Instruct-ReID++: Towards Universal Purpose Instruction-Guided Person Re-identification

Weizhen He*, Yiheng Deng*, Yunfeng Yan, Feng Zhu, Yizhou Wang, Lei Bai, Qingsong Xie, Donglian Qi, Wanli Ouyang, *Senior Member, IEEE*, Shixiang Tang✉

***Abstract*—Human intelligence can retrieve any person according to both visual and language descriptions. However, the current computer vision community studies specific person re-identification (ReID) tasks in different scenarios separately, which limits the applications in the real world. This paper strives to resolve this problem by proposing a novel instruct-ReID task that requires the model to retrieve images according to the given image or language instructions. Instruct-ReID is the first exploration of a general ReID setting, where existing 6 ReID tasks can be viewed as special cases by assigning different instructions. To facilitate research in this new instruct-ReID task, we propose a large-scale OmniReID++ benchmark equipped with diverse data and comprehensive evaluation methods *e.g.,* task-specific and task-free evaluation settings. In the task-specific evaluation setting, gallery sets are categorized according to specific ReID tasks. We propose a novel baseline model, IRM, with an adaptive triplet loss to handle various retrieval tasks within a unified framework. For task-free evaluation setting, where target person images are retrieved from task-agnostic gallery sets, we further propose a new method called IRM++ with novel memory bank-assisted learning. Extensive evaluations of IRM and IRM++ on OmniReID++ benchmark demonstrate the superiority of our proposed methods, achieving state-of-the-art performance on 10 test sets. The datasets, the model, and the code will be available at https://github.com/hwz-zju/Instruct-ReID.**

***Index Terms*—Person Re-identification, Multitask Person Retrieval, Benchmark, General Foundation Model**

I. Introduction

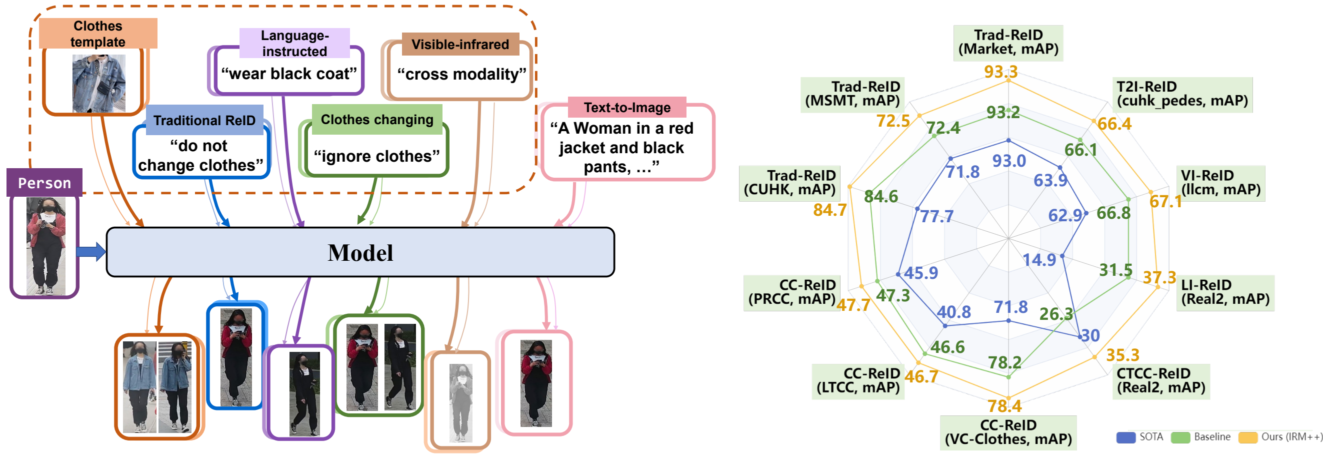
PERSON re-identification (ReID) aims to retrieve the target person from surveillance videos or images across locations and time. This task is challenging due to various viewpoints, illumination changes, unconstrained poses, occlusions, heterogeneous modalities, background clutter, and more. Therefore, recent studies in person re-identification carefully designed task settings and developed state-of-the-art models to tackle every specific scenario, such as [1–5] for traditional ReID (Trad-ReID), [6–10] for clothes-changing ReID (CC-ReID), [11, 12] for clothes template based clothes-changing ReID (CTCC-ReID), [13–17] for visible-infrared ReID (VI-ReID) and [18–22] for text-to-image ReID (T2I-ReID).

Despite the success of these person re-identification methods, they still have inherent limitations in practical applications due to their focus on specific and fixed scenarios. First, customers must deploy distinct models to retrieve persons according to the person retrieval tasks, *e.g.,* 5 models for 5 re-identification settings, significantly increasing the model training and deployment cost. Second, focusing on a single scenario task limits the model performance. ReID models utilize biological appearance cues to retrieve images with the same identity as the query image. A model trained across diverse task scenarios is expected to have a deeper understanding of person re-identification, *e.g.,* the traditional ReID task requires the model to possess the ability to grasp characteristics such as body posture and appearance to compare identities, and clothes-changing ReID can assist the model in extracting non-clothing information for identity determination. Other tasks, *e.g.,* text-to-image ReID and visible-infrared ReID, also help the model learn to extract human identity information across different data modalities. The methods designed for specific scenarios are limited to their particular training data and ignore the benefits of other ReID tasks, leading to limited model performance.

The limitations of addressing each ReID task independently motivate us to design a general approach towards distinct ReID tasks. In this paper, we propose a novel multi-purpose *instruct-ReID* task where **6** existing ReID settings, *i.e.,* Trad-ReID, CC-ReID, CTCC-ReID, VI-ReID, T2I-ReID, and language-instructed ReID (LI-ReID) can be formulated as its special cases. This unified ReID task involves training various ReID tasks using a single model, thereby minimizing training costs and fostering mutual benefits among diverse ReID tasks. Specifically, the instruct-ReID task takes query images and multimodality instructions as the model inputs and requires the model to retrieve the same identity images from the gallery following the instructions. As shown in Fig. 1(a), with certain language or image instructions, the instruct-ReID task can be specialized to existing ReID tasks. For example, the clothes-changing ReID can be viewed as using the instruction "Ignore clothes" to retrieve and the traditional ReID can be seen as utilizing the instruction "Do not change clothes" to retrieve images of individuals wearing the same clothing. As another example, clothes template-based clothes-changing ReID can utilize a clothes template image as instruction. The proposed instruct-ReID task offers three significant advantages: easy deployment, improved performance, and easy extension to new ReID tasks. First, it enables cost-effective and convenient deployment in real-world applications. Unlike existing ReID approaches limited to specific tasks, instruct-ReID allows for utilizing a single model for various ReID scenarios, which

*Weizhen He and Yiheng Deng contributed equally. ✉Shixiang Tang is the corresponding author. This work was done when Weizhen He was an intern at SenseTime. Weizhen He, Yiheng Deng, Donglian Qi and Yunfeng Yan are with the College of Electrical Engineering, Zhejiang University, Hangzhou, 310027, China. Feng Zhu is with SenseTime Group Limited, China. Shixiang Tang, Yizhou Wang and Wanli Ouyang are with Shanghai AI Laboratory, Shanghai, 200232, China. Qingsong Xie is with Shanghai Jiao Tong University, Shanghai, 200240, China.

Manuscript received April 19, 2021; revised August 16, 2021.



(a) One model for diverse ReID **(b) Improved Performance by OmniReID++ benchmark**

Fig. 1. We proposed a new instruct-ReID task that unites various ReID tasks. *Traditional ReID:* The instruction may be "Do not change clothes". *Clothes-changing ReID:* The instruction may be "Ignore clothes". *Clothes template based clothes-changing ReID:* The instruction is a cropped clothes image and the model should retrieve the same person wearing the provided clothing. *Language-instructed ReID:* The instruction is several sentences describing pedestrian attributes. The model is required to retrieve the person described by the instruction. *Visible-Infrared ReID:* The instruction can be "Cross modality". *Text-to-image ReID:* The model retrieves images according to the description sentence.

is more practical in real applications. Second, as all existing ReID tasks can be considered special cases of instruct-ReID, we can unify the training sets of these tasks to exploit the benefits of more data and diverse annotations across various tasks – leading to enhanced performance. Third, instruct-ReID introduces a new language-instructed ReID task which requires the model to retrieve persons following language instructions. This setting is practical for real-world applications, enabling customers to retrieve specific images through language descriptions. For example, customers can retrieve a woman wearing a black coat using one of her pictures and language instructions "Wear black coat".

To facilitate research in Instruct-ReID task, we introduce a new benchmark called OmniReID, derived from 12 datasets [1] representing 6 distinct ReID tasks as described in our previous conference paper. In this paper, we extend OmniReID to OmniReID++, which has extra training data with 99,174 images, 5,221 identities, and 128,413 instructions. Compared to the OmniReID benchmark, the OmniReID++ has the following three improvements:

- The OmniReID++ benchmark emphasizes richer ***diversity*** instruction annotations by incorporating images from a broader range of domains, which extends the scenarios from surveillance and synthetic games to include scenes from movies and internet videos. As shown in Tab. I, OmniReID++ benchmark supplements instructions from winter environments and provides annotations for personnel across seasons and years. OmniReID++ expands the annotated scenes to include mountains, rivers, parks, and movie scenes. The diversity ensures that the trained models are robust and can effectively handle ReID tasks in various real-world scenarios.
- The OmniReID++ benchmark achieves improved ***comprehensiveness*** by offering evaluation datasets to assess various ReID tasks. Specifically, in this paper, we present two evaluation settings along with the large-scale OmniReID++ benchmark, which facilitates evaluating the generalization ability of diverse ReID methods and can be outlined as follows:
  <u>*Task-specific evaluation setting.*</u> In existing benchmarks, the test datasets are designed for specific task scenarios, *e.g.,* Market1501 [23] for Trad-ReID and PRCC [24] for CC-ReID, evaluating the performance of the model on the corresponding task. The OmniReID++ adopts this approach as a task-specific evaluation setting and utilizes 10 publicly available test sets covering 6 ReID tasks. As shown in Fig. 2(a), both images in the query and gallery are associated with a task-specific instruction description, the model is evaluated separately on each task dataset and requires repeated testing with multiple datasets to evaluate the performance on multiple tasks.
  <u>*Task-free evaluation setting.*</u> When evaluating the generalization ability of the model across multiple retrieval tasks jointly, task-specific instructions for each gallery image are discarded. Therefore, we propose a novel task-free evaluation setting as an extension of our previous conference paper. As shown in Fig 2 (b), the test set maintains a consistent gallery input for all tasks. Each image in the gallery set is task-agnostic and only the image feature is extracted for various ReID tasks. The gallery features can be fixed and extracted only once in applications, thus saving more computational resources.
- The OmniReID++ benchmark proposes enhanced ***exhaustiveness*** metric by introducing a novel ***mAP***$\mathcal{T}$, which evaluates both the correctness of identity and the consistency with the instructions. The mAP$\mathcal{T}$ provides a measurement to evaluate the alignment of retrieval results with the finer-grained retrieval intentions, which is different from the metrics in previous ReID tasks, *e.g.,*

[1]The discussion on ethical risks is provided in supplementary materials.

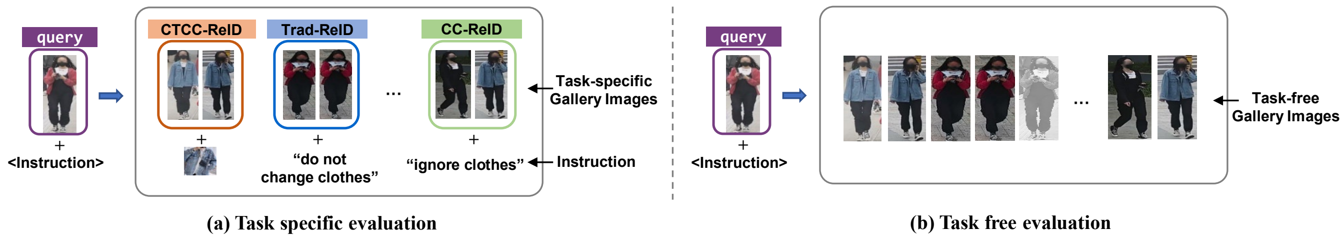


Fig. 2. The comparison between task-specific and task-free evaluation settings. Task-free evaluation setting offers a more uniformly flexible evaluation approach.

TABLE I
COMPARISON OF OMNIREID AND OMNIREID++. CONSIDERING THE ASPECTS OF INSTRUCTION SEASON, TIME SPAN, AND SCENE DIVERSITY, OMNIREID++ OFFERS A RICHER VARIETY COMPARED TO OMNIREID.

Dataset Feature	Attributes	OmniReID	OmniReID++
Season	Spring	✓	✓
	Summer	✓	✓
	Autumn	✓	✓
	Winter		✓
Time Span	Cross day	✓	✓
	Cross month	✓	✓
	Cross season		✓
	Cross year		✓
Scene	Indoor	✓	✓
	Street	✓	✓
	Mountain		✓
	River		✓
	Park		✓
	Movie		✓
	Synthetic	✓	✓

rank-1 accuracy (R1) and mean average precision (mAP), that only measure identity correctness of retrieval images. Specifically, retrieval results with both the correct identity and its similarity with the given instruction exceeding the threshold τ will be considered as the correct retrieval, and the mean average precision is computed based on the refined rank list.

To address the instruct-ReID task, we present an Instruct-ReID Model (IRM) in our previous conference version that can, for the first time, handle the major 6 ReID tasks listed in Fig 1 (a) with a dual-branch framework. Unlike the typical triplet loss [25] that only defines positive/negative pairs by identities, our IRM incorporates a novel adaptive triplet loss to learn a metric space that preserves identity and instruction similarities. Specifically, we design an adaptive margin between two query-instruction pairs based on instruction similarities to pull features with similar instructions close and push features with different instructions apart. Although the previous IRM has achieved significant success in unifying the multiple ReID tasks within one framework, it can not be applied in the newly proposed task-free evaluation setting for the following two reasons. First, the gallery features of various ReID tasks are extracted from different modules of IRM, *e.g.,* Trad-ReID from the output of the attention module and T2I-ReID from the image encoder. However, in the task-free evaluation setting, all images in the gallery set are task-agnostic, making it impractical to extract retrieval features from different model parts based on the task type of gallery images. Second, IRM extracts gallery images and specific instruction features for retrieval comparison, while these task-specific instructions are unavailable in the task-free evaluation setting. To address the challenges, we introduce a new method called IRM++, which employs simple and unified gallery features for various ReID tasks. Concretely, unlike IRM, IRM++ obtains the gallery features by only extracting the image features from diverse tasks using an image encoder, without introducing task-specific multimodality instruction inputs. Furthermore, IRM++ introduces two memory banks in a contrastive learning manner to effectively supervise the model learning. Unlike triplet loss-based methods that include only one negative sample per triplet, memory bank contrastive learning offers multiple negative samples for each iteration, enhancing the discriminative representation learning of the data.

