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

A. Datasets

To verify the effectiveness and superiority of ICL, we exploits four widely-used TIReID datasets to conduct experiments, including three coarse-grained datasets and one fine-grained dataset. A detailed introduction of these datasets is given as follows:

- **CHUK-PEDES** [3] serves as the inaugural large-scale benchmark specifically focused on TIReID. This dataset comprises 40,206 images and 80,412 accompanying text descriptions for 13,003 unique identities. In our experiments, we adhere to the official data split: 11,003 identities are used for training, 1,000 identities for validation, and the remaining 1,000 identities for testing. The average length of all texts is 23.5.
- **ICFG-PEDES** [1] is a widely recognized benchmark derived from the MSMT17 dataset [6]. It comprises 54,522 images from 4,102 unique persons, with each image accompanied by a textual description. In line with the data split adopted by most TIReID methods [2, 5], we use a training set containing 3,102 identities and a validation set with 1,000 identities. Due to the absence of a test set, we use the validation performance as the reported performance. The average length of all texts is 37.2.
- **RSTPReid** [7] is an innovative benchmark dataset derived from the MSMT17 dataset [6], designed specifically for TIReID. This dataset consists of 20,505 images from 4,101 unique identities, with each person has five distinct images. Each image is further paired with two descriptive text annotations. Following the official data partitioning, we utilize 3,701 identities for training, 200 identities for validation, and the remaining 200 identities for testing. The average length of all texts is 26.5.
- **UFine6926** [8] is a recently constructed high-quality fine-grained TIReID dataset, which contains 26,206 images from 6,926 identities, and each image has two ultra-granular text descriptions. The average length of all texts is 80.8, which is much larger than that of the three existing coarse-grained datasets. In our experiments, we follow the official data split, *i.e.*, 4,926 identities for training and 2,000 for testing.

B. Training with RDA

To verify the effectiveness of our ICL, we utilize RDE [4] as our TIReID investigated method. In this section, we detail how we use the RDA strategy to enhance the training of RDE, improving cross-modal learning. We briefly review RDE to illustrate our training process. RDE is a robust method for TIReID that mitigates the negative impact of noisy correspondences in training datasets through dual-granular embeddings and robust similarity learning. First, RDE uses the proposed Confident Consensus Division (CCD) mechanism to partition the correspondence labels of each training image-text pair, which we denote as $\hat{l}_{q,v}$ for image v and text q. Then, RDE guides the training through the Triplet Alignment Loss (TAL), *i.e.*, the loss $\mathcal{L}_m$ is:

$$\mathcal{L}_m = \sum_{i=1}^{K} \hat{l}_{q_i,v_i} \left(\mathcal{L}^b(v_i, q_i) + \mathcal{L}^t(v_i, q_i) \right), \tag{1}$$

where $\mathcal{L}^b(v_i, q_i)$ and $\mathcal{L}^t(v_i, q_i)$ are the TAL losses for dual-granular embeddings, *i.e.*, basic global Embedding and token selection embedding. To apply RDA to RDE, we reorganize the augmented text ($\check{q}$) for each text (q) on each training epoch. Then, we compete the augmentation loss $\mathcal{L}_a$ using augmented texts, *i.e.*,

$$\mathcal{L}_a = \sum_{i=1}^{K} \hat{l}_{\check{q}_i,v_i} \left(\mathcal{L}^b(v_i, \check{q}_i) + \mathcal{L}^t(v_i, \check{q}_i) \right). \tag{2}$$

Finally, the loss for mixed training is

$$\mathcal{L} = \mathcal{L}_m + \gamma \mathcal{L}_a, \tag{3}$$

where $\gamma = \max(0.1, 1 - 0.03 \times e)$ is a balance factor used to control the contribution of enhanced text. The decreasing strategy is to let the model gradually reduce the attention paid to the augmentation text during training, which prevents the model from focusing too much on the augmented texts, thus affecting the learning of the original text domain. Note that, apart from adding the loss for augmentation text, we don't modify any settings of RDE.

```
System instruction:
You are a helpful assistant.
Prompt:
Can this text accurately describe the image?

Text: {Caption}

Answer "Yes" or "No".
Input:
Caption; Image encoded in base64.
```

