Human-centered Interactive Learning via MLLMs for Text-to-Image Person Re-identification

Yang Qin[1] Chao Chen[2] Zhihang Fu[2] Dezhong Peng[1,4,5] Xi Peng[1,3] Peng Hu[1*]
[1]College of Computer Science, Sichuan University [2]Independent Researcher
[3]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, Sichuan University
[4]Sichuan National Innovation New Vision UHD Video Technology Co., Ltd [5]Tianfu Jincheng Laboratory

Abstract

*Despite remarkable advancements in text-to-image person re-identification (TIReID) facilitated by the breakthrough of cross-modal embedding models, existing methods often struggle to distinguish challenging candidate images due to intrinsic limitations, such as network architecture and data quality. To address these issues, we propose an **I**nteractive **C**ross-modal **L**earning framework (ICL), which leverages human-centered interaction to enhance the discriminability of text queries through external multimodal knowledge. To achieve this, we propose a plug-and-play Test-time Humane-centered Interaction (THI) module, which performs visual question answering focused on human characteristics, facilitating multi-round interactions with a multimodal large language model (MLLM) to align query intent with latent target images. Specifically, THI refines user queries based on the MLLM responses to reduce the gap to the best-matching images, thereby boosting ranking accuracy. Additionally, to address the limitation of low-quality training texts, we introduce a novel Reorganization Data Augmentation (RDA) strategy based on information enrichment and diversity enhancement to enhance query discriminability by enriching, decomposing, and reorganizing person descriptions. Extensive experiments on four TIReID benchmarks, i.e., CUHK-PEDES, ICFG-PEDES RSTPReid, RSTPReid, and UFine6926, demonstrate that our method achieves remarkable performance with substantial improvement. Code is available at https://github.com/QinYang79/ICL.*

1. Introduction

Recently, text-to-image person re-identification (TIReID) [3–5, 7, 33, 34, 41] has made great progress in aligning text descriptions and person images, enabling high-accuracy person search and identification. Different from traditional image-to-image re-identification [32, 39, 42, 43], TIReID, as a rising task in the multimodal community [19, 20, 22, 23, 27], retrieves person images using user-customized text queries, which is more practical in many scenarios where an image query is unavailable, such as in video surveillance systems [2] or crowd management [6]. A primary challenge in TIReID is learning to associate text queries with the corresponding person images at a fine-grained level, bridging the modality gap for accurate similarity measurement.

To overcome this challenge, recent efforts have developed various strategies to improve cross-modal associations, including local attribute modeling [37, 44], loss function designing [12, 21], ReID-domain pre-training [28, 38], *etc*. Although these methods achieve promising perfor-

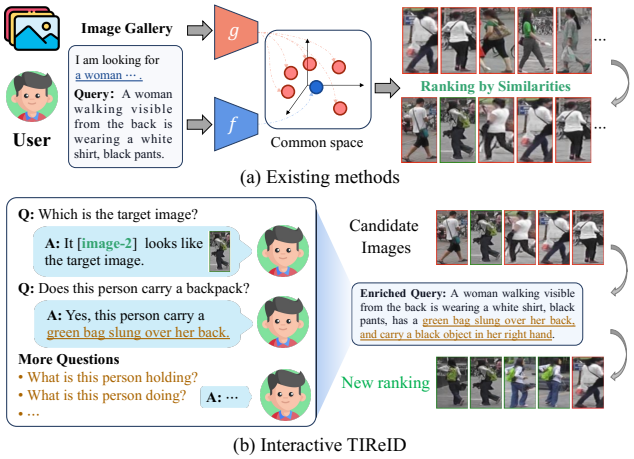


Figure 1. The illustration of our motivation. When performing text-based person re-identification, (a) existing methods commonly exploit cross-modal models to calculate similarity (such as IRRA [12] and RDE [21]) and then obtain the candidate person images by ranking. However, due to the intrinsic limitations of models and training data, they may not be able to distinguish challenging candidate images well enough to obtain satisfactory results. (b) Our motivation is to interact with the system for external guidance like a human and gradually refine the user's query from the candidate items shown in (a) by multiple rounds of question-answering, ultimately improving the overall ranking.

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

mance, they are limited by the inherent defect of the offline models and training data, making it hard to handle dynamic user query inputs and leading to poor generalization. In practice, however, different users tend to input concise, vague, and diverse text queries based on their memories to the target person, thereby raising challenges for the offline models to understand fine-grained human characteristics from the queries. For example, as illustrated in Figure 1 (a), the user text query "*A woman walking visible from the back is wearing a white shirt, and black pants.*" fails to capture some important contextual details like handheld objects or background, resulting in inaccurate candidate images. In brief, it is hard for offline models to handle dynamic and challenging queries solely relying on their learned internal knowledge. To break through this bottleneck, we could seek to align queries with image candidates with the help of external knowledge, thereby reranking the retrieved images and boosting the identification accuracy.

Recent studies [8, 13, 17] have explored integrating external knowledge from pre-trained models or (multimodal) large language models to guide various downstream tasks. For instance, Li *et al*. [17] leveraged multimodal knowledge of pre-trained vision-language models [24] to enhance unimodal image representations for external-guided clustering. However, this approach cannot achieve post-hoc improvement for offline models, leading to infeasibility in practice due to the prohibitive cost of fine-tuning or even retraining. In cross-modal retrieval, Han *et al*. [8] exploited the excellent comprehension and generation capabilities of multimodal large language models (MLLMs) to design an interactive re-ranking pipeline for text-to-video retrieval. However, Han *et al*. [8] introduces label leakage by using the ground truth video as the video in mind for the MLLM answerer agent, which is not applicable in actual testing.