We validate the effectiveness of IRM and IRM++ using ReID-specific pretraining [26] and OmniReID++ benchmark training data without finetuning, demonstrating consistent improvement over previous models across 10 datasets covering 6 ReID tasks. For task-specific evaluation setting, IRM improves **+7.7%**, **+0.6%**, **+0.5%** mAP on CUHK03, MSMT17, Market1501 for traditional ReID, **+6.4%**, **+11.2%**, **+7.1%** mAP on PRCC, LTCC, VC-Clothes for clothes-changing ReID when using RGB images only, **+11.7%** mAP on COCAS+ real2 for clothes template based clothes changing ReID, **+4.3%** mAP on LLCM for visible-infrared ReID, **+2.6%** mAP on CUHK-PEDES for text-to-image ReID, **+24.9%** mAP on COCAS+ real2 for the new language-instructed ReID. For task-free evaluation setting, IRM++ achieves improvements of **+0.3%**, **+0.7%**, and **+7.0%** mAP on Market1501, MSMT17, and CUHK03 for traditional ReID tasks, **+1.8%**, **+6.6%**, and **+5.9%** mAP on PRCC, VC-Clothes, and LTCC for clothes-changing ReID. Additionally, for clothes template-based clothes-changing ReID and our newly introduced language-instructed ReID task, we observe a **+5.3%** mAP and a **+22.4%** mAP enhancement on COCAS+ real2 dataset. Furthermore, our model achieves **+4.2%** mAP on LLCM for visible-infrared ReID and **+2.5%** mAP on CUHK-PEDES for text-to-image ReID.

In summary, the contributions of this paper are three folds. (1) We propose a new instruct-ReID task, where existing traditional ReID, clothes-changing ReID, clothes template-based clothes-changing ReID, visible-infrared ReID, text-to-image ReID, and language-instructed ReID can be viewed as special cases. (2) To facilitate research on instruct-ReID, we establish a large-scale and comprehensive benchmark: OmniReID++, covering the existing 6 ReID tasks. (3) We propose an adaptive triplet loss in IRM to supervise the feature distance of two query-instruction pairs to consider identity and instruction alignments. We utilize memory bank assisted learning in IRM++ to introduce powerful supervision in model training, obtaining better performance for the task-free evaluation setting.

Difference from the conference paper. A preliminary version of this paper is presented in the Instruct-ReID Model (IRM) [27]. This paper extends the previous study with four major improvements. 1) We extend OmniReID to OmniReID++, leveraging more diverse data sources. This expansion places a wider variety of domains, ranging from surveillance and synthetic games to movies and internet videos. This broader dataset ensures that the models trained on OmniReID++ are more adaptable to real-world scenarios. 2) We propose a new task-free evaluation setting that is more aligned with real-world inference scenarios. The novel evaluation setting unifies the gallery retrieval features of various ReID tasks and avoids introducing task-specific instructions for gallery images, which addresses the requirements of retrieving from task-agnostic gallery images and further enhances the comprehensiveness of the evaluation methodology. 3) We introduce a novel evaluation metric, mAPτ, to provide a more precise assessment of whether retrieval results meet the retrieval target. The novel mAPτ measures both the correctness of identity and the consistency of retrieval results with the instructions in the rank list, thereby integrating a finer discrimination method beyond identity criteria. 4) We propose IRM++ with memory bank assisted learning for instruct-ReID task and conduct additional experiments to evaluate the effectiveness of the IRM++. Experiments on the task-free evaluation setting show that IRM++ achieves significant improvements based on OmniReID++ benchmark, *e.g.,* **+9.0%** and **+5.8%** mAP on CTCC-ReID and LI-ReID tasks compared to IRM.

II. Related Work

A. Person Re-identification

Person re-identification aims to retrieve the same images of the same identity with the given query from the gallery set. To support the ReID task on all-weather application, various tasks are conducted on the scenarios with changing environments, perspectives, and poses [2, 19, 20, 28, 29]. Traditional ReID mainly focuses on dealing with indoor/outdoor problems when the target person wears the same clothes. Recently, to extend the application scenarios, clothes-changing ReID (CC-ReID) [11] and clothes template based clothes-changing ReID (CTCC-ReID) [12] are proposed. While CC-ReID forces the model to learn clothes-invariant features [6, 7], CCTC-ReID further extracts clothes template features [12] to retrieve the image of the person wearing template clothes. To capture person's information under low-light environments, visible-infrared person ReID (VI-ReID) methods [14–16] retrieve the visible (infrared) images according to the corresponding infrared (visible) images. In the absence of the query image, GNA-RNN [21] introduced the text-to-image ReID (T2I-ReID) task, which aims at retrieving the person from the textual description. However, existing tasks and methods focus on a single scenario, making it difficult to address the demands of cross-scenario tasks. In this paper, we introduce a new Instruct-ReID task, which can be viewed as a superset of the existing ReID tasks by incorporating instruction information into identification.

B. Loss Function in ReID

Mulitple loss function have been extensively studied during the development of the ReID research [28]. Depending on the structure of the method design, there are three typical types of loss function: **(1) Identity loss** proposed in [30–35] treats each identity as a distinct class, thus turning the training phase of ReID as a classification process. Identity loss using cross-entropy loss to compute the difference between the labels and the predicted probability. It's convenient for training but easy to overfit when the labels are over-confident annotated [36]. **(2) Contrastive loss** [37–40] has been widely used in diverse ReID scenarios. It utilizes Euclidean distance to compute the pairwise distance of two input sample features. Contrastive loss encourages features of the same identities to come closer together while pushing features from different identities farther apart. The Contrastive loss can alleviate the impact of wrong classifications and noise during the clustering stage. Although contrastive loss can be sensitive to the choice of positive and negative samples during training, it faces the challenge of lacking a reference distance between samples in ReID task applications. **(3) Triplet loss** [41–44] is also an extensively explored feature clustering method. It uses a retrieval ranking perspective to address the training process of the ReID model [28]. Triplet loss takes one anchor sample, one positive sample, and one negative sample as input; based on the assumption that the distance between the anchor and the positive sample should be smaller than the distance between the anchor and the negative sample by a predefined margin, triplet loss optimizes the distance within the triplet samples during the training phase.

Different loss functions are often combined to achieve better performance. For instance, a common approach involves linearly combining the identity loss with the triplet loss [45]. In our paper, we employ identity loss, contrastive loss, and triplet loss to provide better optimization for instruct-ReID task. While triplet loss has achieved success in traditional ReID tasks, it fails to distinguish the two positive samples with distinct instructions. Different from previous methods using the vanilla triplet loss with a fixed margin, we proposed an adaptive triplet loss with a variable margin that takes the similarity of sample instructions into consideration. Our approach offers a new solution to the instruction-based retrieval task.

C. Multimodal Retrieval

Multimodel retrieval is a widely used approach that aligns information from multiple modalities to broaden the scope of model applications. In multimodel retrieval, unimodal encoders always encode different modalities for retrieval tasks. For instance, CLIP [46], VideoCLIP [47], COOT [48] and MMV [49] utilize contrastive learning to align the multimodal features for pre-training. Other techniques like HERO [50], Clipbert [51], Vlm [52], bridgeformer [53] and UniVL [54] focus on merging different modalities for retrieval tasks to learn a generic representation. Although there have been numerous studies on multimodal retrieval, most are concentrated on language-vision pretraining or video retrieval, leaving the potential of multimodal retrieval for person ReID largely unexplored. This paper aims to investigate this underexplored area to retrieve anyone with information extracted from multiple modalities.

D. Memory Bank

Memory bank is a non-parametric buffer that stores the key features during training. Memory bank and memory-based learning have been widely explored in supervised learning [55], semi-supervised and unsupervised learning [56, 57]. [56] introduced a memory bank in instance discrimination pretext task, storing the representations of all samples in the dataset, providing sufficient negative samples to the feature clustering. It has also been used in various tasks, such as few-shot video classification [58], semantic segmentation [59], video deblurring [60] and so on. MAUM [13] adopts memory bank in cross-modality person re-identification by leveraging model drift to provide hard positive references, thereby improving robustness against modality discrepancy and imbalance. In IRM++, we introduce a dual-branch memory bank to store the features for both instructions and images for each identity. Different from former MAUM using memory bank for better cross-modality association, IRM++ utilizes memory bank to enhance the model's capacity for instruction following.

III. OmniReID++ Benchmark

To facilitate research on the Instruct-ReID task, we propose the OmniReID++ benchmark including a large-scale pretraining dataset based on 13 publicly available datasets with visual and language annotations. The comparison with the existing ReID benchmark is illustrated in Tab. II. The OmniReID++ has a larger dataset compared to existing benchmarks and the previous version of OmniReID benchmark. While MALS [61] has more identities, its data is synthetic, and our OmniReID++ dataset shows greater diversity.

TABLE II
Comparison of training subsets of different ReID datasets.

dataset	image	ID	domain
MSMT17 [3]	30,248	1,041	indoor/outdoor
Market1501 [23]	12,936	751	outdoor
PRCC [24]	17,896	150	indoor
COCAS+ Real1 [12]	34,469	2,800	indoor/outdoor
LLCM [62]	30,921	713	indoor/outdoor
LaST [63]	71,248	5,000	indoor/outdoor
MALS [61]	1,510,330	1,510,330	synthesis
LUPerson-T [64]	957,606	-	indoor/outdoor
OmniReID++	5,072,218	333,825	indoor/outdoor/synthetic

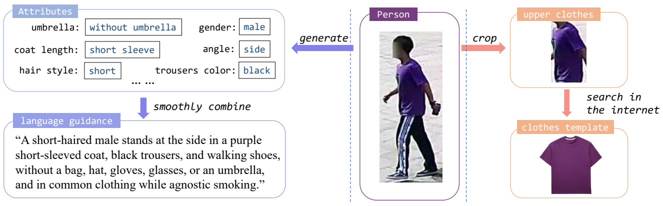
Data Collection. To enable all-purpose person ReID, we collect massive public datasets from various domains and use their training subset as our training subset, including Market1501 [23], MSMT17 [3], CUHK03 [65] from traditional ReID, PRCC [24], VC-Clothes [66], LTCC [67], LaST [63], NKUP [68] and NKUP+ [69] from clothes-changing ReID, LLCM [62] from visible-infrared ReID, CUHK-PEDES [21], SYNTH-PEDES [70] from text-to-image ReID, COCAS+ Real1 [62] from clothes template based clothes-changing ReID and language-instructed ReID, forming 5,072,218 images and 333,825 identities. To fairly compare our method with state-of-the-art methods, the trained models are evaluated on LTCC, PRCC, VC-Clthoes, COCAS+ Real2, LLCM, CUHK-PEDES, Market1501, MSMT17, and CUHK03 test subsets without fine-tuning. We present all dataset statistics of OmniReID++ in the supplementary materials.

Language Annotation Generation. To generate language instruction for language-instructed ReID, we annotate COCAS+ Real1, LTCC, PRCC, LaST, NKUP, NKUP+ and COCAS+ Real2 with language description labels. Similar to Text-to-Image ReID datasets, *e.g.,* CUHK-PEDES [21], language annotations in OmniReID++ are several sentences that describe the visual appearance of pedestrians. We divide our annotation process into *pedestrian attribute generation* and *attribution-to-language transformation*.

Pedestrian Attribute Generation. To obtain a comprehensive and varied description of an individual, we employ an extensive collection of attribute words describing a wide range of human visual characteristics. The collection contains 20 attributes and 92 specific representation words, including full-body clothing, hair color, hairstyle, gender, posture, and accessories such as umbrellas or satchels. Professional annotators manually label all the pedestrian attributes. We provide a practical illustration with the attribute collection in Fig. 3(a). By utilizing instructions on the well-defined attribute combination, models can further enhance ability to identify the target person.

Attribute-to-Language Transformation. Compared with discrete attribute words, language is more natural for consumers. To this end, we transform these attributes into multiple sentences using the Alpaca-LoRA [71] large language model. Specifically, we ask the Alpaca-LoRA with the following sentences: "Generate sentences to describe a person. The above sentences should contain all the attribute information I gave you in the following." The generated annotations are carefully checked and corrected manually to ensure the correctness of the language instructions. All detailed pedestrian attributes and more language annotations are presented in the supplementary materials.

Visual Annotation Generation. Visual annotations are images that describe the characteristics of pedestrians. In this paper, we select clothes as the visual annotations because they are viewed as the most significant visual characteristics of pedestrians. To get high-quality visual annotations, we first crop the upper clothes from the source images (Clothes Copping) and search on the internet to get the most corresponding clothes-template images (Clothes-template Crawling) as visual annotations. Since each person wears the same clothes in



(a) language guidance generation **(b) clothes template collection**

Fig. 3. **(a)** We generate attributes for a person and then transform attributes into sentences by a large language model. **(b)** We crop upper clothes and search them online for clothes templates.

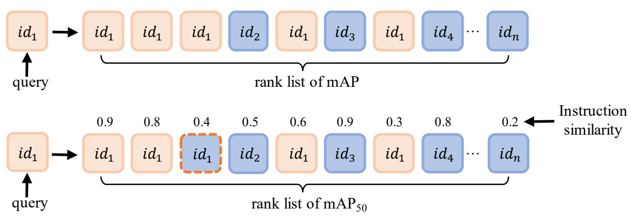


Fig. 4. Illustration of the difference between metric mAP and mAPτ. Light orange in the rank list indicates positive and negative samples in light blue. In the example of mAPτ, we set the threshold to 0.5 and mark with an orange dashed box where the identity matches the query but the instruction similarity does not reach the threshold and is corrected to a negative sample.

traditional ReID datasets, we annotate the clothes-changing PRCC, LTCC and LaST datasets where each person wears multiple clothes to ease the burden of annotations. Fig. 3 (b) shows the detailed process.

Clothes Cropping. We use a human parsing model SCHP [72] to generate the segmentation mask of the upper clothes and then crop the corresponding rectangle upper clothes patches from the original images. These bounding boxes of upper clothes are then manually validated.

Clothes-template Crawling. Given all cropped upper clothes from images in OmniReID datasets, we crawl the templates of these clothes from shopping websites[2]. The top 40 matching clothes templates are downloaded when we crawl each cropped upper clothes. The one with the highest image quality is manually selected.

Training Setting To enable all-purpose person ReID, we perform two training scenarios based on this built benchmark: (1) Single-task Learning (STL): Every task is trained and tested individually using the corresponding dataset. (2) Multi-task Learning (MTL): To acquire one unified model for all tasks, the model is optimized by joint training of the six ReID tasks with all the training datasets. The trained network is then evaluated for different tasks on various datasets.

Evaluation Setting To validate the effectiveness of ReID methods, we propose two testing methodologies along with the OmniReID++ benchmark: (1) Task-specific evaluation: We independently utilize each test dataset included in the OmniReID++ benchmark to validate the performance of a ReID model on that specific task and dataset. When extracting gallery features for person retrieval, each image in the gallery set is associated with a task-specific instruction. (2) Task-free evaluation: In this setting, we consolidate all test datasets from the OmniReID++ benchmark into a single test dataset, which includes query images and instructions representing different task retrieval targets. During inference, the gallery datasets from various sub-datasets are original images without distinction. This approach better simulates the inference scenario of the model in real-world applications.

