Supplementary Material for "Cross-modal Uncertainty Modeling with Diffusion-based Refinement for Text-based Person Retrieval"

Anonymous Author(s)
Submission Id: 5847

ACM Reference Format:
Anonymous Author(s). 2024. Supplementary Material for "Cross-modal Uncertainty Modeling with Diffusion-based Refinement for Text-based Person Retrieval". In *Proceedings of the 47th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 OVERVIEW

In this supplementary information section, we delve into additional analyses of our proposed CUMDR method. The provided supplementary material, outlined below, offers a more comprehensive exploration of our approach, which couldn't be accommodated in the main paper due to space constraints.

- We present additional comparisons with state-of-the-art methods in the new metrics and noisy setting on CUHK-PEDES [1] and RSTPReid [2] datasets for a comprehensive evaluation.
- We conduct additional ablation studies on RSTPReid to demonstrate the effectiveness of our proposed Retrieval-augmented In-context Constructor.
- We conduct additional ablation studies on CUHK-PEDES and ICFG-PEDES to comprehensively assess the significance of each module within our CUMDR method.
- The complete training procedure of CUMDR is outlined, providing the details of our methodology.
- We conduct a more extensive qualitative analysis, incorporating supplementary retrieved results and a thorough examination of the impact of our proposed CUMDR.

For enhanced clarity in presenting our method, we also delineate the training steps in Algorithm 1.

2 FURTHER QUANTITATIVE ANALYSIS

2.1 Additional Evaluation Metrics

We augment our experimental findings presented in Figure 1, incorporating additional metrics such as R@5 and N@5, to underscore the superiority of our proposed method, CUMDR.

Across the results obtained from three benchmark datasets, CUMDR consistently demonstrates more precise retrieval outcomes compared to the current state-of-the-art method, APTM, on all datasets. This consistent improvement is indicative of the robustness of our model across various metrics. Notably, CUMDR outperforms APTM on N@5, emphasizing the advantageous capabilities of our proposed modules in effectively mitigating the influence of inherent data noise, thereby enhancing retrieval performance.

Furthermore, we extend our experiments to CUHK-PERES (Figure 2), utilizing additional metrics, including N@5 and R^2@5, to further highlight the superiority of our CUMDR method. The results clearly illustrate that CUMDR outperforms existing methods, underscoring its superior ability in modeling cross-modal uncertainty and preventing overfitting to a singular target.

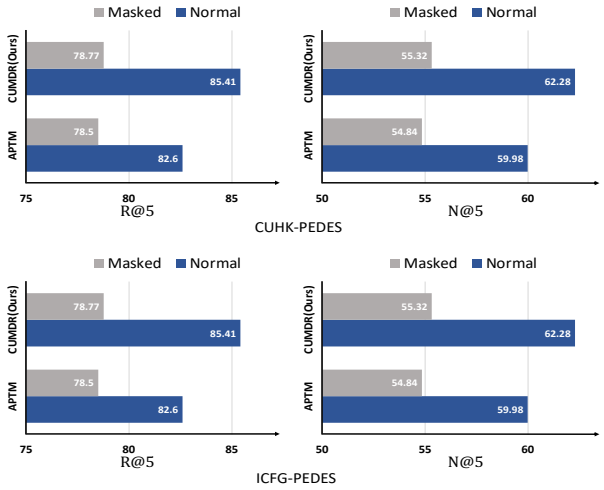

Figure 1: The comparison of our CUMDR method with APTM on CUHK-PEDES and ICFG-PEDES.

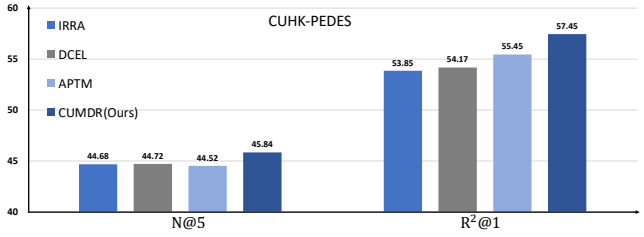

Figure 2: The comparison of our CUMDR method with other three methods on CUHK-PEDES.

