A Additional Ablation Studies

A.1 Comparison with Lightweight Tuning Methods.

Table A.1: Comparison of lightweight tuning methods on RSTPReid(Zhu et al., 2021) benchmark.

Method	Tuning Layers	R@1	R@5	R@10	mAP
Baseline	–	60.64	77.50	83.26	35.54
CoOp*(Zhou et al., 2022)	Prompt Embedding	58.60	79.65	87.50	43.65
Prefix-Tuning*(Li & Liang, 2021)	Prefix Embedding	26.25	51.45	63.75	23.14
LoRA*(Hu et al., 2022)	LoRA Matrix	49.80	73.80	82.70	38.61
Ours	Normalization Layer	61.85	81.40	88.40	46.37

We compare our method with lightweight tuning methods in Table A.1. Baseline is LuPerson-HAM (Jiang et al., 2025) and * means that we try different hyperparameters *i.e.*learning rate, number of virtual tokens, rank of LoRA(Hu et al., 2022) etc., for lightweight tuning methods and selected the best result. CoOp(Zhou et al., 2022), which is a prompt learning method and belong to few-shot learning, fails with adaptation objective of entropy minimization. It can be interpreted that additional prompt token embeddings need labeled data to mimic natural language feature in the token embedding space, instead of adjusting the cross-modal feature distribution in latent space. Additionally, Parameter Efficient Fine-Tuning (PEFT)(Han et al., 2024) provides a practical solution by efficiently adjusting the large models over the various downstream tasks. We also evaluated two representative PEFT methods, *i.e.*, Prefix-Tuning (Li & Liang, 2021) and LoRA (Hu et al., 2022), for test-time adaptation on the RSTPReid benchmark, however, these approaches proved ineffective in our experiments. Although the trainable parameters in PEFT are lightweight, Entropy Minimization fails to provide sufficient supervision for learning discriminative representations.

A.2 More Ablation Study on RSTPReid.

Table A.2: **More Ablation of Bidirectional Top-K Retrieval Consistent Sample Selection K on RSTPReid.** K denotes the mutual top range of bidirectional retrieval. We obtain the best result at $K = 3$, but the number of ground-truth image in RSTPReid is 5, as it represent the borderline of true and false positives. We choose $K = 5$ with gray line as default setting.

Table A.3

K	R@1	R@5	R@10	mAP
1	61.60	81.10	**88.75**	45.90
2	60.90	81.30	**88.75**	46.14
3	**61.90**	81.20	88.05	**46.39**
4	61.85	81.30	88.15	46.38
5	61.85	**81.40**	88.40	46.37
6	61.05	80.95	88.30	46.17
8	61.30	81.10	87.85	46.04
10	61.05	80.80	88.00	45.93
∞	61.55	80.75	88.20	45.96

Effect of K Mutual Neighbours. Despite the physical meaning of top-K, we compare more K in Table A.3. To some extent, results indicate the generalization performance of different K.

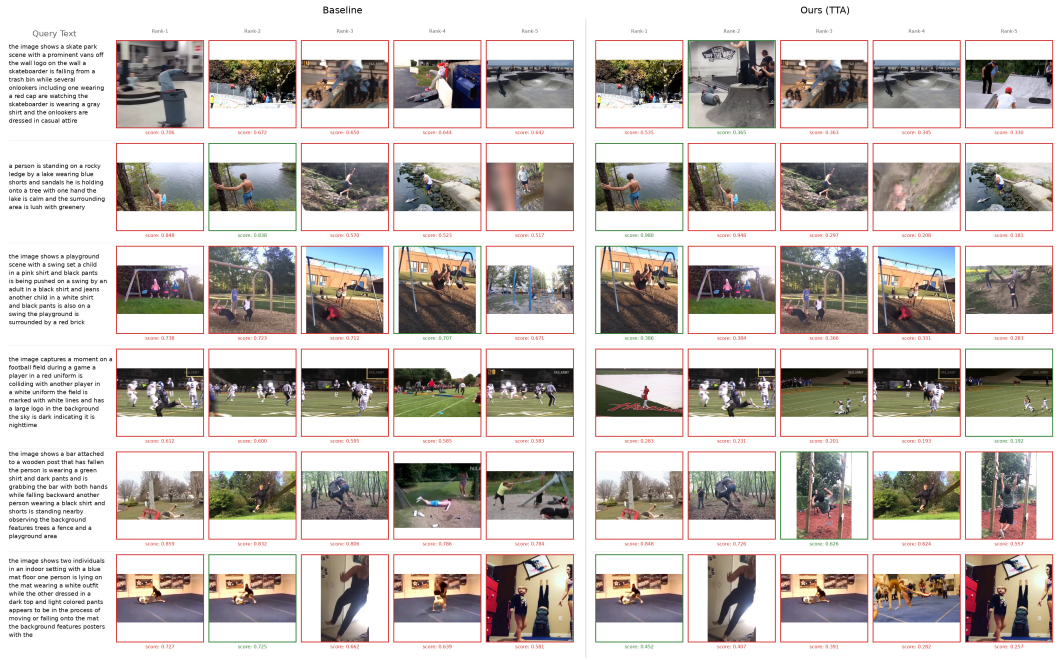

Figure B.1: **Top-5 Text-based Person Search Results on PAB.** The figure displays the Top-5 retrieval results for six representative text queries, with the confidence score for each rank provided below the corresponding image. The matched person images are annotated in green boxes, and the false ones are in red.