Evaluation Metric OmniReID++ employs the commonly used CMC (Cumulative Matching Characteristic) metric along with mAP (mean Average Precision) as the traditional evaluation metrics. Besides, we also introduce a novel mAPτ metric that provides a method to evaluate both the accuracy of person retrieval identity and the compliance with instruction target for instruct-ReID task. The calculation of mAPτ is defined as

$$\mathbf{mAP}\tau = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{k=1}^{n}\frac{\Psi^{\tau}(l_q, l_k^q)}{kn_q}\sum_{i=1}^{k}\Psi^{\tau}(l_q, l_i^q) \tag{1}$$

where Q is the number of query images, n is the maximum computation length set in the q-th image retrieval rank list, n_q is the correct returned results number, $\Psi^{\tau}(\cdot)$ denotes the matching evaluator and the calculation process for the q-th query l_q with a k-th retrieval result l_k^q can be defined as

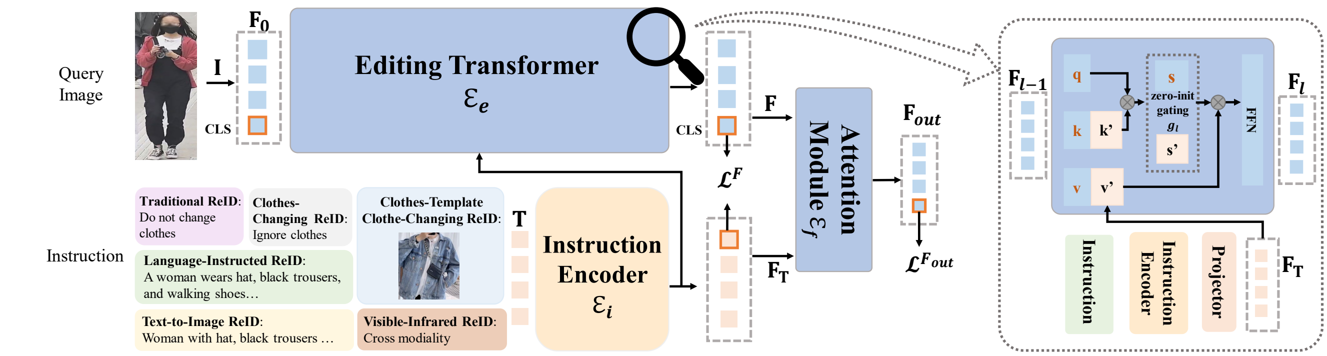
$$\Psi^{\tau}(l_q, l_k^q) = \mathbb{I}\left(y_q = y_k\right) \cdot \mathbb{I}\left(\mathbf{Cos}\left\langle I_q, I_k^q\right\rangle \geq \tau\right) \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, y_q or y_k is the identity label, I_q or I_k^q is instruction and $\mathbf{Cos}\left\langle\cdot\right\rangle$ is the cosine similarity function. The matching evaluator is 1 if the k-th or i-th prediction is correct for the q-th query l_q after applying the threshold τ and 0 for otherwise. The illustration and distinction between mAP and mAPτ are described in Fig. 4. In the calculation process of mAP, images in the retrieval rank list are categorized as positive or negative depending on whether their identities match that of the query image. However, in the proposed instruct-ReID task, we search the target images with both query images and the instructions. The retrieved images with the same identity may not necessarily match the instruction target. Hence, the new proposed mAPτ computes the similarity between each instruction of the gallery image and the query image, and samples in the rank list with instruction similarity below a predefined threshold τ are considered negative. We take the third sample (red dash boundary sample in Fig. 4) in the mAP_{50} rank list for example, where mAP_{50} denotes a threshold τ set at 0.5. Despite sharing the same identity as the query image, its instruction similarity is below 0.5, leading to its classification as a negative sample for calculation purposes. Compared to the traditional mAP metric, the mAPτ metric provides a more precise reflection of whether retrieval results align with the instruction description, even in individual text-to-image tasks.

IV. Instruct ReID Methodology

In this section, we present our proposed instruct-ReID model, which includes IRM for the task-specific evaluation setting and IRM++ for the task-free evaluation setting. We begin by introducing the instruction generation (Sec. IV-A) for different ReID tasks which provides specific instructions for query retrieval guidance for both IRM and IRM++. Then

[2]https://world.taobao.com/, https://www.17qcc.com/



(a) Instruct-ReID Model

(b) l-th editing transformer

Fig. 5. The overall architecture of the proposed method. The instruction is fed into the instruction encoder to extract instruction features (a). The features are then propagated into the editing transformer (b) to capture instruction-edited features. We exploit adaptive triplet loss and identification loss to train the network. For the testing stage, we use the CLS token for retrieval.

the architecture of the proposed Instruct ReID model IRM is discussed in (Sec. IV-B). The overview of IRM is shown in Fig. 5, which consists of three parts: an editing transformer $\mathcal{E}_e$ (Sec. IV-B1) and an instruction encoder $\mathcal{E}_i$ which is a visual language model, and an attention module $\mathcal{E}_f$. Instead of traditional triplet loss, IRM adopts a newly proposed adaptive triplet loss $\mathcal{L}_{atri}$ (Sec. IV-B2) to supervise the model training and the overall loss function is introduced for all ReID tasks in (Sec. IV-B3). Finally, we describe IRM++ in (Sec. IV-C) which also employs an image encoder $\mathcal{E}_e$, an instruction encoder $\mathcal{E}_i$ and an attention module $\mathcal{E}_f$. IRM++ only extracts images features by the image encoder $\mathcal{E}_e$ as the general gallery features retrieval for various ReID tasks. Besides, we present a memory bank contrastive learning method to assist IRM++ learning, achieving better performance in the task-free evaluation setting than the triplet loss based training methods.

A. Instruction Generation

In our proposed instruct-ReID task, the model must retrieve the images that describe the same person following the instructions. By designing different instructions, our instruct-ReID can be specialized to existing ReID tasks, *i.e.,* traditional ReID, clothes-changing ReID, clothes template based clothes-changing ReID, visual-infrared ReID, text-to-image ReID, and language-instructed ReID. We show the current instructions and leave the exploration of better instructions for instruct-ReID to future research.

Traditional ReID: Following Instruct-BLIP [73], we generate 20 instructions[3] from GPT-4 and randomly select one, *e.g.,* "Do not change clothes.", during training. The model is expected to retrieve images of the same person without altering image attributes, such as clothing.

> **### Query image**:{Query image}
> **### Instruction**: "Do not change clothes."[3]
> **### Target image**:{Output}

Clothes-changing ReID: Simiar to Traditional ReID, the instruction is the sentence chosen from a collection of 20 GPT-4 generated sentences[3], *e.g.,* "With clothes changed". The model should retrieve images of the same person even when wearing different outfits.

> **### Query image**:{Query image}
> **### Instruction**: "With clothes changed."[3]
> **### Target image**:{Output}

Clothes template based clothes-changing ReID: The instruction is a clothes template for a query image while a cropped clothes image for a target image. The model should retrieve images of the same person wearing the provided clothes. We provide more examples for training and test in the supplementary materials.

> **### Query image**:{Query image}
> **### Instruction**:{Any clothes template image}
> **### Target image**:{Output}

Visual-Infrared ReID: The instruction is the sentence chosen from a collection of 20 GPT-4 generated sentences[3], *e.g.,* "Retrieve cross modality image". The model should retrieve visible (infrared) images of the same person according to the corresponding infrared (visible) images.

> **### Query image**:{Query image}
> **### Instruction**: "Retrieve cross modality image."[3]
> **### Target image**:{Output}

Text-to-Image ReID: The instruction is the describing sentences, and both images and text are fed into IRM during the training process. While in the inference stage, the image features and instruction features are extracted separately. [4]

> **### Image**:{Image}[4]
> **### Instruction**:{Sentences describing pedestrians}[4]
> **### Target image**:{Output}

Language-instructed ReID: The instruction is several sentences describing pedestrian attributes. We randomly select the description languages from the person images in gallery and

[3]See all available instruction sentences in supplementary materials.

[4]Image and instruction features are extracted separately in test stage.

provide to query images as instruction. The model is required to retrieve images of the same person described in the provided sentences. We provide more examples for training and test in the supplementary materials.

> **### Query image**:{Query image}
> **### Instruction**:{Sentences describing pedestrians}
> **### Target image**:{Output}

B. Model Architecture of IRM

As illustrated in Fig. 5 (a), the pipeline of IRM can be summarized as follows: given an instruction $\mathbf{T}$ associated with a query image $\mathbf{I}$, IRM obtains instruction features $\mathbf{F_T}$ using the instruction encoder $\mathcal{E}i$ which is a visual language model. These extracted instruction features $\mathbf{F_T}$, along with query image tokens, are fed into the designed editing transformer $\mathcal{E}_e$ to obtain features $\mathbf{F}$ edited by instructions. Moreover, both the instruction features $\mathbf{F_T}$ and $\mathbf{F}$ are fed into an attention module $\mathcal{E}_f$ to efficiently integrate features of the query image and instruction using the attention mechanism within the stacked transformer layers. Finally, the output embeddings $\mathbf{F}_{out}$ of the attention module are utilized for person image retrieval.

1) Editing Transformer: The editing transformer consists of L zero-init transformer layers $\mathcal{E}_e = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_L\}$, which can leverage the instruction to edit the feature of query images. Given the l-th zero-init transformer, the output feature $\mathbf{F}_l$ can be formulated as

$$\mathbf{F}_l = \mathcal{F}_l(\mathbf{F}_{l-1}, \mathbf{F_T}), \tag{3}$$

where $\mathbf{F_T} = \mathcal{E}_i(\mathbf{T})$ is the instruction feature extracted by the instruction encoder $\mathcal{E}_i$ and $\mathbf{F}_{l-1}$ is the output feature of $(l{-}1)$-th zero-init transformer layer. Here, we define the initial input $\mathcal{F}_0$ of the first layer as $\mathbf{F}_0 = [\mathbf{f}_0^{\text{CLS}}, \mathbf{f}_0^1, \mathbf{f}_0^2, ..., \mathbf{f}_0^N]$, where $\mathbf{f}_0^{\text{CLS}}$ is the appended [CLS] token, $(\mathbf{f}_0^1, \mathbf{f}_0^2, ..., \mathbf{f}_0^N)$ are the patch tokens of the query image and N is the number of patches of the query image.

We show the detailed structure of each layer in the editing transformer in Fig. 5 (b). Given the features $\mathbf{F}_{l-1}$ and instruction features $\mathbf{F_T}$, the attention map $\mathbf{M}_l$ is defined as

$$\mathbf{M}_l = \left[\text{Softmax}(\mathbf{S}_l), g_l \times \text{Softmax}(\mathbf{S'}_l)\right], \tag{4}$$

where g_l is the gating parameters initialized by zero. Here, $\mathbf{S}_l$ is the attention map between queries and keys of input features and $\mathbf{S}_l'$ is the attention map between queries of input features and keys of instruction features. Mathematically,

$$\mathbf{S}_l = \mathbf{Q}_l\mathbf{K}_l^\top/\sqrt{C}, \mathbf{S}_l' = \mathbf{Q}_l\mathbf{K'}_l^\top/\sqrt{C}, \tag{5}$$

where a linear projection derives queries and keys, *i.e.,* $\mathbf{Q}_l = \text{Linear}_q(\mathbf{F}_{l-1})$, $\mathbf{K}_l = \text{Linear}_k(\mathbf{F}_{l-1})$ and $\mathbf{K}_l' = \text{Linear}_{k'}(\mathbf{F_T})$, respectively. C is the feature dimension of query features. Finally, we calculate the output of the l-th layer by

$$\mathbf{F}_l = \text{Linear}_o(\mathbf{M}_l\left[\mathbf{V}_l, \mathbf{V}_l'\right]), \tag{6}$$

where Linear_o is the feed-forward network after the attention layer in each transformer block, $\mathbf{V}_l$ and $\mathbf{V}_l'$ are the values calculated by $\mathbf{V}_l = \text{Linear}_v(\mathbf{F}_{l-1})$ and $\mathbf{V}_l' = \text{Linear}_{v'}(\mathbf{F_T})$. We use the [CLS] token in the output feature of L-th transformer layer for computing losses and retrieval, *i.e.,* $\mathbf{F} = \mathbf{f}_L^{\text{CLS}}$, where $\mathbf{F}_L = (\mathbf{f}_L^{\text{CLS}}, \mathbf{f}_L^1, \mathbf{f}_L^2, ..., \mathbf{f}_L^N)$ and N is patch number of query images.

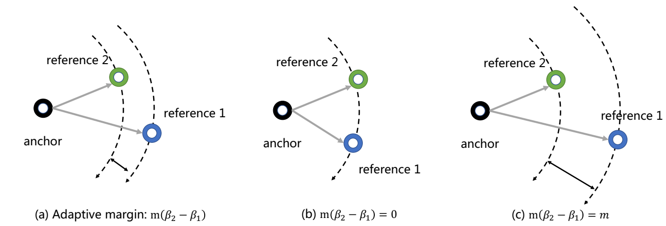


Fig. 6. Illustration of adaptive triplet loss. Unlike the traditional triplet loss where the margin is fixed, the margin in our adaptive triplet loss is defined by the instruction similarity for the two query-instruction pairs that describe the same person. The features associated with similar instructions are pulled to be closer.

2) Adaptive Triplet Loss: Unlike typical triplet loss that defines positive and negative samples solely based on identities, instruct-ReID requires distinguishing images with different instructions for the same identity. Intuitively, an adaptive margin should be set to push or pull samples based on the instruction difference. Let $(\mathbf{F}_i^a, \mathbf{F}_i^{r_1}, \mathbf{F}_i^{r_2})$ be the i-th triplet in the current mini-batch, where $\mathbf{F}_i^a$ is an anchor sample, $\mathbf{F}_i^{r_1}$ and $\mathbf{F}_i^{r_2}$ are reference samples. We propose an adaptive triplet loss as

$$\begin{aligned}\mathcal{L}_{atri} = \frac{1}{N_{tri\dagger}} \sum_{i=1}^{N_{tri\dagger}} \{&\mathbf{Sign}(\beta_1 - \beta_2)[d(\mathbf{F}_i^a, \mathbf{F}_i^{r_1}) + \\ &(\beta_1 - \beta_2)\, m - d(\mathbf{F}_i^a, \mathbf{F}_i^{r_2})]\}_+\end{aligned} \tag{7}$$

where $N_{tri\dagger}$ and m denote the number of triplets and a hyper-parameter for the maximal margin, respectively. d is a Euclidean distance function, *i.e.,* $d(\mathbf{F}_i^a, \mathbf{F}_i^r) = ||\mathbf{F}_i^a - \mathbf{F}_i^r||_2^2$. β_1 and β_2 are relatednesses between the anchor image and the corresponding reference image that consider the identity consistency and the instruction similarity for the adaptive margin. Mathematically, they can be defined as

$$\beta_j = \mathbb{I}\left(y_a = y_{r_j}\right) \mathbf{Cos}\left\langle \mathbf{F_T}^a, \mathbf{F_T}^{r_j} \right\rangle, \tag{8}$$

where y_a and y_{r_j} are the identity labels of the anchor image and the reference image, $\mathbb{I}(\cdot)$ is the indicator function, and $j = \{1, 2\}$ denotes the index of reference samples.

