Pretrain-then-Adapt: Uncertainty-Aware Test-Time Adaptation for Text-based Person Search

Jiahao Zhang
University of Macau
Macau, China
yc57963@um.edu.mo

Shaofei Huang
University of Macau
Macau, China
nowherespyfly@gmail.com

Yaxiong Wang
Hefei University of Technology
Hefei, China
wangyx@hfut.edu.cn

Zhedong Zheng*
University of Macau
Macau, China
zhedongzheng@um.edu.mo

Abstract

Text-based person search faces inherent limitations due to data scarcity, driven by stringent privacy constraints and the high cost of manual annotation. To mitigate this, existing methods usually rely on a **Pretrain-then-Finetune** paradigm, where models are first pretrained on synthetic person-caption data to establish cross-modal alignment, followed by fine-tuning on labeled real-world datasets. However, this paradigm lacks practicality in real-world deployment scenarios, where large-scale annotated target-domain data is typically inaccessible. In this work, we propose a new **Pretrain-then-Adapt** paradigm that eliminates reliance on extensive target-domain supervision through an offline test-time adaptation manner, enabling dynamic model adaptation using only unlabeled test data with minimal post-train time cost. To mitigate overconfidence with false positives of previous entropy-based test-time adaptation, we propose an Uncertainty-Aware Test-Time Adaptation (UATTA) framework, which introduces a bidirectional retrieval disagreement mechanism to estimate uncertainty, *i.e.*, low uncertainty is assigned when an image-text pair ranks highly in both image-to-text and text-to-image retrieval, indicating high alignment; otherwise, high uncertainty is detected. This indicator drives offline test-time model recalibration without labels, effectively mitigating domain shift. We validate UATTA on four benchmarks, *i.e.*, CUHK-PEDES, ICFG-PEDES, RSTPReid, and PAB, showing consistent improvements across both CLIP-based (one-stage) and XVLM-based (two-stage) frameworks. Ablation studies confirm that UATTA outperforms existing offline test-time adaptation strategies, establishing a new benchmark for label-efficient, deployable person search systems. Our code is available at https://github.com/nkuzjh/UATTA.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; **Similarity measures**; • **Computing methodologies** → **Transfer learning**.

Keywords

Person Search, Cross-Modal Retrieval, Domain Gap, Test-Time Adaptation, Uncertainty

*Corresponding author.

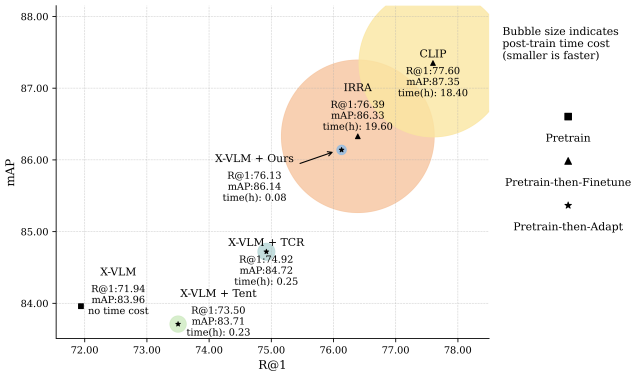


Figure 1: Accuracy vs. efficiency trade-off on the PAB benchmark [52]. Built upon a pretrained X-VLM [57] backbone, our method follows the Pretrain-then-Adapt paradigm and substantially improves retrieval performance with minimal adaptation cost (*i.e.*, post-train GPU time). Compared with Pretrain-then-Finetune approaches, our method reduces post-train GPU time of adaptation by 99.6% while achieving competitive performance, approaching or even matching strong Pretrain-then-Finetune methods such as IRRA [19]. We present Pretrain-then-Finetune results (*i.e.*, CLIP [39] and IRRA [19]) as an upper bound on achievable performance under extensive finetuning. In contrast, our method reaches a favorable operating point close to this upper bound with significantly lower post-train time cost, highlighting its improved accuracy vs. efficiency trade-off under the Pretrain-then-Adapt setting.

ACM Reference Format:
Jiahao Zhang, Shaofei Huang, Yaxiong Wang, and Zhedong Zheng. 2026. Pretrain-then-Adapt: Uncertainty-Aware Test-Time Adaptation for Text-based Person Search. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809598

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809598

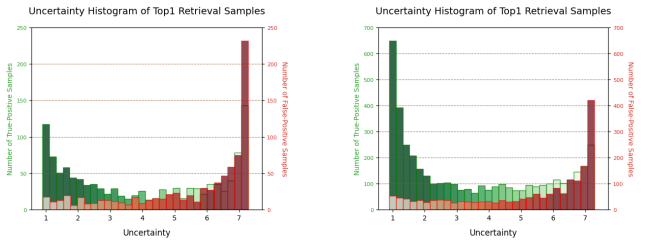
1 Introduction

Text-based person search [7, 29, 66], which involves matching natural language descriptions to specific individuals within large-scale image galleries, is a critical task with applications ranging from locating missing persons [3] to enhancing smart city management [22, 62]. Unlike conventional image-based person re-identification [61, 64], the incorporation of text modality offers a more intuitive and accessible query interface for system operators [63].

Despite its practical advantages, the efficacy of current methods is severely hampered by the domain shift problem, where models trained in controlled settings exhibit significant performance degradation when deployed in unseen, real-world environments. State-of-the-art approaches typically attempt to mitigate this challenge through a **Pretrain-then-Finetune** paradigm [19, 20, 35, 40, 44]. This involves first pretraining on large-scale, often synthetic, person-caption datasets to establish preliminary cross-modal alignments, followed by fine-tuning on domain-specific annotated datasets such as CUHK-PEDES [29]. However, the reliance on labeled data for fine-tuning renders this paradigm impractical for many real-world deployments. In practice, target domain labels are typically unavailable due to stringent privacy regulations [10] and prohibitive annotation costs [40].

To address this limitation, we introduce the source-free offline test-time adaptation (TTA) [8, 46] to the cross-modal retrieval task, formulating a **Pretrain-then-Adapt** paradigm, leveraging to adapt a pretrained model to a new target domain using only unlabeled test samples. Such strategy directly performs tailored adaptation to the specific data distribution of the current test set, therefore alleviating the reliance of labeled target domain data. As depicted in Fig. 1, the Pretrain-then-Adapt paradigm shows superior efficiency and competitive performance compared to traditional Pretrain-then-Finetune paradigm, owing to its independence of fine-tuning on domain specific labeled data. Consequently, this paradigm demonstrates superior efficiency requiring orders-of-magnitude lower adaptation cost in contrast to traditional Pretrain-then-Finetune paradigm. A prevailing practice within this paradigm involves adapting the model via entropy minimization [46, 54], a TTA strategy widely adopted in image classification. By minimizing prediction entropy in an online or offline manner, the model is forced to sharpen its decision boundaries and increase its confidence in unlabeled target samples. However, this approach presents a significant risk of error accumulation where the model can be overconfident in its own erroneous predictions, reinforcing them during adaptation and converging to a suboptimal state [60]. In the context of retrieval, this implies that the model treats false positives as reliable as true positives, thereby amplifying the impact of wrong supervision signals. This raises a crucial research question: *How can we mitigate the risk of overconfident, erroneous adaptation in cross-modal person retrieval, a task that demands fine-grained matching?*

We argue that the key to mitigating the issue of overconfident false positives hinges on reliable uncertainty calibration. As illustrated in Fig. 2, samples exhibiting high uncertainty are predominantly concentrated among false positives. This suggests that high uncertainty serves as an effective proxy for identifying false positives. Consequently, we introduce the **Uncertainty-Aware Test-Time Adaptation (UATTA)** framework, which leverages prediction uncertainty to re-calibrate the offline adaptation process on the entire test set. However, since legitimate uncertainty is intractable to estimate directly without ground truth, we propose *bidirectional retrieval disagreement* as a tractable proxy. A high-uncertainty match will exhibit incongruity between the text-to-image and the corresponding image-to-text retrieval directions, whereas a confident, low-uncertainty match will show symmetric alignment. We provide a theoretical justification that this metric effectively gauges prediction uncertainty. This principle allows us to identify and down-weight potential false positives to avoid overconfidence during adaptation.



(a) RSTPReid **(b) CUHK-PEDES**

Figure 2: Statistical Overview of our Uncertainty Indicator on (a) RSTPReid [66] and (b) CUHK-PEDES [29]. We count the True Positives (TP) and False Positives (FP) sample number in the initial ranking list before adaptation according to the proposed uncertainty score. TP samples consistently cluster in the low-uncertainty region, while FP samples concentrate in the high-uncertainty region across both benchmarks. Therefore, it could serve as the indicator for the test-time adaptation.

