STaG: Human-Annotation-Free Unsupervised Domain Adaptation for Text-Based Person Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Real-world deployment of Text-Based Person Retrieval (TBPR) is bottlenecked by the absence of textual annotations in new domains. Existing Unsupervised Domain Adaptation (UDA-TBPR) methods still assume access to target-domain human-written annotations (albeit unpaired with images), limiting their practicality. We consider a more practical setting: human-annotation-free UDA-TBPR, with labeled image–text data in the source domain and annotation-free, unlabeled images in the target domain. This setting introduces two core challenges: the absence of cross-modal alignment in the target domain, and visual domain shifts from source to target. To tackle these, we propose STaG, a dual-path framework comprising Style-Agnostic Text Alignment (SATA) and Heterogeneous Graph Memory Alignment (HGMA). Specifically, SATA leverages source-domain data to decouple identity-discriminative semantics from machine-generated stylistic artifacts, establishing robust cross-modal alignment for the target domain. Meanwhile, HGMA exploits cross-domain semantic neighborhoods via a memory-based graph to learn domain-invariant visual features. Ultimately, STaG achieves state-of-the-art performance across three benchmarks under this realistic setting.

1 Introduction

Text-Based Person Retrieval (TBPR) (Li et al., 2025; Bai et al., 2023a; Jiang & Ye, 2023) aims to retrieve images of a specific individual from a gallery given a textual description. Compared to general image–text retrieval (Radford et al., 2021; Li et al., 2021a; Faghri et al., 2018), which typically involves matching across diverse semantic categories, TBPR is uniquely challenging as it requires fine-grained discrimination among highly similar human identities. This task is complicated by significant intra-class variation (*e.g.*, changes in pose, clothing, or lighting for the same person) and limited inter-class distinction (*e.g.*, visual similarity between different individuals). Despite these challenges, TBPR has attracted growing interest (Zhang & Lu, 2018; Jiang & Ye, 2023; Qin et al., 2024; Tan et al., 2024) due to its high practical value in public safety scenarios, including assisted searches for suspects or missing persons in surveillance footage.

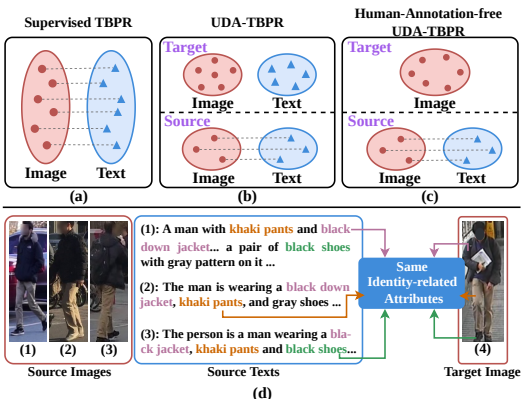


Figure 1: Comparison of different supervision paradigms in TBPR. Our setting relies strictly on unlabeled target images.

Most existing approaches to TBPR operate in a fully supervised setting, relying on large-scale datasets comprising collected images paired with carefully curated human annotations (Fig. 1 (a)). However, in practical deployment scenarios, while image galleries are readily available from surveillance streams, acquiring aligned human-written text annotations for each identity is prohibitively expensive and labor-intensive. This reliance on paired supervision severely hinders the scalability of current methods to new environments and poses a significant barrier to real-world adoption. A promising direction is to introduce unsupervised domain adaptation into TBPR (UDA-TBPR),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where cross-modal knowledge is transferred from a labeled source domain to the unlabeled target domain. However, existing UDA-TBPR methods (Jing et al., 2020; Luo et al., 2025; Niu et al., 2026b) typically assume that human-written textual descriptions for the target domain are available, only unpaired with images (Fig. 1 (b)). This assumption, however, is often unrealistic. In real-world scenarios, person-specific textual descriptions do not exist independently; they are derivative of visual inspection, explicitly generated by annotators observing specific identities. Therefore, the existence of target-domain human texts implies that a latent image-text correspondence has already been established manually. Consequently, *these methods still inevitably incur prohibitive manual annotation costs, failing to provide a truly human-annotation-free solution.* This dependence on manual annotation remains a critical bottleneck that impedes practical deployment.

Motivated by this observation, we consider a more practical, human-annotation-free UDA-TBPR setting: the source domain provides paired image-text data, while the target domain contains only unlabeled images (as shown in Fig. 1 (c)). This setting, while realistic, introduces two substantial challenges: ❶ **Lack of Cross-modal Alignment.** The absence of explicit textual anchors in the target domain leaves the joint embedding space severely under-constrained, making it exceptionally difficult to associate target visual identities with textual queries. ❷ **Visual Domain Shift.** Differences in imaging conditions, backgrounds, resolution, and camera styles make representations overfit to source-specific biases, leading to a catastrophic drop in retrieval accuracy when deployed in new environments.

To address these challenges, we propose STaG (Style-agnostic Text and Graph memory alignment), a novel framework that reconstructs cross-modal semantic associations and mitigates cross-domain distribution shifts by leveraging labeled source-domain data and unlabeled target-domain images. Specifically, STaG integrates two complementary components. ❶ **Style-Agnostic Text Alignment (SATA).** To enable explicit cross-modal alignment in the target domain, a prevailing approach is to utilize MLLMs to synthesize pseudo-descriptions that compensate for the missing modality (Tan et al., 2024; Bai et al., 2023b). However, relying exclusively on these generated texts often degrades retrieval robustness, as their stylistic divergence from human-written queries causes the model to overfit to rigid, machine-specific phrasing (as shown in Fig. 4). To mitigate this, SATA introduces a style-invariance learning paradigm that exploits the source domain to teach the model how to decouple underlying identity semantics from superficial textual styles. In this way, the model develops a style-agnostic representation space that enables the framework to utilize target-domain pseudo-texts without being corrupted by generation artifacts. ❷ **Heterogeneous Graph Memory Alignment (HGMA).** To mitigate visual domain shift, we emphasize domain-invariant person cues by leveraging a semantic similarity prior: visually similar images tend to share identity-related attributes (Fig. 1 (d)). Guided by this prior, we can dynamically inject source-domain visual features into target-domain representations to bridge the domain gap. To enhance representational robustness, we introduce a graph structure to capture broader neighborhood topologies simultaneously. Distinct from prior memory-based and graph-based methods (Khasahmadi et al., 2020; Du et al., 2023; Luo et al., 2025) that operate within a single modality or require target-domain human-annotations, HGMA constructs a heterogeneous graph that links each target image to its most relevant cross-domain neighbors, allowing the target representation to overcome domain-specific biases through aggregated neighbor support. Furthermore, to shield this aggregation from noisy neighbors, we introduce a Confidence-Aware Entropy regularizer (CAE-Loss). Rather than minimizing entropy over single-modality classifier logits (Yue et al., 2026; Yang et al., 2023; Zhang et al., 2022), CAE sharpens cross-modal matching distributions to ensure high-quality image-text alignments.