SIGIR'24, July 14 - July 18, 2024, Washington D.C., USA

ACM ISBN 978-1-4503-XXXX-X/23/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Effect of the Retrieval-augmented In-context Constructor.

No.	Components		RSTPReid			
	RIC	RUCM	R@1	R@5	N@5	R^2@1
0	-	-	72.65	88.80	51.53	53.75
1	-	✔	74.85	89.20	52.03	52.30
3	✔	✔	**76.20**	**91.85**	**53.32**	**53.75**

2.2 Effect of the Retrieval-augmented In-context Constructor

To further explore the significance of our proposed Retrieval-augmented In-context Constructor (RIC), we conduct the corresponding ablation study on the RSTPReid dataset. The experimental outcomes are detailed in Table 2.

From the garnered results, several noteworthy observations emerge: 1) The RIC manifests a substantial enhancement in terms of all the standard metrics. This improvement suggests the fact that the RIC facilitates generating proper captions for each image without excessive hallucination, improving the quality of the generated annotations. 2) With regard to the novel metrics N@5 and R^2@1, our proposed RIC can effectively improve the retrieval performance. This signifies that our CUMDR model identifies more accurate images in the query results compared to other methods. Consequently, our CUMDR demonstrates superior capabilities in matching multiple target images.

Table 2: Ablation studies with respect to model components on CUHK-PEDES.

No.	Components			CUHK-PEDES		
	CUM	RUCM	DAR	R@1	R@5	R@10
0	-	-	-	75.34	89.18	93.18
1	-	-	✔	76.25	90.08	94.01
2	-	✔	-	77.03	91.11	94.91
3	✔	-	-	76.42	90.10	94.10
4	-	✔	✔	77.18	91.26	95.28
5	✔	-	✔	76.83	90.27	94.23
6	✔	✔	-	77.37	91.35	95.31
7	✔	✔	✔	**77.81**	**91.85**	**95.78**

2.3 Additional Ablation Study on CUHK-PEDES

To comprehensively assess the significance of each constituent module within our CUMDR method, we conducted a series of ablation studies targeting specific components, namely Cross-modal Uncertainty Modeling (CUM), Retrieval-augmented Uncertainty-aware Complementary Matching (RUCM), and Diffusion-based Alignment Refinement (DAR). The insights derived from our ablation experiments, as succinctly presented in Table 1, are delineated below:

(1) A comparative analysis of the performance between N0.0 and No.2 reveals that our proposed RUCM module significantly mitigates data noise, leading to an improvement in retrieval performance. This underscores that aligning images with more precise annotations for pairs characterized by high uncertainty effectively reduces the impact of inherent data noise.

(2) Notably, No.4 outperforms both No.2 and No.0. This indicates that incorporating joint distribution from the diffusion-based denoiser is more conducive to generalization when contrasted with a discriminative framework.

(3) Insights drawn from the comparison between No.7 and No.4 suggest that the CUM module efficiently models cross-modal uncertainty. This not only prevents the model from overfitting to a singular target but also enhances its understanding of the intrinsic semantic relationships within text-image pairs. Consequently, this improvement aids the RUCM module, leading to enhanced performance.

Algorithm 1 Overall training procedure of CUMDR.

Input: $\mathcal{X} = \{(I_i, T_i)\}_{i=1}^{B}$, batch size B, temperature parameter τ, learning rate ζ and the pretrained model parameters θ_t. $\gamma = -0.005$ and $\gamma_1 = 300$ are predefined hyper-parameters indecating negative scale factor and margin in $\mathcal{L}_d$ respectively.
Output: Model Parameters θ_t.

```
Load the saved model parameters theta_t.
repeat
    Sample a batch of queries and targets.
    Build distribution for textual descriptions and images.
    Generate joint distribution using DAR.
    Compute the objective L_cdm, L_fdm, L_d, L_wcm and L_dar.
    Update θ_t using AdamW optimizer, learning rate = 1e-5, step learning rate decay.
    θ_t ← θ_t ← ζ∇_{θ_t} L.
until Reach maximum iterations.
Take trained CUMDR to conduct text-based person retrieval.
```

3 FURTHER QUALITATIVE ANALYSIS

3.1 Additional Retrieved Results

We present additional qualitative results of our CUMDR method on three benchmark datasets, RSTPReid, CUHK-PEDES and ICFG-PEDES, illustrated in Figure 4, 3 and Figure 5. Correct targets are highlighted with green boxes, while textual queries are enclosed in yellow boxes. The top-6 retrieved results from various methods demonstrate our model's proficiency in accurately retrieving target images.