Specifically, for a one-stage retrieval model based on CLIP [39], we quantify bidirectional retrieval disagreement uncertainty using the relative disparity between mutual retrieval probabilities derived from the Image-Text Contrastive (ITC) loss. This bidirectional retrieval disagreement uncertainty measure is then used to rectify the entropy minimization objective by re-weighting it with the reciprocal of the uncertainty. For two-stage retrieval architectures like XVLM [57], we apply the same principle to modulate the entropy of the fine-grained predictions from the Image-Text Matching (ITM) module. In both architectures, this uncertainty-aware rectification acts as a dynamic filter, effectively suppressing gradients from overconfident false positives to prevent error accumulation of vanilla entropy minimization. Consequently, by ensuring adaptation is solely based on reliable alignments, UATTA bridges the gap between unsupervised adaptation and supervised fine-tuning. As shown in Fig. 1, our approach realizes a superior accuracy and efficiency trade-off, delivering performance competitive with expensive fine-tuning methods while maintaining the operational efficiency and flexibility of the Pretrain-then-Adapt paradigm.

Our primary contributions are as follows:

- **A Practical Paradigm for Test-Time Adaptation on Text-based Person Search.** We explore a **Pretrain-then-Adapt** paradigm for text-based person search that alleviates the need for labeled target-domain data. This framework offers a practical alternative to the standard Pretrain-then-Finetune pipeline, enhancing deployability in real-world scenarios where data annotation is infeasible.
- **An Uncertainty-Guided Adaptation Method.** We propose an Uncertainty-Aware Test-Time Adaptation (UATTA) framework designed to address domain shift under unsupervised conditions. The method introduces a **bidirectional retrieval disagreement** mechanism to estimate prediction uncertainty. This signal is used to guide the adaptation, aiming to curb error accumulation from overconfident false positive predictions during the offline test-time optimization process.

- **Comprehensive Empirical Evaluation.** We conduct extensive experiments on four challenging benchmarks (CUHK-PEDES [29], ICFG-PEDES [7], RSTPReid [66], and PAB [52]). Our results show that UATTA achieves consistent performance improvements over baseline methods across different model architectures. The findings validate the efficacy of our uncertainty-guided approach and suggest it is a promising direction for label-free adaptation in this domain.

2 Related Work

Text-based Person Search. Text-based person search aims to find the target person of interest via a text query. Different from image-based search [15], the text query is more intuitive for users. A typical dataset is CUHK-PEDES [29]. To align person images and text, recent works usually adopt a pretrain-then-finetune paradigm, in which models first establish cross-modal alignment on synthetic person-caption data and then fine-tune on limited real-world annotations. [40] apply CLIP [39] with a novel divide-conquer-combine strategy to automatically annotate pseudo-text descriptions for a large-scale person re-identification image dataset [9], which reduces human labor and cost. With the help of image generative models, [53] collect a new large-scale cross-modal dataset MALS [53], containing real-world text descriptions and corresponding generated person images with multiple attributes, providing an alternative for real-world person privacy via automatic image generation and attribute extraction. Following this synthetic-pretrain and real-world-finetune approach, [20, 44] boost text-based person search performance by exploiting Multi-modal Large Language Models to obtain text descriptions with various language structures and styles. Existing test-time inference pipelines of this paradigm can be divided into one-stage CLIP-based [39] and XVLM-based [57] frameworks. The former [5, 19, 20, 40, 44] extracts vision and language features independently via separate single-modal models and predicts alignment based on Image-Text Contrastive (ITC) similarity [39]. The latter [25–27, 38, 41, 48, 52, 53, 57] employs an additional fine-grained cross-modal interaction module to exploit Image-Text Matching (ITM) learning and predict binary matching results to rectify top-K results from the first stage. In this paper, we propose a universal Pretrain-then-Adapt paradigm that is not constrained by the scarcity of annotated labels for both one-stage and two-stage frameworks.

Test-Time Adaptation. Test-time Adaptation (TTA) has emerged as a promising paradigm that dynamically aligns the model with the specific test distribution during inference, effectively mitigating domain shift without source data access. Parameter-metric approaches [46, 54] minimize prediction entropy through lightweight parameter updates, e.g., BatchNorm [17] statistics. However, these approaches suffer from confirmation bias as domain shift induces high-confidence errors, a phenomenon exacerbated in cross-modal retrieval where false positives deteriorate performance [60]. Memory-based approaches [18, 58] maintain feature banks for pseudo-label refinement but introduce prohibitive computational overhead for memory indexing and require structural modifications incompatible with frozen VLM backbones. Recent works [36, 37, 43] attempt to reduce overhead through sample selection, but these strategies focus on a small number of high-confidence samples, which induces catastrophic forgetting by overfitting and deviates from pretrained feature manifolds [23]. Notably, our work reformulates offline test-time adaptation through uncertainty-weighted entropy minimization on the whole test set, which suppresses overconfidence on false positives while preserving frozen VLM backbones. By leveraging global domain statistics and filtering unreliable signals via cycle consistency, our approach avoids suboptimal convergence, achieving a superior balance between accuracy and efficiency for real-world deployment.

Uncertainty in Cross-Modal Retrieval. Uncertainty quantification has gained traction in cross-modal retrieval [6, 31, 47]. Generally, uncertainty can be quantified as the discrepancy of representation between different modalities, which is more pronounced under domain gaps [49]. [55] integrate fine- and coarse-grained retrieval with different fluctuations to model uncertainty and rectify the matching objective. Furthermore, Li *et al.* [28] leverage subjective logic to select reliable cross-modal pairs and masked modeling to capture cross-modal relations, and also exploit multi-grained uncertainty-based alignments to mitigate domain shifts. With the help of an extra large vision-language model, Zhao *et al.* [60] use CLIP to reflect the uncertainty of input pairs and boost zero-shot performance via an uncertainty-aware reward feedback mechanism. Li *et al.* [24] optimize the robustness of test-time adaptation via candidate selection, inter-modal gap learning, and intra-modal uniformity learning, yet are constrained to query modal shifts. Through a novel design of probabilistic distance metrics and hierarchical learning objectives, [45] explicitly model uncertainty at multi-grained levels, enabling more nuanced and robust composed image retrieval that can handle polysemy and ambiguity in search intentions. Recent cross-modal retrieval uncertainty estimation methods, whether multi-grained or contrastive, optimize representation similarity via explicit feature-space constraints while neglecting retrieval trajectory consistency. UATTA implicitly optimizes the embedding space by leveraging the inherent consistency of correct retrieval. Specifically, our bidirectional retrieval disagreement mechanism formulates uncertainty estimation with the inherent retrieval trajectory-symmetric nature.

3 Method

In this section, we introduce the proposed Uncertainty-aware Test-Time Adaptation (UATTA) framework for text-based person search, as illustrated in Fig. 3. Firstly, we introduce a dynamic sample selection strategy based on the cycle-consistency to select reliable samples where the original text query can be successfully recovered. Based on the selected samples, we perform the uncertainty estimation. Finally, we integrate estimated uncertainty into test-time adaptation via entropy recalibration resulting in mitigating the adverse effects of erroneous gradients induced by overconfident false positives. It is noted that UATTA applies seamlessly in both CLIP-based one-stage and XVLM-based two-stage models.

3.1 Similarity Matrix Generation.

For a given text query t_q, text-to-image retrieval aims to select corresponding most similar images from the image gallery set $\mathcal{G}_I = \{i_g\}_g, g \in [1, N_I]$. Within our pretrain-then-adapt paradigm,

