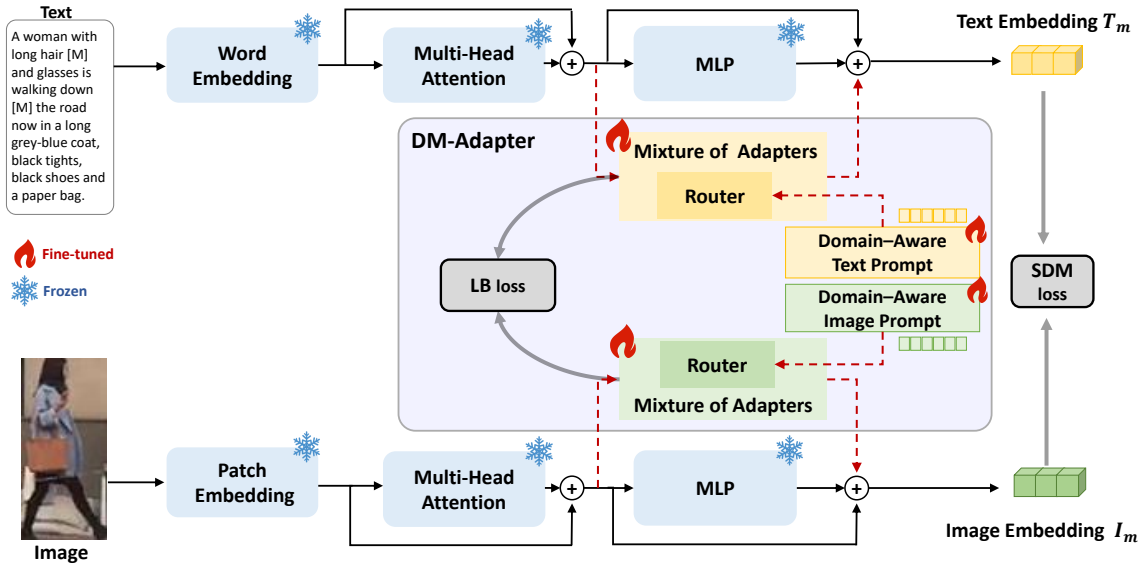

Figure 3: The overall framework of the proposed method. We adopt CLIP (ViT-B/16) as backbone, and design Domain-Aware Mixture-of-Adapters spanning MLP layer. The full parameters of vanilla CLIP are frozen during training phase. Only a fewer of parameters in DM-Adapter are trainable. The overall optimization objective incorporates SDM loss and LB auxiliary loss.

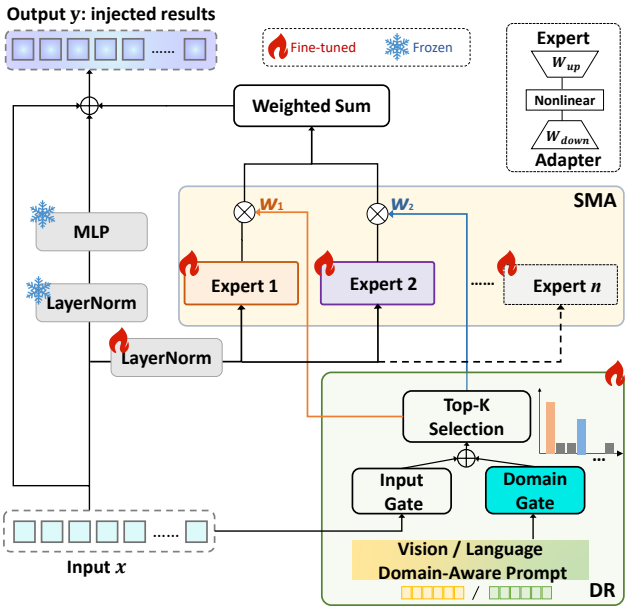

Figure 4: Architecture of DM-Adapter. DM-Adapter is mainly composed of SMA and DR. DR inserts novel domain-aware prompts on the input tokens, and designs an domain gating to capture these prompts.

for an input x is formulated as:

$$h_o = x + MLP(LN(x))$$
$$y = h_o + \sum_{i=0}^{n-1} \text{Softmax}(\text{TopK}\,(x \cdot W))_i \cdot Adapter_i(x), \tag{4}$$

where h_o is the original output of MLP, and y is the final output with expert adapters.