In this paper, we present a novel Interactive Cross-modal Learning framework (ICL) that exploits the multimodal knowledge implicit in MLLMs to enhance the alignment between text queries and target person images. Specifically, we first propose a plug-and-play Test-time Human-centered Interaction (THI) module that refines text queries through interactions with an MLLM, enhancing the post-hoc ability of trained models to distinguish challenging candidates. THI identifies the latent target images through multi-round interactions with MLLMs, asking human-centered questions to the MLLM for fine-grained answers about the images, which refine the query texts to strengthen the alignment with the images, ultimately address the inherent limitations of the input queries and enhancing ranking accuracy. To address the intrinsic limitations in the trained models, we propose a Reorganization Data Augmentation strategy (RDA) to enrich and diversify pedestrian texts for training enhancement, thereby transferring external knowledge in MLLMs into the model. To enrich the texts, RDA applies visual question answering via the MLLM on each image-text pair to supplement fine-grained person characteristics. Moreover, to enhance the diversity of texts, RDA presents a decomposition-reorganization strategy to decompose person descriptions into attribute-specific sub-sentences (*e.g*., clothes, pants, shoes, *etc*.), and rewrite them with MLLM into multiple sentences with the same meaning. Subsequently, the rewritten sentences are randomly reordered and recombined to generate varied augmented texts, thus boosting data diversity and enhancing the generalization of models. Our main contributions are as follows:

- We introduce human-centered interaction to TIReID, proposing a novel MLLM-driven Interactive Cross-modal Learning framework (ICL), which leverages external knowledge to overcome the inherent limitations of existing offline methods in handling dynamic queries.
- A plug-and-play Test-time Human-centered Interaction module (THI) is presented to align query intent with latent target images through multi-round interactions with MLLMs, improving the ranking quality.
- An effective Reorganization Data Augmentation strategy (RDA), applying MLLMs for text decomposition and recombination, is developed to generate discriminative and diverse training texts, improving cross-modal learning.
- Extensive experiments on four text-to-image person re-identification benchmarks verify the effectiveness and superiority of our method, which achieves promising performance and superior generalization.

2. Related Work

2.1. Text-to-Image Person Re-identification

As a challenging topic in multimodal learning, text-to-image person re-identification (TIReID) aims to search the image of the target person with a given natural language query. Existing methods [1, 12, 21, 28, 41] can be roughly divided into three categories according to the used backbone type: unimodal backbones, general multimodal pre-trained backbones, and ReID-domain multimodal pre-trained backbones. The early TIReID methods [33, 35] commonly use modality-specific unimodal backbones to encode text or images for cross-modal alignment, *e.g*., ResNet [10], BERT [30], *etc*. Recently, benefiting from the rapid development of visual-language pre-training models [14, 24], more and more researchers [3, 12, 21] have begun to use general pre-trained models as the backbone networks for solving TIReID, hoping that the pre-trained alignment knowledge can improve modality representation and cross-modal alignment. However, the performance improvement is still limited due to the domain gap between the general pre-training data and TIReID tasks. To this end, some attempts [28, 38] exploit image-caption models or MLLMs to annotate images, thus obtaining a large number of image-

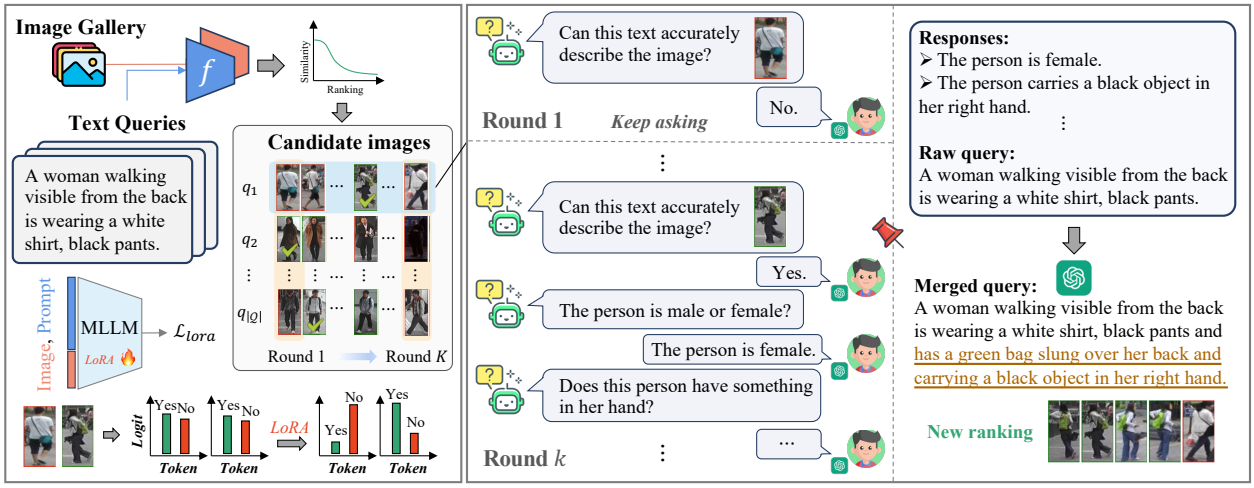


Figure 2. The illustration of our Test-time Human-centered Interaction (THI) module. THI includes K rounds of interactions to align query intention with the latent target image by external guidance, where in each round, we perform human-centered visual question answering around fine-grained person attributes to enhance the semantic consistency between the query and the intended person image, and then improve the final ReID performance on the large-scale evaluation through efficient re-ranking. Besides, we perform supervised fine-tuning via LoRA to inspire the discriminate ability of MLLM for ReID domain images and better align queries with latent target images.

text pairs for ReID-domain pre-training. Although promising performance has been achieved, they still cannot get rid of the inherent defect of offline models during handling dynamic queries. In this paper, we propose a novel interactive cross-modal learning framework based on MLLMs, which exploits dynamic interactions for external guidance to improve the generalization ability.

2.2. Visual Interactive Learning

With the booming development of multimodal large language models (MLLMs) [15, 18, 29], exploiting visual interaction to improve downstream tasks has attracted increasing attention from researchers. We collectively refer to these as visual interaction learning and divide them into offline interaction [28, 31] and online interaction [8, 9, 13] groups by engaging in specific task periods. The former aims to use interactions to improve the training/pre-training data, such as information richness and diversity. It is usually static and separated from users, which is the mainstream of existing TIReID methods, *e.g.*, Tan *et al*. [28] apply MLLMs to conduct interactions with multiple prompt templates for better diversity. Unlike them, online interactions are designed to be able to use real-time feedback to improve training or testing. For example, Levy *et al*. [13] propose to gradually clarify user intention through dialogue to improve image retrieval performance. Likewise, Han *et al*. [8] develop a rerank pipeline based on LLM-based iterative navigation. However, these methods cannot be directly used for TIReID effectively due to the differences in tasks and data domains. In this paper, we develop a test-time module based on MLLMs to conduct interactions for external guidance, overcoming the limitations of the intra-model knowledge in TIReID offline models.