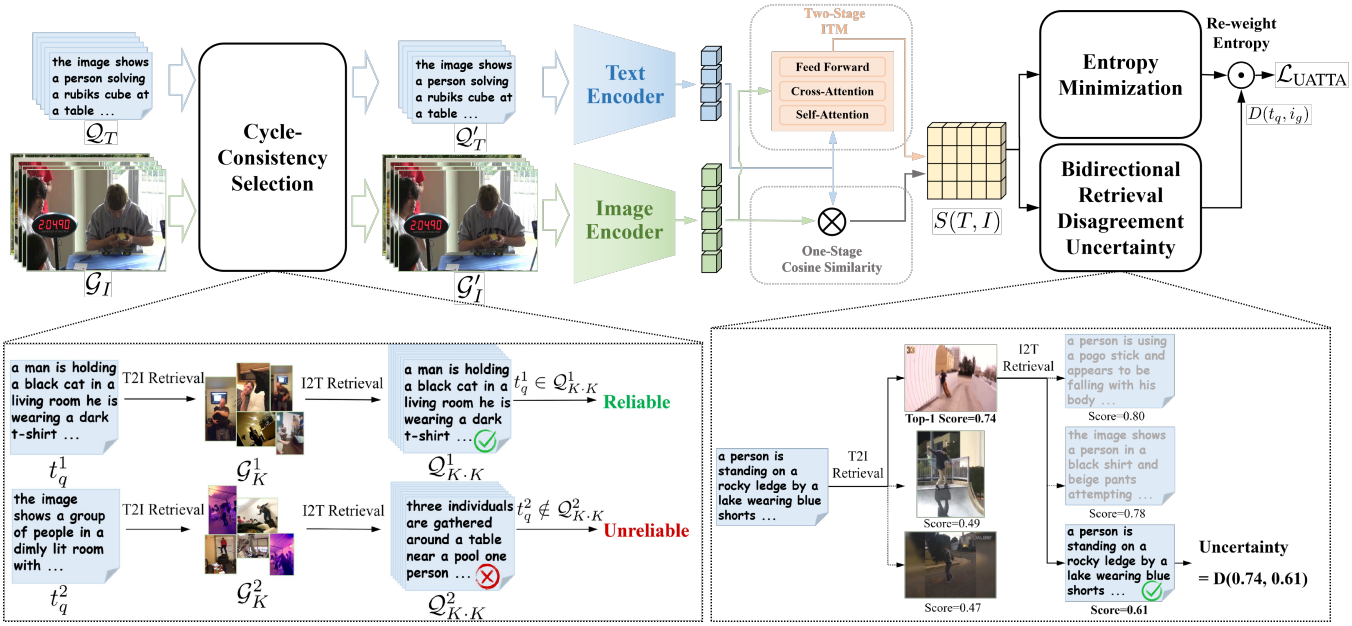


Figure 3: Uncertainty-aware Test-Time Adaptation Framework(UATTA). Given the image gallery set $\mathcal{G}_I$ and text query set $\mathcal{Q}_T$ in the test set, we first select reliable samples by Cycle-Consistency Selection and obtain reliable text query set $\mathcal{Q}'_T$ and reliable image gallery set $\mathcal{G}'_I$. Then we compute the similarity matrix $S(T,I)$ for every pairs to calculate uncertainty. Finally we re-weight the entropy minimization objective with calculated uncertainty. As shown in Cycle-Consistency Selection stage, we sort samples who are mutual top-K neighbors, in which the initial given text t_q^1 is supposed to be inversely found in $\mathcal{G}^1_{K\cdot K}$ by images in retrieval result set $\mathcal{G}^1_K$ of t_q^1. Conversely, text t_q^2 is unreliable as image set $\mathcal{G}^2_{K\cdot K}$ do not contain itself. After selection, based on the reliable images and texts, we further exploits the Bidirectional Retrieval Disagreement mechanism to estimate uncertainty with both text-to-image top-1 retrieval probability and inverse image-to-text retrieval probability, as detailed in Bidirectional Retrieval Disagreement Uncertainty stage. This uncertainty signal is calculated by $D(t,i)$, as detailed in Eq. 9 and dynamically re-weights the entropy minimization objective to uncertainty-weighted gradient re-calibration loss $\mathcal{L}_{\text{UATTA}}$, as detailed in Eq. 10. Our UATTA framework mitigates domain gaps with minimal adaptation cost and zero extra architecture.

the whole text query set $\mathcal{Q}_T$ is accessible during adaptation. We employ the encoders of cross-modal retrieval model to map images in $\mathcal{G}_I$ and texts in $\mathcal{Q}_T$ into a shared embedding space. Subsequently, we compute the similarity scores $s(t_q, i_g)$ between pairs, forming similarity matrix $S(T,I)$. For CLIP-based one-stage retrieval models, the similarity scores is computed using cosine similarity, $s(t_q, i_g) = \cos\left(\mathcal{E}_{\text{T}}(t), \mathcal{E}_{\text{I}}(i)\right)$, where $\mathcal{E}_{\text{T}}, \mathcal{E}_{\text{I}}$ are modality-specific encoders. For XVLM-based two-stage retrieval models, the matching score $s(t_q, i_g) = \mathcal{E}_{\text{ITM}}(\mathcal{E}_{\text{T}}(t), \mathcal{E}_{\text{I}}(i))$ is obtained through an additional image-text matching (ITM) module $\mathcal{E}_{\text{ITM}}$.

3.2 Cycle-Consistency Selection

During our pretrain-then-adapt paradigm, we first introduce the Cycle-Consistency Selection (CCS) to select reliable samples, identifying those queries that fall within the mutual top-K rankings. Given the text query t_q, we first retrieve the top-K most similar images with it to form $\mathcal{G}_K$. Subsequently, for each image in $\mathcal{G}_K$, we perform reverse retrieval to obtain its top-K text candidates. Together, these candidates form the set $\mathcal{Q}_{K\cdot K}$. We consider t_q a reliable sample if and only if it is present in $\mathcal{Q}_{K\cdot K}$. Formally, we define the reliability indicator $r(t_q) \in \{0,1\}$ as:

$$r(t_q) = \begin{cases} 1, \text{if } t_q \in \mathcal{Q}_{K\cdot K}, \\ 0, \text{if } t_q \notin \mathcal{Q}_{K\cdot K}. \end{cases} \tag{1}$$

Then we define $\mathcal{Q}'_T = \{t'_q\}$ as the reliable text query set, where $r(t'_q) = 1, t'_q \in \mathcal{Q}_T$, and define $\mathcal{G}'_I = \{i'_g\}$ as the reliable image gallery set, where i'_g is the retrieval candidates of t'_q. This sample selection retains samples with good retrieval cycle-consistency, guaranteeing them to act as reliable anchors for the subsequent adaptation, which are useful for generalization in adaptation, and discards highly inconsistent pairs, which are harmful false positives otherwise introducing detrimental noise into the optimization process. Interpretation of K.

Generally, K controls the trade-off between reliability and selectivity, in which a larger K provides more stable cycle consistency by incorporating a broader set of candidates, while a smaller K enforces stricter selection, reducing the influence of noisy matches. Empirically, we observe that the optimal K often correlates with the number of ground-truth positives per query and it can be interpreted as an approximation of the local neighborhood size in the embedding space. We further find that performance remains stable

within a reasonable range of K, indicating that the method is not overly sensitive to this hyperparameter.

3.3 Bidirectional Retrieval Disagreement Uncertainty

Uncertainty form a Bayesian Perspective. Uncertainty is typically delineated into aleatoric (data) and epistemic (model) components within a Bayesian framework [21]. Drawing upon this taxonomy, we propose to quantify retrieval uncertainty through behavioral observation, which captures these inherent ambiguities. We define the model's uncertainty as the variance of its parameters, $\mathrm{Unc}(\theta) := \mathrm{Var}(\theta)$, a standard definition from a Bayesian perspective where parameters are treated as random variables. A large $\mathrm{Var}(\theta)$ signifies high uncertainty in the learned weights. Higher parameter variance correlates with elevated uncertainty in learned representations, enabling principled uncertainty-aware adaptation through gradient reweighting.

However, directly computing the parameter variance is computationally intractable in deep neural networks. To address this, we propose a tractable proxy named **Bidirectional Retrieval Disagreement**, denoted as $D(t_q, i_g)$. We posit that the epistemic uncertainty of a retrieval model can be effectively quantified by measuring the inconsistency between its multi-modal encoders. Concurrently, given a pair (t_q, i_g), the bidirectional retrieval disagreement is defined as the difference between the distinct text-to-image retrieval probability p_{T2I} and image-to-text retrieval probability p_{I2T}:

$$D(t_q, i_g) := \left\| p_{T2I}(y|t_q, i_g, \theta) - p_{I2T}(y|t_q, i_g, \theta) \right\|, \tag{2}$$

y is a latent matching variable, which can not be observed during adaptation, denoting whether i_g and t_q is truly-matched or not. Importantly, y is not required in practice, and the formulation is only used for conceptual explanation. Then in this context, $p_{T2I}(y|t_q, i_g, \theta)$, $p_{I2T}(y|t_q, i_g, \theta)$ denote the probabilities of text-to-image and image-to-text search predictions. To operationalize this metric, we instantiate probability as temperature-scaled softmax of similarity scores, which are parameterized by the model weights θ, and compute over the top-K retrieved matches to focus on hard candidates, as follows:

$$\begin{aligned} p_{T2I}(y|t_q, i_g) &= \frac{\exp(s'(t_q, i_g))}{\sum_{j=1}^{K} \exp(s'(t_q, i_j))}, \\ p_{I2T}(y|t_q, i_g) &= \frac{\exp(s'(i_g, t_q))}{\sum_{j=1}^{K} \exp(s'(i_g, t_j))}, \end{aligned} \tag{3}$$

where $s'(t_q, i_g)$ denotes top-K similarity matrix derived from $s(t_q, i_g)$, $s'(i_g, t_q)$ denotes top-K similarity matrix from its transpose $s(t_q, i_g)^\top$. We adopt a standard softmax formulation with the temperature fixed to 1 (*i.e.*no additional scaling), and thus omit the temperature term in the formulation. While the functions p_{T2I} and p_{I2T} depend on different subsets of parameters (*e.g.*, separate modal prediction modules), our analysis considers uncertainty over the entire parameter vector θ.