Domain-Aware Router. The general gating function is typically determined by the input tokens like Equation (3) and (4). Here, MOE places no prior constraints on the router, which can easily result in imbalanced routing. Meanwhile, existing routing ignores domain information when transferring foundation models to the specific TPR task.

To supplement domain-specific person knowledge, we propose a **D**omain-Aware **R**outer (DR), which incorporates domain information by several tunable prompts p embedded in vision and language, and designs a domain-aware gating function $p \cdot W_d$ based on these prompts. Thus, the item $x \cdot W$ in Equation (3) and (4) is then modified to $x \cdot W + p \cdot W_d$. The output of DM-Adapter can be reformulated as:

$$y = h_o + \sum_{i=0}^{n-1} \text{Softmax}\,(\text{TopK}\,(x \cdot W + p \cdot W_d))_i \cdot Adapter_i(x). \tag{5}$$

Load-Balancing loss. To ensure the effectiveness of our designed domain-aware router, an auxiliary loss is utilized to balance routing in mixture-of-adapters, which encourages experts to receive roughly equal numbers of training samples, and avoids concentrating the load on a single expert. Inspired by (Chen et al. 2024) and (Shazeer et al. 2017), we design Top-K **L**oad-**B**alancing (**LB**) loss to balance the average weights of the selected K experts:

$$\mathcal{L}_{aux} = \alpha \cdot \sum_{i}^{n} f_i \times p_i, \tag{6}$$

where f_i is the fraction of tokens assigned to the ith expert under the Top-K mechanism, p_i represents the average routing weight for the ith expert, and α is a hyperparameter.

Optimization and Inference

A parameter-free loss function is adopted in training phase termed as Similarity Distribution Matching (SDM) (Jiang and Ye 2023), which integrates cosine similarity distributions of the $N \times N$ embeddings for image-text pairs into the KL divergence to build up the connection of two modalities.

$$\mathcal{L}_{i2t} = KL\left(\mathbf{p_i} \| \mathbf{q_i}\right) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{i,j} \log \left(\frac{p_{i,j}}{q_{i,j} + \epsilon} \right), \tag{7}$$

where $p_{i,j}$ is the probability denoting the similarity between image-text pairs and $q_{i,j}$ is the true matching probability. Considering the SDM loss from text to image, the bi-directional SDM loss is formulated as:

$$\mathcal{L}_{sdm} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}. \tag{8}$$

Our framework is trained in an end-to-end manner, and the overall optimization objective incorporating auxiliary loss in Equation (6) from vision and language is defined as:

$$\mathcal{L} = \mathcal{L}_{sdm} + \alpha \cdot \left(\mathcal{L}^{I}_{aux} + \mathcal{L}^{T}_{aux} \right) \tag{9}$$

During inference, the trained network incorporating DM-Adapter calculates the similarities between text and image embeddings. The Top-K candidates are then processed to derive the relevant evaluation metrics for each query.

Experimental Results

Experimental Setup

Datasets. **CUHK-PEDES** (Li et al. 2017) as the most commonly used dataset, contains 40,206 images and 80,412 textual descriptions for 13,003 identities. The training set consists of 11,003 identities with 34, 054 images and 68, 126 texts. Both the validation set and test set have 1,000 identities. **ICFG-PEDES** (Ding et al. 2021) contains 54,522 images for 4,102 identities. Each image corresponds to one description. The training and test sets contain 3,102 identities and 1,000 identities respectively. **RSTPReid** (Zhu et al. 2021) as a newly released dataset contains 20,505 images of 4,101 identities. Each image has 2 descriptions. The training, validation and test sets contain 3701 identities with 18505 images, 200 identities with 1000 images, and 200 identities with 1000 images respectively.

Evaluation Measures. Rank-k metrics (k=1,5,10) are adopted as the primary evaluation metrics, which denote the probability of finding at least one person image matching within the Top-K candidates when given a textual description. Additionally, we adopt the mean Average Precision (mAP) as a comprehensive retrieval criterion. The higher Rank-k, mAP indicates better performance.