B More Qualitative Results

B.1 Qualitative Analysis of Person Search Performance

To qualitatively validate the effectiveness of our Uncertainty-Aware Test-Time Adaptation (UATTA), we present a visual comparison of retrieval results between the Baseline and UATTA on the PAB dataset in Figure B.1. The visualization effectively showcases two key strengths of UATTA: First, in challenging cases (Rows 1, 4, 5) where the Baseline fails due to overly high confidence in false positives, UATTA successfully rectifies the score distribution by mitigating this over-confidence, leading to the correct identification of the ground-truth image. Second, for scenarios requiring fine-grained semantic distinction (Rows 2, 3, 6), UATTA leverages the bidirectional retrieval disagreement proxy to effectively disambiguate subtle differences between the text and image modalities. This mechanism allows UATTA to elevate the correct match from a low rank to Rank-1, significantly outperforming the Baseline. Overall, the qualitative results confirm that UATTA achieves a sharper, more robust, and accurate confidence distribution, validating its superiority in handling both retrieval ambiguity and fine-grained visual differences.

B.2 Visualization of Feature Space Shifts

In Figure B.2, T-SNE visualization provides an intuitive illustration of the impact of Test-Time Adaptation (TTA) on Feature Space. The visualization is focused on a representative subset of the Top-15 most frequent person identities to ensure clarity and showcase the adaptation effects vividly. We notice that the initial spread of original Query features (circles) demonstrates the significant domain gap and feature ambiguity present before adaptation, justifying the necessity of TTA. After TTA, regions circled by dotted ellipses indicate that query features, post-TTA (diamonds), are effectively adapted to align more closely with their respective gallery feature (squares) clusters. This convergence demonstrates the efficacy of TTA in reducing feature disparity and enhancing matching performance. While the majority of person identities show strong alignment, we observe that

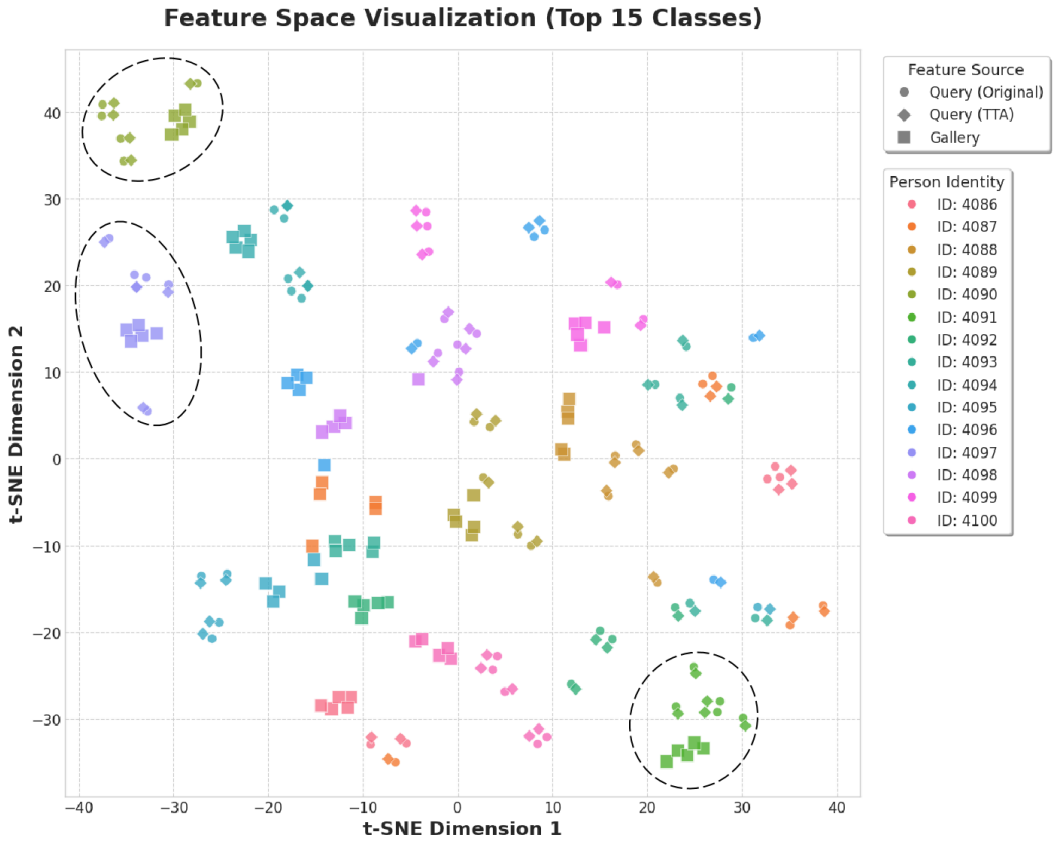

Figure B.2: **T-SNE Visualization of Feature Space Shifts on RSTPReid.** The three distinct point types represent: original query features (circles) before Test-Time Adaptation (TTA), query features after TTA (diamonds), and gallery features (squares). Different colors distinguish individual person identities.

some identities still exhibit residual ambiguity after TTA, suggesting potential avenues for future improvement in feature consolidation.

REFERENCES

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. Modeling thousands of human annotators for generalizable text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9220–9230, 2025.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

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*, pp. 209–217, 2021.