Theoretical Justification. We now provide a theoretical sketch to justify the proportionality between the parameter variance and our proposed proxy, *i.e.*, $\mathrm{Var}(\theta) \propto D(t_q, i_g)$. The proof is based on the principle of symmetric consistency. An idealized model with zero uncertainty ($\mathrm{Var}(\theta) = 0$) can be represented by a single set of optimal parameters, $\theta_0 = E[\theta]$. Such a deterministic model, if well-trained, should exhibit symmetric predictions, meaning the probability of retrieving i_g from t_q is consistent with retrieving t_q from i_g. Consequently, for this ideal model, the prediction disagreement is negligible:

$$D(t_q, i_g)|_{\theta=\theta_0} = \left\| p_{T2I}(y|t_q, i_g, \theta_0) - p_{I2T}(y|t_q, i_g, \theta_0) \right\| \approx 0. \tag{4}$$

In a realistic model, however, uncertainty implies that $\mathrm{Var}(\theta) > 0$. The parameters θ are subject to perturbations around their mean θ_0. These parameter perturbations disrupt the model's symmetric consistency, as they affect the distinct computational paths of p_{T2I} and p_{I2T} differently.

To formalize the relationship between parameter variance and prediction disagreement, we analyze the effect of these perturbations using a first-order Taylor expansion of the prediction functions around θ_0 (here we omit t and i for simplicity):

$$p_{T2I}(y|\theta) \approx p_{T2I}(y|\theta_0) + (\theta - \theta_0)^T \nabla_\theta p_{T2I}(y|\theta_0), \tag{5}$$

$$p_{I2T}(y|\theta) \approx p_{I2T}(y|\theta_0) + (\theta - \theta_0)^T \nabla_\theta p_{I2T}(y|\theta_0). \tag{6}$$

By substituting these into the definition of $D(t, i)$, we obtain:

$$\begin{aligned} D(t_q, i_g) \approx \Big\| &(p_{T2I}(y|\theta_0) - p_{I2T}(y|\theta_0)) \\ &+ (\theta - \theta_0)^T (\nabla_\theta p_{T2I}(y|\theta_0) - \nabla_\theta p_{I2T}(y|\theta_0)) \Big\|. \end{aligned} \tag{7}$$

Applying the symmetric consistency assumption, where $p_{T2I}(y|\theta_0) - p_{I2T}(y|\theta_0) \approx 0$, the expression simplifies to:

$$D(t_q, i_g) \approx \left\| (\theta - \theta_0)^T (\nabla_\theta p_{T2I}(y|\theta_0) - \nabla_\theta p_{I2T}(y|\theta_0)) \right\|. \tag{8}$$

This result demonstrates that the magnitude of the prediction disagreement $D(t_q, i_g)$ is directly dependent on the parameter deviation $(\theta - \theta_0)$. Since $\mathrm{Var}(\theta) = E[(\theta - \theta_0)^2]$ measures the expected squared magnitude of this deviation, a larger parameter variance will lead to a larger expected prediction disagreement. This establishes the proportionality $\mathrm{Var}(\theta) \propto D(t_q, i_g)$, validating the use of prediction disagreement as a computationally efficient and theoretically grounded proxy for model uncertainty.

3.4 Uncertainty-aware Test-Time Adaptation

From the preceding Cycle-Consistency Selection quantification procedure, we obtain a reliable subset with good retrieval cycle consistency. Leveraging this curated sample set, we perform adaptation through entropy minimization with uncertainty-weighted gradients, effectively instantiating the principle of input-dependent loss attenuation in Bayesian framework [21]. This strategy aligns the model's feature distribution to the target domain while preserving cross-modal consistency. Consequently, we bridge the synthetic-to-real domain gap without requiring labeled target-domain supervision. Empirically, we find that raw probability differences defined in Eq. 2 are insufficient to capture bidirectional disagreement. When both p_{T2I} and p_{I2T} approach zero, their absolute difference is negligible, falsely implying high consistency. Therefore, in practice, we normalize the absolute difference by the mean value to penalize low-confidence pairs while preserving consistency for high-confidence matches. Furthermore, we employ exponential amplification to accentuate the discriminative contrast between asymmetric matches (one high probability, one low) and symmetric high-confidence

matches, modifying $D(t_q, i_g)$ as:

$$D(t_q, i_g) := \exp\left(\frac{|p_{T2I}(y|t_q, i_g) - p_{I2T}(y|t_q, i_g)|}{\frac{p_{T2I}(y|t_q,i_g)+p_{I2T}(y|t_q,i_g)}{2}}\right). \tag{9}$$

Normalization prevents degenerate cases where both probabilities are small, while exponential amplification enhances the contrast between asymmetric matches and symmetric high-confidence matches. Further ablation studies analyzing the contribution of each component are provided in Sec. 4.3. This design is consistent with common practices in uncertainty calibration, where normalization and scaling are used to improve discriminative behavior.

To date, previous TTA method[46] employs entropy minimization objective $\mathcal{L}_{\text{Tent}} = -\sum p \log(p)$ for classification adaptation, where p denotes the model's prediction probability distribution, suffering from overconfident predictions on false-positive samples[46, 60]. We thus far reformulate the bidirectional adaptation objective combined with Eq. 9 through uncertainty-weighted gradient re-calibration:

$$\mathcal{L}_{\text{UATTA}} = \sum_{i_g \in \mathcal{G}'_I, t_q \in \mathcal{Q}'_T} \left(\frac{-p_{T2I}(y|t_q, i_g) \log(p_{T2I}(y|t_q, i_g))}{D(t_q, i_g)} + \frac{-p_{I2T}(y|t_q, i_g) \log(p_{I2T}(y|t_q, i_g))}{D(t_q, i_g)} \right), \tag{10}$$

where $\mathcal{G}'_I, \mathcal{Q}'_T$ are the reliable image gallery set and text query set obtain through Cycle-Consistency Selection as described in Sec.3.2.

Analysis. The Bidirectional Retrieval Disagreement $D(t_q, i_g)$ serves as an uncertainty-weighted recalibration mechanism, adaptively modulating the contribution of each text-image pair to the entropy minimization objective. Specifically, low-uncertainty pairs, which predominantly correspond to true positives as illustrated in Fig. 2, receive amplified gradient updates that strengthen cross-modal alignment. Conversely, high-uncertainty pairs undergo gradient suppression, preventing error propagation from ambiguous or false matches. This dual consistency constraint, which requires cycle-consistent retrieval from both text-to-image and image-to-text directions, naturally partitions samples into confident matches and uncertain candidates without auxiliary supervision. Remarkably, UATTA achieves effective label-free adaptation under domain shift through implicit embedding space optimization, with minimal adaptation cost and zero architectural modifications to pretrained vision-language models.

4 Experiment

4.1 Experiment Setting

We conduct experiments on two distinct frameworks for text-based person search: a one-stage retrieval framework and a two-stage retrieve-and-match framework. These choices allow us to evaluate our approach on tasks with varying complexity, from standard retrieval to fine-grained matching. **(1) CLIP-based One-Stage Framework.** For the standard person retrieval task, we adopt the state-of-the-art LuPerson-HAM model as our baseline. Our experiments are conducted on three real-world benchmarks: RSTPReid, CUHK-PEDES, and ICFG-PEDES. A key challenge is that LuPerson-HAM is pre-trained on synthetic annotations, which creates a significant domain gap compared to the human-annotated captions in the test sets. Our test-time adaptation method is designed to bridge this gap. **(2) XVLM-based Two-Stage Framework.** For the more complex person anomaly search task, which requires both coarse-grained retrieval and fine-grained matching, we follow the state-of-the-art CMP model. This model, based on the XVLM architecture, is evaluated on the PAB benchmark. Similar to the one-stage setup, PAB's training data is synthetically generated, while its test data consists of real-world images with human-corrected captions, presenting a clear domain gap that motivates our approach.

Table 1: Quantitative comparison of our proposed Pretrain-then-Adapt paradigm with state-of-the-art methods on Text-based Person Anomaly Search benchmark PAB [52]. The gpu used for post-training is NVIDIA GeForce RTX 3090 GPU. Best results are bold. Second best results are underlined.