3. Methodology

In this section, we introduce an Interactive Cross-modal Learning framework (ICL), which consists of two core comments to address the inherent challenges in offline models and training data, *i.e.*, Test-time Human-centered Interaction module (THI) and Reorganization Data Augmentation (RDA). In Section 3.1, we provide the necessary definitions to facilitate the study. Then, we outline the details of our THI and RDA in Sections 3.2 and 3.3, respectively.

3.1. Problem and Symbol Statement

Suppose that we have the query text $q \in \mathcal{Q}$, a pedestrian image $v \in \mathcal{V}$, where $\mathcal{Q}$ and $\mathcal{V}$ are the text query set and image gallery. The purpose of TIReID is to utilize the text query to retrieve the ideal images from the gallery, thus achieving person searching by identities of retrieved images. To achieve this, existing methods usually train an offline cross-modal model $f_{\text{cross}} = \{f, g\}$ to measure the similarities between text queries and pedestrian images. For a text q and an image v, the similarity to measure the matching degree can be represented as $S_{q,v} \equiv \text{Sim}(f(q), g(v))$, where $\text{Sim}(*)$ is the similarity function, f and g are the text and image encoders, respectively. Then, we can utilize q to search relevant candidate images $\hat{v}$ from $\mathcal{V}$ as follows:

$$\hat{\mathcal{V}}(q) = \{\hat{v}_k\}_{k=1}^{K} = \text{Top-}K_{v \in \mathcal{V}}(S_{q,v}), \tag{1}$$

where $\hat{\mathcal{V}}(q)$ is the candidate set for q and K is the number of candidate images. In addition, our ICL also involves MLLMs to conduct interactions and training augmentation,

which we denote as $\mathcal{M}$, and the prompt template function for $\mathcal{M}$ is denoted by the symbol $\mathcal{T}$. Based on the above definitions, we will elaborate on ICL in the following sections.

3.2. Test-time Humane-center Interaction

Anchor Localization. Due to the model bottleneck, it is hard for existing methods to distinguish challenging candidate images relying solely on internal knowledge. Thus, we exploit the fine-grained image understanding capability of MLLMs to conduct external guidance. We first propose asking questions such as "*Can this text accurately describe the image?*" to let MLLMs explicitly tell us the latent target (anchor) image. As shown in Figure 2, given a text query q, we first get the corresponding candidate set $\hat{\mathcal{V}}(q)$ and then perform multiple (K) rounds of interactions according to the ranking until the ideal image is determined. For the k-th round, the answer for interaction can represented as:

$$a^q_{\hat{v}_k} = \mathcal{M}(\mathcal{T}_{\text{loc}}(q, \hat{v}_k)), \tag{2}$$

where $k \in \{1, \cdots, K\}$, $a^q_{\hat{v}_k}$ is the answer of 'Yes' or 'No', $\hat{v}_k$ is the Top-k candidate image for query q, and $\mathcal{T}_{\text{loc}}$ is the prompt template function for anchor localization. However, since there is a domain gap between ReID images and the generic images used for pre-training or instruction-tuning MLLMs, the answers of MLLMs are often unreliable. To solve this, we exploit LoRA [11] to perform supervised fine-tuning (SFT) on MLLMs to inspire the fine-grained ability to identify the person image and the SFT loss is:

$$\mathcal{L}_{lora} = -\sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log(p_{\mathcal{M}}(y_t|x, y < t)), \tag{3}$$

where x is the input `Pormpt` and y is the output `Response`, $\mathcal{Z} = \{\mathcal{Z}^+, \mathcal{Z}^-\}$ is the SFT dataset. For $\mathcal{Z}^+$, we select part of training texts ($\{q_i\}_{i=1}^{N_l}$) and the corresponding ground-truth images ($\{v_i^+\}_{i=1}^{N_l}$) from the training set to construct the input `Pormpts`, *i.e.*, $\{\mathcal{T}_{\text{loc}}(q_i, v_i^+)\}_{i=1}^{N_l}$, expecting MLLMs to output the `Response` of 'Yes'. Similarly, for $\mathcal{Z}^-$, we also use the text queries and the corresponding negative images from the training set to construct the input `Pormpts`, expecting MLLMs to output the `Response` of 'No'. To make $\mathcal{Z}^-$ more discriminative, we use a TIReID pre-trained cross-modal model to obtain the Top-10 image with different person ID as input negative sample image for each text query by similarity ranking.

Human-centered VQA. For each round, once we confirm the response of 'Yes', we will ask the anchor image ($\bar{v}$) a series of questions to learn more details about pedestrian characteristics. These details can help us alleviate the discrepancy between the user query and the information within the anchor image, thus improving overall ranking by refining the query. We call this process human-centered visual question answering (VQA), which is expressed as:

$$r_{\bar{v}} = \mathcal{M}(\mathcal{T}_{\text{vqa}}(\{c_i\}_{i=1}^{N_q}, \bar{v})), \tag{4}$$

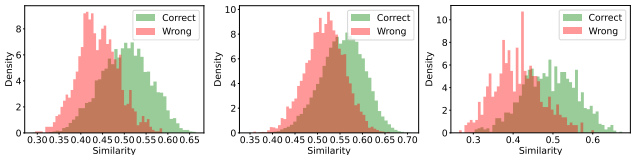
where $\{c_i\}_{i=1}^{N_q}$ are N_q questions directed at the detail characteristics (*e.g.*, gender, hair, upper body, lower body, shoes, *etc.*) of the pedestrian image, $r_{\bar{v}}$ is the answer to $\{c_i\}_{i=1}^{N_q}$ one by one, and $\mathcal{T}_{\text{vqa}}$ is the prompt template function for human-centered VQA. To improve the consistency between the image and the user query, a natural strategy is to concatenate these one-by-one answers after the original text. However, due to the limitation of the maximum text length that can be processed by the text model (*e.g.*, it is usually 77 for CLIP), this will destroy the sequence structure. To this end, we recommend merging these answers and the original text by MLLMs to obtain a fluent and concise text, *i.e.*,

$$\hat{q} = \mathcal{M}(\mathcal{T}_{\text{aggr}}(r_{\bar{v}}, q)), \tag{5}$$

where $\hat{q}$ is the merged text query and $\mathcal{T}_{\text{aggr}}$ is is the prompt template function for text aggregation.

