# Instruct-ReID: A Multi-purpose Person Re-identification Task with Instructions

**Weizhen He**[1†] **Yiheng Deng**[1] **Shixiang Tang**[2,3*] **Qihao Chen**[4]
**Qingsong Xie**[5] **Yizhou Wang**[2] **Lei Bai**[2] **Feng Zhu**[3] **Rui Zhao**[3,6]
**Wanli Ouyang**[2] **Donglian Qi**[1] **Yunfeng Yan**[1*]
[1]Zhejiang University [2]Shanghai AI Laboratory
[3]SenseTime Research [4]Liaoning Technical University [5]Shanghai Jiao Tong University
[6]Qing Yuan Research Institute, Shanghai Jiao Tong University
hewz@zju.edu.cn, yvonnech@zju.edu.cn

## 1. More results on different pre-trained models

In this section, we provide more results about IRM on the publicly released pre-trained models *i.e.,* DeiT [12], ALBEF [6], HAP [16], and PASS [19] in Tab. 1. We can see two conclusions from Tab. 1. First, we can see models pre-trained on human-centric images, *i.e.,* HAP [16] and PASS [19], can naturally increase the performance of person-retrieval tasks. Second, although ALBEF [6] 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 (384) image resolution using the PASS pre-trained model. The results show a further improvement in performance compared to the 256×128 results, which indicates that increasing the image size can contribute to achieving better retrieval results.

## 2. Visualizaton

We visualize the retrieval results of CTCC-ReID, LI-ReID, CC-ReID and Trad-ReID tasks in Fig. 1 and VI-ReID, T2I-ReID in Fig. 2. 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 4th and 5th 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 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 3rd and row 2 person 4th images in Fig 2(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.

## 3. Details of OmniReID

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.

### 3.1. Dataset Statistics of OmniReID

OmniReID collects 12 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. 2, 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-

Table 1. More results of IRM on different pre-trained models. We train and test using the default image resolution of 256×128 (256), with (384) indicating an 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 [8] | 14.9 | 63.9 [1] | 65.8 [17] | 62.9 [17] | 40.8 [3] | 45.9 [5] | 71.8 [4] | 93.0 [19] | 71.8 [19] | 77.7 [11] |
| DeiT [12] | STL(256) | 31.5 | 30.1 | 64.2 | 65.1 | 62.3 | 46.3 | 43.1 | 82.1 | 87.9 | 67.9 | 77.5 |
| | MTL(256) | 40.2 | 38.7 | 65.7 | 66.2 | 62.9 | 53.2 | 46.8 | 76.3 | 90.0 | 68.9 | 78.7 |
| ALBEF [6] | STL(256) | 30.6 | 28.7 | 67.7 | 61.2 | 58.8 | 40.2 | 42.1 | 67.1 | 78.9 | 60.3 | 69.2 |
| | MTL(256) | 37.2 | 34.5 | **68.3** | 62.1 | 60.1 | 41.2 | 44.1 | 60.5 | 79.7 | 62.1 | 69.9 |
| HAP [16] | STL(256) | 30.8 | 31.1 | 63.2 | 66.2 | 62.9 | 45.3 | 44.7 | 79.2 | 90.2 | 70.3 | 79.6 |
| | MTL(256) | 39.2 | 39.8 | 65.1 | 67.1 | 64.3 | 49.2 | 48.7 | 77.3 | 91.7 | 72.3 | 82.2 |
| PASS [19] | STL(256) | 32.2 | 30.7 | 65.3 | 66.6 | 64.5 | 46.7 | 46.0 | 80.1 | 92.3 | 71.9 | 83.3 |
| | MTL(256) | 41.7 | 39.8 | 66.5 | 68.5 | 67.2 | 52.0 | 52.3 | 78.9 | 93.5 | 72.4 | 85.4 |
| | STL(384) | 33.7 | 33.8 | 65.9 | 68.9 | 64.7 | 48.9 | 49.3 | **82.5** | 92.7 | 73.7 | 84.1 |
| | MTL(384) | **42.5** | **42.3** | 67.2 | **71.2** | **67.3** | **54.1** | **58.4** | 81.8 | **93.9** | **75.4** | **88.7** |


Figure 1. Illustration of all tasks retrieval results. We visualize the task-specific instructions on three people as examples. Green and red boxes mean true and false matches.