Implementation Details. The framework consists of a pre-trained image encoder, *i.e.*, CLIP-ViT-B/16, a pre-trained text encoder, *i.e.*, CLIP text Transformer, and PETL modules with mixture-of-adapters. The image is resized to 384 × 128, and the length of textual token sequence is 77. The model is trained using Adam optimizer for 60 epochs, with a batch size of 128 and an initial learning rate 3×10^{-4}. We utilize the reduction parameter 8 representing the bottleneck dimension in adapter as CSKT (Liu et al. 2024a). Top-K is set to 2, and the number of experts is 6. The hyper-parameter α that indicates the auxiliary loss is set to 0.5. We perform experiments on a single NVIDIA 4090 24GB GPU.

Performance and Memory Efficiency

We categorize existing methods into those based on CLIP and those based on other architectures. The primary baseline is recent PETL-based method CSKT (Liu et al. 2024a) based on CLIP backbone.

Results on CUHK-PEDES. As shown in Table 1, on the most common benchmark CUHK-PEDES, our DM-Adapter outperforms the PETL-based method CSKT across three Rank-k metrics and mAP by a large margin, with +2.47%, +1.82%, +1.05% and +2.16%, respectively . Meanwhile, as shown in Table 2, DM-Adapter is comparable with IRRA, given that IRRA with 195M trainable parameters, integrates a complex implicit reasoning module and sophisticated loss functions. In contrast, DM-Adapter with only a few 16M trainable parameters, achieves the trade-off between performance and costs.

Results on ICFG-PEDES and RSTPReid. We then perform experiments on the ICFG-PEDES dataset in Table 3 and the newly released RSTPReid dataset in Table 4, which demonstrate the similar comparison results to those in CUHK-PEDES when DM-Adapter is compared with IRRA and CSKT. DM-Adapter exceeds CSKT by +2.25% on R@1, +0.8% on R@5, and +0.94% on mAP. Furthermore, we observe that for the most complex CLIP-based method CFine, which incorporates multiple explicit granularity modules, has the worse overall performance on RSTPReid. DM-Adapter outperforms CFine by an absolute margin of +9.45% on R@1. We infer that the full-model model with a larger number of training parameters CFine (Yan et al. 2023) is prone to overfitting on the smallest RSTPReid dataset, compared to the PETL-based methods such as DM-Adapter and CSKT, resulting in poor generalization.

Overall, DM-Adapter reliably delivers the better trade-off between performance and computation costs across all three benchmark datasets, highlighting the generalization and robustness of our proposed approach.

Ablation Study

A comprehensive ablation study for components of DM-Adapter is presented in Table 5, including the most critical accuracy metric R@1 and the average metric on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets. The results in No.1 serve as the backbone baseline by zero-shot CLIP, where inference is performed directly on the original frozen CLIP model without adding any additional trainable modules. In No.2, the sub-module adapter in CSKT based on the full-model frozen CLIP for TPR task is reproduced by spanning single adapter on MLP layer of CLIP, and the R@1 metric is enhanced to 71.00% by following the proper initialization (Gao et al. 2023).