Method	Type	Post-train	R@1	R@5	R@10	mAP
Pure pretraining (no adaptation / finetuning)						
MRA [51]	Pretrain	—	9.91	23.66	31.45	17.15
RaSa [1]	Pretrain	—	21.74	27.30	27.96	24.35
WoRA [42]	Pretrain	—	22.25	45.91	53.54	33.39
APTM [53]	Pretrain	—	22.90	45.80	52.38	33.56
CAMeL [56]	Pretrain	—	24.47	50.00	58.75	36.75
IRRA [19]	Pretrain	—	30.59	59.61	68.91	44.41
CLIP [39]	Pretrain	—	47.57	81.55	89.03	62.73
X-VLM [57]	Pretrain	—	71.94	97.78	98.99	83.96
Pretrain-then-Finetune (Pre-FT)						
MRA [51]	Pre-FT	1.06h (4 GPUs)	70.53	94.69	97.47	81.59
APTM [53]	Pre-FT	0.51h (4 GPUs)	72.14	95.30	97.17	82.78
CAMeL [56]	Pre-FT	1.01h (4 GPUs)	74.30	96.79	98.84	84.20
WoRA [42]	Pre-FT	0.88h (4 GPUs)	74.47	96.82	98.48	84.60
IRRA [19]	Pre-FT	19.6h (4 GPUs)	76.39	97.62	99.14	86.33
CLIP [39]	Pre-FT	18.4h (4 GPUs)	77.60	98.84	99.75	87.35
RaSa [1]	Pre-FT	0.74h (4 GPUs)	80.79	98.89	99.65	89.20
X-VLM [57]	Pre-FT	40.5h (4 GPUs)	81.95	98.84	99.19	89.86
X-VLM + CMP [52]	Pre-FT	48.1h (4 GPUs)	84.93	99.09	99.75	91.66
Pretrain-then-Adapt (Pre-Adp)						
X-VLM + SAR [37]	Pre-Adp	0.38h (1 GPU)	73.20	97.87	99.09	84.58
X-VLM + Tent [46]	Pre-Adp	0.23h (1 GPU)	73.50	95.65	97.57	83.71
X-VLM + SHOT [34]	Pre-Adp	0.26h (1 GPU)	73.66	95.80	97.82	83.97
X-VLM + READ [50]	Pre-Adp	0.23h (1 GPU)	74.62	96.00	98.18	84.61
X-VLM + TCR [24]	Pre-Adp	0.25h (1 GPU)	74.92	96.15	97.97	84.72
X-VLM + Ours	Pre-Adp	**0.08h (1 GPU)**	**76.13**	**98.02**	**99.09**	**86.14**

Implementation Details. During test-time adaptation, we optimize only the affine parameters (γ and β) of the Layer Normalization layers within the final six layers of the text encoder. This specific choice is made to maintain consistency with the CMP baseline, where these last six layers correspond to the cross-modal attention blocks essential for image-text matching. We adopt the AdamW optimizer for all experiments. For the LuPerson-HAM baseline, the learning rate is set to $1e-3$, with a query texts number of 32 and a positive-to-negative image sample ratio of 1:3. For the XVLM baseline, the learning rate is $1e-4$, the number of query texts is 16, and the sample ratio is 1:7. The batch size is maintained at a constant 128, configured jointly by the number of query texts and the specified positive-to-negative ratio. The number of adaptation rounds is adjusted based on the test set size, *i.e.*, 50 for PAB and RSTP-Reid, and 10 for ICFG-PEDES and CUHK-PEDES.

Table 2: Quantitative comparison of our proposed Pretrain-then-Adapt paradigm with state-of-the-art direct transfer models and other existing Test-Time-Adaptation, Semi-supervised and Unsupervised methods on real-world text-based person search benchmarks [7, 29, 66]. Best results are bold. Second best results are underlined.

Method	RSTPReid				CUHK-PEDES				ICFG-PEDES			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
Pure pretraining (no adaptation / finetuning)												
CLIP [39]	12.65	27.16	–	11.15	6.67	17.91	–	2.51	13.45	33.85	–	10.31
LuPerson-T [40]	22.40	–	–	17.08	21.88	–	–	19.96	11.46	–	–	4.56
SYNTH-PEDES [67]	42.69	–	–	31.18	57.58	–	–	52.45	57.08	–	–	32.06
LuPerson-MLLM [44]	51.65	74.20	82.85	38.31	38.29	56.60	64.56	20.43	57.61	75.99	82.76	51.45
LuPerson-HAM [20]	59.50	80.05	87.05	44.11	70.59	86.89	91.78	63.39	60.64	**77.50**	**83.26**	35.54
Unsupervised Domain Adaptation												
GAAP [33]	44.45	65.15	75.30	31.21	47.64	67.79	76.08	41.28	27.12	44.91	53.56	11.43
GTR [2]	46.65	70.70	80.65	34.95	48.49	68.88	76.51	43.67	29.64	47.23	55.54	14.20
PSPD [4]	48.50	69.95	78.50	34.83	53.47	72.81	76.57	46.41	38.49	53.40	60.35	16.49
MUMA [32]	54.35	76.05	83.65	40.50	59.52	77.79	84.65	52.75	38.11	56.01	63.96	19.02
Semi-supervised Domain Adaptation												
CMMT [59]	–	–	–	–	57.10	78.14	85.23	–	–	–	–	–
Generation-then-Retrieval [11]	56.45	–	–	44.45	63.87	–	–	57.18	46.46	–	–	26.90
TextReID [13]	–	–	–	–	64.40	81.27	87.96	61.19	–	–	–	–
ECCA [12]	–	–	–	–	68.13	**87.26**	91.88	-	–	–	–	–
Pretrain-then-Adapt (Pre-Adp)												
LuPerson-HAM + CoOp [65]	58.60	79.65	87.50	43.65	70.09	86.48	91.32	63.10	60.28	76.24	82.31	35.16
LuPerson-HAM + SAR [37]	59.55	80.05	87.00	44.12	70.63	86.87	91.79	63.40	60.64	**77.50**	83.25	35.54
LuPerson-HAM + Tent [46]	59.65	79.75	87.30	44.24	70.30	87.02	91.74	63.26	59.59	76.89	82.85	34.86
LuPerson-HAM + READ [50]	59.80	79.90	87.30	44.37	70.06	86.98	91.82	63.12	60.31	77.09	82.96	35.27
LuPerson-HAM + SHOT [34]	60.10	79.85	87.10	44.46	70.43	86.90	91.99	63.30	60.31	76.95	82.86	35.10
LuPerson-HAM + TCR [24]	61.00	80.85	88.35	45.94	70.66	87.21	**92.13**	**63.60**	59.32	75.63	81.63	35.13
LuPerson-HAM + **Ours**	**61.85**	**81.40**	**88.40**	**46.37**	**70.92**	86.89	91.86	63.50	**62.15**	77.31	82.95	**36.11**

4.2 Comparison with State-of-the-arts

Comparison with Pretrain Models. We compare our method with state-of-the-art methods on multiple benchmarks. As shown in Table 1, our method significantly improves +4.19% R@1 compared to pretrained XVLM, which proves the capacity of our Pretrain-then-Adapt paradigm on mitigating domain gaps between unrelated pretrained data and specific person anomaly search data. Notably, our pretrain-then-adapt paradigm achieves significant efficiency gains with merely 0.08 hours of adaptation time. The process of adaptation operates directly on unlabeled test data of target domain, while others need finetuning on labeled train data of target domain, consuming additional post-train burden. Although some models, benefiting from lightweight fine-tuning modules, reduce post-train time from dozens of hours to approximately one hour, they still require 4 NVIDIA GeForce RTX 3090 GPU whereas only single 3090 GPU for ours. The efficiency gains become particularly significant when considering practical deployment constraints in privacy-sensitive and resource-constrained environments. We observe a similar improvement on three text-based person search benchmarks in Table 2. The results show that the R@1 score increases 2.35%, 0.33% and 1.51% on RSTPReid, CUHK-PEDES and ICFG-PEDES respectively, and the mAP score is improved by 2.26, 0.29 and 0.57. These boosts underscore the efficacy of our proposed bidirectional retrieval disagreement uncertainty and sample selection in mitigating the impact from false positives, which generally refines model to be overconfident in traditional entropy minimization test-time adaptation methods.