ReID task, we label LTCC with clothes image instructions and employ COCAS+ Real1, LTCC for training, while COCAS+ Real2 serves as the test set. For LI-ReID, we annotate the COCAS+ Real1 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 4973044 images for training and 183038 images for test, which unites all-purpose person retrieval into one instruct-ReID task.

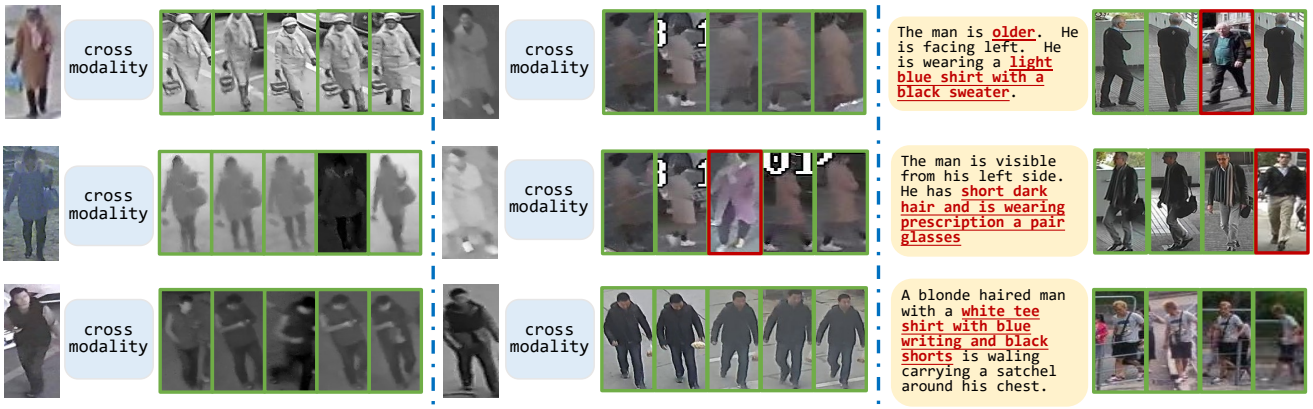
### 3.2. Discussion of Ethical Issues

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.




**(a) VI-ReID: VIS-to-IR** **(b) VI-ReID: IR-to-VIS** **(c) T2I-ReID**

Figure 2. 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.

Table 2. 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 LI-ReID task, we do not calculate the same image twice when calculating the total number of images in Multi-task Learning. ‡means the LTCC with extra clothes templates for CTCC task.

| **Single-task learning** | **Training** | | **Test** | |
|---|---|---|---|---|
| | Dataset | Number of images | Dataset | Number of images |
| Trad-ReID | MSMT17 [14] | 30,248 | MSMT17 [14] | 93,820 |
| Trad-ReID | CUHK03 [9] | 7,365 | CUHK03 [9] | 6,732 |
| Trad-ReID | Market1501 [18] | 12,936 | Market1501 [18] | 23,100 |
| LI-ReID | COCAS+ Real1 [8] | 34,469 | COCAS+ Real2 [8] | 14,449 |
| CC-ReID | PRCC [15] | 17,896 | PRCC [15] | 10,800 |
| CC-ReID | VC-Clothes [13] | 9,449 | VC-Clothes [13] | 9,611 |
| CC-ReID | LTCC [10] | 9,423 | LTCC [10] | 7,543 |
| CTCC-ReID | COCAS+ Real1 [8] | 34,469† | COCAS+ Real2 [8] | 14,449† |
| VI-ReID | LLCM [17] | 30,921 | LLCM [17] | 13,909 |
| T2I-ReID | CUHK-PEDES [7] | 28,566 | CUHK-PEDES [7] | 3,074 |
| **Multi-task learning** | **Training** | | **Test** | |
| | Dataset | Number of images | Dataset | Number of images |
| OmniReID | MSMT17 [14] +CUHK03 [9] +Market1501 [18] +COCAS+ Real1 [8] +PRCC [15] +VC-Clothes [13] +LTCC [10] +COCAS+ Real1 [8] +LLCM [17] +CUHK-PEDES [7] +LTCC‡ [10] +SYNTH-PEDES [20] | 4,973,044 | MSMT17 [14] CUHK03 [9] Market1501 [18] COCAS+ Real2 [8] PRCC [15] VC-Clothes [13] LTCC [10] COCAS+ Real2 [8] LLCM [17] CUHK-PEDES [7] - - | 93,820 6,732 23,100 14,449 10,800 9,611 7,543 14,449† 13,909 3,074 - - |