Efficient Re-ranking. To apply the above process to large-scale evaluation, efficiency is a factor that needs to be considered in the retrieval task, since introducing MLLMs to interactions for external guidance inevitably brings additional inference costs. To improve the interaction efficiency, we recommended adopting different strategies for different rounds of interactions. We observe that the retrieval quality is positively correlated with the similarity between the query and the Top-1 candidate image, as shown in Figure 3. To this end, in the first round, we only interact with images whose cross-modal similarity is greater than a threshold ξ and the answer of anchor localization is 'Yes'. This is because the retrievals with low similarities are more likely to be wrong items. After the first round, we only interact with the image that is all considered 'No' in the previous rounds of anchor localization and whose cross-modal similarity is less than the threshold ξ. This allows us to reduce unnecessary interactions and find text queries that really require interactions, making our method more efficient on large-scale evaluation. Given a query q, if the above constraints are met, the interaction is carried out and we can get merged text $\hat{q}$ after human-centered VQA, the re-ranking similarity $\hat{S}_{q,v}$ of any image v in $\mathcal{V}$ is:

$$\hat{S}_{q,v} = \lambda S_{q,v} + (1-\lambda)\bar{S}_{\hat{q},v}, \tag{6}$$



(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 3. The similarity statistics of the Top-1 retrieved items in the test sets of the three benchmarks. It is obvious that the items with higher similarity are more likely to be correct retrievals.

where $\bar{S}_{\hat{q},v} \equiv 1$ if v is $\hat{v}_1$, otherwise, $\bar{S}_{\hat{q},v} \equiv S_{\hat{q},v}$, and $\lambda \in [0,1]$ is a hyperparameter to balance the contribution of the raw query and the refined text. To make our approach clearer, we describe the detailed algorithm process of our THI in Algorithm 1. Due to space limitations, all prompt templates can be found in the supplementary material.

Algorithm 1 The interaction process of our THI

Input: The query set $\mathcal{Q}$, the image gallery $\mathcal{V}$, the offline model f_{cross}, the MLLM $\mathcal{M}$, the similarity threshold ξ, the number of interaction rounds K;

Obtain candidate sets $\{\hat{\mathcal{V}}(q_i)\}_{i=1}^{|\mathcal{Q}|}$ for all queries in $\mathcal{Q}$ via Equation (1);
for $k = 1, 2, \cdots, K$ **do**
  for $i = 1, 2, \cdots, |\mathcal{Q}|$ **do**
    Conduct anchor localization via Equation (2) and output the answer of $a_{\hat{v}_k^i}^{q_i}$ based on the k-th candidate image $\hat{v}_k^i$ in $\hat{\mathcal{V}}(q_i)$;
    if $a_{\hat{v}_k^i}^{q_i}$ shows 'Yes', $k = 1$, and $S_{q_i,\hat{v}_1^i} > \xi$ **then**
      Conduct human-centered VQA via Equations (4) and (5) to get the refined query $\hat{q}_i$;
      Compute the re-ranking similarities between query q_i and all images via Equation (6);
    end if
    if $\{a_{\hat{v}_j^i}^{q_i}\}_{j=1}^{k-1}$ all show 'No', $a_{\hat{v}_k^i}^{q_i}$ shows 'Yes', $k > 1$, and $S_{q_i,\hat{v}_1^i} \leq \xi$ **then**
      Conduct human-centered VQA via Equations (4) and (5) to get the refined query $\hat{q}_i$;
      Compute the re-ranking similarities between query q_i and all images via Equation (6);
    end if
  end for
end for
Re-ranking based on similarities;

Output: The new candidate images.

3.3. Reorganization Data Augmentation

Although THI can exploit interactions to provide external knowledge during test time to improve retrieval quality, the cross-modal embedding model is still a bottleneck that limits its further performance improvement. To this end, as shown in Figure 4, we introduce a new Reorganization Data Augmentation (RDA) strategy by enriching, decomposing, and reorganizing person descriptions to improve the discriminability and diversity of training data. We first apply Human-centered VQA in THI to obtain relevant characteristic descriptions and merge them with the original text, thus obtaining a text with richer information, *i.e.*, $\hat{q} = \mathcal{M}(\mathcal{T}_{\text{aggr}}(\mathcal{M}(\mathcal{T}_{\text{vqa}}(\{c_i\}_{i=1}^{N_q}, v)), q))$. Then, we apply MLLMs to decompose enriched text $\hat{q}$ into multiple independent sub-sentences that describe individual attributes without interfering with each other, *i.e.*, $\{\tilde{q}_i\}_{i=1}^{n} = \mathcal{M}(\mathcal{T}_{\text{dec}}(\hat{q}))$, where $\{\tilde{q}_i\}_{i=1}^{n}$ are the n sub-sentences and $\mathcal{T}_{\text{dec}}$ is the prompt template function for text decomposition. The purpose of decomposition is to be able to reorganize the text in a different order later. To increase diversity, we rewrite each sub-sentence into multiple sentences with different styles but the same meaning. For sub-sentence $\tilde{q}$, we conduct rewriting by $\mathcal{R} = \{\breve{q}_j\}_{j=1}^{m} = \mathcal{M}(\mathcal{T}_{\text{rwt}}(\tilde{q}))$, where $\mathcal{R}$ is the set with m rewritten sentences and $\mathcal{T}_{\text{rwt}}$ is the prompt template function for text rewriting.

