ITSELF: Attention Guided Fine-Grained Alignment for Vision–Language Retrieval Supplementary Material

Tien-Huy Nguyen[1,2,*] Huu-Loc Tran[1,2*] Thanh Duc Ngo[1,2]
[1] University of Information Technology, Ho Chi Minh City, VIETNAM
[2] Vietnam National University, Ho Chi Minh City, VIETNAM

1. Experimental Details

Datasets. We conduct experiments on three widely used text-to-image person retrieval benchmarks.

1. CUHK-PEDES [4] provides 40,206 pedestrian images paired with 80,412 textual descriptions corresponding to 13,003 identities, with splits of 11,003 for training, 1,000 for validation, and 1,000 for testing.
2. ICFG-PEDES [2] consists of 54,522 image-text pairs from 4,102 individual IDs, which are split into 34,674 and 19,848 for training and testing, respectively.
3. RSTPReid [7] contains 20,505 images of 4,101 individual IDs, with each ID having 5 images and each image associated with the corresponding two annotated text descriptions.

Implementation Details For a fair comparison with prior work, we initialize our modality-specific encoders using the pre-trained CLIP-ViT/B-16 [6] model, the same version used by IRRA [3]. To increase data diversity, we apply random horizontal flipping, random cropping, and erasing for images, along with random masking, replacement, and removal for text tokens. Input images are resized to 384 × 128, and the maximum text length is set to 77 tokens. We train the model for 60 epochs using the Adam optimizer with a learning rate initialized to 1×10^{-5} and a cosine learning rate scheduler. The batch size is 256 and temperature parameter τ is set to 0.015. The hyperparameter for MARS is Middle and Late Layer and the discard ratio is 0.25. Normalization Strategy is L1 Normalization. For ATS, p_start and p_end value are 0.65 and 0.5, respectively. We set t_small value equal to the time when the epoch that baseline achieves the best results.

2. More Quantitative Results

To further validate our approach, we conduct extensive quantitative experiments and ablation studies. We analyze the impact of our Adaptive Token Scheduler (ATS), demonstrating in **Tab. 1** that a step-level application yields the best results. We also assess the generalizability of our method with different CLIP backbones in **Tab. 2**, confirming consistent performance gains over the baseline. Finally, we provide a detailed comparison of our MARS selection strategy against several heuristic-based alternatives in **Tab. 3**, which confirms the superiority of our proposed method.

Setting	R@1	R@5	R@10	mAP
Baseline	65.00	85.45	90.15	52.26
Epoch (ATS)	65.25	83.85	89.95	51.39
Step (ATS)	**65.95**	**85.70**	90.10	**52.71**

Table 1. Ablation study on the effect of the Adaptive Token Scheduler (ATS). The baseline does not use the scheduler, while ATS is applied either at the epoch or step level. The results show that step-level scheduling achieves the best performance, yielding improvements in both R@1, R@5 and mAP compared to the baseline.

Setting	R@1	R@5	R@10	mAP
using ViT/B-16 CLIP as backbone				
Baseline	61.30	80.85	87.65	49.12
Ours	**67.30**	**85.60**	**90.50**	**53.05**
using ViT/B-32 CLIP as backbone				
Baseline	59.65	79.35	86.35	47.40
Ours	**64.50**	**84.10**	**90.40**	**50.28**

Table 2. Ablation study on different CLIP backbones. Our method consistently improves performance over the baseline when applied to both ViT/B-16 and the lightweight ViT/B-32 backbone. Notably, even with the smaller ViT/B-32, our approach achieves clear gains in R@1 and mAP, demonstrating its effectiveness across different model capacities.

*Authors contributed equally to this paper.

Setting	R@1	R@5	R@10	mAP
A	60.25	79.70	88.10	48.27
B	66.25	84.45	90.20	52.29
C	60.95	81.40	87.60	48.99
D	65.65	84.00	89.95	51.58
MARS	**66.95**	**85.15**	**90.40**	**53.05**

Table 3. Ablation study of different strategies for selecting top-K patches based on attention statistics. We compare our method (MARS) against four baseline strategies: selecting patches with the minimum mean attention (A), maximum mean attention (B), minimum standard deviation of attention (C), and maximum standard deviation of attention (D). The results clearly demonstrate that our MARS method outperforms all baseline approaches across every evaluation metric. MARS achieves the highest performance with R@1 of 66.95% and mAP of 53.05%. This underscores the effectiveness of our selection strategy compared to simpler heuristics based only on the mean or standard deviation of attention scores.