Our contributions can be summarized as follows:

❶ We introduce a practical, human-annotation-free UDA-TBPR setting. By entirely eliminating the reliance on target-domain human annotations, this flexible paradigm offers a more realistic solution for real world deployment.

❷ We propose STaG, a novel framework tailored for our new challenging setting. It incorporates SATA to establish robust cross-modal associations and HGMA to mitigate visual domain shifts, achieving robust alignment without target-domain human annotations.

❸ Extensive experiments across three standard benchmarks show that STaG outperforms state-of-the-art methods and effectively narrows the gap to fully supervised upper bounds, validating its robustness entirely without target-domain human annotations.

2 Related Work

2.1 Supervised TBPR

TBPR aims to retrieve images of a target person from a gallery given a natural-language description. Since the task was introduced by (Li et al., 2017) together with the CUHK-PEDES benchmark, TBPR has evolved into a representative problem at the intersection of person re-identification and vision-language retrieval. Most supervised TBPR methods can be grouped into two lines. ❶ Cross-modal alignment methods learn a shared embedding space for images and texts. Early approaches primarily relied on global feature alignment (Li et al., 2017; Zheng et al., 2020b; Farooq et al., 2020), while subsequent work explored fine-grained correspondences across parts, phrases, or multiple granularities (Chen et al., 2021; Suo et al., 2022; Wang et al., 2024; Qi et al., 2025; Li et al., 2024a). Many of these methods introduce structural priors, such as human parsing, pose estimation, and linguistic toolkits (e.g., NLTK (Loper & Bird, 2002)), to better associate local visual cues with textual entities (Loper & Bird, 2002; Wang et al., 2020; Jing et al., 2019; Zheng et al., 2020a). ❷ Representation learning methods target modality-invariant identity features and robustness under nuisance factors. Typical strategies include auxiliary objectives (e.g., attribute/color reasoning) and debiasing modules (e.g., reducing over-reliance on color) to improve cross-modal discriminability and generalization (Wu et al., 2021; Wang et al., 2022), often coupled with adaptive alignment mechanisms to handle complex scenes (Li et al., 2021b; Gao et al., 2021; 2022; Ji et al., 2022). Recent studies further leverage pretrained vision-language models (VLMs), such as CLIP (Radford et al., 2021), and multimodal LLMs via prompt learning, feature calibration, or knowledge transfer to enhance retrieval performance (Jiang & Ye, 2023; Han et al., 2021; Yan et al., 2022; Liu et al., 2027; Cao et al., 2023; Lu et al., 2025; Bai et al., 2023a). Most recently, the community has explored generative and interactive paradigms. For instance, DSEE (Chen et al., 2026b) employs generative-driven semantic enhancement to enrich cross-modal semantic diversity. Other work incorporates interactive querying and contextual cues to refine query-image matching (Qin et al., 2025; Bai et al., 2025; Xu et al., 2025). Despite these advances, most existing approaches (including VLM-based variants) still rely heavily on paired image–text supervision during training and/or adaptation. When the target domain provides only unlabeled images and no textual descriptions, target images cannot directly participate in cross-modal alignment, making such methods difficult to deploy under a more practical human-annotation-free UDA-TBPR setting.

2.2 Unsupervised Domain Adaptation for TBPR

To mitigate cross-domain distribution shifts, several works have explored Unsupervised Domain Adaptation (UDA) for TBPR. (Niu et al., 2019) proposed fusing direction-specific adaptation strategies for real-world person search; MAN (Jing et al., 2020) addressed the domain gap by aligning feature distributions through moment matching; and GCKD (Luo et al., 2025) constructed cross-modal graphs to propagate relational structures and distill knowledge across domains. However, this graph construction is similarly predicated on the availability of target-domain descriptions to establish the necessary nodes and edges for message passing. Crucially, these methods assume access to target-domain human annotations during adaptation. Such annotations are costly and often impractical to collect in real-world surveillance, motivating a strictly human-annotation-free paradigm.

2.3 Low Resource TBPR

Beyond UDA-TBPR, low-resource TBPR addresses scenarios where the target domain itself is data-scarce and lacks a designated source domain. These works mainly fall into two categories. Weakly supervised and semi-supervised TBPR reduce identity-level supervision by assuming paired image–text data without identity labels, or a small number of labeled pairs together with abundant unlabeled data (Zhao et al., 2021; 2025; Fu et al., 2025; Gong et al., 2024; Zhang et al., 2025; Gao et al., 2024). Despite reduced labeling requirements, these methods still fundamentally rely on paired cross-modal data and typically require textual descriptions to enable effective cross-modal alignment. Fully unsupervised TBPR further eliminates the need for both identity labels and explicit image–text pairings and instead relies on pseudo texts generated or rewritten by VLMs/MLLMs and then learning with noise-robust objectives (Bai et al., 2023b; Shao et al., 2023; Li et al., 2024b; Mu et al., 2025; Ye et al., 2026; Niu et al., 2026a; Chen et al., 2026a). While reducing annotation cost, performance is often constrained by pseudo-text quality and noise in the induced supervision. Different from these works, we explicitly leverage the rich, well-annotated resources of the source domain to enhance TBPR performance in the target domain, thereby bridging the gap between cross-domain knowledge transfer and low-resource retrieval.

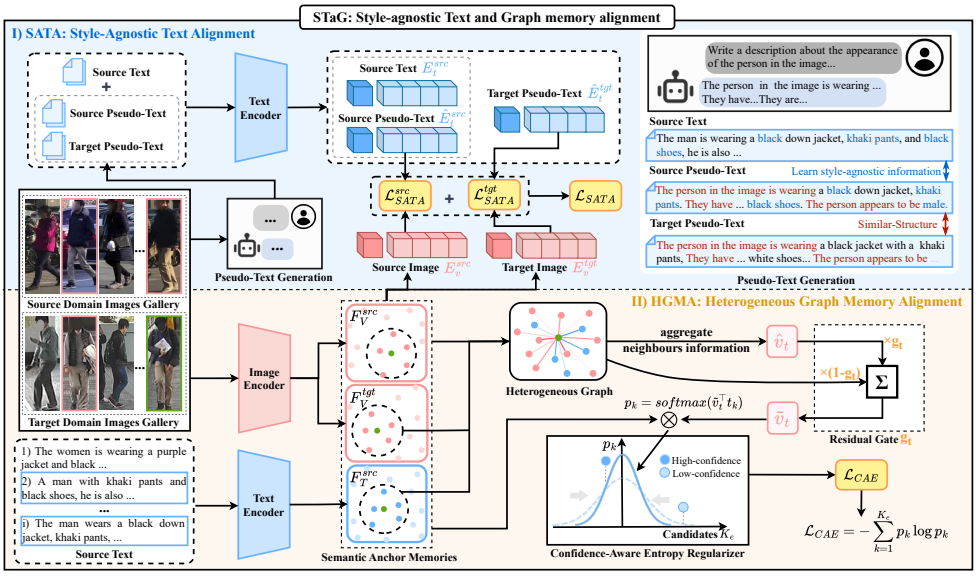


Figure 2: Overview of the proposed STaG. It consists of SATA and HGMA. **SATA** learns style-agnostic text alignment by enforcing identity-level consistency between human-written source descriptions and model-generated pseudo descriptions, producing reliable pseudo-text supervision for target-domain retrieval. **HGMA** mitigates visual domain shift by constructing a heterogeneous graph over cross-domain nearest neighbors and aggregating neighborhood information via semantic anchor memories, with residual gating g_t and a confidence-aware entropy regularizer $\mathcal{L}_{CAE}$ to improve representation robustness.