The concept of adaptive triplet loss is described by Fig. 6(a). We discuss the adaptive loss in two cases. First, as shown in Fig. 6(b), the margin is set to zero if the triplet has the same identity and the instructions of the two reference samples are equally similar to the instruction of the anchor sample. This makes the distances between the anchor and two reference points the same. Second, as shown in Fig. 6(c), when there is a significant difference in instruction similarities, the margin distance between the anchor and two references becomes closer to the maximum value m, forcing the model to learn discriminative features. Adaptive triplet loss makes features from the same person become distinctive based on the similarity of instructions, which helps to retrieve images that align the requirements of given instructions in the CTCC-ReID and LI-ReID tasks.

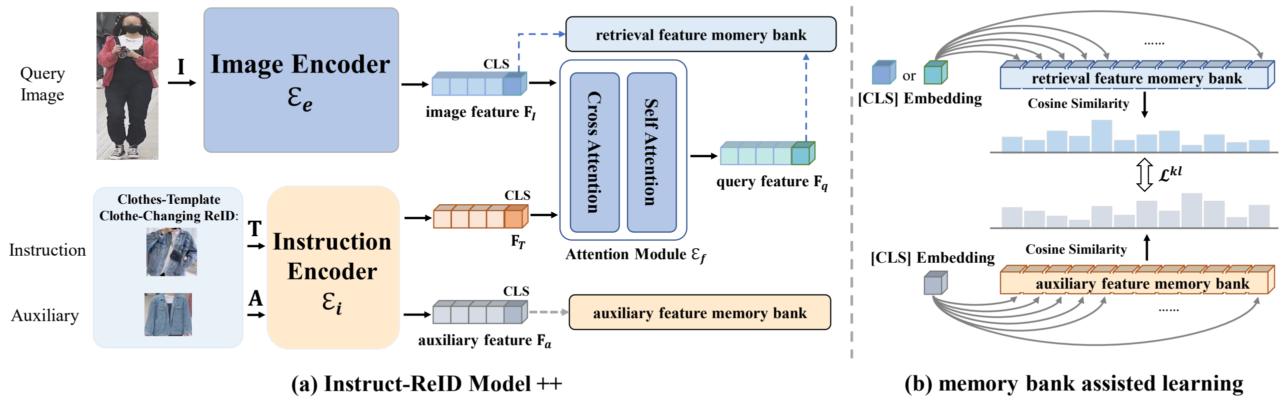


Fig. 7. (a) The structure of the instruct-ReID++ Model. Auxiliary is employed only in training step, to endow model with powerful supervision (b) We propose a novel learning scheme, using two memory banks to guiding feature alignment.

3) Overall Loss Function: We impose an identification loss $\mathcal{L}_{id}$, which is the classification loss on identities, and an adaptive triplet loss $\mathcal{L}_{atri}$ on $\mathbf{F}$ and the final fusion features $\mathbf{F_{out}}$ to supervise the model for training Trad-ReID, CC-ReID, VI-ReID, CTCC-ReID, and LI-ReID tasks. The overall loss is combined as

$$\mathcal{L} = \alpha \mathcal{L}_{atri}(F) + \mathcal{L}_{id}(F) + \alpha \mathcal{L}_{atri}(F_{out}) + \mathcal{L}_{id}(F_{out}) \tag{9}$$

where $\mathbf{F}$ and $\mathbf{F}_{out}$ indicate the source of features used in calculating the loss and α is a hyperparameter to adjust the weight of different losses.

For the T2I-ReID task, we adopt a contrastive loss $\mathcal{L}_{cl}$ to align the image features $\mathbf{F}$ and text features $\mathbf{F_T}$ to enable text-based person retrieval. We also employ a binary classification loss $\mathcal{L}_{match}$ to learn whether an inputted image-text pair is positive or negative, defined as:

$$\mathcal{L} = \mathcal{L}_{cl}(F) + \mathcal{L}_{match}(F_{out}) \tag{10}$$

where, $\mathcal{L}_{match}(F_{out}) = \mathcal{C}_e(\hat{y}, F_{out}) = \mathcal{C}_e[\hat{y}, \mathcal{E}_f(\mathbf{F}, \mathbf{F_T})]$, respectively. $\mathcal{C}_e$ represents a binary cross-entropy function, $\hat{y}$ is a 2-dimension one-hot vector representing the ground-truth label *i.e.,* $[0, 1]^\intercal$ for the positive pair and $[1, 0]^\intercal$ for the negative pair, which is formed by matching text features with corresponding image features before inputting into the attention module $\mathcal{E}_f$.

C. Model Architecture of IRM++

The overview of the proposed IRM++ is shown in Fig. 7 (a), which employs a pretrained ViT encoder $\mathcal{E}_e$ for image encoding and utilizes a vision-language model, *e.g.,* ALBEF [74], as the instruction encoder $\mathcal{E}_i$. Our IRM++ processes the outputs of the query and gallery features in two basic formats. For the *query feature*, the query image $\mathbf{I}$ and the paired instruction $\mathbf{T}$ are separately input to the image encoder and instruction encoder to extract image features $\mathbf{F}_I = \mathcal{E}_e(\mathbf{I})$ and instruction features $\mathbf{F}_T = \mathcal{E}_i(\mathbf{T})$. These features are then combined by an attention module $\mathcal{E}_f$ to integrate the information from the query image and retrieval intent instruction, ultimately outputting the query feature $\mathbf{F}_q = \mathcal{E}_f(\mathbf{F}_I, \mathbf{F}_T)$. For the *gallery feature*, gallery images from various ReID tasks are input to the image encoder of IRM++, and the extracted image features are used as the gallery features for retrieval, *e.g.,* $\mathbf{F}_g = \mathbf{F}_I = \mathcal{E}_e(\mathbf{I})$. IRM++ removes the task-specific instruction inputs and provides a more general way of extracting gallery features, which is suitable for the task-free evaluation setting. For clarity, in the subsequent discussion, we assign both the query feature $\mathbf{F}_q$ and the gallery feature $\mathbf{F}_g$ to the retrieval feature $\mathbf{F}_r$.

To facilitate IRM++ in the task-free setting, we utilize two memory banks, shown in Fig. 7 (b), and implement a contrastive learning method. In the training phase, we utilize the instruction annotation as the auxiliary data $\mathbf{A}$ and the auxiliary feature $\mathbf{F}_a = \mathcal{E}_i(\mathbf{A})$ is derived from the instruction encoder. We designate images in the training set as query type or gallery type. For instance, in CTCC-ReID, images paired with template clothes image instructions are denoted as query type, while images wearing target clothes are classified as gallery type. Similarly, in LI-ReID, images paired with language description instructions are designated as query type, and those that match the retrieval target are classified as gallery type. We extract the query feature $\mathbf{F}_q$ or the gallery feature $\mathbf{F}_g$ based on the corresponding query or gallery type of the training set images, and assign them as the retrieval feature $\mathbf{F}_r$. Based on the retrieval feature $\mathbf{F}_r$ and auxiliary feature $\mathbf{F}_a$, we maintain the retrieval feature memory bank $\mathbf{M}_r = \{m_1^r, m_2^r, ..., m_N^r\}$ and auxiliary feature memory bank $\mathbf{M}_a = \{m_1^a, m_2^a, ..., m_N^a\}$, where N represents the number of identities in the training set. Both ranks are initialized randomly, during the training process, the retrieval feature $\mathbf{F}_r$ associated with identity i will be stored in m_i^r, replacing the original feature inside it. Similarly, the feature m_i^a will be replaced by auxiliary feature $\mathbf{F}_a$ that associates with identity i. For each iteration, we compute the cosine similarity of $\mathbf{F}_r$ with each element in $\mathbf{M}_r$ and obtain the similarity score $\mathbf{S}_r = \text{Similarity}(\mathbf{F}_r, \mathbf{M}_r) = \{s_1^r, s_2^r, ..., s_N^r\}$. The calculation process for each specific s_i^r can be defined as

$$s_i^r = \mathbf{Cos}\left\langle \mathbf{F_r}, \mathbf{m_i^r} \right\rangle. \tag{11}$$

We do the same way to the similarity score of auxiliary, getting $\mathbf{S}_a = \text{Similarity}(\mathbf{F}_a, \mathbf{M}_a) = \{s_1^a, s_2^a, ..., s_N^a\}$. In this way, the similarity score of the auxiliary can be seen as a soft label, while the ground truth identity can be seen as a hard label. We utilize a contrastive learning loss to guide the learning with soft label, which is defined as

$$\mathcal{L}_{soft} = \mathcal{L}_{kl}(\mathbf{S}_a, \mathbf{S}_r) = \frac{1}{N_{kl}} \sum_{i=1}^{N_{kl}} \mathcal{P}(s_i^a) \mathbf{log} \frac{\mathcal{P}(s_i^r)}{\mathcal{P}(s_i^a) + \xi}, \quad (12)$$

where N_{kl} and ξ denote the number of contrastive learning samples and a smoothing factor to prevent division by zero, respectively. $\mathcal{P}(\cdot)$ is a softmax function. We turned ground truth into a $\mathbf{N}$-length one-hot label $\mathbf{S}_{gt} = \{s_1^{gt}, s_2^{gt}, ..., s_N^{gt}\}$, with the same shape to the similarity score. $s_i^{gt} = 1$ if $\mathbf{F}_r$ with identity i, instead $s_i^{gt} = 0$. We also use contrastive loss to guide the supervision of this hard label and similarly by

$$\mathcal{L}_{hard} = \mathcal{L}_{kl}(\mathbf{S}_{gt}, \mathbf{S}_r) = \frac{1}{N_{kl}} \sum_{i=1}^{N_{kl}} \mathcal{P}(s_i^{gt}) \mathbf{log} \frac{\mathcal{P}(s_i^r)}{\mathcal{P}(s_i^{gt}) + \xi}. \quad (13)$$

The soft label provides detailed and fine-grained instruction supervision and the hard label ensures the model learns reliable identity information. By combining the contrastive loss and the identification loss $\mathcal{L}_{id}$, the overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{id}(F) + \mathcal{L}_{soft}(F) + \beta \times \mathcal{L}_{hard}(F), \quad (14)$$

where β is a hyperparameter to adjust the loss weights.

V. Experiments

A. Datasets and Evaluation Metric

1) Training datasets: Using the proposed **OmniReID++** benchmark, we train IRM and IRM++ under two training settings: (1) Single-task Learning (STL): The model is trained separately on each dataset corresponding to the individual task. (2) Multi-task Learning (MTL): The model is trained jointly using all datasets included in OmniReID++.

2) Evaluation Datasets: To assess the efficacy of our proposed method across diverse ReID scenarios, we conduct the following tests according to the OmniReID++ benchmark settings: (1) task-specific evaluation setting: we independently evaluate the performance of IRM on each test dataset included in the OmniReID++ benchmark. We then compare our results with existing state-of-the-art methods. (2) task-free evaluation setting: we consolidate all test datasets from the OmniReID++ benchmark into a unified test set and present the evaluation results of IRM++ under this configuration. Additionally, we assess the performance of IRM++ on each sub-test dataset to compare with state-of-the-art methods. We refer to sub-test set and joint test set evaluations as *Single Test Set Evaluation* and *Joint Test Set Evaluation*, respectively.

3) Evaluation Metric: In the single test set evaluation setting, we use metrics such as the Cumulative Matching Characteristic (CMC) curve at top-1 and mAP (mean Average Precision) to assess ReID performances quantitatively. For the joint test set evaluation setting, in addition to the mAP metric, we also utilize our proposed mAP$\mathcal{T}$ evaluation metric to validate the performance of the metrics at different thresholds.

TABLE III
THE PERFORMANCE OF CLOTHSE-CHANGING REID OF OUR METHOD AND THE STATE-OF-THE-ART METHODS. MEAN AVERAGE PRECISION (MAP) AND TOP1 ARE USED TO QUANTIFY THE ACCURACY. † DENOTES THAT THE MODEL IS TRAINED WITH MULTIPLE DATASETS. * DENOTES THAT THE MODEL IS PRE-TRAINED ON 4 MILLION IMAGES.

Methods	**LTCC**		**PRCC**		**VC-Clothes**	
	mAP	Top1	mAP	Top1	mAP	Top1
HACNN [75]	26.7	60.2	-	21.8	-	-
RGA-SC [76]	27.5	65.0	-	42.3	67.4	71.1
PCB [77]	30.6	65.1	38.7	41.8	62.2	62.0
IANet [78]	31.0	63.7	45.9	46.3	-	-
CAL [7]	40.8	74.2	-	-	-	-
TransReID [79]	-	-	-	44.2	71.8	72.0
IRM (STL)	46.7	66.7	46.0	48.1	**80.1**	**90.1**
IRM (MTL)†	**52.0**	**75.8**	**52.3**	**54.2**	78.9	89.7

B. Implementation Details

For the editing transformer, we use the plain ViT-Base with the ReID-specific pretraining [26]. We adopt ALBEF [74] as our instruction encoder. All images are resized into 256×128 for training and testing. We use the AdamW optimizer with a base learning rate of 1e-5 and a weight decay of 5e-4. We linearly warmup the learning rate from 1e-7 to 1e-5 for the first 1000 iterations. Random cropping, flipping, and erasing are used for data augmentation during training. For each training batch, we randomly select 32 identities with 4 image samples for each identity. The Single-task Learning takes one NVIDIA V100 GPU and the Multi-task Learning takes 32 NVIDIA V100 GPUS for training, respectively.

C. Task-specific Evaluation Setting Results

In this subsection, we validate the effectiveness of the proposed IRM under the task-specific setting. All results are obtained based on the OmniReID++ benchmark, and we compare with current state-of-the-art methods on the six tasks, using commonly used metrics such as mAP and CMC (rank1)

Clothes-Changing ReID (CC-ReID). As shown in Tab. III, IRM outperforms all state-of-the-art methods on LTCC, PRCC and VC-Clothes, showing that the model can effectively extract clothes-invariant features following the instructions, *e.g.,* "Ignore clothes". Specifically, on STL, IRM improves CAL [7], TransReID [79] by **+5.9%** mAP and **+8.3%** mAP on LTCC and VC-Clothes, respectively. Using multi-task training further improves the performance of IRM to **52.0%** mAP on LTCC, and reaches a new state-of-the-art on PRCC with a **52.3%** mAP. While multi-task learning leads to slightly lower performance than single-task learning on VC-Clothes, IRM still achieves a higher Top-1 than TransReID. We analyze that this drop is due to the domain gap between VC-Clothes (Synthetic) and datasets (Real) and leave it for future work.