Comparison with other TTA methods. We modify others test-time adaptation methods, *i.e.*, Tent[46], SHOT[34], SAR[37], READ[50], TCR[24] from fully Test-Time Adaptation paradigm[46] to our Pretrain-then-Adapt paradigm on RSTPReid, CUHK-PEDES and ICFG-PEDES. As shown in Table 1, Our method demonstrates superior performance and efficiency, achieving gains of 1.21% in R@1 and 1.42% in mAP over all compared baselines, with 0.15 fewer hours of adaptation time on 1 gpu. As shown in Table 2, it is evident on ICFG-PEDES that all test-time adaptation methods fail and our method outperforms baseline by 1.51% R@1 and 0.57% mAP, but our method also has performance degradation at R@5 and R@10,

Table 3: Comparison of Uncertainty Formulations on RSTPReid [66] benchmark. p_{T2I} is the text-to-image retrieval probability, p_{I2T} is the inverse retrieval probability, which uses the gallery image from p_{T2I} to retrieve the corresponding query text. N_{T2I} is the size of image gallery per identity. N_{I2T} is the size of text query per identity. ϵ is a small constant to prevent divided by zero. $\mathbf{s}_{\text{T2I}}^{\text{top-}K}$ denotes the top-K similarity matrix of text-to-image retrieval. Equally, $\mathbf{s}_{\text{I2T}}^{\text{top-}K}$ denotes the top-K similarity matrix of inverse directional image-to-text retrieval. The similarity matrix is transformed by a softmax function to obtain retrieval probabilities for the top-K results. Best results are bolded.

Uncertainty Formulation	Bidirectional Retrieval Probability	RSTPReid R1	R5	R10	mAP
$\exp\left(1-\frac{p_{\text{T2I}}+p_{\text{I2T}}}{2}\right)$	$p_{\text{T2I}}=\text{softmax}(\mathbf{s}_{\text{T2I}}^{\text{top-}K})$ $p_{\text{I2T}}=\text{softmax}(\mathbf{s}_{\text{I2T}}^{\text{top-}K})$	61.40	80.50	87.75	46.16
$\lvert\log(p_{\text{T2I}}+\varepsilon)-\log(p_{\text{I2T}}+\varepsilon)\rvert$	same as above	61.75	81.30	88.40	45.88
$\exp\left(\frac{\lvert p_{\text{T2I}}-p_{\text{I2T}}\rvert}{\frac{p_{\text{T2I}}+p_{\text{I2T}}}{2}}\right)$	same as above	**61.85**	81.40	88.40	46.37
$\exp\left(\frac{\lvert p_{\text{T2I}}\cdot N_{\text{T2I}}-p_{\text{I2T}}\cdot N_{\text{I2T}}\rvert}{\frac{p_{\text{T2I}}\cdot N_{\text{T2I}}+p_{\text{I2T}}\cdot N_{\text{I2T}}}{2}}\right)$	same as above	61.75	81.35	88.60	**46.58**
$\exp\left(\frac{\lvert p_{\text{T2I}}-p_{\text{I2T}}\cdot\frac{N_{\text{I2T}}}{N_{\text{T2I}}}\rvert}{\frac{p_{\text{T2I}}+p_{\text{I2T}}\cdot\frac{N_{\text{I2T}}}{N_{\text{T2I}}}}{2}}\right)$	same as above	61.70	81.45	88.60	46.57
$\exp\left(\frac{\lvert p_{\text{T2I}}\cdot N_{\text{T2I}}-p_{\text{I2T}}\cdot N_{\text{I2T}}\rvert}{\frac{p_{\text{T2I}}\cdot N_{\text{T2I}}+p_{\text{I2T}}\cdot N_{\text{I2T}}}{2}}\right)$	$p_{\text{T2I}}=\text{softmax}(\mathbf{s}_{\text{T2I}}^{\text{top-}K}\cdot N_{\text{T2I}})$ $p_{\text{I2T}}=\text{softmax}(\mathbf{s}_{\text{I2T}}^{\text{top-}K}\cdot N_{\text{I2T}})$	61.75	**81.90**	**88.90**	46.47
$\exp\left(\frac{\lvert p_{\text{T2I}}/N_{\text{T2I}}-p_{\text{I2T}}/N_{\text{I2T}}\rvert}{\frac{p_{\text{T2I}}/N_{\text{T2I}}+p_{\text{I2T}}/N_{\text{I2T}}}{2}}\right)$	same as above	61.30	81.40	88.55	46.04

because the bidirectional retrieval disagreement mechanism is designed to rectify the harmfulness from top-1 false positives and neglects possible potential true positives in top-2 to top-10 range. This is a future direction for us to explore smooth utilization of these potential true positives in edge zone. Similar situation occurs on CUHK-PEDES, as our method achieves best R@1 of 70.92% but is inferior to TCR on R@5, R@10 and mAP. On RSTPReid, our method surpasses other existing methods.

Comparison with other Semi-supervised and Unsupervised Methods. Generally, unsupervised [2, 4, 32, 33] and semi-supervised [11–13, 59] paradigms for text-based person search leverage advanced VLMs to synthesize pseudo-annotations, serving as proxies for supervised image-text pairs. However, this reliance on synthetic data inevitably introduces intrinsic domain shifts. In contrast, our approach performs direct adaptation on the test data. Despite the absence of ground-truth pairings, the textual descriptions remain aligned with the target domain. Consequently, our method focuses on mitigating the distribution shifts of the pretrained model, avoiding the noisy discriminative supervision characteristic of prior approaches. As evidenced in Table 2, existing unsupervised and semi-supervised methods struggle to fully leverage the pretrained model's capacity, often compromising representation quality due to label noise, thereby limiting their practical generalization potential in complex real-world deployment scenarios.

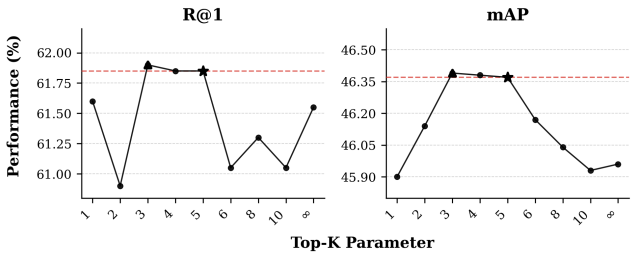


Figure 4: Ablation study of bidirectional Top-K retrieval consistent sample selection on RSTPReid. K denotes the mutual top range in bidirectional retrieval. Best performance is achieved at $K = 3$. Since each identity in RSTPReid contains 5 ground-truth images, we adopt $K = 5$ as the default setting to represent the borderline of true and false positives.

4.3 Ablation Studies and Further Discussion

Effect of Uncertainty Formulation. We present an ablation study on the formulation of uncertainty in Table 3. The core thought is to assign a lower uncertainty on both top retrieval directions and a higher uncertainty while only uni-directional retrieval works. Formulation 1 in Table 3 only considers the higher similarity of true positives but ignore the difference between TP and FP. At a opposite perspective, formulation 2 focuses on the difference neglecting the absolute numerical magnitude. Combining with two views, formulation 3 achieves best score on RSTPReid, while the others are scaled version based on formulation 3 to balance the number of positive samples in the two retrieval directions. The extreme amplifications and balances destroy the suitable consistent distribution of TP and FP, then consequently weaken performance at R@1 score, which is the primary standard we use to choose formulation.

Effect of K Mutual Neighbours. We conduct an ablation study on the hyper-parameter K in the Cycle-Consistency Selection, as shown in Fig. 4. Based on empirical results, we adopt $K = 5$ as the default setting without dataset-specific tuning, which achieves a favorable balance between performance and stability.

Specifically, smaller values such as $K = 1$ restrict the adaptation process to only highly confident image-text pairs, limiting the diversity of selected samples and reducing generalization ability. In contrast, larger values (e.g., $K = \infty$) include all candidate pairs without selection, introducing a substantial number of false positives with high uncertainty, which negatively impacts adaptation performance. Intermediate values of K allow the model to incorporate both low-uncertainty pairs and moderately uncertain pairs, enabling uncertainty to play an effective role in modulating the adaptation process.

From a methodological perspective, K controls the locality of cycle-consistency: smaller K enforces stricter agreement, while larger K allows greater tolerance to retrieval noise. From a geometric viewpoint, K can be interpreted as approximating the size of the local neighborhood in the embedding space. Empirically, we observe that the optimal K correlates with the number of semantically

Table 4: Ablation of the ratio between positive and negatives on RSTPReid benchmarks. We apply different ratio of positive and negatives to compute entropy. The ratio of 1 : 3 improves the stability in test-time adaptation. Our default setting is in gray.

Ratio	R@1	R@5	R@10	mAP
1 : 1	60.70	81.20	88.50	46.01
1 : 2	61.15	80.75	88.10	45.92
1 : 3	**61.85**	**81.40**	88.40	**46.37**
1 : 5	60.95	80.85	88.05	46.10
1 : 7	61.55	81.35	**88.75**	46.32
1 : 15	60.80	81.00	88.50	46.24

Table 5: Comparison of other prevailing lightweight tuning methods on RSTPReid[66] benchmark.