3 METHOD

In this section, we present our framework STaG under a more practical UDA-TBPR setting, where the source domain contains paired image–text data and the target domain contains only unlabeled images. Let the source domain be denoted as $D^s = \{(I_i^s, T_i^s)\}_{i=1}^{N_s}$ and the target domain as $D^t = \{I_i^t\}_{i=1}^{N_t}$, where I is a pedestrian image and T is its textual description. As illustrated in Fig. 2, STaG comprises two complementary components: Style-Agnostic Text Alignment (SATA) and Heterogeneous Graph Memory Alignment (HGMA).

3.1 PRELIMINARY

We adopt a CLIP-style dual-encoder as the default retrieval backbone; it can also be replaced with other backbones as verified in Section 4.1. The backbone comprises an image encoder $f_\theta(\cdot)$ and a text encoder $g_\phi(\cdot)$. Given a mini-batch of M paired samples $\{(I_i, T_i)\}_{i=1}^{M}$, we compute ℓ_2-normalized embeddings as follows:

$$v_i = \frac{f_\theta(I_i)}{\|f_\theta(I_i)\|_2}, \qquad t_i = \frac{g_\phi(T_i)}{\|g_\phi(T_i)\|_2}. \tag{1}$$

A symmetric contrastive loss (InfoNCE) is employed to pull matched pairs together and push unmatched pairs away:

$$\mathcal{L}_{i\to t} = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{\exp(v_i^\top t_i/\tau)}{\sum_{j=1}^{M}\exp(v_i^\top t_j/\tau)}, \quad \mathcal{L}_{t\to i} = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{\exp(t_i^\top v_i/\tau)}{\sum_{j=1}^{M}\exp(t_i^\top v_j/\tau)}, \tag{2}$$

where τ is a temperature coefficient. The overall contrastive training objective is formulated as:

$$\mathcal{L}_{itc} = \frac{1}{2}\left(\mathcal{L}_{i\to t} + \mathcal{L}_{t\to i}\right). \tag{3}$$

3.2 SATA: STYLE-AGNOSTIC TEXT ALIGNMENT

In the absence of target-domain textual descriptions, target images cannot participate in explicit image–text alignment signals. While generating pseudo-descriptions for target images offers a poten-

tial remedy, machine-generated texts often exhibit a severe stylistic divergence compared to human-written queries (as shown in Fig. 4(b)). This style mismatch can introduce bias and ultimately degrade retrieval robustness.

To address this challenge, SATA leverages source-domain data to establish a style-invariance constraint. The core objective is to enforce representational consistency between human-written queries and machine-generated descriptions. By aligning both the original human annotation and a generated pseudo-description to the identical source visual anchor, the model is inherently penalized for relying on surface-level phrasing (Fig. 4(b)). Instead, it is compelled to distill style-invariant, identity-discriminative semantic features. Crucially, since the pseudo-texts in both domains are synthesized using the same generator $G(\cdot)$ and prompt template (Fig. 4(a)), these machine-generated descriptions remain strictly consistent across domains. Consequently, the style invariance calibrated on the source domain seamlessly transfers to the target domain, enabling robust cross-modal alignment even in the absence of human annotations. Specifically, using a unified prompt template (detailed in Appendix B), we synthesize pseudo-descriptions for both domains as $\hat{T}_i^s = G(I_i^s)$ and $\hat{T}_j^t = G(I_j^t)$. We then construct augmented datasets to facilitate joint optimization. For the source domain, we pair each image with both its original human-written text T^s and its generated pseudo-text $\hat{T}^s$. For the target domain, we utilize the generated pairs, yielding the following sets for contrastive training:

$$\hat{\mathcal{D}}^s = \mathcal{D}^s \bigcup \left\{ \langle I_i^s, \hat{T}_i^s \rangle \right\}_{i=1}^{N_s}, \quad \hat{\mathcal{D}}^t = \left\{ \langle I_i^t, \hat{T}_i^t \rangle \right\}_{i=1}^{N_t}, \tag{4}$$

where $\hat{\mathcal{D}}^s$ denotes the augmented source dataset which combines the ground-truth pairs in $\mathcal{D}^s$ with generated image-pseudo-text pairs $\langle I, \hat{T} \rangle$ and $\hat{\mathcal{D}}^t$ represents the target dataset constructed entirely from such generated pairs. The retrieval backbone is then optimized using the SATA objective:

$$\mathcal{L}_{\text{SATA}} = \mathcal{L}_{\text{SATA}}^{src} + \mathcal{L}_{\text{SATA}}^{tgt} = \frac{1}{|\hat{\mathcal{D}}^s|} \sum_{(I,T) \in \hat{\mathcal{D}}^s} \mathcal{L}_{itc}(I,T) + \frac{1}{|\hat{\mathcal{D}}^t|} \sum_{(I,T) \in \hat{\mathcal{D}}^t} \mathcal{L}_{itc}(I,T), \tag{5}$$

where $\mathcal{L}_{itc}$ is the symmetric image-text contrastive loss defined in Eq. 3. By minimizing $\mathcal{L}_{\text{SATA}}^{src}$, the shared visual anchor acts as a critical semantic filter, decoupling style-invariant semantic representations from rigid sentence structures. Crucially, since the target-domain pseudo-texts contain these exact same generation artifacts, optimizing $\mathcal{L}_{SATA}^{tgt}$ allows the pre-calibrated text encoder to intrinsically filter out the machine-style biases. Rather than overfitting to the generated phrasing, $\mathcal{L}_{SATA}^{tgt}$ effectively projects the target images into the unified, style-agnostic semantic space, achieving robust adaptation despite the complete lack of human annotations. To empirically substantiate this decoupling mechanism, we provide extensive visual analyses in Appendices G.1 and J.2.

Notably, since the target domain strictly lacks human annotations, synthesizing pseudo-descriptions via MLLM serves as a critical enabler for explicit cross-modal alignment. Crucially, this generation is strictly a one-time, offline preprocessing step decoupled from the training loop, introducing zero additional latency. A detailed cost analysis in Appendix E.1 demonstrates that STaG is a practical and scalable solution highly suitable for real-world deployment.