Clothes Template Based Clothes-Changing ReID (CTCC-ReID). Our method achieves desirable performance on the CTCC-ReID task in Tab. V, which shows that a fixed instruction encoder is enough for this tough task. Concretely, when only trained on COCAS+ Real1, IRM outperforms BC-Net [11] and DualBCT-Net [12], both of which learn an independent clothes branch, by **+9.6%** mAP and **+2.2%** mAP,

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON VISIBLE-INFRARED REID AND TEXT-TO-IMAGE REID. THE VI-REID SETTING IS *VIS-to-IR* AND *IR-to-VIS* IN LLCM. † DENOTES THE MODEL IS TRAINED WITH MULTIPLE DATASETS.* DENOTES THE MODEL IS TRAINED UNDER THE SAME IMAGE SHAPE AS IRM, *i.e.,* 256×128.

Methods	T2I-ReID		VI-ReID: LLCM			
	CUHK-PEDES		VIS-to-IR		IR-to-VIS	
	mAP	Top1	mAP	Top1	mAP	Top1
ALBEF [74]	56.7	60.3	-	-	-	-
CAJ [80]	-	-	59.8	56.5	56.6	48.8
MMN [81]	-	-	62.7	59.9	58.9	52.5
DEEN [62]	-	-	65.8	62.5	62.9	54.9
SAF [82]	58.6	64.1	-	-	-	-
PSLD [83]	60.1	64.1	-	-	-	-
RaSa* [18]	63.9	69.6	-	-	-	-
IRM (STL)	65.3	72.8	66.6	66.2	64.5	64.9
IRM (MTL)†	**66.5**	**74.2**	**67.5**	**66.7**	**67.2**	**65.7**

respectively. By integrating the knowledge on other instruct ReID tasks during multi-task learning, we are able to further improve the performance of IRM, achieving an mAP of **41.7%** and pushing the performance limits on CTCC-ReID.

Visible-Infrared ReID (VI-ReID). We validate the performance of IRM on Visible-Infrared ReID datasets LLCM, which is a new and challenging low-light cross-modality dataset and has a more significant number of identities and images. From Tab. IV, we can see that the results on the two test modes show that the proposed IRM achieves competitive performance against all other state-of-the-art methods. Specifically, for the VIS-to-IR mode on LLCM, IRM achieves **67.5%** mAP and exceeds previous state-of-the-art methods like DEEN [62] by **+1.7%**. For the IR-to-VIS mode on LLCM, IRM achieves **65.7%** Rank-1 accuracy and **67.2%** mAP, which is a new state-of-the-art result. The results validate the effectiveness of our method.

Text-to-Image ReID (T2I-ReID). As shown in Tab. IV, IRM shows competitive performance with a mAP of **66.5%** on the CUHK-PEDES [21], which is **+2.6%**, **+6.4%**, **+7.9%** higher than previous methods RaSa [18] (63.9%), PSLD [83] (60.1%), SAF [82] (58.6%). Because a few images in CUHK-PEDES training set are from the test sets of Market1501 and CUHK03, we filtered out duplicate images from the test sets during the multi-task learning (MTL) process. Following the common research works, the testing is conducted on uniformly resized images with a resolution of 256×128.

Language-Instructed ReID (LI-ReID). As a novel setting, no previous works have been done to retrieve a person using several sentences as the instruction, therefore, we compare IRM with a straightforward baseline. In the baseline method, only a ViT-Base is trained on COCAS+ Real1 images without utilizing the information from language instruction, leading to poor person re-identification ability. As shown in Tab. V, IRM improves the baseline by **+15.8%** mAP, because IRM can integrate instruction information into identity features. With MTL, IRM achieves extra **+9.1%** performance gain by using more images and general information in diverse ReID tasks.

Traditional ReID (Trad-ReID). IRM also shows its effectiveness on Trad-ReID tasks in Tab. V. Specifically, when trained on a single dataset, compared with PASS [26], IRM achieves comparable performance on Market1501, MSMT17 and **+6.9%** performance gain on CUHK03 compared with SAN [84]. With multi-task training, IRM can outperform the recent multi-task pretraining HumanBench [85] and self-supervised pretraining [26]. We do not compare with SOLDIER [86] because it only reports ReID results with the image size of 384×192 instead of 256×128 in our method. More importantly, SOLDIER primarily focuses on pretraining and is evaluated only on Trad-ReID, while the claimed contribution of IRM is to tackle multiple ReID tasks with one model.

D. Task-free Evaluation Setting Results

In this subsection, we validate the effectiveness of the proposed IRM++ under the task-free setting. All results are obtained based on the OmniReID++ benchmark. Specifically, we first evaluate the performance of IRM++ on the six tasks in the Single Test Set Evaluation Setting, using metrics such as mAP and CMC (rank1). Then, we provide the mAP$\mathcal{T}$ results of IRM++ and current state-of-the-art methods under the Joint Test Set Evaluation Setting for reference, aiming to facilitate research in this field.

Single Test Set Evaluation. We evaluate the performance of IRM++ on each single test set, which is a subset of the task-free evaluation setting. As shown in Tab. VI, for clothes template based clothes-changing ReID (CTCC-ReID), language-instructed ReID (LI-ReID), clothes-changing ReID (CC-ReID), visible-infrared ReID (VI-ReID), text-to-image ReID (T2I-ReID), and traditional ReID (Trad-ReID), under Multi-task learning (MTL) training, our IRM++ achieves state-of-the-art (SOTA) results, surpassing the best-performing methods by **+0.3%** to **+22.4%** mAP. Experimental results demonstrate that IRM++ effectively extracts generic gallery features and compares them with features extracted from query images and instructions for accurate retrieval. Besides, we conduct a baseline method IRM++* by employing a triplet loss as the supervision function. For CTCC-ReID and LI-ReID tasks, IRM++ surpasses the baseline method by **+10.0%** and **+9.0%** mAP (STL), **+7.1%** and **+5.8%** mAP (MTL), which indicates that the memory bank assisted contrastive learning utilized in IRM++ further enhances the model performance compared to optimization with triplet loss on these two ReID tasks. We analyze that the memory bank-assisted contrastive loss employs multiple negative samples for each iteration, providing stronger supervision during training compared to triplet-based optimization methods, which only utilize one negative sample per iteration. For CC-ReID, VI-ReID, and Trad-ReID, our IRM++ and baseline method achieve similar results, with experimental variations within **0.3%** mAP under Single-task learning (STL) training.

Joint Test Set Evaluation. As shown in Tab. VII, we unify all testing set data into a joint set and provide benchmark testing results on the proposed dataset. We also perform inference with existing state-of-the-art methods on this testing set for comparison. Since DualBCT Net [12], RaSa [18], PASS [26], CAL [7] and DEEN [62] cannot receive multi-modal instruction inputs, *e.g.*, clothes template images and

TABLE V

Performance comparison with the state-of-the-art methods on Clothes -Template Clothes-Changing ReID, Language-Instructed ReID, and Traditional ReID. † denotes that the model is trained with multiple datasets. * denotes that the model is pretrained on 4 million pedestrian images. The size of the input images used in the table is 256x128.

Methods	CTCC-ReID		LI-ReID		Trad-ReID					
	COCAS+ Real2		COCAS+ Real2		Market1501		MSMT17		CUHK03	
	mAP	Top1	mAP	Top1	mAP	Top1	mAP	Top1	mAP	Top1
Baseline	-	-	14.9	31.6	-	-	-	-	-	-
TransReID [79]	5.5	17.5	-	-	86.8	94.4	61.0	81.8	-	-
BC-Net [11]	22.6	36.9	-	-	-	-	-	-	-	-
DualBCT-Net [12]	30.0	48.9	-	-	-	-	-	-	-	-
SAN [84]	-	-	-	-	88.0	96.1	-	-	76.4	80.1
HumanBench† [85]	-	-	-	-	89.5	-	69.1	-	77.7	-
PASS* [26]	-	-	-	-	93.0	96.8	71.8	88.2	-	-
IRM (STL)	32.2	54.8	30.7	60.8	92.3	96.2	71.9	86.2	83.3	86.5
IRM (MTL)†	**41.7**	**64.9**	**39.8**	**71.6**	**93.5**	**96.5**	**72.4**	**86.9**	**85.4**	**86.5**

TABLE VI

The performance comparison of IRM++ with the state-of-the-art methods on 6 instruct-ReID tasks and all results are presented based on task-free evaluation setting. † denotes that the test mode is VIS-to-IR and ‡ denotes IR-to-VIS mode on LLCM. * denotes that the model is trained by triplet loss as a baseline method. The size of the input images used in the table is 256x128.

Methods	CTCC-ReID	LI-ReID	T2I-ReID	VI-ReID		CC-ReID			Trad-ReID		
	Real2	Real2	CUHK.	LLCM†	LLCM‡	LTCC	PRCC	VC-Clo.	Market1501	MSMT17	CUHK03
SOTA	30.0 [12]	14.9	63.9 [18]	62.9 [62]	65.8 [62]	40.8 [7]	45.9 [78]	71.8 [79]	93.0 [26]	71.8 [26]	77.7 [85]
IRM++* (STL)	22.2	28.7	64.2	64.6	66.5	46.1	46.3	**80.4**	92.4	72.0	83.5
IRM++* (MTL)	26.3	31.5	66.1	66.0	66.8	46.6	47.3	78.2	93.2	72.4	84.6
IRM++ (STL)	32.2	35.8	64.3	64.5	66.3	46.2	46.5	80.3	92.1	72.2	83.3
IRM++ (MTL)	**35.3**	**37.3**	**66.4**	**66.1**	**67.1**	**46.7**	**47.7**	78.4	**93.3**	**72.5**	**84.7**

description texts, and the instruction similarity is unavailable, we only measure the traditional mAP metric on the Joint Test dataset. We also provide the inference performance of the baseline method IRM++* on the newly proposed testing set. Under the traditional mAP evaluation metric, our IRM++ achieves the best result of **65.3%** mAP, surpassing the baseline method and existing methods by **9.1%** mAP and **22.1%** mAP, respectively. The results demonstrate that IRM++ exhibits outstanding performance in the task-free evaluation setting. Based on instruction similarity and our proposed mAP$\mathcal{T}$ evaluation metric, we provide evaluation results at three different thresholds of 0.25, 0.50, and 0.75. Compared to the baseline method, IRM++ achieves better results of **+8.9%** (mAP$_{25}$), **+6.7%** (mAP$_{50}$), and **+6.3%** (mAP$_{75}$). We observe that under the mAP$\mathcal{T}$ evaluation metric, the test results of our proposed IRM++ show a decreasing trend as the threshold increases. This indicates that with increasing thresholds, more gallery samples with the same identity as the query image but not matching the instruction attributes are corrected back to negative samples, aligning with our proposed mAP$\mathcal{T}$ evaluation metric. The selection of thresholds for different models and scenarios is an area worthy of future research.

TABLE VII

The performance of Joint Test Set Evaluation Setting results includes the $mAP\mathcal{T}$ results at three thresholds. * denotes that the model is trained by triplet loss as a baseline method.

Method	mAP	mAP$_{25}$	mAP$_{50}$	mAP$_{75}$
DualBCT-Net [12]	37.1	-	-	-
RaSa [18]	43.2	-	-	-
PASS [26]	42.2	-	-	-
CAL [7]	37.3	-	-	-
DEEN [62]	29.4	-	-	-
IRM++*	56.2	55.8	55.4	54.2
IRM++	**65.3**	**64.7**	**62.1**	**60.5**

E. Ablation Study

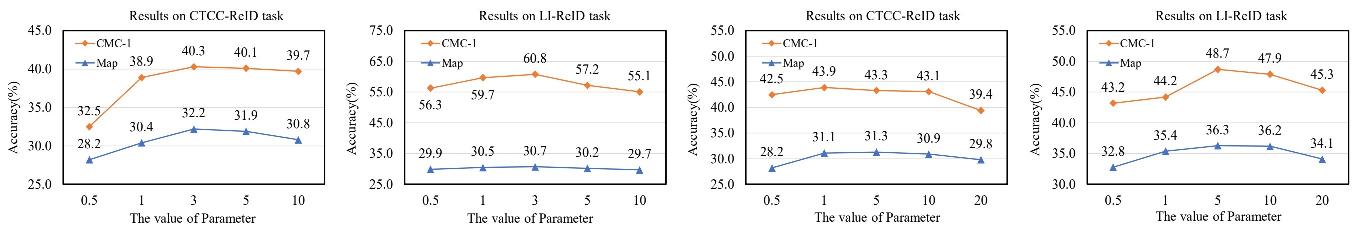
Editing Transformer. To verify the effectiveness of the instruction integrating design in the editing transformer, we compare it with the traditional ViT base model, where the image features are extracted without fusing information from instruction features. Results in Tab. VIII show that adopting the editing transformer leads to **-0.6%** mAP performance drop in the MTL scenario. Consistent results can be observed in STL, indicating the effectiveness of the instruction integrating design in the editing transformer.

Adaptive Triplet Loss. Tab. VIII shows that adaptive triplet loss in MTL outperforms the traditional triplet loss by **+0.3%** mAP on average, indicating that the proposed loss boosts the model to learn more discriminative features following different instructions. In the STL scenario, for CC-ReID and Trad-ReID, the instructions are fixed sentences leading to the same performance of adaptive/traditional triplet loss. However, in CTCC-ReID and LI-ReID, where instructions vary among samples, using adaptive triplet loss brings about **+0.7%**, **+0.5%** mAP gain, which shows the effectiveness of adaptive triplet loss in learning both identity and instruction similarity.

Ablation studies on hyper-parameters. As depicted in Fig. 8

TABLE VIII
ABLATION STUDY. THE PERFORMANCE COMPARISON OF OUR EDITING TRANSFORMER AND USING VIT BASE TRANSFORMER (W/O EDITING), AND THE COMPARISONS WITH TRIPLET LOSS (W/O $\mathcal{L}_{atri}$) AND THE PROPOSED ADAPTIVE TRIPLET LOSS IN TERMS OF MAP. † DENOTES THAT THE TEST MODE IS VIS-TO-IR AND ‡ DENOTES IR-TO-VIS MODE ON LLCM.