3. More Qualitative Results

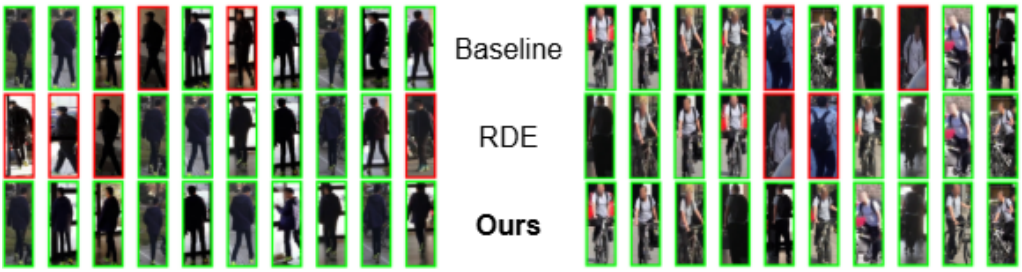
More Retrieval Results. Fig. 1, Fig. 2, Fig. 3 and Fig. 4 provide additional qualitative comparisons across the CUHK-PEDES, ICFG-PEDES, and RSTPReid benchmarks. The examples consistently demonstrate that our method retrieves visually and semantically accurate matches, even in challenging cases involving fine-grained attributes, small accessories, and visually similar distractors. Compared to the baseline, RDE [5], and other strong methods such as IRRA [3] and TBPSCLIP [1], our approach shows superior robustness in capturing subtle cues like clothing textures, color combinations, and carried objects (e.g., backpacks, purses, or phones). Notably, our model remains reliable under domain shifts, handling diverse scenarios ranging from crowded street scenes to low-light images. These results further validate the effectiveness of our framework in producing more discriminative and generalizable text-image alignments.

More Attention Map Visualization. Fig. 5 presents additional qualitative comparisons of attention maps between our method and RDE on the RSTPReid benchmark. The results show that our model consistently attends to more discriminative and semantically relevant regions described in the text queries, such as specific clothing colors, accessories, and carried items (e.g., backpacks, bags, or bicycles). In contrast, RDE often produces diffuse or misaligned attention, failing to capture fine-grained cues. These visualizations further highlight the effectiveness of our approach in leveraging textual guidance to localize meaningful visual regions, thereby enabling more accurate text-based person retrieval.

Figure 1. More examples on CUHK-PEDES benchmark

A middle-aged man with short black hair is wearing a **black insulated jacket with a hood**. He is also wearing **black fitted pants and neon green sneakers with a black Nike logo** on the sides.

A man in his mid 40's having a **bald patch**. He is wearing an off **white t-shirt and black trousers**. He is also holding a **black backpack.**



A man in his fifties with medium-length with a receding hair is wearing a **black bomber jacket**. He is also wearing a pair of **green chinos pants and blue sneakers.**

A young woman is wearing a persian **blue down jacket with white black checkered print** at the bottom and a hood having a fawn furry lining. She is also wearing **slim-fit black pants with dark and light blue shaded running shoes having a black lining** at the bottom and soles in white light blue colour.

A man with black short hair is wearing a **blue puffer jacket with a hood and a pair of fitted black pants**. He is wearing **green running shoes with white details** and his hands are inside the pocket.

A boy with black medium length hair is wearing a **dark blue hoodie that has green sleeves**. He is also wearing **grey denim loose fitted jeans with dark blue running shoes**. He has a **black and orange school bag** that has Mickey Mouse on it strapped on the shoulder.