3.3 HGMA: Heterogeneous Graph Memory Alignment

Due to inherent visual domain shifts, an encoder trained only on the source domain inevitably produces shifted embeddings for target images, degrading cross-domain matching performance. To mitigate this, we propose the Heterogeneous Graph Memory Alignment (HGMA) module. HGMA leverages the insight that visually similar samples share identity-related semantics regardless of their domain. Consequently, it constructs a heterogeneous graph to aggregate semantic cues from both cross-domain and intra-domain neighbors, yielding a more robust target representation. Distinct from prior memory- and graph-based methods (Khasahmadi et al., 2020; Du et al., 2023; Luo et al., 2025) that either propagate evidence within a single modality or require target-domain human-annotations, both of which are structurally inapplicable under our human-annotation-free setting.

Given a target image feature v_t, relying solely on candidate anchors within a single mini-batch is insufficient. To expand the semantic search space, we maintain three fixed-size First-In-First-Out (FIFO) Semantic Anchor Memories (SAMs): $\{\mathcal{M}_v^s, \mathcal{M}_t^s, \mathcal{M}_v^t\}$, which cache the encoded features of source images, source texts, and target images, respectively. These memories are updated dynam-

ically during training via an enqueue-dequeue mechanism [1]. Leveraging these SAMs, we retrieve the most relevant semantic anchors to populate the graph nodes. Specifically, the intra-domain neighbor set $\mathcal{N}_v^t$ is formulated as the optimal subset of size K_g that maximizes the cosine similarity score:

$$\mathcal{N}_v^t = \underset{\mathcal{S}\subset\mathcal{M}_v^t, |\mathcal{S}|=K_g}{\arg\max} \sum_{m\in\mathcal{S}} \frac{v_t^\top m}{\|v_t\|\|m\|}. \tag{6}$$

Ideally, $\mathcal{N}_v^t$ captures the local manifold structure in the target domain. Similarly, we independently obtain cross-domain neighbor sets $\mathcal{N}_v^s$ and $\mathcal{N}_t^s$ from $\mathcal{M}_v^s$ and $\mathcal{M}_t^s$, respectively. The final neighborhood $\mathcal{N} = \{\mathcal{N}_v^s, \mathcal{N}_t^s, \mathcal{N}_v^t\}$ comprises a total of $3 \times K_g$ anchors, serving as the basis for graph aggregation. Following neighbor retrieval, we compute an attention-based aggregation for each set. For any $\mathcal{N}_x \in \{\mathcal{N}_v^s, \mathcal{N}_t^s, \mathcal{N}_v^t\}$, we define:

$$\mathrm{Agg}(v_t; \mathcal{N}_x) = \sum_{n\in\mathcal{N}_x} \underbrace{\frac{\exp(v_t^\top n/\tau_g)}{\sum_{n'\in\mathcal{N}_x}\exp(v_t^\top n'/\tau_g)}}_{w_{t,n}^{(x)}} \cdot n, \tag{7}$$

where $w_{t,n}^{(x)}$ denotes the normalized attention weight assigned to neighbor n. We then fuse v_t with the aggregated messages via set-specific linear projections:

$$\hat{v}_t = \sigma\big(W_0 v_t + \sum_{\mathcal{N}_x} W_{\mathcal{N}_x}\,\mathrm{Agg}(v_t; \mathcal{N}_x)\big), \tag{8}$$

where $\sigma(\cdot)$ is the ReLU activation function, W_0 and $W_{\mathcal{N}_x}$ are learnable linear projections. The use of separate projections $W_{\mathcal{N}_x}$ allows the model to account for heterogeneous statistics across the different memory sources while fusing them into a unified embedding space.

To prevent negative transfer from weakly related neighbors, we apply a confidence-based residual gate $g_t \in [0, 1]$. We define confidence as the mean similarity between $\hat{v}_t$ and all retrieved neighbors, mapped to $[0, 1]$ via a sigmoid function $\phi(\cdot)$:

$$g_t = \phi\left(\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} \hat{v}_t^\top n\right), \quad \tilde{v}_t = \mathrm{norm}\left((1-g_t)v_t + g_t\hat{v}_t\right), \tag{9}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization, resulting in the corrected target embedding $\tilde{v}_t$. We further introduce a Confidence-Aware Entropy (CAE) regularizer that sharpens a cross-modal matching distribution over retrieved source-text anchors, distinct from classifier-entropy minimization in test-time adaptation, which operates on closed-set single-modality logits(Yue et al., 2026; Yang et al., 2023; Zhang et al., 2022). Specifically, given $\tilde{v}_t$, we retrieve its Top-K_e nearest source-text embeddings $\{t_k\}_{k=1}^{K_e}$ from $\mathcal{M}_t^s$, compute a matching distribution p_k, and minimize its entropy:

$$p_k = \frac{\exp(\tilde{v}_t^\top t_k/\tau_e)}{\sum_{j=1}^{K_e}\exp(\tilde{v}_t^\top t_j/\tau_e)}, \quad \mathcal{L}_{\mathrm{CAE}} = -\sum_{k=1}^{K_e} p_k \log p_k. \tag{10}$$

where τ_e is a temperature parameter. Minimizing $\mathcal{L}_{\mathrm{CAE}}$ sharpens the matching distribution, thereby encouraging high-confidence image-text associations and reducing ambiguous alignments. Our overall training objective is then given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SATA}} + \mathcal{L}_{\mathrm{CAE}}. \tag{11}$$

4 Experiments

In this section, we evaluate the effectiveness of the proposed STaG framework. Comprehensive details regarding the datasets, protocols, and implementation details are provided in Appendix C.

4.1 Comparisons with State-of-The-Art Methods

To evaluate the effectiveness of the proposed STaG framework, we conduct the comparison with state-of-the-art methods under the human-annotation-free UDA-TBPR setting. Notably, existing methods are primarily designed for either fully supervised or human-annotation-available UDA-TBPR settings. To ensure a fair and meaningful comparison under our human-annotation-free

[1] Bounded by a constant capacity M, the space complexity of each SAM is strictly $\mathcal{O}(M \times D)$ (where D is the feature dimension). This decouples the memory overhead from the overall dataset scale, keeping the memory cost bounded during training. For a detailed complexity analysis, see Appendix E.2.

Table 1: **Comparison with state-of-the-art methods under the human-annotation-free UDA-TBPR setting.** We evaluate performance across four transfer scenarios, using CUHK-PEDES and ICFG-PEDES as source domains, respectively. The table compares other methods (top) with our **STaG** module integrated into three backbones (bottom). Absolute performance Improvements (Δ) are highlighted in **bold**. Note: all compared methods were originally designed for data-rich settings that assume access to textual annotations. For fair comparison, we reproduced them under our more practical, human-annotation-free UDA-TBPR setting.