Methods	**CTCC-ReID**	**LI-ReID**	**T2I-ReID**	**VI-ReID**		**CC-ReID**			**Trad-ReID**			**Avg.**
	Real2	Real2	CUHK.	LLCM†	LLCM‡	LTCC	PRCC	VC-Clo.	Market1501	MSMT17	CUHK03	
IRM (STL)	32.2	30.7	65.3	66.6	64.5	46.7	46.0	80.1	92.3	71.9	83.3	61.8
w/o editing	32.6	30.5	65.7	66.2	65.1	46.2	45.7	77.8	92.5	71.4	83.2	61.5
w/o $\mathcal{L}_{atri}$	31.5	30.2	-	66.6	64.5	46.7	46.0	80.1	92.3	71.9	83.3	61.3
IRM (MTL)	41.7	39.8	66.5	67.5	67.2	52.0	52.3	78.9	93.5	72.4	85.4	65.1
w/o editing	41.0	38.8	66.1	67.3	65.2	51.2	51.0	78.6	93.0	71.8	85.1	64.5
w/o $\mathcal{L}_{atri}$	40.7	39.2	65.2	68.4	66.3	52.0	52.4	77.9	92.9	72.0	85.5	64.8



(a) Weight α of triplet loss **(b) Weight β of kl loss**

Fig. 8. Ablation experiments on the CTCC-ReID and LI-ReID tasks to test the effects of different combinations of loss weights in the loss function.

TABLE IX
MORE RESULTS OF IRM ON DIFFERENT PRE-TRAINED MODELS. ALL RESULTS ARE OBTAINED ON OMNIREID++ BENCHMARK TASK-SPECIFIC EVALUATION SETTING. WE TRAIN AND TEST USING THE DEFAULT IMAGE RESOLUTION OF 256×128, AND PROVIDE ADDITIONAL RESULTS OF IMAGE RESOLUTION OF 384×192. † DENOTES THAT THE TEST MODE IS VIS-TO-IR AND ‡ DENOTES IR-TO-VIS MODE ON LLCM.

IRM		**CTCC-ReID**	**LI-ReID**	**T2I-ReID**	**VI-ReID**		**CC-ReID**			**Trad-ReID**		
		Real2	Real2	CUHK.	LLCM†	LLCM‡	LTCC	PRCC	VC-Clo.	Market1501	MSMT17	CUHK03
SOTA	-	30.0 [12]	14.9	63.9 [18]	65.8 [62]	62.9 [62]	40.8 [7]	45.9 [78]	71.8 [79]	93.0 [26]	71.8 [26]	77.7 [85]
DeiT [87]	STL(256×128)	31.5	30.1	64.2	65.1	62.3	46.3	43.1	82.1	87.9	67.9	77.5
	MTL(256×128)	40.2	38.7	65.7	66.2	62.9	53.2	46.8	76.3	90.0	68.9	78.7
ALBEF [74]	STL(256×128)	30.6	28.7	67.7	61.2	58.8	40.2	42.1	67.1	78.9	60.3	69.2
	MTL(256×128)	37.2	34.5	**68.3**	62.1	60.1	41.2	44.1	60.5	79.7	62.1	69.9
HAP [88]	STL(256×128)	30.8	31.1	63.2	66.2	62.9	45.3	44.7	79.2	90.2	70.3	79.6
	MTL(256×128)	39.2	39.8	65.1	67.1	64.3	49.2	48.7	77.3	91.7	72.3	82.2
PASS [26]	STL(256×128)	32.2	30.7	65.3	66.6	64.5	46.7	46.0	80.1	92.3	71.9	83.3
	MTL(256×128)	41.7	39.8	66.5	68.5	67.2	52.0	52.3	78.9	93.5	72.4	85.4
	STL(384×192)	33.7	33.8	65.9	68.9	64.7	48.9	49.3	**82.5**	92.7	73.7	84.1
	MTL(384×192)	**42.5**	**42.3**	67.2	**71.2**	**67.3**	**54.1**	**58.4**	81.8	**93.9**	**75.4**	**88.7**

(a), we perform ablation experiments to analyze the effect of loss weights on the identity loss and adaptive triplet loss components within the total loss during the training of the IRM model. The best results are achieved when the triplet loss weight α is set to 3. Similarly, Fig. 8 (b) illustrates the impact of different weights on the contrastive loss components within the total loss during the training of the IRM++ model. We vary the hard label contrastive loss weight coefficient from 0.5 to 20. Notably, when the weight β is set to 5, except the rank-1 result of CTCC-ReID is marginally lower compared to the β 1, all other experiments get the best results. Based on these experimental findings, we establish the default values of the triplet loss weight as 3 and the hard label contrastive loss weight as 5 for experiments across each ReID task.

Ablation on different pre-trained models To investigate the impact of different pre-trained models on model performance, we provide more results about IRM and IRM++ on the publicly released pre-trained models *i.e.,* DeiT [87], ALBEF [74], HAP [88], and PASS [26]. We validate IRM based on the task-specific evaluation setting as shown in Tab. IX and test IRM++ on the task-free evaluation setting as shown in Tab. X. We can see two conclusions from the results. First, we can see models pre-trained on human-centric images, *i.e.,* HAP [88] and PASS [26], can naturally increase the performance of person-retrieval tasks. Second, although ALBEF [74] exhibits lower performance on Trad-ReID and other image-based ReID tasks, it achieves the highest performance on the T2I-ReID task due to its vision-language pretraining knowledge. Additionally, we validate the effectiveness of IRM on 384×192 image resolution using the PASS pre-trained model. The results show a further performance improvement compared to the 256×128 results, which indicates that increasing the image size can contribute to better retrieval results.

TABLE X
MORE RESULTS OF IRM++ ON DIFFERENT PRE-TRAINED MODELS. ALL RESULTS ARE OBTAINED ON OMNIREID++ BENCHMARK TASK-FREE EVALUATION SETTING. WE TRAIN AND TEST USING THE DEFAULT IMAGE RESOLUTION OF 256×128, AND PROVIDE ADDITIONAL RESULTS OF IMAGE RESOLUTION OF 384×192. † DENOTES THAT THE TEST MODE IS VIS-TO-IR AND ‡ DENOTES IR-TO-VIS MODE ON LLCM.

IRM++		CTCC-ReID	LI-ReID	T2I-ReID	VI-ReID		CC-ReID			Trad-ReID		
		Real2	Real2	CUHK.	LLCM†	LLCM‡	LTCC	PRCC	VC-Clo.	Market1501	MSMT17	CUHK03
SOTA	-	30.0 [12]	14.9	63.9 [18]	65.8 [62]	62.9 [62]	40.8 [7]	45.9 [78]	71.8 [79]	93.0 [26]	71.8 [26]	77.7 [85]
DeiT [87]	STL(256×128)	29.3	27.9	63.1	64.4	62.1	46.1	42.9	80.1	87.2	67.7	77.4
	MTL(256×128)	32.8	34.7	66.9	66.1	62.3	52.8	45.9	75.3	89.8	68.1	78.2
ALBEF [74]	STL(256×128)	27.6	25.4	66.9	61.1	58.7	40.1	42.2	66.7	78.7	60.1	68.8
	MTL(256×128)	29.9	30.2	**68.1**	63.0	60.3	41.1	44.1	59.2	79.3	62.0	69.4
HAP [88]	STL(256×128)	28.9	30.4	63.7	65.9	62.8	44.1	43.2	79.3	90.4	70.4	78.9
	MTL(256×128)	31.2	32.1	65.0	66.2	63.9	46.3	47.6	77.1	91.2	72.4	81.9
PASS [26]	STL(256×128)	32.2	35.8	64.3	64.5	66.3	46.2	46.5	80.3	92.1	72.2	83.3
	MTL(256×128)	35.3	37.3	66.4	66.1	67.1	46.7	47.7	78.4	93.3	72.5	84.7
	STL(384×192)	32.5	36.2	64.6	64.8	67.1	46.7	46.6	**80.2**	92.7	73.2	84.0
	MTL(384×192)	**35.7**	**37.9**	67.7	**66.6**	**67.5**	**46.9**	**48.1**	78.8	**93.7**	**74.5**	**86.1**

Fig. 9. Illustration of all tasks retrieval results. We visualize the task-specific instructions on three people as examples. To clearly describe the retrieval results, we use green boxes to mark true retrieval samples and red boxes mean false matches.

F. Visualization

Visualization of retrieval results. We visualize the retrieval results of CTCC-ReID, LI-ReID, CC-ReID and Trad-ReID tasks in Fig. 9 and VI-ReID, T2I-ReID in Fig. 10. Given a query image, IRM not only retrieves the right person from the gallery but also finds specific target images of the person following the instruction. Concretely, for CTCC-ReID, IRM retrieves images of query persons wearing instructed clothes as shown in the first row. For LI-ReID, IRM effectively parses information from languages such as bag condition (*e.g.*, row 2 person #1) and clothes attribute (*e.g.*, row 2 person #2,3) to find the correct image. For CC-ReID, with the "Ignore clothes" instruction, our method focuses on biometric features and successfully retrieves the person in the case of changing clothes, *e.g.* the 4^{th} and 5^{th} images of row 3 person #1, which the clothes are different from the query image. For Trad-ReID shown in row 4, "do not change clothes" instructs IRM to pay attention to clothes, a main feature of a person's image. In this case, IRM retrieves images with the same clothes in the query image. For failed cases, *e.g.*, 5^{th} image row 1 and 4^{th} image row 2,3 of person #2, the similar features like hairstyle, body shape, and posture between retrieved images and the query image cause the mismatch of IRM. For VI-ReID, we visualize the retrieval results for both VIS-to-IR and IR-to-VIS modes. IRM can retrieve the correct cross-modality images even in low visual conditions. For T2I-ReID, when there are images of different identities in the gallery but with representations consistent with the text description (*e.g.*, row 1 3^{rd} and row 2 person 4^{th} images in Fig 10(c)), it may introduce some noise to IRM. However, IRM is still capable of indexing most of the correct results based on the given descriptions.

Visualization of attention maps. We visualize attention feature maps to understand what IRM has learned across the 6 ReID tasks. As illustrated in Fig. 11, we conduct inference for all 6 ReID settings on the same image, demonstrating that IRM learns to attend to different regions based on the instruction. The feature maps of Trad-ReID mainly emphasize holistic features such as the face and upper clothing, *e.g.*, 1^{st} to 3^{rd} attention map images of row one. For CC-ReID, the feature maps focus on clothing-irrelevant features to mitigate interference from inconsistent clothes, *e.g.*, 4^{th} to 6^{th} attention map images. In VI-ReID, the attention maps exhibit a balanced focus on human features, yet still capture regions with more

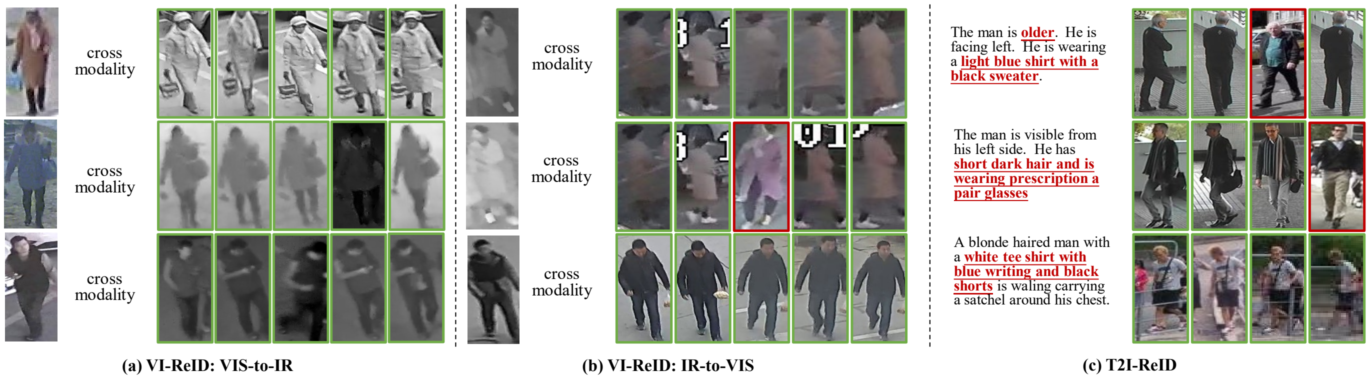


Fig. 10. Illustration of VI-ReID and T2I-ReID tasks retrieval results. We visualize both VIS-to-IR and IR-to-VIS mode results on three people in the LLCM dataset. There are about four images for each person in CUHK-PEDES, we visualize the top 4 results for T2I-ReID on the CUHK-PEDES dataset as examples. Green and red boxes mean true and false matches.

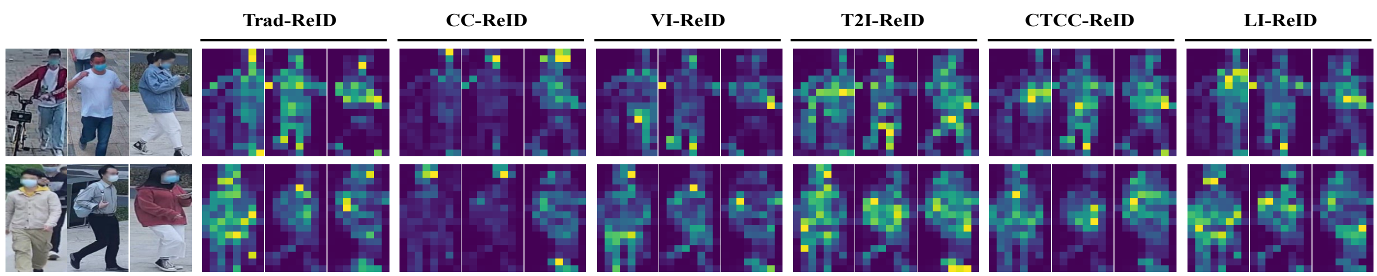


Fig. 11. Illustration of the attention maps for instruct-ReID. We conduct inference for Trad-ReID, CC-ReID, VI-ReID, T2I-ReID, CTCC-ReID, and LI-ReID tasks on the same image using our proposed IRM model. Visualizing the attention maps helps us explore how well the model comprehends various tasks.

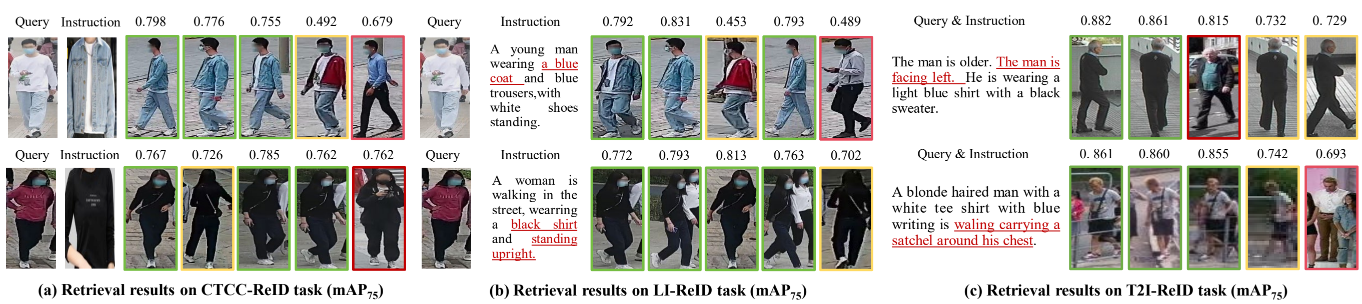


Fig. 12. Visualization of retrieval results for CTCC-ReID, LI-ReID, and T2I-ReID using the mAPτ metric with a threshold set at 0.75. Positive samples are highlighted in green boxes, negative samples in red boxes, and retrieval results with correct identities but instruction similarity below the threshold, corrected to negative samples, are marked with yellow boxes.

prominent characteristics, *e.g.*, the 1^{st} attention map image of VI-ReID example row one. In T2I-ReID and LI-ReID, the introduction of specified language descriptions leads the model to highlight certain attributes, such as backpacks and trousers, *e.g.*, 2^{nd} attention map image of LI-ReID example row two. For CTCC-ReID, the model emphasizes the upper clothing, indicating potential clothes template changing. These visualizations of attention maps qualitatively demonstrate the efficacy of our method in addressing various ReID tasks.