1) Leveraging the proposed CUM, CUMDR excels at finding more correct images compared to APTM. This underscores our model's enhanced ability to model uncertainty.

2) Through the integration of external matching in RUCM, CUMDR effectively reduces excessive uncertainty, directing the model's focus towards the text query and enabling discrimination of detailed features like *"skinny pants"* or *"bag over her right shoulder"*.

3) Notably, our model demonstrates the capability to retrieve correct target images from galleries of varying sizes, attributed to the DAR module. DAR contributes to enhanced retrieval performance by framing the cross-modal relationship as a joint probability.

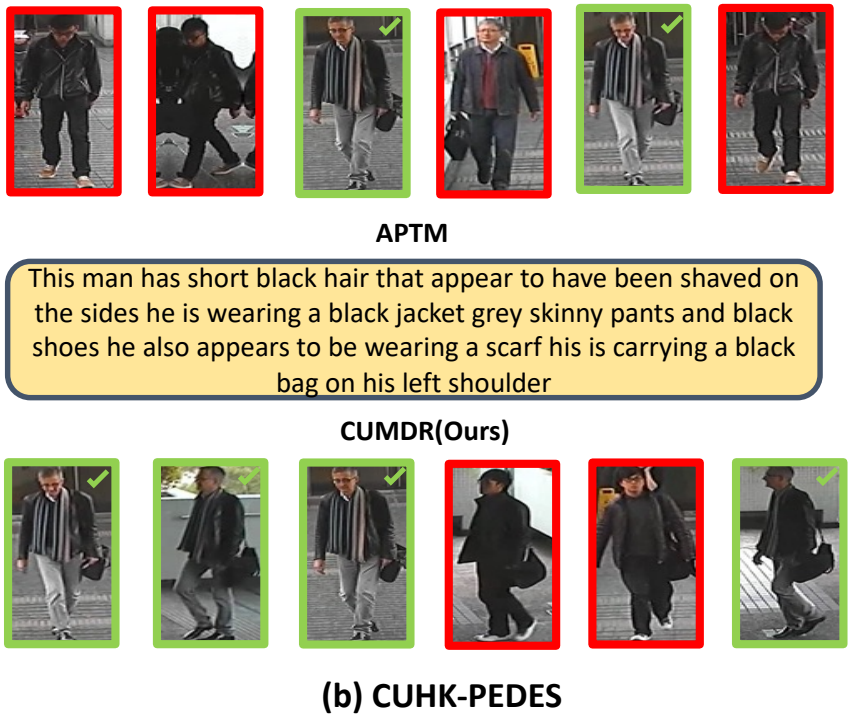

Figure 3: Qualitative results for textual queries on CUHK-PEDES. We show the top 6 retrieved results for each query.

3.2 Analysis on Distribution Representations Visualization

To further underscore the advantages of our CUMDR model, we visualize the textual and visual representations of the top-10 retrieved results in Figure 6 and Figure 7. Each ellipse represents the 50% confidence region for visual representations, with the textual query highlighted by a white ellipse with a 90% confidence region. Ground-truth targets are indicated by green boxes.

The textual query, as denoted by the white ellipses, effectively spans the majority of ground truth images in the distribution embedding space, while excluding negative samples. Notably, images with common features tend to cluster together. For instance, in Figure 7, two overlapped ellipses representing two negative retrieval results (images with pink and red boxes) exhibit high similarity simultaneously. We hypothesize that this phenomenon is indicative of the fact that the span and shape of each ellipse convey semantic information. Intuitively, these results suggest that implicitly modeling cross-modal uncertainty through distribution representations empowers the model to convey rich semantic information and intricate relationships, thereby comprehensively enhancing performance.

