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

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, Peng Hu*
College of Computer Science, Sichuan University

GitHub: https://github.com/QinYang79/ICL

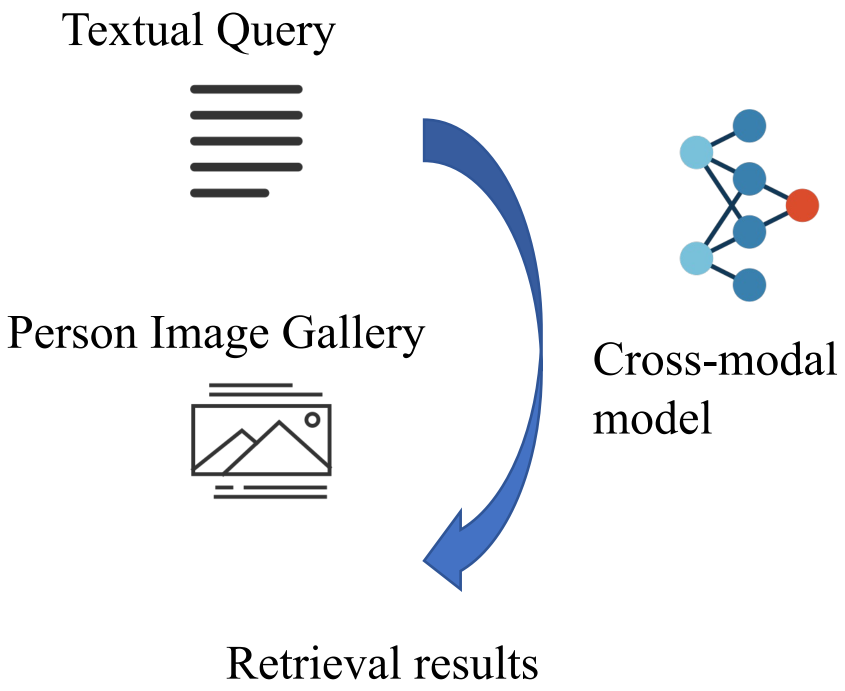
CVPR Nashville JUNE 11-15, 2025

Background

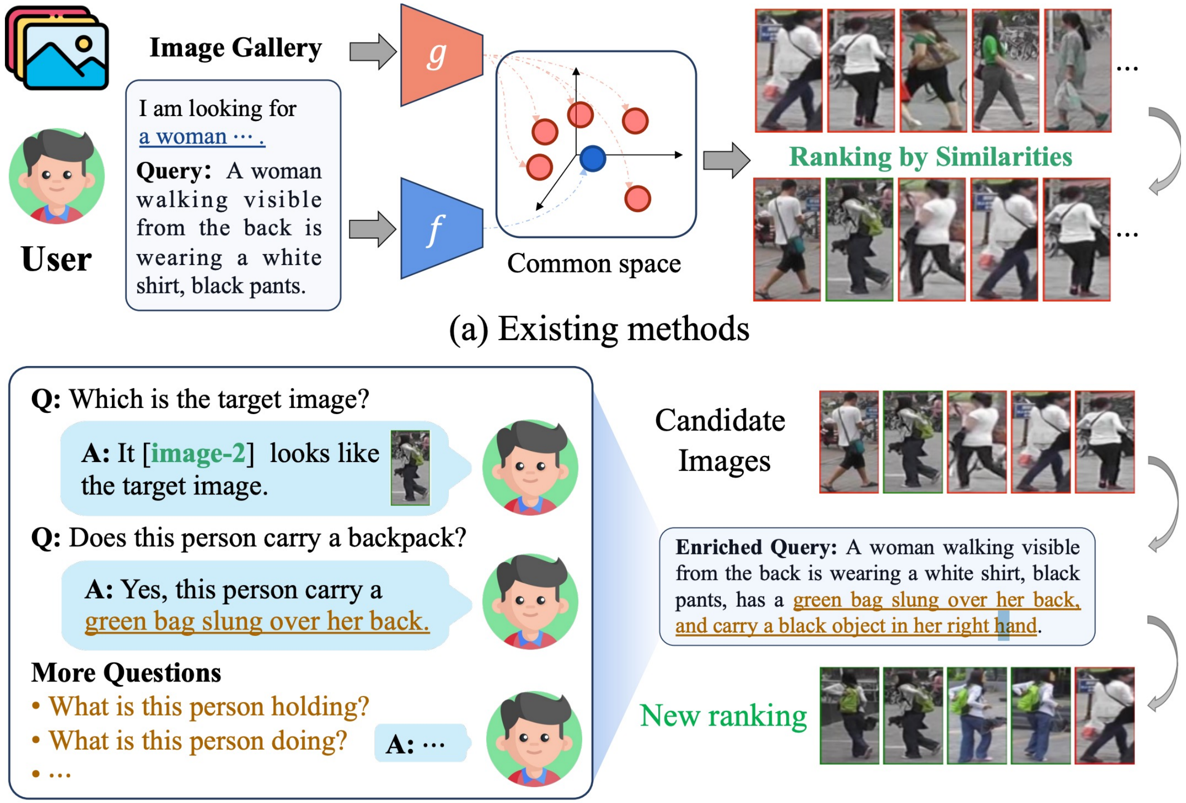
Basical definition for Text-to-Image Person Re-identification (TIReID)