Visualization of mAPτ. In this subsection, we visually validate the improvement of the proposed evaluation metric over the traditional mAP metric. We visualize the indexing results on the CTCC-ReID, LI-ReID, and T2I-ReID tasks, setting the threshold to 0.75. The visualization demonstrates that mAPτ provides a more accurate measure of both identity correctness and retrieval description consistency in the retrieval results. The green box indicates correctly indexed positive samples, the red box indicates incorrectly indexed negative samples and the yellow box indicates indexed results with correct IDs but with instruction similarity below the set threshold, which are then relabeled as negative samples. For instance, in Fig 12(a) for person 1, the 4^{th} image, and in Fig 12(b) for person 1, the 3^{rd} image, although the model successfully indexed images with the same ID as the query image, it did not find the person wearing the blue jacket as required by the query instruction. In these scenarios, the proposed mAPτ evaluation metric can successfully filter out retrieval results that do not match the query instruction description. In addition to assessing whether clothing features match, this evaluation metric can also validate the ability of the ReID model to recognize fine-grained features such as body posture and accessories. For example, in Fig 12(b) for person 2, the 5^{th} image (body orientation does not match the retrieval target), and in Fig 12(c) for person 2, the 5^{th} image (unable to discern the status of the carried accessory). Setting the threshold for mAPτ is a crucial consideration, as seen in Fig 12(c) for person 1, the 4^{th} and 5^{th} images, where an excessively large threshold may classify some true positive samples as negative samples. Choosing an appropriate threshold is a topic worthy of future research.

VI. Conclusion

This proposes one unified instruct-ReID task to jointly tackle existing traditional ReID, clothes-changing ReID, clothes template based clothes-changing ReID, language-instruct ReID, visual-infrared ReID, and text-to-image ReID tasks, which holds great potential in social surveillance. To tackle the instruct-ReID task, we build a large-scale and comprehensive OmniReID++ benchmark and a generic framework with an adaptive triplet loss. We hope our OmniReID++ can facilitate future works such as unified network structure design and multi-task learning methods on a broad variety of retrieval tasks.

References

[1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *CVPR*, 2015.

[2] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017.

[3] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[4] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8371–8381, October 2021.

[5] Jianing Li, Shiliang Zhang, Qi Tian, Meng Wang, and Wen Gao. Pose-guided representation learning for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):622–635, 2019.

[6] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *CVPR*, 2021.

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022.

[8] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, 2022.

[9] Xiujun Shu, Ge Li, Xiao Wang, Weijian Ruan, and Qi Tian. Semantic-guided pixel sampling for cloth-changing person re-identification. *SPL*, 2021.

[10] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, 2021.

[11] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020.

[12] Shihua Li, Haobin Chen, Shijie Yu, Zhiqun He, Feng Zhu, Rui Zhao, Jie Chen, and Yu Qiao. Cocas+: Large-scale clothes-changing person re-identification with clothes templates. *TCSVT*, 2022.

[13] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022.

[14] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2843–2851, 2022.

[15] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022.

[16] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022.

[17] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2299–2315, 2024.

[18] 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. *arXiv preprint arXiv:2305.13653*, 2023.

[19] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.

[20] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *TOMM*, 2020.

[21] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017.

[22] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. *arXiv preprint arXiv:1712.01493*, 2017.

[23] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015.

[24] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[26] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European Conference on Computer Vision*, pages 198–214. Springer, 2022.

[27] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, Donglian Qi, and Yunfeng Yan. Instruct-reid: A multi-purpose person re-identification task with instructions, 2023.

[28] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021.

[29] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[30] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.
[31] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 3800–3808, 2017.
[32] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.
[33] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *CVPR*, 2018.
[34] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5735–5744, 2019.
[35] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1770–1782, 2019.
[36] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.
[37] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 135–153. Springer, 2016.
[38] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.
[39] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Learning invariance from generated variance for unsupervised person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[40] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin Hu. A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6649–6666, 2021.
[41] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
[42] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.
[43] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 732–748. Springer, 2016.
[44] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4894–4912, 2021.
[45] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):902–917, 2019.
[46] 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 *ICML*, 2021.
[47] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
[48] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *NeurIPS*, 2020.
[49] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020.
[50] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.
[51] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021.
[52] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021.
[53] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridgeformer: Bridging video-text retrieval with multiple choice questions. *arXiv preprint arXiv:2201.04850*, 2022.
[54] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.
[55] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020.
[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.
[57] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.
[58] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):273–285, 2020.
[59] Junsong Fan and Zhaoxiang Zhang. Memory-based cross-image contexts for weakly supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):6006–6020, 2022.
[60] Bo Ji and Angela Yao. Multi-scale memory-based video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1919–1928, 2022.
[61] 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.
[62] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023.

[63] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.

[64] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023.

[65] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[66] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPR Workshops*, 2020.

[67] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020.

[68] Kai Wang, Zhi Ma, Shiyan Chen, Jinni Yang, Keke Zhou, and Tao Li. A benchmark for clothes variation in person re-identification. *International Journal of Intelligent Systems*, 35(12):1881–1898, 2020.

[69] Mengmeng Liu, Zhi Ma, Tao Li, Yanfeng Jiang, and Kai Wang. Long-term person re-identification with dramatic appearance change: Algorithm and benchmark. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6406–6415, 2022.

[70] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*, 2023.

[71] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[72] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *TPAMI*, 2020.

[73] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[74] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[75] Wei Li and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[76] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020.

[77] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[78] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019.

[79] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *CVPR*, 2021.

[80] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.

[81] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021.

[82] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022.

[83] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021.

[84] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, 2020.

[85] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. *arXiv preprint arXiv:2303.05675*, 2023.

[86] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15050–15061, June 2023.

[87] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021.

[88] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. Hap: Structure-aware masked image modeling for human-centric perception. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

APPENDIX

VII. DETAILS OF OMNIREID++ BENCHMARK

In the main text, we briefly introduce the number of images and number of tasks of OmniReID++. For the evaluation of OmniReID++, we introduce the evaluation scenario and evaluation protocols. In this section, we present detailed information on the training dataset and evaluation dataset and discuss the ethical issues of these datasets.

A. Dataset Statistics of OmniReID++

OmniReID++ collects 15 publicly available datasets of 6 existing ReID tasks, including Traditional ReID (Trad-ReID), Clothes-Changing ReID (CC-ReID), Clothes Template Based Clothes-Changing ReID (CTCC-ReID), Visual-Infrared ReID (VI-ReID), Text-to-Image ReID (T2I-ReID) and Language-Instructed ReID (LI-ReID). As shown in Tab. XI, we perform two training scenarios based on the built benchmark **OmniReID++**: 1) Single-task Learning (STL): Every dataset is treated as a single task, which is trained and tested individually. 2) Multi-task Learning (MTL): The model is optimized by joint training of all the ReID tasks with all the training datasets. The trained model is then evaluated on different tasks with various datasets. In Trad-ReID, we utilize the widely-used MSMT17, CUHK03, and Market1501 datasets. For the CC-ReID task, we choose the widely-used PRCC, LTCC, and VC-Clothes datasets, considering the diversity of domains. Regarding the CTCC-ReID task, we label LTCC, PRCC, LaST with clothes image instructions and employ COCAS+ Real1, LTCC, PRCC, LaST for training, while COCAS+ Real2 serves as the test set. For LI-ReID, we annotate the COCAS+ Real1, PRCC, LaST, NKUP, NKUP+ and COCAS+ Real2 datasets with language instructions for training and test. In VI-ReID, we select a new and challenging low-light cross-modality dataset called LLCM. Finally, for T2I-ReID, we opt for the widely-used CUHK-PEDES and SYNTH-PEDES datasets for training, with CUHK-PEDES serving as the test set. OmniReID++ forms a total of 5,072,218 images for training and 183,038 images for test, which unites all-purpose person retrieval into one instruct-ReID task.

B. Discussion of Ethical Issues

The usage of OmniReID++ might bring several risks, such as privacy, and problematic content. We discuss these risks and their mitigation strategies in this subsection.

First, we conduct a thorough review of each dataset and guarantee that none of the ReID tasks used in our paper are withdrawn. The demographic makeup of the datasets used is not representative of the broader population but these datasets can be used for scientific experimentation.

Second, we adopt the following measures to mitigate potential security risks while adhering to the copyright policies of each dataset:

- We will NOT re-release these public datasets but will only provide the download links or webpages when we release the dataset. We do not claim copyright ownership of the original data, and anyone who wants to use the dataset should still be approved by the original assignee.
- We will NOT modify these datasets but exclusively provide visual caption annotation files for publicly available datasets. We confirm these annotations do NOT contain identification information.
- We obtained explicit permission through email correspondence for annotating the public datasets.
- Access to our annotation file links is contingent upon adherence to our scholarly and research-oriented guidelines.

Also, we provide an agreement for anyone who wants to use our OmniReID++ benchmark.

OmniReID++ Agreement

This Agreement outlines the terms and conditions governing the use of OmniReID++. By signing this agreement, the Recipient agrees to the following terms:

- The Recipient agrees the OmniReID++ does not claim Copyrights of Market1501, CUHK03, MSMT17, PRCC, VC-Clothes, LTCC, LaST, NKUP, NKUP+, COCAS+, LLCM, CUHK-PEDES, SYNTH-PEDES and they SHOULD obtain these public datasets from these data providers.
- The Recipient agrees they must comply with ALL LICENSEs of Market1501, CUHK03, MSMT17, PRCC, VC-Clothes, LTCC, LaST, NKUP, NKUP+, COCAS+, LLCM, CUHK-PEDES, SYNTH-PEDES when they use OmniReID++.
- The Recipient agrees the demographic makeup of OmniReID++ is not representative of the broader population.
- The Recipient agrees OmniReID++ should only be available for non-commercial research purposes. Any other use, in particular any use for commercial purposes, is prohibited.
- The Recipient agrees not to use the data for any unlawful, unethical, or malicious purposes.
- The Recipient agrees not to further copy, publish or distribute any portion of the OmniReID++.
- The Recipient agrees [OUR INSTITUTE] reserves the right to terminate the access to the OmniReID++ at any time.
  Name:
  Organization/Affiliation:
  Position:
  Email:
  Address:
  Address (Line2):
  City: Country:
  Signature:
  Date:

VIII. DETAILS OF LANGUAGE ANNOTATION GENERATION

In this section, we provide more details about the generation of language annotation in OmniReID++. To depict pedestrian

TABLE XI

DATASET STATISTICS OF OMNIREID++. WE SUMMARIZE BOTH THE TRAINING AND TEST SET IMAGES BASED ON THE BUILT BENCHMARK. ⋆ MEANS THE IMAGES IS THE SAME AS THE CTCC-REID TASK, WE DO NOT CALCULATE THE SAME IMAGE TWICE WHEN CALCULATING THE TOTAL NUMBER OF IMAGES IN MULTI-TASK LEARNING. †MEANS THE DATASETS ARE ANNOTATED WITH EXTRA CLOTHES TEMPLATES AND TEXT DESCRIPTIONS FOR CTCC-REID AND LI-REID TASKS. ‡MEANS THE DATASETS ARE ANNOTATED WITH EXTRA TEXT DESCRIPTIONS FOR LI-REID TASKS

Single-task learning	**Training**		**Test**	
	Dataset	Number of images	Dataset	Number of images
Trad-ReID	MSMT17 [3]	30,248	MSMT17 [3]	93,820
Trad-ReID	CUHK03 [65]	7,365	CUHK03 [65]	6,732
Trad-ReID	Market1501 [23]	12,936	Market1501 [23]	23,100
LI-ReID	COCAS+ Real1 [12]	34,469	COCAS+ Real2 [12]	14,449
CC-ReID	PRCC [24]	17,896	PRCC [24]	10,800
CC-ReID	VC-Clothes [66]	9,449	VC-Clothes [66]	9,611
CC-ReID	LTCC [67]	9,423	LTCC [67]	7,543
CTCC-ReID	COCAS+ Real1⋆ [12]	34,469⋆	COCAS+ Real2⋆ [12]	14,449⋆
VI-ReID	LLCM [62]	30,921	LLCM [62]	13,909
T2I-ReID	CUHK-PEDES [21]	28,566	CUHK-PEDES [21]	3,074
Multi-task learning	**Training**		**Test**	
	Dataset	Number of images	Dataset	Number of images
OmniReID++	MSMT17 [3] +CUHK03 [65] +Market1501 [23] +COCAS+ Real1† [12] +PRCC† [24] +VC-Clothes [66] +LTCC† [67] +LLCM [62] +CUHK-PEDES [21] +LAST† [63] +NKUP‡ [68] +NKUP+‡ [69] +SYNTH-PEDES [70]	5,072,218	MSMT17 [3] CUHK03 [65] Market1501 [23] COCAS+ Real2 [12] PRCC [24] VC-Clothes [66] LTCC [67] COCAS+ Real2⋆ [12] LLCM [62] CUHK-PEDES [21] - - -	93,820 6,732 23,100 14,449 10,800 9,611 7,543 14,449⋆ 13,909 3,074 - - -

images with more detailed descriptions, we first manually annotate OmniReID++ with local description attribute words, including clothes color, accessory, pose, etc. From a human perspective, language descriptions with sentences are much effective in describing person than simply listing attribute. Based on the attribute annotations we collect, we merge them into language annotations to provide a more comprehensive description of the person.

A. *Pedestrian Attribute Generation*

We annotate 20 attributes and 92 specific representation in words for OmniReID++, as shown in Tab. XII. The attributes are carefully selected considering a wide range of human visual characteristics from the datasets, including full-body clothing, hair color, hairstyle, gender, age, actions, posture, and accessories such as umbrellas or satchels. Though there might be more than one representation for attributes such as coat color and trousers color, only one representation corresponding to the image is selected for annotation. The attributes are annotated by professional annotators in the image level, thus the annotation file will contain more accurate and detailed description. For example, in Fig. 13 (a), although the two images are of the same identity, the difference of pedestrian angle is also annotated.