- 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, 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, 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:

# 4. Details of Language Annotation Generation

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

## 4.1. Pedestrian Attribute Generation

We annotate 20 attributes and 92 specific representation in words for OmniReID, as shown in Tab. 3. 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. 3 (a), although the two images are of the same identity, the difference of pedestrian angle is also annotated.

## 4.2. 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. 3 (a) presents the examples of the same identity and Fig. 3 (b) presents the examples of the transformation with different domains and identities.

# 5. 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.

## 5.1. Instructions on traditional ReID, cloth-changing ReID and visual-infrared ReID

Following the methods in Instruct-BLIP [2], which uses GPT-4 to generate 20 different instructions for traditional ReID (*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

Table 3. 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” |

expressions such as ’fetch images across different modalities’ and so on) as shown in Tab. 4 in detail. We randomly choose one from these instructions when training each mini-batch and evaluating every instruction for testing the model.

### 5.2. Text to image ReID

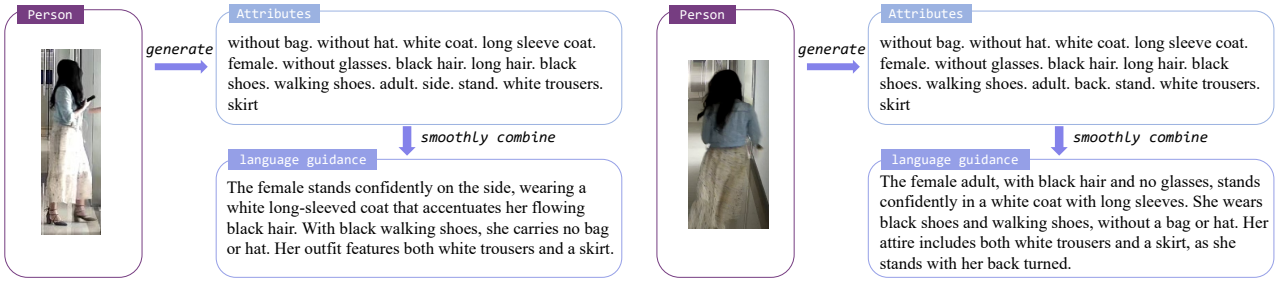
In T2I-ReID scenario, during the training process, both images and responding description texts are fed into IRM. We adopt a contrastive loss to align the image features and text features. To further enhance the retrieval capability of the model, we employ a classifier to determine whether an inputted 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.