(a) A woman walking visible from the back is wearing a white shirt, black pants and has a green bag slung over her back and carrying a black object in her right hand.

(b) The pedestrian with long, dark hair carries a backpack. She wears a loose top, denim bottoms, and sandals.



Observation



(a) Existing methods

(b) Interactive TIReID

- Offline model
- Knowledge limitations

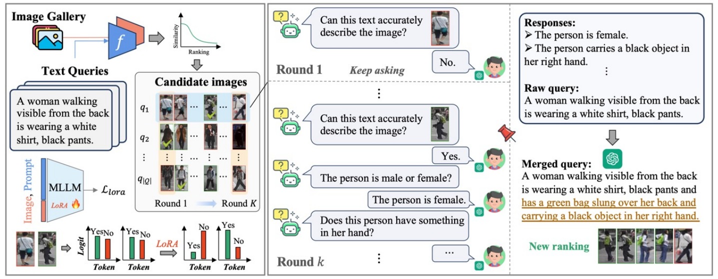
Difficulty distinguishing challenging candidate images

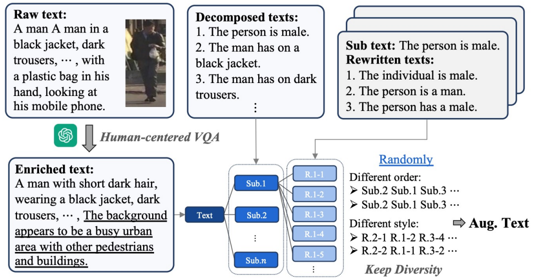
- Transfer external knowledge of MLLM into offline models;
- Empower existing methods to handle dynamic queries

Method

To overcome intrinsic limitations, we propose an Interactive Cross-modal Learning framework (ICL), which leverages human-centered interaction to enhance the discriminability of text queries through external multimodal knowledge. ICL consists of two core components

- **Test-time Humane-centered Interaction (THI):** THI performs visual question answering focus ed on human characteristics, facilitating multi-round inter-actions with a multimodal large langu-age model (MLLM) to align query intent with latent target images.
- **Reorganization Data Augmentation (RDA):** RDA is proposed based on information enrichme-nt and diversity enhancement to enhance query discrimen-ability by enriching, decomposing, and reorganizing text descriptions.





Method-THI

Steps:

1. **Anchor localization**

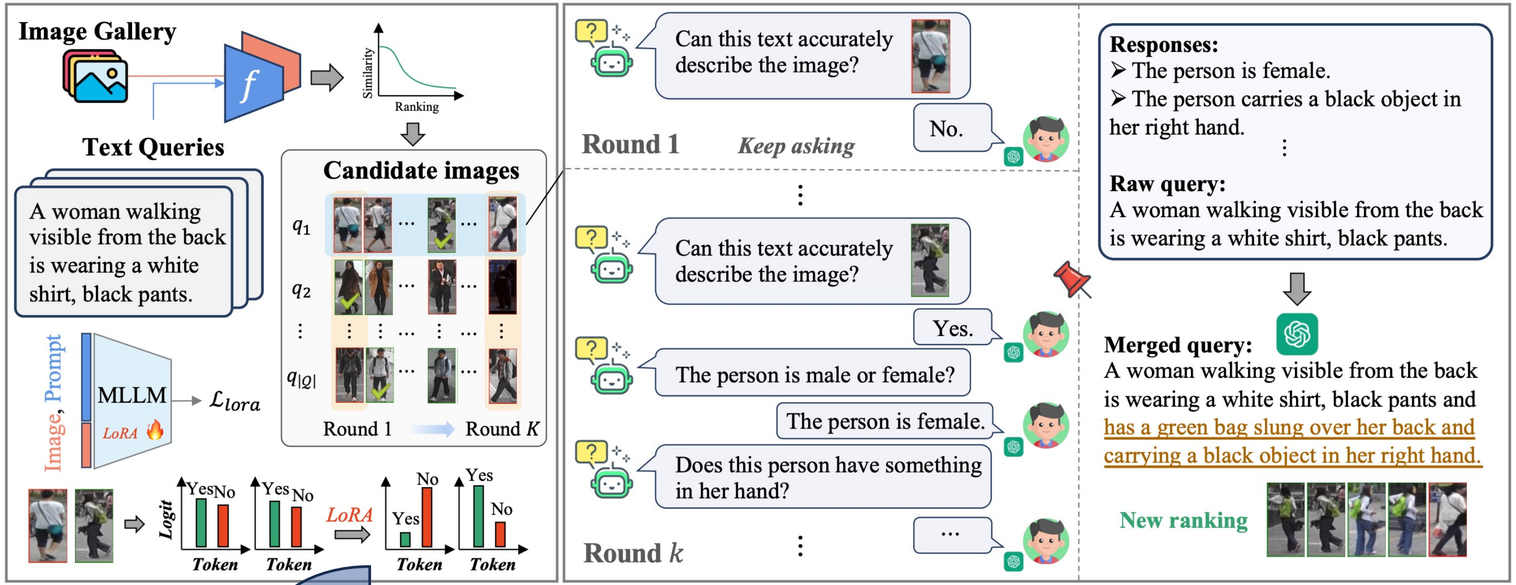
$$a^{q}_{\hat{v}_k} = \mathcal{M}(\mathcal{T}_{\text{loc}}(q, \hat{v}_k))$$