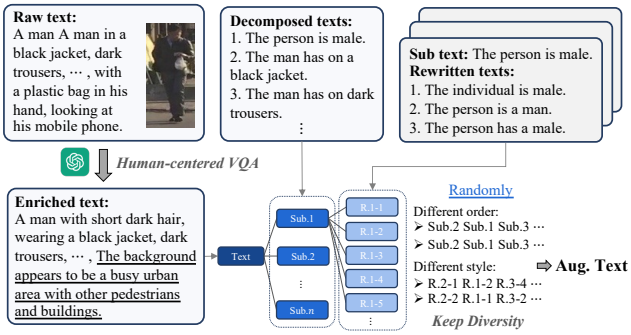


Figure 4. The illustration of our RDA. The purpose of RDA is to supplement more details to the original training texts through human-centered VQA, improving the discriminability of texts. In addition, to enhance diversity, RDA maximizes diversity through the Decomposition-Rewriting-Reorganization strategy.

Finally, for a text q, we can obtain the set $\{\mathcal{R}_i\}_{i=1}^{n}$ that contains a large number of style sub-sentences. So far, we can obtain augmented texts with different sub-sentence orders and different style combinations as shown in Figure 4, which we represent as $\breve{q}$. During training, we can mix the augmentation texts ($\breve{q}$) with the original texts (q) for cross-modal learning, thereby improving model-intra knowledge and generalization. Due to space limitations, more training details can be found in the supplementary material.

4. Experiments

In this section, we conduct extensive experiments to verify the effectiveness, superiority, and generalization of the proposed ICL on four public benchmark datasets.

4.1. Datasets and Evaluation Protocols

Datasets. In our experiments, we use three coarse-grained benchmarks, CHUK-PEDES [16], ICFG-PEDES [5], and RSTPReid [41], and one fine-grained benchmark, UFine6926 [44], to evaluate our ICL. Compared with the coarse-grained benchmark, the fine-grained benchmark has richer texts with fine-grained information. For all datasets, we follow their official settings for data partitioning. More details are provided in the supplementary materials.

Evaluation Protocols. To measure performance, we utilize the widely accepted Rank-K (1,5,10) metrics to mea-

Methods	Image Enc.	Text Enc.	CUHK-PEDES Rank-1	Rank-5	Rank-10	mAP	mINP	ICFG-PEDES Rank-1	Rank-5	Rank-10	mAP	mINP	RSTPReid Rank-1	Rank-5	Rank-10	mAP	mINP
❶ *VL-Backbones w/o ReID-domain pre-training*																	
IVT [26]	ViT-Base	BERT	65.69	85.93	91.15	60.66	-	56.04	73.60	80.22	-	-	46.70	70.00	78.80	-	-
LCR^{2}S [36]	RN50	BERT	67.36	84.19	89.62	59.20	-	57.93	76.08	82.40	38.21	-	54.95	76.65	84.70	40.92	-
CFine [37]	CLIP-ViT	BERT	69.57	85.93	91.15	-	-	60.83	76.55	82.42	-	-	50.55	72.50	81.60	-	-
RaSa [1]	Swin-B	BERT	76.51	90.29	94.25	69.38	-	65.28	80.40	85.12	41.29	-	66.90	86.50	91.35	52.31	-
IRRA [12]	CLIP-ViT	CLIP-X.	73.38	89.93	93.71	66.13	50.24	63.46	80.25	85.82	38.06	7.93	60.20	81.30	88.20	47.17	25.28
TBPS [3]	CLIP-ViT	CLIP-X.	73.54	88.19	92.35	65.38	49.25	65.05	80.34	85.47	39.83	7.87	62.10	81.90	87.75	48.00	25.86
CFAM [44]	CLIP-ViT	CLIP-X.	75.60	90.53	94.36	67.27	-	65.38	81.17	86.35	39.42	-	62.45	83.55	91.10	49.50	-
RDE [21]	CLIP-ViT	CLIP-X.	75.94	90.14	94.12	67.56	51.44	67.68	82.47	87.36	40.06	7.87	65.35	83.95	89.90	50.88	28.08
Our ICL	CLIP-ViT	CLIP-X.	76.41	90.48	94.33	68.04	51.99	68.11	82.59	87.52	40.81	8.18	67.70	86.05	91.75	52.62	29.36
Our ICL⋆	CLIP-ViT	CLIP-X.	77.91	90.27	94.14	69.13	53.40	69.02	82.45	87.36	41.21	8.30	70.55	85.95	91.65	53.68	30.13
❷ *VL-Backbones with ReID-domain pre-training*																	
IRRA$^\flat$ [12]	CLIP-ViT	CLIP-X.	74.05	89.48	93.64	66.57	-	64.37	80.75	86.12	38.85	-	61.90	80.60	89.30	48.08	-
APTM [38]	Swin-B	BERT	76.53	90.04	94.15	66.91	-	68.51	82.99	87.56	41.22	-	67.50	85.70	91.45	52.56	-
NAM$^\natural$ [28]	CLIP-ViT	CLIP-X.	77.47	90.84	94.67	69.43	54.08	66.76	82.02	87.17	41.45	9.53	67.15	86.55	91.90	52.00	28.46
Our ICL	CLIP-ViT	CLIP-X.	78.18	91.63	94.83	69.58	53.48	69.22	83.49	88.06	42.34	9.01	70.00	86.60	91.70	54.16	30.93
Our ICL⋆	CLIP-ViT	CLIP-X.	79.06	91.26	94.72	70.44	54.70	70.05	83.35	87.91	42.70	9.13	72.55	86.60	91.30	55.19	31.72

Table 1. Performance on the three coarse-grained benchmarks. The results with THI are marked with ⋆. Note that IRRA$^\flat$ means using the pre-trained Backbones with MALS [38] and the results of NAM$^\natural$ are reproduced by us.