Method	Tuning Layers	R@1	R@5	R@10	mAP
Baseline	–	60.64	77.50	83.26	35.54
CoOp*[65]	Prompt Emb.	58.60	79.65	87.50	43.65
Prefix-Tuning*[30]	Prefix Emb.	26.25	51.45	63.75	23.14
LoRA*[16]	LoRA Matrix	49.80	73.80	82.70	38.61
Ours	Norm. Layer	**61.85**	**81.40**	**88.40**	**46.37**

similar instances per query, and values within the range $K \in [3, 8]$ consistently yield stable performance across datasets. Compared to the baseline performance of 58.50% in R@1 and 44.11% in mAP, our method consistently improves performance across all choices of K. Specifically, R@1 varies only within a narrow range of 60.90% to 61.90% (a fluctuation of 1.0% absolute point), and mAP varies from 45.90% to 46.40% (a fluctuation of 0.5% absolute point). Notably, this variation is significantly smaller than the overall performance gain over the baseline (+2.40% to +3.40% in R@1 and +1.79% to +2.29% in mAP), indicating that the improvement is robust and not sensitive to the specific choice of K. Importantly, our method is not sensitive to the exact choice of K within this range, as demonstrated in Fig. 4. This further supports that the choice of K does not require careful tuning in practice. Although the ablation is conducted on RSTPReid, we apply the same default setting across all datasets and observe consistent performance improvements, suggesting that the choice of K generalizes well in practice. This behavior can be attributed to the fact that the local neighborhood structure in the embedding space is relatively stable across datasets. The effectiveness of moderate K values is also consistent with our uncertainty formulation, which benefits from a balance between confident and moderately uncertain pairs.

Effect of Negative Samples. In Table 4, we compare several experiments of the ratio between positive and negatives for one query. The optimal performance is presented with configuration of 1 : 3 on RSTPReid. This suggests that a suitable choice of ratio enhances the adaptation process with softmax entropy based on a formulation akin to multiple classification.

Comparison with Lightweight Tuning Methods. We compare our method with lightweight tuning methods in Table 5. Baseline is LuPerson-HAM [20] and * means that we try different hyperparameters, *i.e.*, learning rate, number of virtual tokens, rank of LoRA[16] etc., for lightweight tuning methods and selected the best result. CoOp[65], which is a prompt learning method and belong to few-shot learning, fails with adaptation objective of entropy minimization. This failure suggests that learnable prompt tokens require labeled data to be grounded in a semantically meaningful embedding space mimicking natural language. In the absence of supervision, the adaptation process merely adjusts the cross-modal feature distribution while disregarding the intrinsic semantic representation. Additionally, Parameter Efficient Fine-Tuning (PEFT) [14] 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 [30] and LoRA [16], 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.

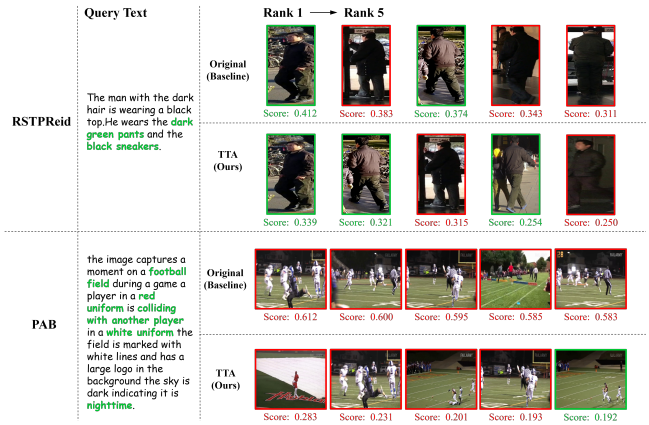


Figure 5: Top-5 Text-based Person Search Results on RSTPReid and PAB. The figure presents the Top-5 retrieval results for representative text queries on the RSTPReid and PAB, where the similarity score of each retrieved image is reported below the corresponding result. Correctly matched person images are highlighted with green bounding boxes, while false matches are indicated in red. On RSTPReid, our method consistently promotes more ground-truth matches to higher ranks, demonstrating improved ranking quality under the text-to-image retrieval setting. In contrast, results on PAB illustrate that our approach effectively mitigates overconfident false positives by re-calibrating retrieval scores, thereby recovering correct matches that are suppressed by the baseline. These observations highlight the robustness of the proposed UATTA across different dataset characteristics.

4.4 Qualitative Results

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 RSTPReid and PAB benchmarks in Fig. 5. The visualization effectively showcases two key strengths of UATTA: Firstly, in challenging cases on RSTPReid 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. Secondly, for scenarios

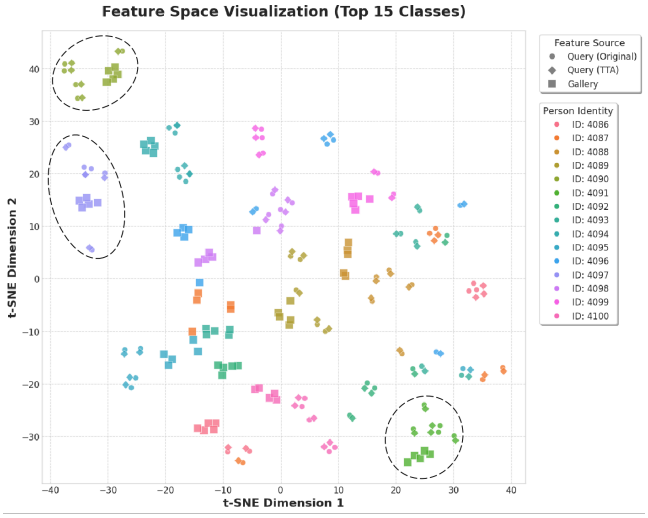


Figure 6: T-SNE Visualization of Feature Space Shifts on RSTPReid. 3 distinct point types represent: original query features (circles) before TTA, query features after TTA (diamonds), and gallery features (squares). Different colors distinguish individual person identities.

requiring fine-grained semantic distinction on PAB, UATTA leverages the bidirectional retrieval disagreement proxy to effectively disambiguate subtle differences between the text and image modalities. This mechanism allows UATTA to promote more ground-truth matches to higher ranks. Overall, the qualitative results confirm that UATTA achieves robust and accurate confidence distribution by re-calibrating retrieval scores, validating its superiority in handling both retrieval ambiguity and fine-grained visual differences across different dataset characteristics.

Visualization of Feature Space Shifts. In Fig. 6, 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 some identities still exhibit residual ambiguity after TTA, suggesting potential avenues for future improvement in feature consolidation.

4.5 Computational Cost Analysis

Complexity. The dominant cost comes from computing the similarity matrix $S(T, I)$, which is $\mathbf{O}(|T||I|)$, required by all retrieval baselines. Bidirectional retrieval introduces only a constant-factor overhead (2× similarity lookup), without extra feature encoding. Since all image and text embeddings are precomputed, reverse retrieval does not require additional forward passes and feature extraction. Therefore, the overhead is negligible compared to feature encoding.

Memory. Since our Pretrain-then-Adapt paradigm performs in an offline test-time adaptation manner which belongs to transductive learning setting. The similarity matrix and Cycle-Consistency Selection are precomputed inevitably on the entire dataset once time before the adaptation, which needs additional memory overhead according to the scale of different dataset. However, in the practical adaptation process, we only employs a limited number of positive and negative matches as shown in Table4 in a batched manner, avoiding full materialization in memory.

5 Discussion and Conclusion

In this work, we introduce a practical and label-free Pretrain-then-Adapt paradigm for text-based person search. We propose Uncertainty-Aware Test-Time Adaptation (UATTA), which leverages unlabeled test data to recalibrate predictions under domain shift. Its core component, Bidirectional Retrieval Disagreement (BRD), estimates uncertainty via discrepancies between text-to-image and image-to-text retrieval probabilities, effectively suppressing overconfident false positives while preserving reliable alignments. Extensive experiments on four benchmarks and both CLIP-based and XVLM-based architectures demonstrate consistent performance gains without requiring target-domain annotations or architectural changes.

Robustness under domain shift and noisy samples. Under ambiguous text or low-quality images, both retrieval directions may become uniformly uncertain, reducing alignment reliability. In such cases, Cycle-Consistency Selection (CCS) may discard hard but correct samples, reflecting a trade-off between noise reduction and sample coverage, and partially explaining the drop in R@5 and R@10. Nevertheless, uncertainty-aware entropy re-calibration mitigates this issue by suppressing unreliable updates, improving robustness under moderate domain shifts.