2. **Human-centered VQA**

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

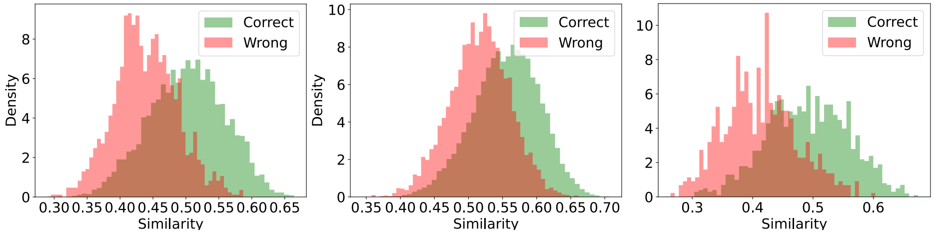
$$\hat{q} = \mathcal{M}(\mathcal{T}_{\text{aggr}}(r_{\bar{v}}, q))$$

3. **Efficient re-ranking**



$$\mathcal{Z} = \{\mathcal{Z}^+, \mathcal{Z}^-\}$$

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



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

Method-THI

Algorithm 1 The interaction process of our THI

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

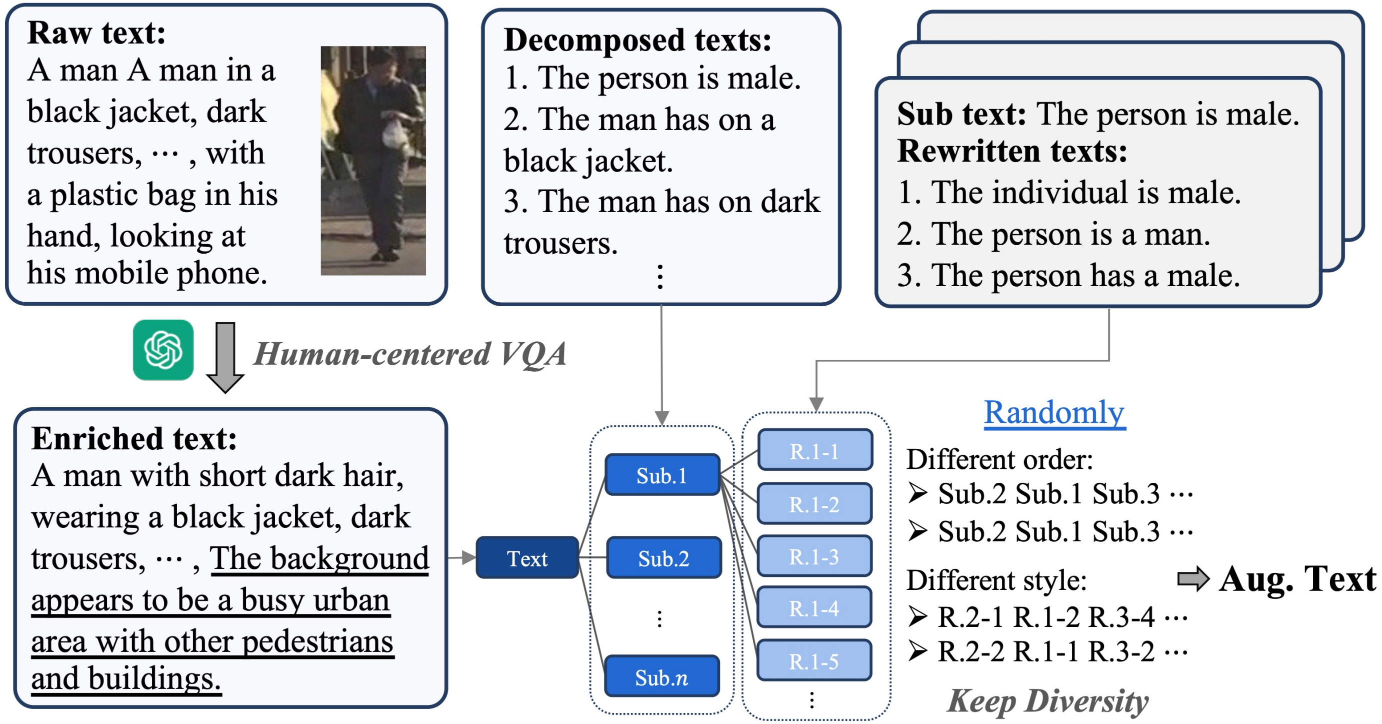
```
Obtain candidate sets {V̂(q_i)}_{i=1}^{|Q|} for all queries in Q via Equation (1);
for k = 1, 2, ··· , K do
    for i = 1, 2, ··· , |Q| do
        Conduct anchor localization via Equation (2) and output the answer of a^{q_i}_{v̂^i_k} based on the k-th candidate image v̂^i_k in V̂(q_i);
        if a^{q_i}_{v̂^i_k} shows 'Yes', k = 1, and S_{q_i, v̂^i_1} > ξ then
            Conduct human-centered VQA via Equations (4) and (5) to get the refined query q̂_i;
            Compute the re-ranking similarities between query q_i and all images via Equation (6);
        end if
        if {a^{q_i}_{v̂^i_j}}_{j=1}^{k-1} all show 'No', a^{q_i}_{v̂^i_k} shows 'Yes', k > 1, and S_{q_i, v̂^i_1} ≤ ξ then
            Conduct human-centered VQA via Equations (4) and (5) to get the refined query q̂_i;