	Method	Type	Ref	Image Enc.	Text Enc.	R@1	R@5	R@10	mAP
w/o CLIP	CMPM/C (Zhang and Lu 2018)	L	ECCV18	RN50	LSTM	49.37	71.69	79.27	-
	ViTAA (Wang et al. 2020)	L	ECCV20	RN50	LSTM	55.97	75.84	83.52	-
	NAFS (Gao et al. 2021)	L	arXiv21	RN50	BERT	59.36	79.13	86.00	54.07
	DSSL (Zhu et al. 2021)	L	MM21	RN50	BERT	59.98	80.41	87.56	-
	SSAN (Ding et al. 2021)	L	arXiv21	RN50	LSTM	61.37	80.15	86.73	-
	SAF (Li, Cao, and Zhang 2022)	L	ICASSP22	ViT-Base	BERT	64.13	82.62	88.40	58.61
	TIPCB (Chen et al. 2022b)	L	Neuro22	RN50	BERT	64.26	83.19	89.10	-
	AXM-Net (Farooq et al. 2022)	L	MM22	RN50	BERT	64.44	80.52	86.77	58.73
	LGUR (Shao et al. 2022)	L	MM22	DeiT-Small	BERT	65.25	83.12	89.00	-
	IVT (Shu et al. 2022)	G	ECCV22	ViT-Base	BERT	65.59	83.11	89.21	-
w/ CLIP	Han et al. (Han et al. 2021)	G	BMVC21	CLIP-RN101	CLIP-Transformer	64.08	81.73	88.19	60.08
	CFine (Yan et al. 2023)	L	TIP23	CLIP-ViT	BERT	69.57	85.93	91.15	-
	IRRA-CLIP (Jiang and Ye 2023)	G	CVPR23	CLIP-ViT	CLIP-Transformer	68.19	86.47	91.47	61.12
	IRRA* (Jiang and Ye 2023)	G	CVPR23	CLIP-ViT	CLIP-Transformer	71.15	87.66	92.58	64.84
	IRRA (Jiang and Ye 2023)	G	CVPR23	CLIP-ViT	CLIP-Transformer	73.38	89.93	93.71	66.13
	CSKT (Liu et al. 2024a)	P+G	ICASSP24	CLIP-ViT	CLIP-Transformer	69.70	86.92	91.80	62.74
	DM-Adapter (Ours)	P+G	-	CLIP-ViT	CLIP-Transformer	72.17	88.74	92.85	64.33

$*$ indicates our replication results after a minor bug correction, which can be regarded as a data augmentation technique in vanilla IRRA.

Table 1: Comparison Performance with other methods on CUHK-PEDES. The left column denotes whether using CLIP. "G" and "L" in "Type" denote global and local matching. "P" stands for the PETL-related methods (such as CSKT and ours).

Method	R@1 ↑	Memory Cost (M) ↓	Trainable # Param (M) ↓
IRRA *	71.15	7034 (28.64%)	195M
IRRA	73.38	7034 (28.64%)	195M
CFine	69.57	13570 (55.24%)	205M
CSKT	69.70	2338 (9.52%)	12M
DM-Adapter (Ours)	72.17	2952 (12.02%)	16M

Table 2: Analysis of Memory Efficiency and Effectiveness on CUHK-PEDES. To ensure a fair comparison of efficiency, batch size for all methods is set to 32.

	Method	R@1	R@5	R@10	mAP
w/o CLIP	CMPM/C (Zhang and Lu 2018)	43.51	65.44	74.26	-
	ViTAA (Wang et al. 2020)	50.98	68.79	75.78	-
	SSAN (Ding et al. 2021)	54.23	72.63	79.53	-
	SAF (Li, Cao, and Zhang 2022)	54.86	72.13	79.13	32.76
	TIPCB (Chen et al. 2022b)	54.96	74.72	81.89	-
	IVT (Shu et al. 2022)	56.04	73.60	80.22	-
	LGUR (Shao et al. 2022)	59.02	75.32	81.56	-
w/ CLIP	CFine (Yan et al. 2023)	60.83	76.55	82.42	-
	IRRA-CLIP (Jiang and Ye 2023)	56.74	75.72	82.26	31.84
	IRRA* (Jiang and Ye 2023)	61.36	78.66	84.60	37.95
	IRRA (Jiang and Ye 2023)	63.46	80.25	85.82	38.06
	CSKT (Liu et al. 2024a)	58.90	77.31	83.56	33.87
	DM-Adapter (Ours)	62.64	79.53	85.32	36.50

Table 3: Comparison on ICFG-PEDES.

To demonstrate the effectiveness of our proposed Sparse Mixture-of-Adapters (SMA), we compare it with single adapter (No.2 *vs.* No.3), and show that SMA achieves a significant improvement by +1.00% on CUHK-PEDES and +0.5% on RSTPReid. It indicates that adopting MOE structure enhances capabilities of individual expert, enabling the more fine-grained feature extraction, where each expert processes input tokens from different specialized perspectives.