Method	Source: CUHK-PEDES								Source: ICFG-PEDES							
	→ ICFG-PEDES				→ RSTPReid				→ CUHK-PEDES				→ RSTPReid			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
IVT (Shu et al., 2022)	27.49	47.80	57.09	16.03	39.35	67.60	76.85	30.70	22.63	42.29	52.36	19.85	43.70	65.10	75.55	37.65
POUF (Tanwisuth et al., 2023)	32.31	54.83	64.13	16.90	39.40	63.25	70.50	15.89	18.21	37.96	57.88	15.98	34.45	57.90	68.10	25.25
ReCLIP (Hu et al., 2023)	37.39	55.90	63.88	18.64	44.40	66.80	76.95	31.36	25.20	46.25	55.96	23.39	40.80	62.35	71.90	32.08
CFine (Yan et al., 2022)	38.81	57.27	65.46	-	48.30	71.90	82.45	-	32.67	54.03	63.68	25.47	47.40	70.60	79.35	42.30
RICA (Bai et al., 2026)	40.01	57.87	64.52	15.89	47.80	70.65	79.65	34.52	34.11	55.99	65.74	30.44	39.95	63.75	73.30	31.63
GCKD (Luo et al., 2025)	42.15	60.36	68.81	24.01	54.00	77.35	85.05	36.62	38.33	59.51	68.30	35.56	49.15	72.35	79.95	38.81
IRRA (Jiang & Ye, 2023)	36.00	58.69	68.88	19.34	46.90	70.40	79.60	35.42	33.43	56.11	66.23	31.38	45.10	69.20	78.75	36.76
+STaG (Ours)	42.59	61.20	69.15	21.38	52.85	78.20	85.25	39.91	44.64	65.76	74.03	38.98	48.60	71.80	80.35	37.43
Improvement Δ	**+6.59**	**+2.51**	**+0.27**	**+2.04**	**+5.95**	**+7.80**	**+5.65**	**+4.49**	**+11.21**	**+9.65**	**+7.80**	**+7.60**	**+3.50**	**+2.60**	**+1.60**	**+0.67**
TBPS-CLIP (Cao et al., 2023)	37.72	56.41	64.27	19.39	48.75	71.35	81.60	35.85	35.99	56.94	66.23	32.07	44.50	67.00	76.85	33.45
+STaG (Ours)	44.79	61.85	69.11	24.21	49.50	72.35	82.20	36.13	44.23	65.09	73.08	38.64	53.70	75.00	82.60	40.15
Improvement Δ	**+7.07**	**+5.44**	**+4.84**	**+4.82**	**+0.75**	**+1.00**	**+0.60**	**+0.28**	**+8.24**	**+8.15**	**+6.85**	**+6.57**	**+9.20**	**+8.00**	**+5.75**	**+6.70**
RaSa (Bai et al., 2023a)	48.14	65.29	71.83	22.65	58.05	79.50	85.60	43.50	48.65	69.90	76.53	42.03	55.00	73.65	81.55	46.18
+STaG (Ours)	50.45	67.24	73.55	24.23	60.75	82.25	87.15	45.88	57.44	76.13	83.25	50.97	55.70	76.95	84.15	46.88
Improvement Δ	**+2.31**	**+1.95**	**+1.72**	**+1.58**	**+2.70**	**+2.75**	**+1.55**	**+2.38**	**+8.79**	**+6.23**	**+6.72**	**+8.94**	**+0.70**	**+3.30**	**+2.60**	**+0.70**

UDA-TBPR protocol, we reproduced representative existing methods using the identical MLLM-generated pseudo-texts $\hat{T}^t$. Quantitative results are summarized in Table 1. Key observations are outlined below: ❶ STaG significantly outperforms all compared methods. Notably, when instantiated with RaSa as the baseline, STaG establishes a new performance upper bound, achieving 50.45% R-1 accuracy on the challenging CUHK-PEDES → ICFG-PEDES. The inferior performance of the compared methods can be attributed to their design assumptions. Originally developed for fully supervised setting or annotation-available UDA-TBPR, these methods lack mechanisms specifically tailored to the more practical and challenging human-annotation-free scenario. Consequently, they struggle to achieve sufficient cross-modal alignment in the absence of target-domain annotations. On the other hand, fully supervised models (IVT, CFine, RICA, IRRA, TBPS-CLIP, and RaSa) are unable to handle the visual domain shift inherent in UDA-TBPR. In contrast, STaG explicitly addresses both challenges through its proposed SATA and HGMA modules. ❷ STaG achieves consistent performance gains across diverse baselines, including IRRA, TBPS-CLIP, and RaSa, showing its broad applicability and strong compatibility. For instance, when applied to the IRRA baseline in ICFG-PEDES → CUHK-PEDES, STaG improves R@1 by 11.21%, underscoring its effectiveness across varied architectural designs. ❸ Furthermore, extended evaluations in Appendix H demonstrate that despite operating under the most stringent human-annotation-free setting, STaG secures state-of-the-art performance across all datasets. Notably, it not only effectively bridges the gap to fully supervised upper bounds but also surpasses traditional text-dependent UDA methods in specific cross-domain scenarios.

4.2 Ablation Study

4.2.1 Effectiveness of the STaG Framework

STaG comprises two components: SATA and HGMA. To assess their individual contributions, we conduct an ablation study, summarized in Table 2. Our observations are as follows: ❶ Activating either module alone consistently outperforms the baseline, demonstrating that both cross-modal alignment (via SATA) and domain-invariant visual learning (via HGMA) are beneficial in the absence of target-domain text annotations; ❷ Combining both modules yields the best performance, confirming that they address complementary challenges. Specifically, SATA reconstructs missing semantic associations between modalities, while HGMA mitigates visual distribution shifts across domains. Their synergistic integration ultimately enables robust cross-domain adaptation under human-annotation-free condition.

4.2.2 Effectiveness of the SATA Component

We further analyze the SATA component to validate its core design motivation: enabling explicit cross-modal alignment in the target domain while mitigating the stylistic discrepancy between

Table 2: **Ablation study of STaG components. Bold** indicates the best performance, and the final configuration (Ours) is highlighted in blue.

Methods	HGMA	SATA	ICFG-PEDES→CUHK-PEDES				ICFG-PEDES→RSTPReid			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Baseline			35.99	56.94	66.23	32.07	44.50	67.00	76.85	33.45
+ HGMA	✓		36.63	58.08	67.46	32.94	47.25	71.10	79.85	38.37
+ SATA		✓	41.03	59.71	67.46	36.11	50.15	72.10	80.90	37.32
+ STaG (Ours)	✓	✓	**44.23**	**65.09**	**73.08**	**38.64**	**53.70**	**75.00**	**82.60**	**40.15**

Table 3: **Ablation of key components in SATA.** Experiments use ICFG-PEDES as the source domain. $\hat{T}^n$, $\hat{T}^s$, and $\hat{T}^t$ denote the use of random-token noise, source pseudo-texts, and target pseudo-texts, respectively. **Bold** indicates the best performance, and our final configuration is highlighted in blue.