            Compute the re-ranking similarities between query q_i and all images via Equation (6);
        end if
    end for
end for
Re-ranking based on similarities;
```

Output: The new candidate images.

Steps:

1. **Anchor localization**
2. **Human-centered VQA**
3. **Efficient re-ranking**

Method-RDA



Steps:

1. **Enrich text via VQA**
2. **Text decomposition**
3. **Diversity rewriting**
4. **Random recombination**

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

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

$$\mathcal{L} = \mathcal{L}_m + \gamma \mathcal{L}_a$$

Experiments

Datasets

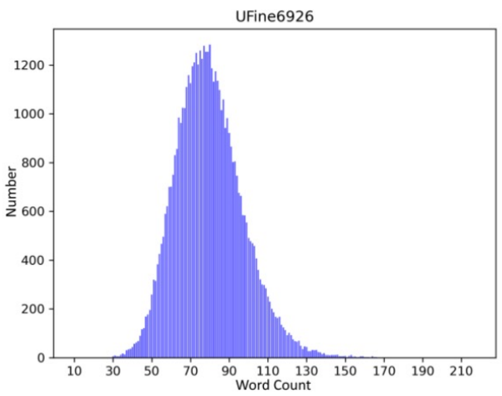
The **CHUK-PEDES, ICFG-PEDES,** and **RSTPReid**, and **UFine6926** datasets

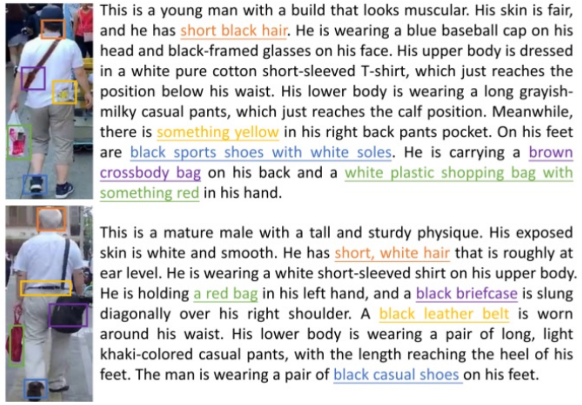
Coarse-grained *fine-grained*

Evaluation Protocols

Rank-K metrics (K=1,5,10) and the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP)

Baselines: IRRA, RDE (SOTA in 2023, 2024) , and so on





The fine-grained dataset UFine6926

Experiments

Comparison with **State-of-the-Arts**

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

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

In group ❶, the Rank-1 scores on the three datasets are improved by 1.50%, 0.91%, and 2.85%, respectively. In addition, mAP and mINP scores have also improved greatly, which indicates that the overall ranking has improved.