To further validate the effectiveness of Load-Balancing loss, we compared the performance of SMA trained with (No.3) and without (No.2) the LB loss. It is evident that LB is valid in alleviating load imbalance, especially on smaller

	Method	R@1	R@5	R@10	mAP
w/o CLIP	DSSL (Zhu et al. 2021)	32.43	55.08	63.19	-
	SSAN (Ding et al. 2021)	43.50	67.80	77.15	-
	SAF (Li, Cao, and Zhang 2022)	44.05	67.30	76.25	36.81
	IVT (Shu et al. 2022)	46.70	70.00	78.80	-
w/ CLIP	CFine (Yan et al. 2023)	50.55	72.50	81.60	-
	IRRA-CLIP (Jiang and Ye 2023)	54.05	80.70	88.00	43.41
	IRRA* (Jiang and Ye 2023)	57.50	80.15	87.05	44.31
	IRRA (Jiang and Ye 2023)	60.20	81.30	88.20	47.17
	CSKT (Liu et al. 2024a)	57.75	81.30	88.35	46.43
	DM-Adapter (Ours)	60.00	82.10	87.90	47.37

Table 4: Comparison on RSTPReid.

datasets such as RSTPReid.

Moreover, we compare the performance of the model whether using our proposed Domain-Aware Router (DR) or original router (No.3 *vs.* No.4). The results show that DR can prompt the router to select experts by supplementing domain-aware information and routing, and thus achieve better performance.

In summary, sparse mixture-of-adapters enhances model capacity and helps to unleash the fine-grained feature extraction power of the pretrained CLIP. It further pushs parameter-efficient transfer learning to the limit for the specific TPR task with domain information prompts. Compared to the original adapter MLP-Adapter (No.2), DM-Adapter (No.4) achieves significant improvements across all three datasets, gaining a substantial increment by +1.05% on the overall average R@1 metric.

Hyper-parameter Analysis

The Number of Experts. As shown in Figure 5 (*upper*), to investigate the impact of the number of experts n, we sample n as 2, 4, 6, 8 and 10 to evaluate the R@1 and trainable parameters under different numbers of experts. A constant Top-K value is fixed to 2 in the overall experiment. When n is less than 6, average R@1 gradually increases with an in-

No.	Methods	Components				CUHK-PEDES	ICFG-PEDES	RSTPReid	Avg.
		MLP-Adapter	SMA	LB	DR				
0	Zero-shot CLIP					12.65	6.66	13.55	10.96
1	+ MLP-Adapter	✓				71.00	62.13	58.55	63.89
2	+ Sparse Mixture-of-Adapters (w/o LB)		✓			72.00	62.13	59.05	64.38
3	+ Sparse Mixture-of-Adapters ((w/ LB))		✓	✓		72.14	62.40	59.40	64.65
4	+ Domain-Aware Router (**DM-Adapter**)		✓	✓	✓	**72.17**	**62.64**	**60.00**	**64.94**

Table 5: Ablation study on R@1 about each component of DM-Adapter. The metric Avg. denotes the average R@1 across three datasets. We reproduce MLP-Adapter (Liu et al. 2024a) by following the proper initialization (Gao et al. 2023).

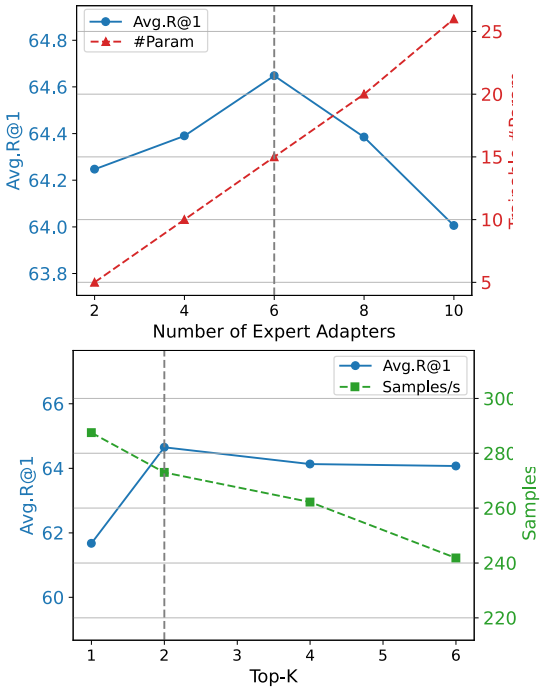

Figure 5: The results of experiments for hyper-parameters.

creasing number of experts. However, when n exceeds 6, larger n leads to a decrease in performance. we observe that although increasing n can proportionally enhance the model's information capacity, a larger n does not necessarily lead to better performance. This suggests that the model's capacity cannot grow indefinitely and should be aligned with the scale of the training data. We ultimately determine $n = 6$ as a practical choice.