Methods	$\hat{T}^n$	$\hat{T}^s$	$\hat{T}^t$	ICFG-PEDES→CUHK-PEDES				ICFG-PEDES→RSTPReid			
				R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Baseline				36.63	58.08	67.46	32.94	47.25	71.10	79.85	38.37
M_0	✓			38.50	60.00	68.47	33.24	48.30	72.20	79.95	38.40
M_1		✓		39.27	61.49	70.52	35.49	50.70	74.00	81.30	38.92
M_2			✓	40.39	62.23	71.09	36.22	51.30	74.55	82.00	39.11
M_3 **(Ours)**		✓	✓	**44.23**	**65.09**	**73.08**	**38.64**	**53.70**	**75.00**	**82.60**	**40.15**

human-annotated and machine-generated descriptions. To this end, Table 3 compares three ablated variants and the full model that employ different forms of textual descriptions for cross-modal alignment. We make the following key observations: ❶ M_0 replaces pseudo-texts with unstructured random-token noise ($\hat{T}^n$). This confirms that SATA's efficacy strictly relies on valid visual-semantic grounding, rather than acting as a generic text-based regularizer. ❷ M_1 shows that using source-domain generated text alone for explicit cross-modal alignment yields only negligible performance gains. This confirms that the stylistic refinement mechanism is effective only when applied to alignment signals derived from the target domain. ❸ Using target pseudo-texts $\hat{T}^t$ alone (M_2) enables explicit cross-modal alignment and improves performance over the baseline. However, the gains remain limited. As highlighted in the introduction, directly relying on model-generated descriptions risks overfitting to surface-level phrasing, leaving the model vulnerable to the distributional mismatch between synthetic pseudo-texts and real human-annotated descriptions. ❹ The full model (M_3) embodies our complete solution. Compared to M_2, it additionally incorporates source pseudo-texts ($\hat{T}^s$), leveraging source images as stylistic references. By enforcing consistency between human-written and model-generated descriptions in the source domain, the model is encouraged to ignore stylistic variations and instead focus on identity-discriminative semantic content. This effectively *calibrates* the encoder, producing pseudo-text supervision that is robust to stylistic shifts and more reliable for cross-modal alignment in the target domain. ❺ Distribution-level analysis and token-level visualizations in Appendices G.1 and J.2 further show that SATA successfully mitigates the model's overfitting to specific sentence structures. ❻ We additionally examine pseudo-text quality, generator choice, attribute hallucinations, and target-domain test-image corruptions in Appendix F, where STaG remains robust across the evaluated textual and visual perturbations. Empirical evidence consistently demonstrates that STaG exhibits remarkable resilience against these generation errors. This exceptional stability is inherently driven by our synergistic architecture: SATA anchors the shared semantic space using clean source annotations, while HGMA effectively dilutes individual target-domain textual errors through structural neighborhood consensus.

4.2.3 Effectiveness of the HGMA Component

We ablate the internal components of HGMA in Table 4 to evaluate the contributions of heterogeneous memory cues and confidence-aware regularization. Our key observations are as follows: ❶ M_0 replaces KNN retrieval with uniformly random neighbors($\mathcal{M}_{rad}$). This confirms HGMA's benefit stems from identity-consistent graph topology rather than generic feature smoothing. ❷ M_1 and M_3 ablate the cross-domain visual anchors ($\mathcal{M}_v^s$) and intra-domain visual anchors ($\mathcal{M}_v^t$), respectively. Omitting either consistently degrades retrieval performance, indicating that source visual anchors bridge the visual domain shift while target visual anchors capture the local manifold structure of the unlabeled target distribution. ❸ M_2 ablates the cross-domain textual anchors ($\mathcal{M}_t^s$),

Table 4: **Ablation of HGMA.** Evaluation of memory configurations and regularization mechanisms across two cross-dataset settings. Bold indicates the best performance, and our final configuration is highlighted in blue.

Methods	ICFG-PEDES→CUHK-PEDES				ICFG-PEDES→RSTPReid			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Baseline	41.03	59.71	67.46	36.11	50.15	72.10	80.90	37.32
M_0 (STaG w/ $\mathcal{M}_{rad}$)	41.57	60.29	67.88	36.40	50.65	72.50	81.00	37.56
M_1 (STaG w/o $\mathcal{M}_v^s$)	43.72	64.89	72.85	38.29	51.55	73.45	81.70	38.83
M_2 (STaG w/o $\mathcal{M}_t^s$)	42.88	63.90	71.91	37.56	51.05	72.85	80.55	37.81
M_3 (STaG w/o $\mathcal{M}_v^t$)	43.44	64.42	72.87	38.56	51.60	74.40	81.20	38.43
M_4 (STaG w/o g_t)	44.07	65.01	72.99	38.61	52.65	74.30	82.00	39.87
M_5 (STaG w/o $\mathcal{L}_{CAE}$)	43.96	64.98	73.02	38.55	52.90	74.55	82.05	39.96
STaG	**44.23**	**65.09**	**73.08**	**38.64**	**53.70**	**75.00**	**82.60**	**40.15**

causing a notable drop on both transfers. This directly substantiates HGMA's central design, that source-text anchors serve as cross-modal references in the absence of target-domain captions. Without them, the heterogeneous graph loses its only direct text-side supervision and the cross-modal alignment signal collapses back to the unimodal regime. Retrieved neighbors from the SAMs are visualized in Appendix J.3. ❹ M_4 removes confidence-based residual gate (g_t), and M_5 removes confidence-aware entropy regularizer ($\mathcal{L}_{CAE}$) that enhances cross-modal matching reliability. The subsequent performance drops emphasize their necessity as structural safeguards. By modulating feature fusion and sharpening the matching distribution, these mechanisms are crucial for suppressing negative transfer from weakly related neighbors and mitigating error propagation when relying on noisy target-domain pseudo-texts. Appendix H.4 further evaluates CAE under direct transfer, with no observed degradation. ❺ Ultimately, integrating all components achieves the highest accuracy, confirming that robust adaptation requires the synergistic combination of accurate neighborhood retrieval, heterogeneous memory cues, and confidence-aware regularization.

4.3 Hyper-parameter Analysis

HGMA has two hyperparameters: K_g (Eq. 6), the number of visual neighbors retrieved per target sample, and K_e (Eq. 10), the number of source text neighbors used in $\mathcal{L}_{\text{CAE}}$. We vary one of the hyperparameters (K_g, K_e) at a time while keeping the other fixed, and report the results in Fig. 3. It can be observed that both excessively small and excessively large values of K_g or K_e lead to suboptimal performance. A value that is too small provides insufficient evidence for propagation or confidence estimation, whereas a value that is too large introduces weakly related candidates. We set K_g=5 and K_e=10 as the default values in our experiments.