In group ❷, our method achieves the best scores on most metrics, especially Rank-1 reached 72.55% on RSTPReid, which is sufficient to verify the superiority.

Experiments

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

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

Text length is over 80

Our method can still achieve excellent performance, with Rank-1 exceeding **91%**. This shows that interaction is also applicable to the fine-grained scenario.

Experiments

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

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

When THI is performed, the cross-domain performance is dramatically improved, for example, from CUHK-PEDES to RSTPReid, THI brings an improvement of more than **4%** on Rank-1.

Experiments

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

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

From the generalization experiments, ICL can also achieve the best cross-domain performance, *e.g.*, compared with the best baseline RDE, from UFine6926 domain to CUHK-PEDES domain, our method improves Rank-1 and mAP by **17.49%** and **10.86%**, respectively, which further verifies the crossdomain generalization of our ICL.

Experiments

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

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

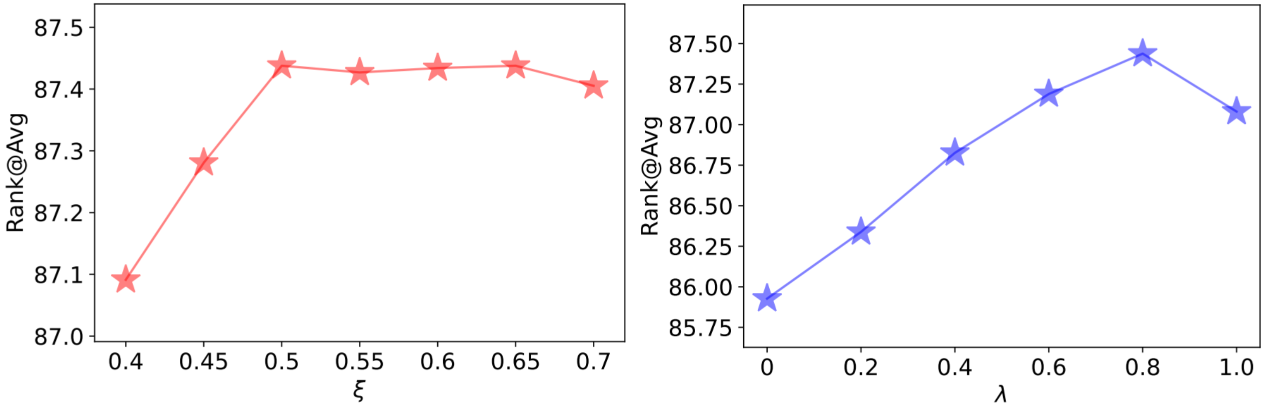
The interactive strategy application can significantly improve Rank-1 and mAP, which shows that the external guidance by interactions via MLLMs can further clarify the text-image alignments and improve the overall ranking.

Experiments

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

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

→ Each module is valid

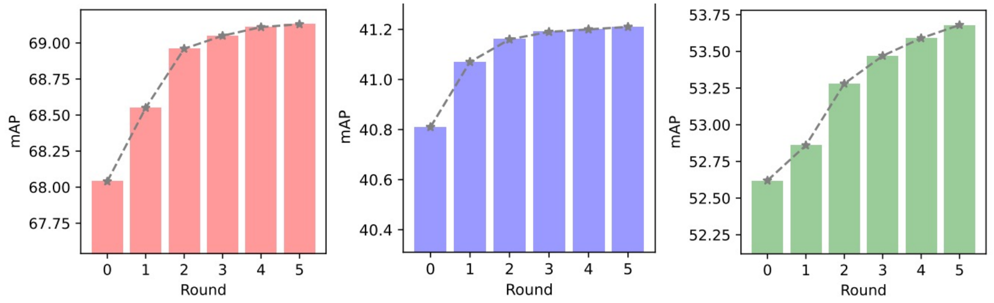


(a) The similarity threshold ξ (b) The balance factor λ

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

→ Set ξ in the range of $0.5 \sim 0.6$ and λ to 0.8

Experiments



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

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

As the interaction progresses, mAP continues to improve.

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

Through interaction, more details can be obtained.

Conclusion

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

Thanks for your attention!

College of Computer Science
Sichuan University