Methods	Rank-1	Rank-5	Rank-10	mAP	mINP
LGUR [25]	70.69	84.57	89.91	68.93	-
SSAN [5]	75.09	88.63	92.84	73.14	-
IRRA [12]	85.02	94.31	96.75	83.91	77.30
RDE [21]	87.60	95.65	97.46	86.10	79.54
CFAM(B/16) [44]	85.55	94.51	97.02	84.23	-
CFAM(L/14) [44]	88.51	95.58	97.49	87.09	-
❶ **Our ICL**	89.17	96.13	97.88	87.49	81.50
❶ **Our ICL⋆**	90.67	95.98	97.86	88.29	82.60
❷ **Our ICL**	91.02	96.98	98.17	89.76	84.70
❷ **Our ICL⋆**	91.78	96.83	98.16	90.33	85.62

Table 2. Performance comparison on the UFine6926 dataset. The results of IRRA and RDE are reproduced by us.

sure the TIReID performance. Like [12, 21], we also report the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) as auxiliary metrics.

4.2. Implementation Details

To achieve interactive TIReID, we choose RDE [21] and Qwen2-VL-7B-Instruct [29] as the investigated TIReID method and MLLM. For the sake of fairness, we do not modify any settings of RDE, including model architecture (CLIP-ViTB/16 [24]), hyperparameters, and training parameters. To be compatible with the fine-grained benchmark, following [44], the maximum length of the textual tokens of CLIP is set to 168 by interpolating the positional embedding layer with an initial learning rate of 5e-5. As for fine-tuning MLLMs, we use the Llama-Factory framework [40] to conduct SFT with LoRA [11] for 2 epochs. The train batch size per device is set to 4, the gradient accumulation steps are set to 16, the initial learning rate is 5e-5, and the hyperparameters α and r of LoRA are set to 16 and 8, respectively. During inference, the temperature is set to 0.01 to keep reproducibility. Note that all experiments can be completed on two GeForce RTX 24GB 3090 GPUs.

4.3. Comparison with State-of-the-Arts

In this section, to verify the superiority of our ICL, we compare our method with more than 15 baselines including recent advanced methods (*e.g.*, RDE (CVPR'24) [21] and NAM (CVPR'24) [28]). Based on the backbone type, we divide the baselines into two groups (❶ and ❷) as shown Table 1, *i.e.*, the baselines using VL-Backbones w/o and with ReID-domain pre-training. For the group ❷, we use the pre-trained weights released by [28] to initialize CLIP for fair comparison. The results are reported in Tables 1 and 2.

Results on coarse-grained datasets. Table 1 report the results evaluated on the coarse-grained datasets. We can see that our method can achieve competitive performance even without THI. By performing THI, the Rank-1 score of our method is greatly improved, *e.g.*, in the group ❶, the Rank-1 scores on the three datasets are improved by 1.50%, 0.91%, and 2.85%, respectively. In addition, mAP and mINP scores have also improved greatly, which indicates that the overall ranking has improved. In group ❷, our method achieves the best scores on most metrics, especially Rank-1 reached 72.55% on RSTPReid, which is sufficient to verify the superiority. However, we found that THI slightly degraded Rank-5 and Rank-10, which is because we only conducted 5 rounds of interaction. As long as the locating anchor image is incorrect, the errors will accumulate, which can be relieved by increasing the rounds of interaction. But in general, THI can significantly improve Rank-1 and the overall ranking (mAP and mINP).

Results on fine-grained dataset. Table 2 shows the results on the fine-grained dataset whose average text length is over 80. Such kind of fine-grained description often has a clear query intention. For better comparison, we provide the performance of IRRA and RDE on the UFine6926 dataset. From the results, our method can still achieve excellent per-

Methods	Training Sets	CUHK-PEDES					ICFG-PEDES					RSTPReid				
		Rank-1	Rank-5	Rank-10	mAP	mINP	Rank-1	Rank-5	Rank-10	mAP	mINP	Rank-1	Rank-5	Rank-10	mAP	mINP
IRRA [12]	CUHK-PEDES	73.38	89.93	93.71	66.13	50.24	42.41	62.11	69.62	21.77	1.95	53.25	77.15	85.35	39.63	16.60
	ICFG-PEDES	33.48	56.29	66.33	31.56	19.20	63.46	80.25	85.82	38.06	7.93	45.30	69.25	78.80	36.82	18.38
	RSTPReid	32.80	55.26	65.81	30.29	17.61	32.30	49.67	57.80	20.54	3.84	60.20	81.30	88.20	47.17	25.28
RDE [21]	CUHK-PEDES	75.94	90.14	94.12	67.56	51.44	48.18	66.30	73.70	25.00	2.33	54.90	77.50	86.50	41.27	17.84
	ICFG-PEDES	38.11	59.24	68.44	34.16	20.44	67.68	82.47	87.36	40.06	7.87	49.25	72.10	80.20	38.46	18.33
	RSTPReid	36.94	58.22	67.58	33.65	20.42	42.17	58.32	65.49	26.37	4.94	65.35	83.95	89.90	50.88	28.08
Our ICL	CUHK-PEDES	76.41	90.48	94.33	68.04	51.99	48.57	66.66	73.75	25.30	2.40	55.80	79.60	87.65	42.09	17.41
	ICFG-PEDES	42.87	64.20	73.44	38.19	23.58	68.11	82.59	87.52	40.81	8.18	52.50	75.05	83.00	41.82	21.14
	RSTPReid	41.31	61.86	70.31	36.78	22.37	45.93	62.70	68.80	28.89	5.63	67.70	86.05	91.75	52.62	29.36
Our ICL⋆	CUHK-PEDES	77.91	90.27	94.14	69.13	53.40	52.80	66.49	73.49	25.60	2.44	61.30	79.25	87.40	43.42	18.01
	ICFG-PEDES	49.29	64.34	73.55	40.82	25.38	69.02	82.45	87.36	41.21	8.30	60.15	75.30	83.15	43.72	22.04
	RSTPReid	47.35	61.45	70.34	38.91	23.68	50.52	61.56	68.57	29.26	5.73	70.55	85.95	91.65	53.68	30.13

Table 3. Comparison of mutual generalization capabilities between coarse-grained datasets.