### 5.3. Clothes Template Based Clothes-Changing ReID

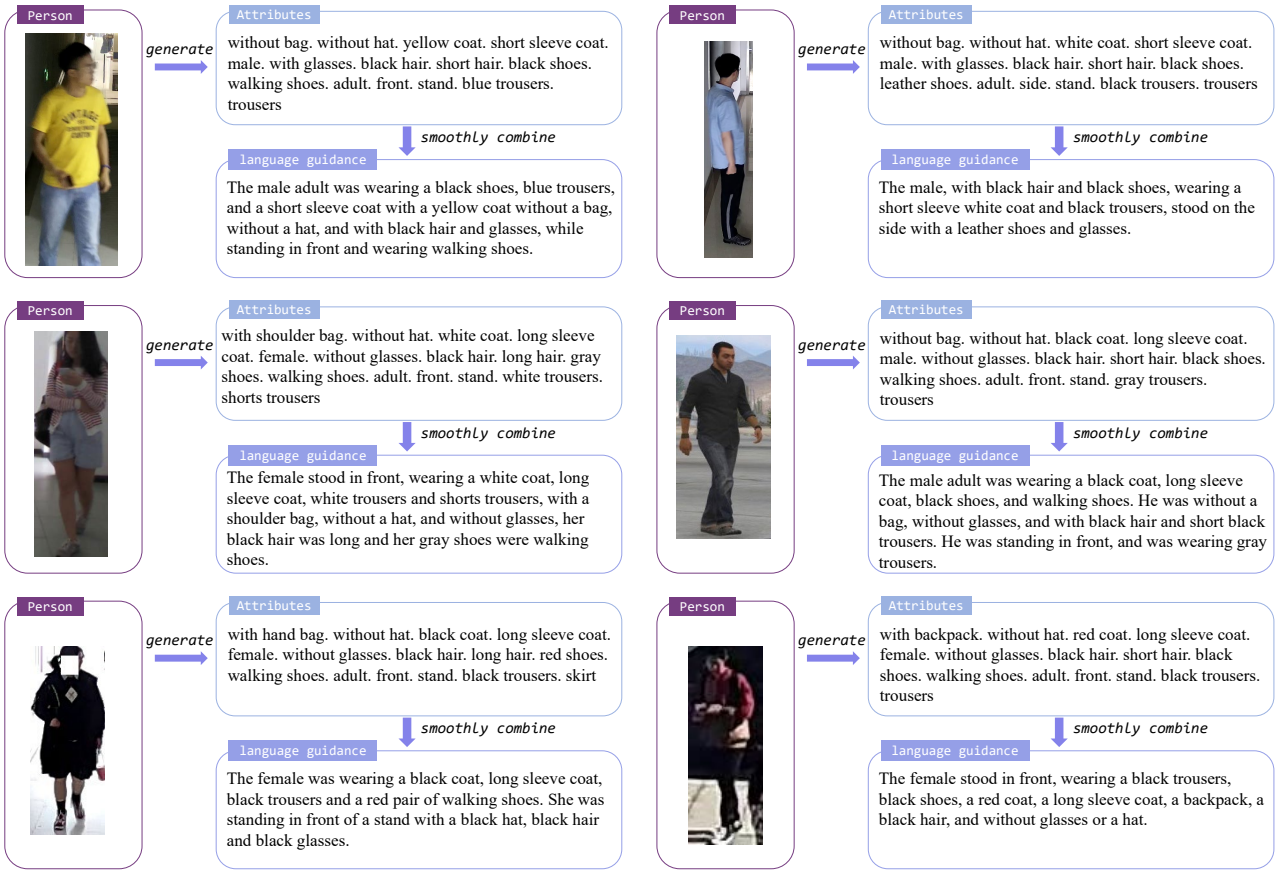
In CCTC-ReID scenario, as shown in Fig. 4, the instruction is a clothes template for a query image, clothes regions cropped by a detector from themselves are treated as instruction for gallery images. During the training process, the biometric feature and clothes feature are extracted from person images and instructions. In the inference stage, the model should retrieve images of the same person wearing the provided clothes.

### 5.4. Language-instructed ReID

We provide more examples for LI-ReID as shown in Fig. 5. Similar to CTCC-ReID, the instruction for gallery images is several sentences describing pedestrian attributes. We randomly select the description languages from the corresponding 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.




(a) same identity in different images




(b) different identities in multi domains

Figure 3. 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.

Table 4. 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" |


Figure 4. 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.


Figure 5. 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.

## References


[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv preprint arXiv:2305.13653*, 2023. 2

[2] 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. 4

[3] 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. 2

[4] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *CVPR*, 2021. 2

[5] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019. 2

[6] 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. 1, 2

[7] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 3

[8] 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. 2, 3

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3

[10] 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. 3

[11] 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. 2

[12] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 1, 2

[13] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPR Workshops*, 2020. 3

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

[15] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 3

[16] 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. 1, 2

[17] 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. 2, 3

[18] 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. 3

[19] 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. 1, 2

[20] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*, 2023. 3