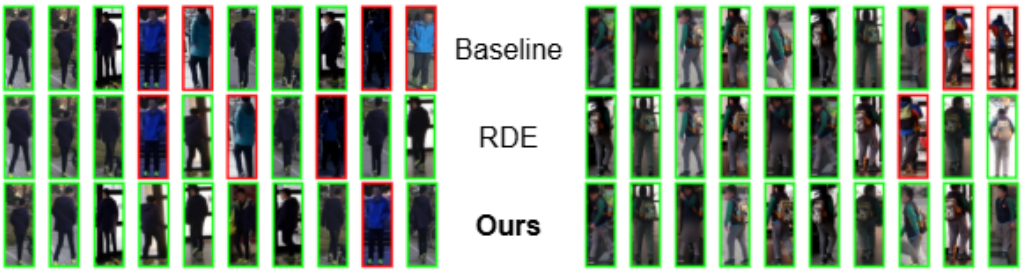


Figure 2. More examples on ICFG-PEDES benchmark

Figure 3. More examples on RSTPReid benchmark

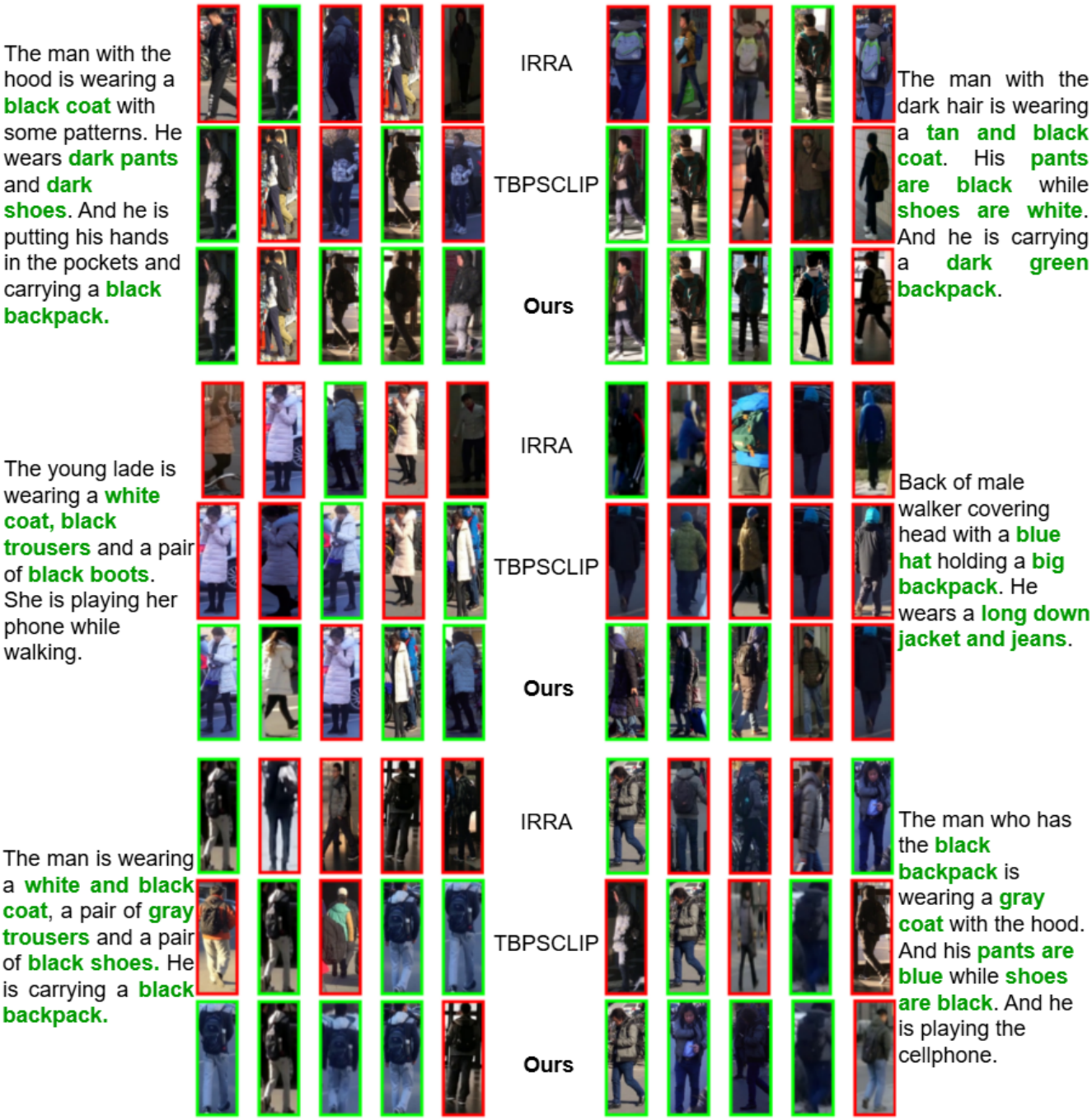


Figure 4. More examples compare with other methods on RSTPReid benchmark

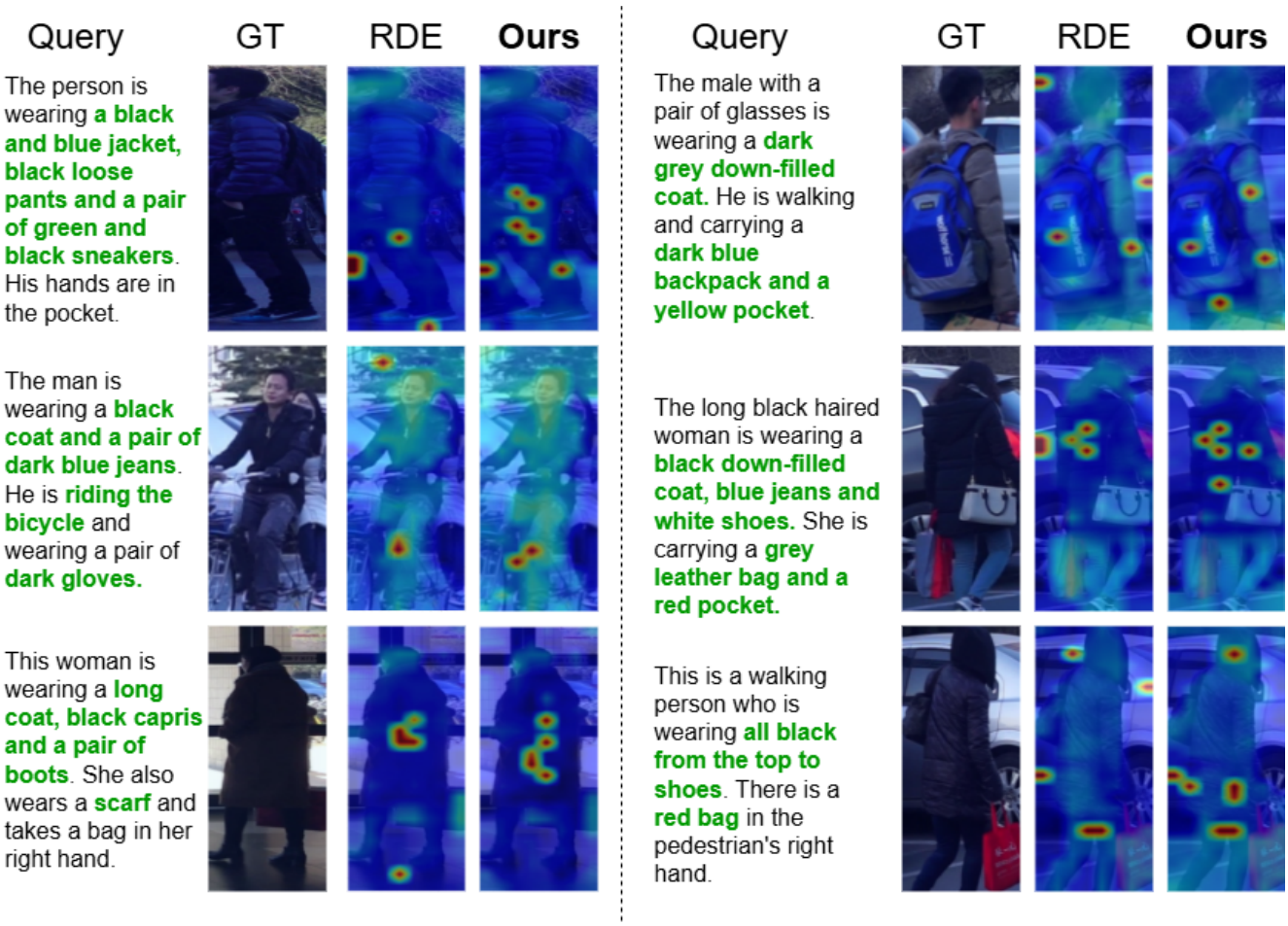


Figure 5. More attention map examples compared with RDE on the RSTPReid benchmark

References

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

[2] 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

[3] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2787–2797, 2023. 1, 2

[4] 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

[5] 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. 2

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