Source → Target	Methods	Rank-1	Rank-5	Rank-10	mAP	mINP
CUHK. → UFine.	IRRA [12]	37.51	54.92	64.29	40.76	34.33
	RDE [21]	40.37	57.49	66.05	42.68	35.78
	Our ICL	46.40	63.55	72.08	48.68	41.56
	Our ICL⋆	**57.76**	**64.13**	**72.81**	**53.97**	**45.64**
ICFG. → UFine.	IRRA [12]	15.02	26.79	33.90	17.10	12.75
	RDE [21]	17.86	31.01	38.56	19.82	14.74
	Our ICL	27.95	44.20	52.20	29.85	23.20
	Our ICL⋆	**36.81**	**44.65**	**52.73**	**34.12**	**26.61**
RSTP. → UFine.	IRRA [12]	13.21	25.67	33.93	15.60	11.09
	RDE [21]	14.00	25.23	32.64	16.22	11.90
	Our ICL	23.89	38.30	46.70	25.54	19.20
	Our ICL⋆	**31.23**	**38.56**	**47.02**	**28.90**	**21.80**
UFine. → CUHK	IRRA [12]	37.74	60.12	70.13	35.94	23.21
	RDE [21]	39.41	61.14	70.11	36.49	23.32
	Our ICL	49.04	70.27	**78.64**	44.54	29.58
	Our ICL⋆	**56.87**	**70.19**	78.53	**47.31**	**31.20**
UFine. → ICFG.	IRRA [12]	34.52	55.41	64.44	17.96	1.95
	RDE [21]	40.37	60.14	68.41	20.54	2.19
	Our ICL	43.10	**62.92**	**70.73**	22.73	2.56
	Our ICL⋆	**47.83**	62.67	70.48	**23.16**	**2.62**
UFine. → RSTP.	IRRA [12]	37.65	63.70	73.00	29.00	11.80
	RDE [21]	39.90	63.50	74.75	29.92	12.43
	Our ICL	48.85	**72.65**	**81.80**	36.91	16.39
	Our ICL⋆	**55.35**	72.40	81.50	**38.64**	**17.23**

Table 4. Generalization capabilities between coarse-grained and fine-grained datasets. The best scores in each task are in **bold**.

formance, with Rank-1 exceeding 91%. This shows that interaction is also applicable to the fine-grained scenario.

4.4. Generalization Study

To evaluate the generalization, Tables 3 and 4 report the cross-domain performance on four datasets, including the coarse-to-fine, coarse-to-fine, and fine-to-coarse generalization experiments. From the results, our method achieves better generalization than RDE even without THI thanks to the training enhancement of RDA. When THI is performed, the cross-domain performance is dramatically improved, for example, from CUHK-PEDES to RSTPReid, THI brings an improvement of more than 4% on Rank-1. This shows that THI is a potential solution to generalization challenges in the future. From the generalization experiments between fine-grained and coarse-grained datasets shown in Table 4, ICL can also achieve the best cross-domain performance, *e.g.*, compared with the best baseline RDE, from UFine6926 domain to CUHK-PEDES domain, our method improves Rank-1 and mAP by 17.49% and 10.86%, respectively, which further verifies the cross-domain generalization of our method.

Methods	THI	CUHK-PEDES		ICFG-PEDES		RSTPReid		ΔAvg
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	
CLIP [24]	✘	71.64	63.92	60.11	34.52	56.55	44.52	**+2.57**
	✔	73.77	65.66	63.57	34.95	61.45	46.25	
IRRA [12]	✘	73.38	66.13	63.46	38.06	60.20	47.17	**+1.86**
	✔	76.06	67.42	65.26	38.58	63.75	48.47	
RDE [21]	✘	75.94	67.56	67.68	40.06	65.35	50.88	**+1.41**
	✔	77.47	68.62	68.72	40.63	68.45	52.01	

Table 5. Transferability results on three coarse-grained benchmarks. ΔAvg represents the average improvement.

4.5. Transferability Study

Our interactive module is separate and independent from the training of existing TIReID methods, so it can be plug-and-play with existing methods to improve ReID performance. To verify the transferability of our THI, we conduct experiments on multiple baselines, and the results are shown in Table 5. Except for the CLIP results we reproduced, all other experiments used public pre-trained model weights. From the results of Table 5, the interactive strategy application can significantly improve Rank-1 and mAP, which shows that the external guidance by interactions via MLLMs can further clarify the text-image alignments and improve the overall ranking. The average improvements on the three datasets are 2.57%, 1.86%, and 1.41%, respectively, which seriously proves the transferability of our THI.

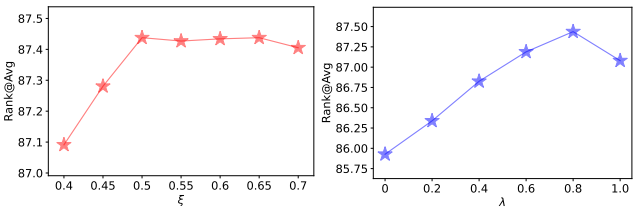
4.6. Ablation and Parameter Study

To explore the effects of each proposed component, *i.e.*, THI and RDA, we first conduct the ablation study on three

coarse-grained datasets as reported in Table 6. From the results, each component can bring performance gains on Rank-1 and mAP, which verifies the reliability and rationality of the method. Especially the introduction of THI has greatly improved the Rank-1 accuracy from 68.95% to 70.55% on the RSTPReid dataset. In addition, our RDA also brings considerable performance gains, especially on the RSTPreid dataset. Also, we conduct the parameter analysis on the CHUK-PEDES dataset on two free hyperparameters, *i.e.*, the similarity threshold ξ and the balance factor λ. The former filters unnecessary retrievals to improve the interaction efficiency, while the latter controls the contribution of refined texts. Based on Figure 5, in our experiments, we set ξ in the range of $0.5 \sim 0.6$ and λ to 0.8, thus mitigating the risk of introducing noisy external knowledge.

No.	THI	RDA	LoRA	CUHK-PEDES Rank-1	CUHK-PEDES mAP	ICFG-PEDES Rank-1	ICFG-PEDES mAP	RSTPReid Rank-1	RSTPReid mAP
#1	✔	✔	✔	**77.91**	**69.32**	**69.02**	**41.21**	**70.55**	**53.68**
#2	✔	✔	✘	76.38	68.59	67.92	41.13	69.00	53.11
#3	✘	✔	✘	76.41	68.04	68.11	40.81	67.70	52.62
#4	✘	✘	✘	75.94	67.56	67.68	40.06	65.35	50.88

Table 6. Ablation studies on CHUK-PEDES, ICFG-PEDES, and RSTPReid datasets. The best scores are in **bold**.