B. *Attribute-to-Language Transformation*

We provide some example from our OmniReID++ that provides sentence descriptions of individuals in images. Compared with discrete attribute words, language is more natural for consumers. To this end, we transform these attributes into multiple sentences using the Alpaca-LoRA large language model. Specifically, we ask the Alpaca-LoRA with the following sentences: "Generate sentences to describe a person. The above sentences should contain all the attribute information I gave you in the following." Annotators carefully check the generated annotations to ensure the correctness of the language instructions. Fig. 13 (a) presents the examples of the same identity and Fig. 13 (b) presents the examples of the transformation with different domains and identities.

IX. INSTRUCTION GENERATION

In our proposed instruct-ReID task, each identity is further split into query and gallery, where query set consists of query person images and clothes templates, and gallery set consists of target person images. In this section, we provide more training and test examples for Clothes template based clothes-changing ReID and Language-instructed ReID scenarios.

A. *Instructions on traditional ReID, cloth-changing ReID and visual-infrared ReID*

Following the methods in Instruct-BLIP [73], which uses GPT-4 to generate 20 different instructions for traditional ReID

TABLE XII
DETAILS OF ATTRIBUTES FROM OMNIREID++ THAT DESCRIBE PERSON IMAGES. WE SELECT 20 ATTRIBUTES AND 92 SPECIFIC REPRESENTATION WORDS CONSIDERING A WIDE RANGE OF HUMAN VISUAL APPEARANCES IN DETAIL.

Attribute	representation in words
coat color	”black coat”, ”blue coat”, ”gray coat”, ”green coat”, ”purple coat”, ”red coat”, ”white coat”, ”yellow coat”
trousers color	”black trousers”, ”blue trousers”, ”gray trousers”, ”green trousers”, ”purple trousers”, ”red trousers”, ”white trousers”, ”yellow trousers”
coat length	”agnostic length coat”, ”long sleeve coat”, ”short sleeve coat”, ” bareback coat”
trousers length	”shorts trousers”, ”skirt”, ”trousers”
gender code	”female”, ”agnostic gender”, ”male”
glass style	”without glasses”, ”with glasses”, ”with sunglasses”
hair color	”black hair”, ”agnostic color hair”, ”white hair”, ”yellow hair”
hair style	”bald hair”, ”agnostic style hair”, ”long hair”, ”short hair”
bag style	”backpack”, ”hand bag”, ”shoulder bag”, ”waist pack”, ”trolley”, ”agnostic style bag”, ”without bag”
cap style	”with hat”, ”without hat”
shoes color	”black shoes”, ”blue shoes”, ”gray shoes”, ”green shoes”, ”purple shoes”, ”red shoes”, ”white shoes”, ”yellow shoes”
shoes style	”boots”, ”leather shoes”, ”sandal”, ”walking shoes”
age	”adult”, ”child”, ”old”
person angle	”back”, ”front”, ”side”
pose	”lie”, ”pose agnostic”, ”sit”, ”stand”, ”stoop”
coat style	”business suit”, ”agnostic style coat”, ”dress”, ”jacket”, ”long coat”, ”shirt”, ”sweater”, ”t-shirt”
glove	”with glove”, ”agnostic glove”, ”without glove”
smoking	”smoking”, ”agnostic smoking”, ”without smoking”
umbrella	”with umbrella”, ”without umbrella”
uniform	”chef uniform”, ”common clothing”, ”firefighter uniform”, ”medical uniform”, ”office uniform”, ”agnostic uniform”, ”worker uniform”

(*e.g.*, We use ’do not change clothes’ to generate equivalent expressions such as ’maintain consistent clothes’ and so on), 20 different instructions for clothes-changing ReID (*e.g.*, We use ’ignore clothes’ to generate equivalent expressions such as ’change your clothes’ and so on), and 20 different instructions for visual-infrared ReID (*e.g.*, We use ’retrieve cross-modality images’ to generate equivalent expressions such as ’fetch images across different modalities’ and so on) as shown in Tab. XIII in detail. We randomly choose one from these instructions when training each mini-batch and evaluating every instruction for testing the model.

B. Text to image ReID

In T2I-ReID scenario, during the training process, both images and responding description texts are fed into IRM. We adopt a contrastive loss to align the image features and text features. To further enhance the retrieval capability of the model, we employ a classifier to determine whether an inputted image-text pair is positive or negative. Specifically, we retain the original image-text pairs as positives and form negative pairs by matching text features with unrelated image features before inputting them into the attention module. In the inference stage, the query is the describing sentences, and the image features and query features are extracted separately. given a query text feature, we rank all the test gallery image features based on their similarity with the text. We select the top 128 image features and pair them with the query text feature. These pairs are then input into the attention module, further utilizing the matching scores to rank these images. The search is deemed to be successful if top-K images contain any corresponding identity.

C. Clothes Template Based Clothes-Changing ReID

In CCTC-ReID scenario, as shown in Fig. 14, the instruction is a clothes template for a query image, clothes regions cropped by a detector from themselves are treated as instruction for gallery images. During the training process, the biometric feature and clothes feature are extracted from person images and instructions. In the inference stage, the model should retrieve images of the same person wearing the provided clothes.

D. Language-instructed ReID

We provide more examples for LI-ReID as shown in Fig. 15. Similar to CTCC-ReID, the instruction for gallery images is several sentences describing pedestrian attributes. We randomly select the description languages from the corresponding

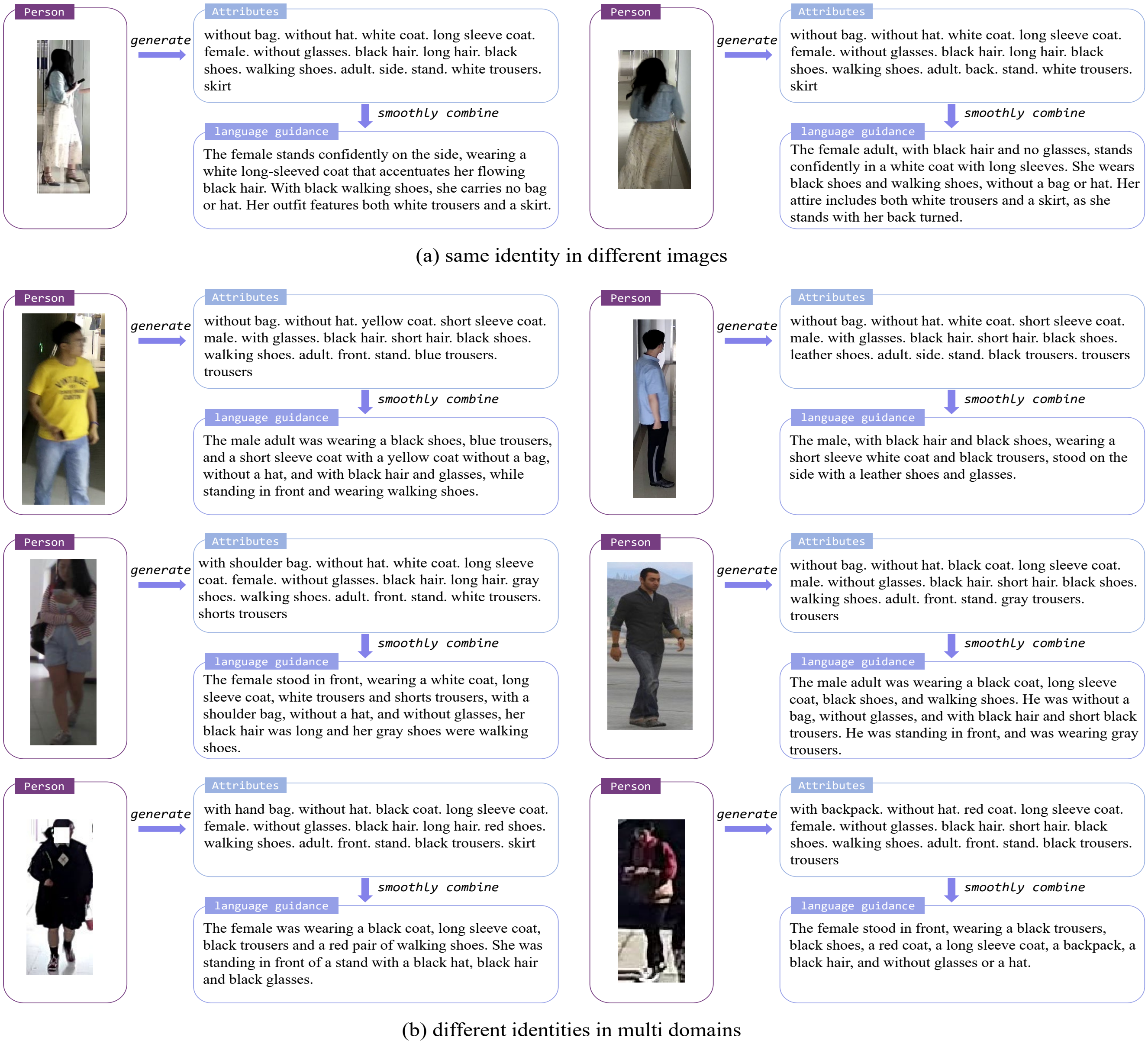


Fig. 13. We first generate attributes for a person and then transform attributes into sentences by a large language model. **(a)** The attributes and language guidance of same identity. **(b)** More instances of attributes-to-language transformation with different domains and identities.

person images in gallery and provide to query images as instruction. The model is required to retrieve images of the same person following the provided sentences.

Fig. 14. Examples of CTCC-ReID task for training and test. Each identity is further split into query and gallery, where query set consists of query person images and clothes templates as instructions, and gallery set consists of target person images and cropped clothes images as instructions.

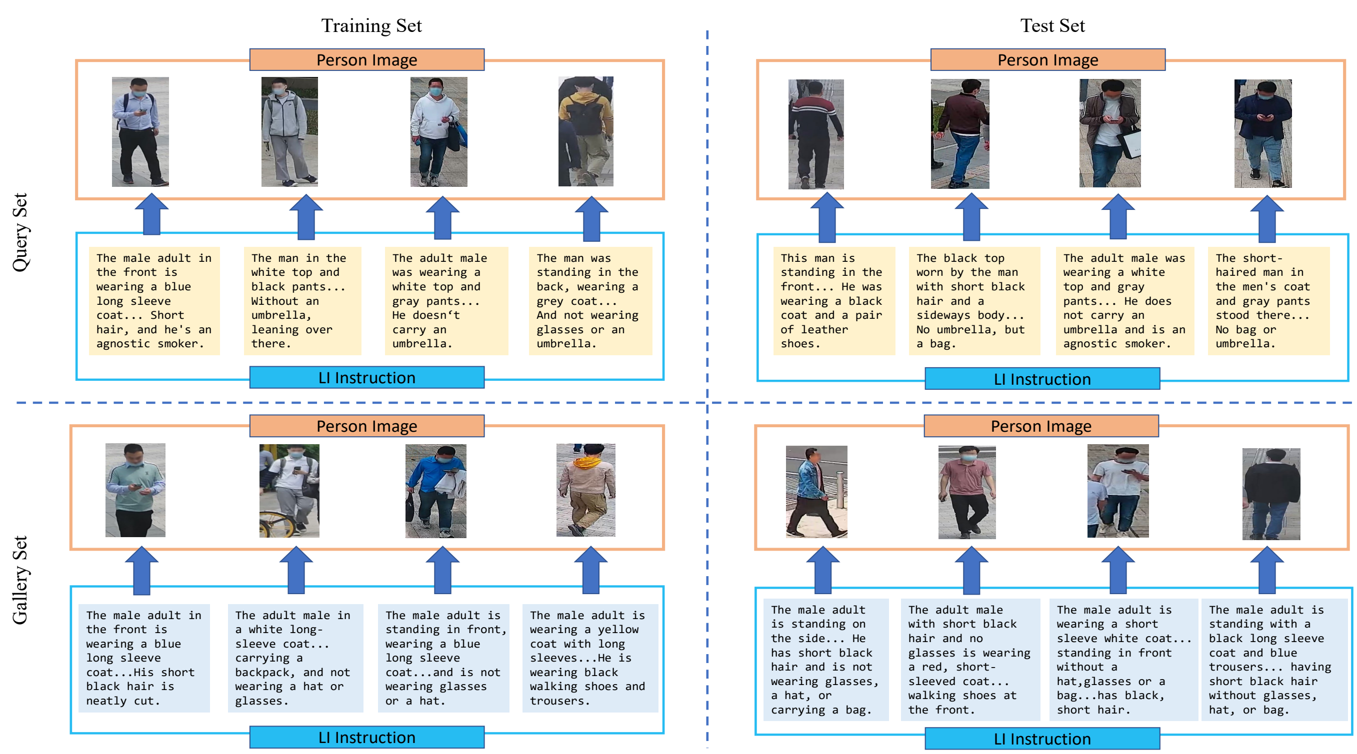


Fig. 15. Examples of LI-ReID task for training and test. Each identity is further split into query and gallery, where query set consists of query person images and language instructions (randomly selected from gallery corresponding to the person), and gallery set consists of target person images and description language for themselves as instructions.

TABLE XIII
DETAILS OF INSTRUCTIONS FOR TRAD-REID, CC-REID, AND VI-REID IN OMNIREID++. WE USE GPT-4 TO GENERATE 20 DIFFERENT EXPRESSIONS FOR 'DO NOT CHANGE CLOTHES' AS INSTRUCTIONS FOR TRADITIONAL REID AND SIMILARLY, WE GENERATED 20 DIFFERENT EXPRESSIONS OF 'IGNORE CLOTHES' FOR CLOTHES-CHANGING REID AND 20 DIFFERENT EXPRESSIONS OF 'RETRIEVE CROSS-MODALITY IMAGES' FOR VISUAL-INFRARED REID.

ReID task	instruction representation
Trad-ReID	"do not change clothes", "maintain consistent clothes", "keep original clothes", "preserve current clothes", "retain existing clothes", "wear the same clothes", "stick with your clothes", "don't alter your clothes", "no changes to clothes", "unchanged outfit", "clothes remain constant", "no clothing adjustments", "steady clothing choice", "clothing remains unchanged", "consistent clothing selection", "retain your clothing style", "clothing choice remains", "don't swap clothes", "maintain clothing selection", "clothes stay the same"
CC-ReID	"change your clothes", "swap outfits", "switch attire", "get into a different outfit", "try on something new", "put on fresh clothing", "dress in alternative attire", "alter your outfit", "wear something else", "don a different ensemble", "trade your garments", "shift your wardrobe", "exchange your clothing", "update your attire", "replace your outfit", "clothe yourself differently", "switch your style", "update your look", "put on a new wardrobe", "ignore clothes"
VI-ReID	"retrieve cross-modality images", "fetch images across different modalities", "collect images from various modalities", "obtain images spanning different modalities", "retrieve images from diverse modalities", "gather images across modalities", "access images across different modalities", "acquire images spanning various modalities", "extract images from different modalities", "retrieve images across multiple modalities", "fetch images from distinct modalities", "collect images across various modalities", "access images from different modalities", "obtain images from diverse modalities", "gather images spanning different modalities", "extract images from various modalities", "retrieve images across varied modalities", "obtain images from distinct modalities", "access images across multiple modalities", "collect images spanning diverse modalities"