Self-consistency vs. correctness. Bidirectional Retrieval Disagreement (BRD) measures model self-consistency rather than correctness and relies on a near-deterministic pretrained model assumption. Our analysis, based on a first-order Taylor approximation, provides an intuitive rather than rigorous guarantee and is validated empirically. The method may fail in consistent-but-wrong scenarios caused by spurious correlations or calibration shifts, which require advances in domain-robust representation learning.

Overall, these limitations define the boundary of applicability but do not affect our main conclusion: uncertainty-aware test-time adaptation is an effective and efficient solution for label-free deployment under realistic domain shifts.

6 Acknowledgement

We acknowledge supports from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, the Jiangsu Provincial Science and Technology Program (Grant No. SBZ20250900116), the University of Macau MYRG-GRG2024-00077-FST-UMDF, and the Macao Science and Technology Development Fund Grant FDCT/0043/2025/RIA1.

References

[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023. Rasa: Relation and sensitivity aware representation learning for text-based person search. *arXiv:2305.13653* (2023).

[2] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. 2023. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*. 757–767.

[3] Maryam Bukhari, Sadaf Yasmin, Sheneela Naz, Muazzam Maqsood, Jehyeok Rew, and Seungmin Rho. 2023. Language and vision based person re-identification for surveillance systems using deep learning with LIP layers. *Image and Vision Computing* 132 (2023), 104658.

[4] Feng Chen, Jielong He, Yang Liu, Heng Liu, Zhe Chen, and Yaxiong Wang. 2025. Unsupervised Cross-Modal Person Search via Progressive Diverse Text Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. Association for Computing Machinery, 6047–6056.

[5] Pengxu Chen, Huazhong Liu, Jihong Ding, Xinghao Huang, Shaojun Zou, and Laurence Tianruo Yang. 2025. Class Activation Values: Lucid and Faithful Visual Interpretations for CLIP-based Text-Image Retrievals. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 844–853.

[6] Weijing Chen, Linli Yao, and Qin Jin. 2023. Rethinking Benchmarks for Cross-modal Image-text Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. 1241–1251.

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

[8] Qichao Dong, Lingyu Liu, Yaxiong Wang, Jason J. R. Liu, and Zhedong Zheng. 2025. Domain-Agnostic Neural Oil Painting via Normalization Affine Test-Time Adaptation. In *ACM Multimedia - BNI Track*.

[9] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. 2021. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14750–14759.

[10] Bipin Gaikwad and Abhijit Karmakar. 2023. Real-time distributed video analytics for privacy-aware person search. *Computer Vision and Image Understanding* 234 (2023), 103749.

[11] Daming Gao, Yang Bai, Min Cao, Hao Dou, Mang Ye, and Min Zhang. 2025. Semi-Supervised Text-Based Person Search. *IEEE Transactions on Image Processing* 34 (jan 2025), 5888–5903.

[12] Tiantian Gong, Junsheng Wang, and Liyan Zhang. 2024. Enhancing cross-modal completion and alignment for unsupervised incomplete text-to-image person retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24)*. Article 88, 9 pages.

[13] Xiao Han, Sen He, Li Zhang, and Tao Xiang. 2021. Text-Based Person Search with Limited Data. In *BMVC*.

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

[15] Bohan Hou, Haoqiang Lin, Xuemeng Song, Haokun Wen, Meng Liu, Yupeng Hu, and Xiangyu Zhao. 2025. FiRE: Enhancing MLLMs with Fine-Grained Context Learning for Complex Image Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 803–812.

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

[17] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. pmlr, 448–456.

[18] Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems* 34 (2021), 2427–2440.

[19] Ding Jiang and Mang Ye. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. 2025. Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9220–9230.

[21] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).

[22] Samee Khan, Tanveer Hussain, Amin Ullah, and Sung Baik. 2021. Deep-ReID: Deep Features and Autoencoder Assisted Image Patching Strategy for Person Re-identification in Smart Cities Surveillance. *Multimedia Tools and Applications* 83 (01 2021). doi:10.1007/s11042-020-10145-8

[23] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. 2024. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *ICLR* (2024).

[24] Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, XitingLiu, and Mouxing Yang. 2025. Test-time Adaptation for Cross-modal Retrieval with Query Shift. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=BmG88rONaU

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[28] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. 2024. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3172–3180.

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

[30] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.

[31] Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. 2025. Revolutionizing Text-to-Image Retrieval as Autoregressive Token-to-Voken Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. 813–822.

[32] Zongyi Li, Li Jianbo, Yuxuan Shi, Jiazhong Chen, Shijuan Huang, Linnan Tu, Fei Shen, and Hefei Ling. 2025. Exploring the Potential of Large Vision-Language Models for Unsupervised Text-Based Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5119–5127.

[33] Zongyi Li, Jianbo Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Runsheng Wang, and Shijuan Huang. 2024. Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization*. 1047–1055.

[34] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *International Conference on Machine Learning (ICML)*. 6028–6039.

[35] Vuong D Nguyen, Samiha Mirza, Abdollah Zakeri, Ayush Gupta, Khadija Khaldi, Rahma Aloui, Pranav Mantini, Shishir K Shah, and Fatima Merchant. 2024. Tackling domain shifts in person re-identification: A survey and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4149–4159.

[36] Kai Niu, Liucun Shi, Ke Han, Qinzi Zhao, Yue Wu, and Yanning Zhang. 2025. Test-Time Adaptation for Text-Based Person Search. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. 2997–3006.

[37] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400* (2023).

[38] Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. 2023. Learnable pillar-based re-ranking for image-text retrieval. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 1252–1261.

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

[40] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11174–11184.

[41] Liangxu Su, Rong Quan, Zhiyuan Qi, and Jie Qin. 2024. MACA: Memory-aided Coarse-to-fine Alignment for Text-based Person Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 2497–2501.

[42] Jintao Sun, Hao Fei, Gangyi Ding, and Zhedong Zheng. 2025. From Data Deluge to Data Curation: A Filtering-WoRA Paradigm for Efficient Text-based Person Search. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. ACM, 2341–2351. doi:10.1145/3696410.3714788

[43] Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. 2025. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[44] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17127–17137.

[45] Haomiao Tang, Jinpeng Wang, Yuang Peng, GuangHao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, and Shu-Tao Xia. 2025. Modeling Uncertainty in Composed Image Retrieval via Probabilistic Embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1210–1222.

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=uXl3bZLkr3c

[47] Junsheng Wang, Tiantian Gong, and Yan Yan. 2024. Semi-supervised Prototype Semantic Association Learning for Robust Cross-modal Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 872–881.

[48] Yaxiong Wang, Lianwei Wu, Lechao Cheng, Zhun Zhong, Yujiao Wu, and Meng Wang. 2025. Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 792–802.

[49] Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. 208–217.

[50] Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. 2024. Test-time adaptation against multi-modal reliability bias. In *The twelfth international conference on learning representations*.

[51] Shuyu Yang, Yaxiong Wang, Yongrui Li, Li Zhu, and Zhedong Zheng. 2026. Minimizing the Pretraining Gap: Domain-Aligned Text-Based Person Retrieval. *Pattern Recognition* (2026).

[52] Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2025. Beyond Walking: A Large-Scale Image-Text Benchmark for Text-based Person Anomaly Search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11720–11730.

[53] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM international conference on multimedia*. 4492–4501.

[54] Tao Yang, Shenglong Zhou, Yuwang Wang, Yan Lu, and Nanning Zheng. 2022. Test-time batch normalization. *arXiv preprint arXiv:2205.10210* (2022).

[55] Chen Yiyang, Zheng Zhedong, Ji Wei, Qu Leigang, and Chua Tat-Seng. 2024. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Yb5KvPkKQg

[56] Hang Yu, Jiahao Wen, and Zhedong Zheng. 2025. CAMeL: Cross-modality Adaptive Meta-Learning for Text-based Person Retrieval. *IEEE Transactions on Information Forensics and Security* (2025).

[57] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:244129883

[58] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023. AdaNPC: Exploring Non-Parametric Classifier for Test-Time Adaptation. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 41647–41676. https://proceedings.mlr.press/v202/zhang23am.html

[59] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. 2021. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11395–11404.

[60] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2024. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. *ICLR* (2024).

[61] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[62] Zhedong Zheng and Liang Zheng. 2024. 2. object re-identification: Problems, algorithms and responsible research practice. *The Boundaries of Data* (2024), 21.

[63] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[64] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*. 3754–3762.

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

[66] 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.

[67] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. 2024. Plip: Language-image pre-training for person representation learning. *Advances in Neural Information Processing Systems* 37 (2024), 45666–45702.