REFERENCES

[1] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5187–5196.

[2] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.

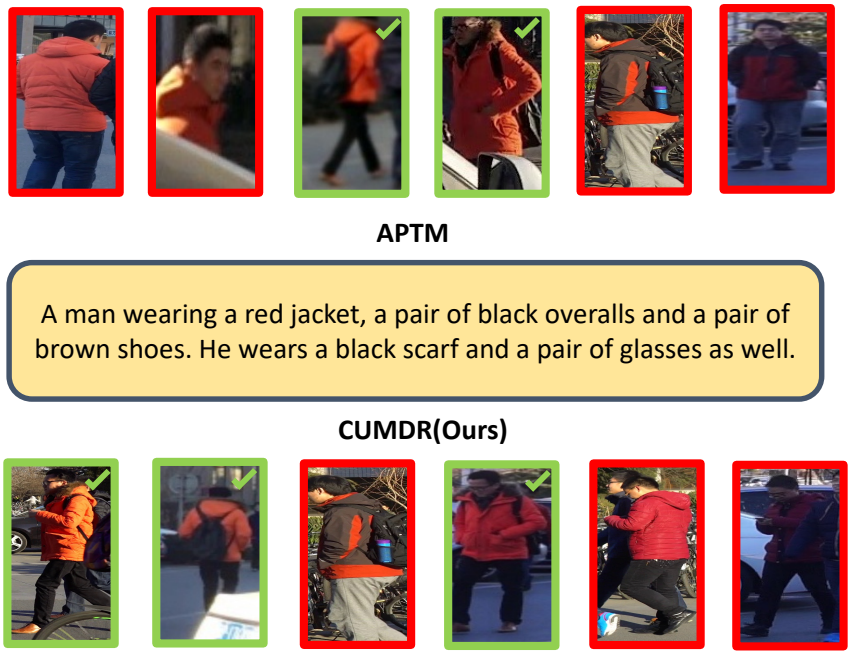

Figure 4: Qualitative results for textual queries on RSTPReid. We show the top 6 retrieved results for each query.

Figure 5: Qualitative results for textual queries on ICFG-PEDES. We show the top 6 retrieved results for each query.