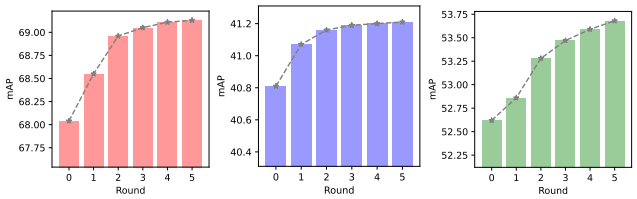


(a) The similarity threshold ξ (b) The balance factor λ

Figure 5. Variation of performance with different ξ and λ.

4.7. Interactive Study

This section further explores the interactive module, *i.e.* THI. We report the performance (mAP) changes after multiple rounds of interactions in Figure 6. The mAP score can reflect the overall retrieval quality. From the results, the performance improvement is obvious in the first few rounds ($\leq$ 2) since the items that meet the query semantics are mostly concentrated at the top of the ranking. However, after $>$ 2 rounds, the performance gain is gradually not obvious as the number of queries requiring interaction decreases. But generally, as the number of rounds increases, the overall performance gradually improves. In all our experiments, we performed 5 rounds of interaction. In addition, we visualize the top-10 retrieved results before and after applying THI in Figure 7. Due to the inherent limitations of the intra-model knowledge, it is difficult to obtain satisfactory results without the help of THI. In contrast, our THI can dynamically enrich queries by interacting with MLLMs to achieve a more reliable ranking. More example results are given in the supplementary material.



(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 6. Performance (mAP) versus rounds on three datasets. Round 0 indicates the setting without using THI.

Figure 7. Top-10 retrieved results on CUHK-PEDES dataset between ICL (the first row) and ICL with THI (the second row).

5. Conclusion

In this paper, we explore interactive text-to-image person re-identification, which aims to improve the alignment between dynamic queries and challenging candidate images by leveraging external guidance from MLLMs. To achieve this, we develop an Interactive Cross-modal Learning (ICL) framework to alleviate the inherent challenges of offline models and training data by, including a plug-and-play Test-time Human-centered Interaction (THI) module and Reorganization Data Augmentation (RDA). Extensive experiments and analysis show that our framework can effectively transfer external knowledge in MLLMs into offline models for guiding re-identification, showing excellent performance and generalization.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grants 62472295, 62372315, 62176171, and U21B2040; in part by Sichuan Science and Technology Planning Projects (2024YFHZ0089, 2024NSFTD0049, 2024YFHZ0144, 2024NSFTD0047, 2024NSFTD0038); in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403 and CJ202303; in part by Chengdu Science and Technology Project under Grant 2023-XT00-00004-GX.

References

[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 555–563, 2023. 2, 6

[2] Maryam Bukhari, Sadaf Yasmin, Sheneela Naz, Muazzam Maqsood, Jehyeok Rew, and Seungmin Rho. Language and vision based person re-identification for surveillance systems using deep learning with lip layers. *Image and Vision Computing*, 132:104658, 2023. 1

[3] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 465–473, 2024. 1, 2, 6

[4] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022.

[5] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. 1, 5, 6

[6] Hiren Galiyawala and Mehul S Raval. Person retrieval in surveillance using textual query: a review. *Multimedia Tools and Applications*, 80(18):27343–27383, 2021. 1

[7] Daming Gao, Yang Bai, Min Cao, Hao Dou, Mang Ye, and Min Zhang. Semi-supervised text-based person search. *arXiv preprint arXiv:2404.18106*, 2024. 1

[8] Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. Merlin: Multimodal embedding refinement via llm-based iterative navigation for text-video retrieval-rerank pipeline. *arXiv preprint arXiv:2407.12508*, 2024. 2, 3

[9] Chen He, Shenshen Li, Zheng Wang, Hua Chen, Fumin Shen, and Xing Xu. Chatting with interactive memory for text-based person retrieval. *Multimedia Systems*, 31(1):31, 2025. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4, 6

[12] Ding Jiang and Mang Ye. 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*, pages 2787–2797, 2023. 1, 2, 6, 7

[13] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[16] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017. 5

[17] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *Forty-first International Conference on Machine Learning*. 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[19] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 1

[20] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in neural information processing systems*, 36:24829–24840, 2023. 1

[21] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 1, 2, 6, 7

[22] Yalan Qin, Chuan Qin, Xinpeng Zhang, and Guorui Feng. Dual consensus anchor learning for fast multi-view clustering. *IEEE Transactions on Image Processing*, 2024. 1

[23] Yalan Qin, Nan Pu, Nicu Sebe, and Guorui Feng. Latent space learning based ensemble clustering. *IEEE Transactions on Image Processing*, 2025. 1

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6, 7

[25] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 6

[26] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer, 2022. 6

[27] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 1

[28] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17127–17137, 2024. 1, 2, 3, 6

[29] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 6

[30] Yansong Wang, Yundong Sun, Yansheng Fu, Dongjie Zhu, and Zhaoshuo Tian. Spectrum-bert: pre-training of deep bidirectional transformers for spectral classification of chinese liquors. *IEEE Transactions on Instrumentation and Measurement*, 2024. 2

[31] Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. Multimodal llm enhanced cross-lingual cross-modal retrieval. *arXiv preprint arXiv:2409.19961*, 2024. 3

[32] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Demo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1

[33] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 402–420. Springer, 2020. 1, 2

[34] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022. 1

[35] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021. 2

[36] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211, 2023. 6

[37] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 2023. 1, 6

[38] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4492–4501, 2023. 1, 2, 6

[39] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17117–17126, 2024. 1

[40] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. 6

[41] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021. 1, 2, 5

[42] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning, 2023. 1

[43] Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. *arXiv preprint arXiv:2409.18569*, 2024. 1

[44] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024. 1, 5, 6