Figure 1. The prompt template $\mathcal{T}_{\text{loc}}$.

```
System instruction:
You are a helpful assistant.
Prompt:
According to the pedestrian image, answer the following questions one by one:

1. Is this person male or female?
2. What hairstyle does the person have, such as hair length and color?
3. What is this person wearing on his upper body? If clearly visible, what are the color,
type, and sleeve length?
4. What are the characteristics of this person's pants? If clearly visible, what are the
color, type, and trouser leg length?
5. Does this person have any patterns on his/her clothes or pants?
6. What are the characteristics of this person's shoes? If clearly visible, what are the
color and style?
7. Does this person wear glasses? If clearly visible, what are the color and style?
8. Is this person wearing a scarf? If clearly visible, what are the color and style?
9. Does this person have something in his/her hand? If so, what is it and what color is
it?
10. Does this person carry a backpack? If clearly visible, what are the color and style?
11. Does this person wear a hat? If clearly visible, what are the color and style?
12. Is this person wearing a belt or waistband?
13. What is this person doing?
14. What is the background?
15. Are there other people in the background of this person?
Input:
Image encoded in base64.
```

Figure 2. The prompt template $\mathcal{T}_{\text{vaq}}$.

C. Prompt templates

In this section, we describe all the prompt templates used by ICL to provide a clear understanding of the inputs to MLLMs. These template prompt functions includes $\mathcal{T}_{\text{loc}}$, $\mathcal{T}_{\text{vqa}}$, $\mathcal{T}_{\text{aggr}}$.

D. SFT for Anchor Location

In this section, we mainly introduce the supervised fine-tuning on MLLMs using LoRA to improve Anchor Location. To save fine-tuning costs, we use the CCD of RDE to obtain convincing data from the training sets of the four benchmarks and randomly select partial data to construct the SFT dataset, *i.e.*, 40,000 pairs for CHUK-PEDES, 40,000 pairs for CHUK-PEDES, 15,000 pairs for ICFG-PEDES, 15,000 pairs for RSTPReid, and 15,000 pairs for UFine6926. Then, for each dataset, we use the pre-trained cross-modal model of RDE to obtain the Top-10 image with different person ID as input negative sample image for each text by similarity ranking. Finally, we get an SFT dataset with 170,000 records. We fine-tune Qwen2-VL-7B-Instruct on two 3090 GPUs, which takes about 11 hours.

E. More Retrieved Results

In this section, we provide more retrieved results for a comprehensive qualitative visualization. These results are obtained from the trained model on the CUHK-PEDES

```
System instruction:
You are a helpful assistant.
Prompt:
Aggregate the following subtexts into continuous and concise text sentences.

Example:
Subtexts: ['The person is wearing a black jacket with a white stripe on the sleeve.', 'The
person is male.', 'The person has short brown hair.', 'The person is wearing a sleeveless
striped shirt and a green tank top.', 'The person is wearing green pants.', 'The person
is wearing green pants.', 'The person is wearing a red scarf around their neck.', 'The
person is wearing a red hat on their head.', 'The background is an outdoor area with some
structures and other people.']
Output: The man has short brown hair and is wearing a black jacket with a white stripe on
the sleeve, a sleeveless striped shirt, a green tank top, green pants, a red scarf around
his neck, and a red hat. The background features an outdoor area with some structures and
other people.

Now let's get started.
Subtexts: [{Raw caption},{LIST[0]},···]

Output aggregated sentences without any explanation.
Input:
Raw caption; LIST of one-by-one answers.
```

Figure 3. The prompt template $\mathcal{T}_{\text{aggr}}$.

dataset. From the examples, as shown in Figure 4, we can see that the external guidance through TUI adds more details to the raw queries, such as background, belongings, details of clothing, *etc*., which can effectively improve the overall ranking for better retrieval. However, it cannot be ignored that TUI may introduce noisy texts due to the hallucinations of MLLMs. For example, in the last example, MLLM thinks the man is wearing a black vest, but it is actually a black backpack. For this reason, while obtaining external guidance through MLLM's interactions, resisting the introduction of noise is also an important point that must be considered in future research. But in general, more details about the person undoubtedly bring better retrieval results.

F. The efficiency analysis

In this section, we report the average in-process time per query to investigate the efficiency of THI, as shown in Table 1. The 'Time' column refers to the average time to process interactions with each query, which is user-friendly and only 0.25∼0.32 seconds per text query.

Benchmarks	CUHK-PEDES	ICFG-PEDES	RSTPReid	UFine6926
Time/query	0.2522	0.2824	0.3135	0.3157

Table 1. The average interaction time required for each query.

References

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

[2] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 1

[3] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017. 1

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

[5] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer, 2022. 1

[6] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 1

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

[8] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-

Raw: The pedestrian with short, dark hair wears a white, short sleeve shirt with gray shorts.
Merged: The man with short, dark hair is wearing a white, short-sleeved polo shirt with light green, knee-length shorts, light-colored, open-toed sandals, and no accessories. He is walking on a sidewalk with a set of stairs in the background.