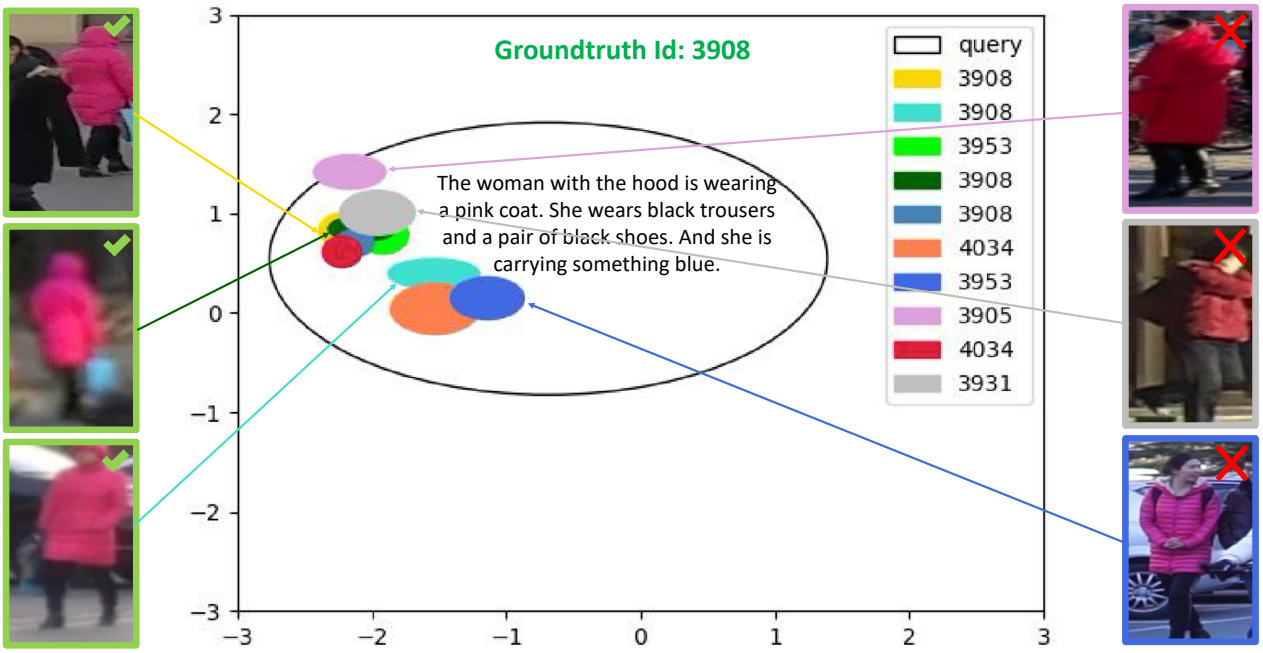

Figure 6: Visualization of the distribution for query (Groundtruth ID: 3908) in RSTPReid

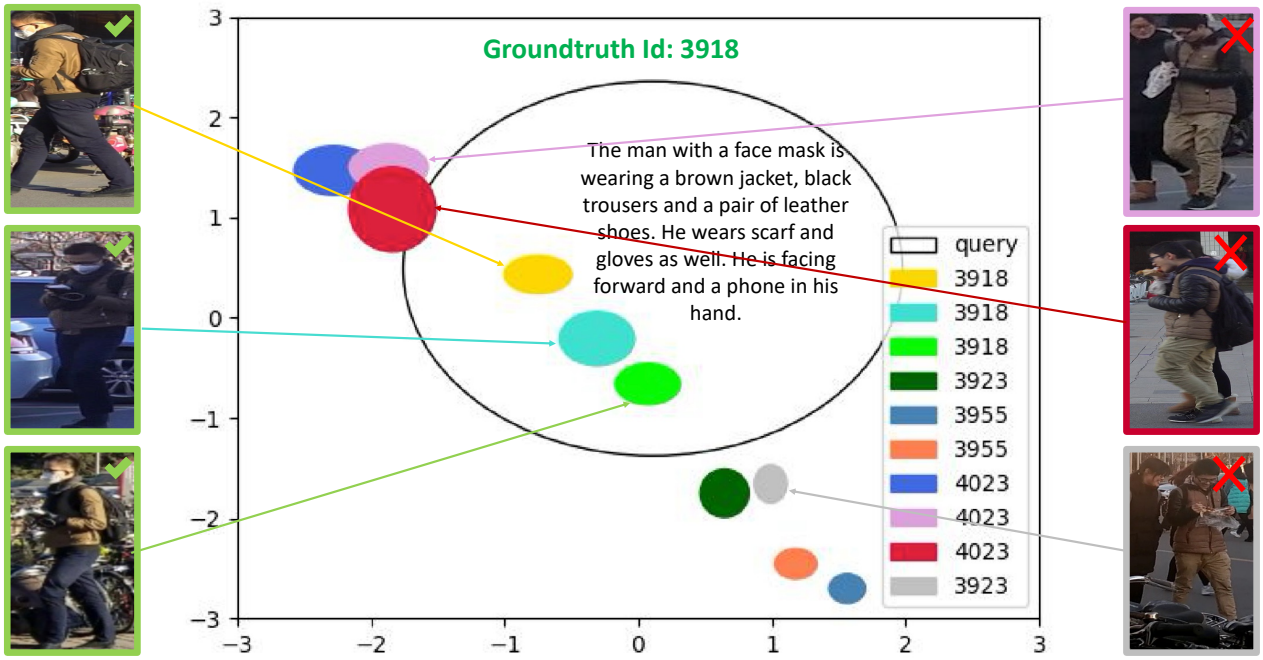

Figure 7: Visualization of the distribution for query (Groundtruth ID: 3918) in RSTPReid