The Number of Top-K. In Figure 5 (*lower*), we set the number of experts $n = 6$ and explore the R@1 accuracy and computational complexity with the change of Top-K. It clearly demonstrates that the sample processing speed deteriorates as Top-K increases. When Top-K is 1, performance shows worse as it would damage the powerful ability of mixture-of-experts and degrades to a single adapter. As Top-K increases further, the model's performance essentially reaches a plateau without additional improvement. Given the need to strike a balance between efficiency and performance, a practical choice for Top-K would be 2.

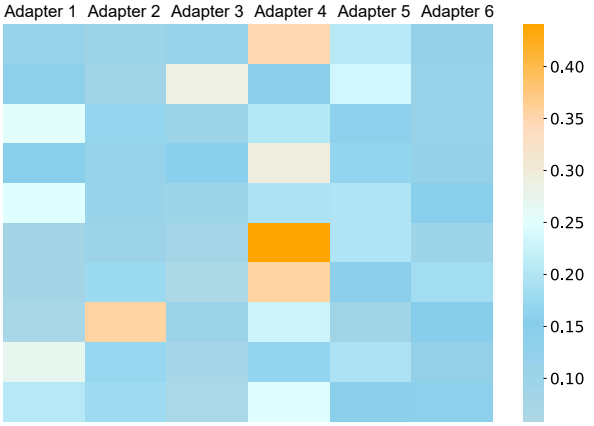

Figure 6: Visualization of Expert Weight. Each row represents the weights assigned to different experts for tokens of a description. Each column indicates a expert adapter.

Visualization of Expert Weight

In Figure 6, we present a weight visualization of the 6 expert adapters for a person description. We analyze the weights of DM-Adapter in the 12*th* layer of CLIP, as it represents the most high-level features. Each row represents the distribution of weights assigned to input tokens across mixture-of-experts, and the sum of the weights equals 1 under normalization. This demonstrates that DM-Adapter can implicitly handle feature granularity more precisely, as different experts specialize in distinct aspects of person knowledge.

Conclusion

In this paper, we present a novel CLIP-based parameter-efficient transfer learning method DM-Adapter to achieve implicit fine-grained knowledge transferring, which freezes the entire CLIP backbone and only trains a few parameters with domain-aware mixture-of-adapters. Extensive experiments demonstrate that our approach achieves the best trade-off between robust performance and parameter efficiency, and outperforms existing the PETL-based methods.

Acknowledgements

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(No.KJZD20231023094700001) and National Natural Science Foundation of China (62402252).

References

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

Chen, J.; Guo, L.; Sun, J.; Shao, S.; Yuan, Z.; Lin, L.; and Zhang, D. 2024. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1110–1119.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022a. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

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

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *arXiv preprint arXiv:2107.12666*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4477–4485.

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

Gao, Q.; Zhao, C.; Sun, Y.; Xi, T.; Zhang, G.; Ghanem, B.; and Zhang, J. 2023. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11483–11493.

Han, X.; He, S.; Zhang, L.; and Xiang, T. 2021. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*.

Han, Z.; Gao, C.; Liu, J.; Zhang, S. Q.; et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

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

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person Search With Natural Language Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Liu, Y.; Li, Y.; Liu, Z.; Yang, W.; Wang, Y.; and Liao, Q. 2024a. CLIP-based Synergistic Knowledge Transfer for Text-based Person Retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7935–7939. IEEE.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024b. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

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

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

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

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

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

Song, Z.; Hu, G.; and Zhao, C. 2024. Diverse Person: Customize Your Own Dataset for Text-Based Person Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4943–4951.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. Vitaa: Visual-textual attributes alignment in person search by natu-

ral language. In *European Conference on Computer Vision*, 402–420. Springer.

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

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Zhao, Z.; Liu, B.; Lu, Y.; Chu, Q.; and Yu, N. 2024. Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-to-Image Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7534–7542.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.