Raw: A young woman in a bright green tshirt and black jeans is walking down the stairs
Merged: The young woman, with long, straight green hair, is wearing a bright green t-shirt and black jeans, walking down the stairs. She has a patterned shirt, gray pants, white shoes with a light-colored sole, and a belt.

Raw: The man is wearing a dark colored shirt white sneakers dark jeans and is carrying a book bag both his arm at down near his sides
Merged: The man is wearing a black shirt with short sleeves, black pants with a regular fit, white sneakers, and carrying a black backpack.

Raw: A woman wearing a white bacon and eggs t-shirt, a floral pring skirt and a pair of shoes.
Merged: The woman has long blonde hair and is wearing a white bacon and eggs t-shirt with a floral print skirt and white shoes.

Raw: The man has white curly hair and is wearing a beige short sleeved collared shirt and dark pants.
Merged: The man has short, light brown hair and is wearing a beige short-sleeved collared shirt and dark pants. He is holding a white plastic bag and walking in an outdoor setting with a metal structure in the background.

Raw: A woman with shoulder length black hair is wearing a white shirt and light blue jeans
Merged: The woman has long black hair and is wearing a light-colored top, likely a cardigan or sweater, with short sleeves, light blue jeans, white shoes, and a black handbag.

Raw: The man is wearing a white t shirt and denim shorts. He has dark hair and is carrying a duffle bag on his back.
Merged: The man is wearing a white t-shirt with a black vest, blue shorts, black and yellow shoes, and a hat. He has short hair, is carrying a black backpack, and is holding a white object in his hand.

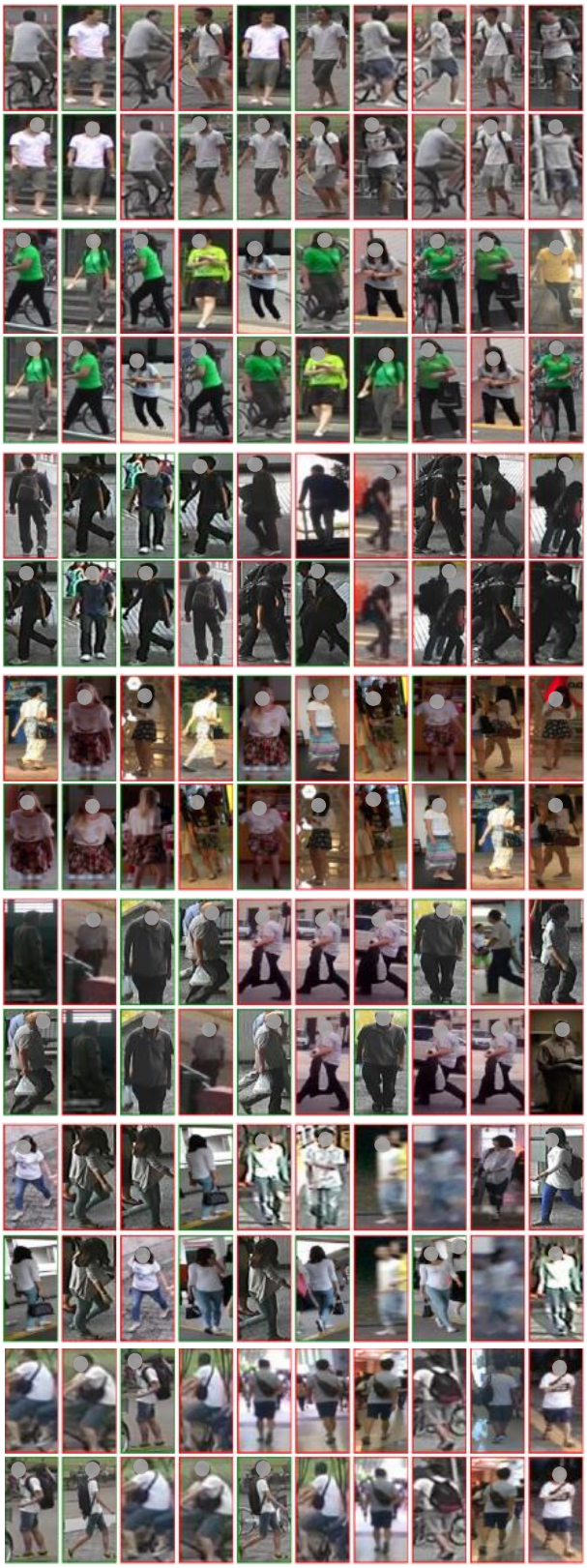

Figure 4. Top-10 retrieved results on CUHK-PEDES dataset between ICL (the first row) and ICL with TUI (the second row). All face areas of people in images are masked for privacy and security.

fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024. 1