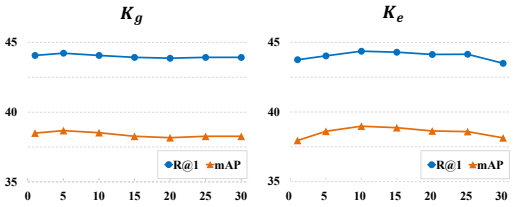


Figure 3: **Ablation of hyperparameters K_g and K_e** using ICFG-PEDES as the source domain and CUHK-PEDES as target domain. (Left) Performance under varying neighbor sizes K_g for graph construction in HGMA. (Right) Performance under varying top-K sizes K_e for the confidence-aware entropy regularization.

5 Conclusion

In this paper, we formally address the highly practical yet under-explored setting of human-annotation-free UDA-TBPR. This paradigm poses significant challenges due to the complete absence of human-written target-domain descriptions and inherent visual domain shifts across datasets. To overcome these bottlenecks, we introduce STaG, a novel framework that synergistically reconstructs cross-modal semantics and mitigates visual domain gaps. Specifically, STaG leverages SATA to decouple identity-discriminative semantics from machine-generated stylistic artifacts, while employing HGMA to establish reliable cross-domain visual neighborhoods via heterogeneous graph consensus. Extensive experiments across multiple benchmarks demonstrate that STaG achieves state-of-the-art performance. Ultimately, by eliminating the reliance on target-domain human annotations, STaG paves the way for the scalable and cost-effective real-world deployment of person retrieval systems.

AI USE STATEMENT

For manuscript preparation, we used generative AI tools to aid writing and polish the language, including grammar, wording, and readability. The authors reviewed and revised all AI-assisted text to ensure that it preserved the intended technical meaning and accurately reflected the methods and results presented in the paper.

Separately, the proposed STaG framework uses a frozen multimodal large language model, InternVL2.5-8B (Chen et al., 2025), to generate image-conditioned pseudo-descriptions as part of the research methodology. This use is distinct from assistance with manuscript preparation and is documented in Section 3.2 and Appendix B and F, including the generation procedure, prompt, and evaluations of pseudo-text quality and robustness. These descriptions are used as synthetic supervision, not as ground-truth human annotations.

The authors take full responsibility for the final content of this work, including the text, scientific claims, experimental results, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses the existing CUHK-PEDES, ICFG-PEDES, and RSTPReid benchmark datasets (Li et al., 2017; Ding et al., 2021; Zhu et al., 2021), without collecting new pedestrian images. The pseudo-descriptions used in our framework are generated from these existing benchmark images. Our research focuses on reducing target-domain annotation requirements for text-based person retrieval. We encourage responsible use of the proposed method with appropriate human oversight in practical applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymous implementation of the core components (SATA and HGMA) of our proposed method at `https://anonymous.4open.science/r/STaG-0485/README.md`.

REFERENCES

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, pp. 555–563. International Joint Conferences on Artificial Intelligence Organization, August 2023a. doi: 10.24963/ijcai.2023/62. URL http://dx.doi.org/10.24963/ijcai.2023/62.

Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data, 2023b. URL https://arxiv.org/abs/2305.12964.

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3952–3962, 2025. doi: 10.1109/CVPR52734.2025.00374.

Yu Bai, Wentao Ma, Shan Zhao, Tianwei Yan, Shezheng Song, Chengyu Wang, and Qian Wan. Rica: Re-ranking with intra-modal and cross-modal alignment for text-based person search. *Expert Systems with Applications*, 306:130931, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.130931. URL https://www.sciencedirect.com/science/article/pii/S0957417425045464.

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search, 2023. URL https://arxiv.org/abs/2308.10045.

Feng Chen, Jielong He, Fuzhong Yang, Jun Huang, Bin Chen, and Wei Huo. Unsupervised cross-modal person search via text style normalization. *Image and Vision Computing*, 171:105999, 2026a. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2026.105999. URL https://www.sciencedirect.com/science/article/pii/S026288562600106X.

Kun Chen, Xi Yang, and Nannan Wang. Toward semantically enhanced representation learning for text-based person retrieval. *IEEE Transactions on Image Processing*, 35:4161–4176, 2026b. doi: 10.1109/TIP.2026.3684811.

Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, Yuhui Zheng, and Ruili Wang. Tipcb: A simple but effective part-based convolutional baseline for text-based person search, 2021. URL https://arxiv.org/abs/2105.11628.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification, 2021. URL https://arxiv.org/abs/2107.12666.

Qianjin Du, Shiji Zhou, Xiaohui Kuang, Gang Zhao, and Jidong Zhai. Joint geometrical and statistical domain adaptation for cross-domain code vulnerability detection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12791–12800, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.788. URL https://aclanthology.org/2023.emnlp-main.788/.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018. URL https://arxiv.org/abs/1707.05612.

Ammarah Farooq, Muhammad Awais, Fei Yan, Josef Kittler, Ali Akbari, and Syed Safwan Khalid. A convolutional baseline for person re-identification using vision and language descriptions, 2020. URL https://arxiv.org/abs/2003.00808.

Ao Fu, Jiaqi Zhao, Yong Zhou, Wenliang Du, Rui Yao, and Abdulmotaleb El Saddik. Similarity regulation and calibration alignment for weakly supervised text-based person re-identification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(3), March 2025. ISSN 1551-6857. doi: 10.1145/3711861. URL https://doi.org/10.1145/3711861.

Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search, 2021. URL https://arxiv.org/abs/2101.03036.

Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Fangzhou Lin, Xing Sun, and Xiang Bai. Conditional feature learning based transformer for text-based person search. *IEEE Transactions on Image Processing*, 31:6097–6108, 2022. doi: 10.1109/TIP.2022.3205216.

Daming Gao, Yang Bai, Min Cao, Hao Dou, Mang Ye, and Min Zhang. Semi-supervised text-based person search, 2024. URL https://arxiv.org/abs/2404.18106.

Tiantian Gong, Junsheng Wang, and Liyan Zhang. Enhancing cross-modal completion and alignment for unsupervised incomplete text-to-image person retrieval. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 794–802. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/88. URL https://doi.org/10.24963/ijcai.2024/88. Main Track.

Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data, 2021. URL https://arxiv.org/abs/2110.10807.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, Cheng-Hao Kuo, and Ram Nevatia. Reclip: Refine contrastive language image pre-training with source free domain adaptation, 2023. URL https://arxiv.org/abs/2308.03793.

Zhong Ji, Junhua Hu, Deyin Liu, Lin Yuanbo Wu, and Ye zhao. Asymmetric cross-scale alignment for text-based person search, 2022. URL https://arxiv.org/abs/2212.11958.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, 2023. URL https://arxiv.org/abs/2303.12501.

Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search, 2019. URL https://arxiv.org/abs/1809.08440.

Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10675–10683, 2020. doi: 10.1109/CVPR42600.2020.01069.

Amir Hosein Khasahmadi, Kaveh Hassani, Parsa Moradi, Leo Lee, and Quaid Morris. Memory-based graph networks, 2020. URL https://arxiv.org/abs/2002.09518.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021a. URL https://arxiv.org/abs/2107.07651.

Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i4.28101. URL https://doi.org/10.1609/aaai.v38i4.28101.

Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search, 2021b. URL https://arxiv.org/abs/2112.06714.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description, 2017. URL https://arxiv.org/abs/1702.05729.

Zongyi Li, Jianbo Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Runsheng Wang, and Shijuan Huang. Cross-modal generation and alignment via attribute-guided prompt for unsupervised text-based person retrieval. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 1047–1055. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/116. URL https://doi.org/10.24963/ijcai.2024/116. Main Track.

Zongyi Li, Li Jianbo, Yuxuan Shi, Jiazhong Chen, Shijuan Huang, Linnan Tu, Fei Shen, and Hefei Ling. Exploring the potential of large vision-language models for unsupervised text-based person retrieval. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i5.32543. URL https://doi.org/10.1609/aaai.v39i5.32543.

Jing Liu, Donglai Wei, Yang Liu, Sipeng Zhang, Tong Yang, Wei Zhou, Weiping Ding, and Victor C.M. Leung. Scmm: Calibrating cross-modal representations for text-based person search. *Pattern Recognition*, 182:114775, February 2027. ISSN 0031-3203. doi: 10.1016/j.patcog.2026.114775. URL http://dx.doi.org/10.1016/j.patcog.2026.114775.

Edward Loper and Steven Bird. Nltk: The natural language toolkit, 2002. URL https://arxiv.org/abs/cs/0205028.

Zefeng Lu, Ronghao Lin, Yap-Peng Tan, and Haifeng Hu. Prompt-guided transformer and mllm interactive learning for text-based pedestrian search. *IEEE Transactions on Information Forensics and Security*, 20:7181–7196, 2025. doi: 10.1109/TIFS.2025.3586488.

Bingjun Luo, Jinpeng Wang, Wang Zewen, Junjie Zhu, and Xibin Zhao. Graph-based cross-domain knowledge distillation for cross-dataset text-to-image person retrieval, 2025. URL https://arxiv.org/abs/2501.15052.

Xiaoxuan Mu, Haoyu Tang, Han Jiang, Tianyuan Liang, Qinghai Zheng, and Jihua Zhu. Face: A dual-template and adaptive curriculum framework for unsupervised text-based person search. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 4107–4116, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755315. URL https://doi.org/10.1145/3746027.3755315.

Kai Niu, Yan Huang, and Liang Wang. Fusing two directions in cross-domain adaption for real life person search by language. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1815–1818, 2019. doi: 10.1109/ICCVW.2019.00225.

Kai Niu, Jiahui Chen, Ke Han, Xinyue Song, and Yanning Zhang. Pseudo sentences evaluation and quality-aware robust learning for unsupervised text-based person search. *IEEE Transactions on Image Processing*, 35:5482–5495, 2026a. doi: 10.1109/TIP.2026.3694187.

Kai Niu, Qinzi Zhao, Jiahui Chen, and Yanning Zhang. Cache-aided cross-modal correlation correction for unsupervised cross-domain text-based person search. *Pattern Recognition*, 172:112521, 2026b. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112521. URL https://www.sciencedirect.com/science/article/pii/S0031320325011847.

Chenghuan Qi, Xi Yang, Nannan Wang, and Xinbo Gao. Granularity-aware hyperbolic representation for text-based person search. *IEEE Transactions on Information Forensics and Security*, 20: 5745–5757, 2025. doi: 10.1109/TIFS.2025.3574970.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification, 2024. URL https://arxiv.org/abs/2308.09911.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification, 2025. URL https://arxiv.org/abs/2506.11036.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification, 2023. URL https://arxiv.org/abs/2309.01420.

Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval, 2022. URL https://arxiv.org/abs/2208.08608.

Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. A simple and robust correlation filtering method for text-based person search. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 726–742, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19832-8. doi: 10.1007/978-3-031-19833-5_42. URL https://doi.org/10.1007/978-3-031-19833-5_42.

Wentao Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid, 2024. URL https://arxiv.org/abs/2405.04940.

Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models, 2023. URL https://arxiv.org/abs/2305.00350.

Haiguang Wang, Yu Wu, Mengxia Wu, Cao Min, and Min Zhang. Laip: Learning local alignment from image-phrase modeling for text-based person search, 2024. URL https://arxiv.org/abs/2406.10845.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language, 2020. URL https://arxiv.org/abs/2005.07327.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 5314–5322. ACM, October 2022. doi: 10.1145/3503161.3548057. URL http://dx.doi.org/10.1145/3503161.3548057.

Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: Language-guided person search via color reasoning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1604–1613, 2021. doi: 10.1109/ICCV48922.2021.00165.

Yingjia Xu, Jinlin Wu, Daming Gao, Zhen Chen, Yang Yang, Min Cao, Mang Ye, and Zhen Lei. Sa-person: Text-based person retrieval with scene-aware re-ranking, 2025. URL https://arxiv.org/abs/2505.24466.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification, 2022. URL https://arxiv.org/abs/2210.10276.

Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, Shangling Jui, and Jian Yang. Trust your good friends: Source-free domain adaptation by reciprocal neighborhood clustering, 2023. URL https://arxiv.org/abs/2309.00528.

Mang Ye, Yucheng Ji, Yang Bai, Min Cao, Siyuan Chai, Bo Du, and Min Zhang. Generative retrieval for unsupervised text-based person search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2026. doi: 10.1109/TPAMI.2026.3713379.

Xia Yue, Anfeng Liu, Ning Chen, Chenjia Huang, Hui Liu, Zhou Huang, and Leyuan Fang. Beyond distribution shifts: Adaptive hyperspectral image classification at test time. *Neurocomputing*, 678:133188, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133188. URL https://www.sciencedirect.com/science/article/pii/S0925231226005850.

Xiaohong Zhang, Yi Chen, Ziyi Shen, Yuming Shen, Haofeng Zhang, and Yudong Zhang. Confidence-and-refinement adaptation model for cross-domain semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9529–9542, 2022. doi: 10.1109/TITS.2022.3140481.

Yafei Zhang, Yongle Shang, and Huafeng Li. Dual-granularity cross-modal identity association for weakly-supervised text-to-person image matching, 2025. URL https://arxiv.org/abs/2507.06744.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *European conference on computer vision*, pp. 707–723. Springer, 2018.

Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11395–11404, October 2021.

Xinpeng Zhao, Yanwei Zheng, Chuanlin Lan, Xiaowei Zhang, Bowen Huang, Jibin Yang, and Dongxiao Yu. Cpcl: Cross-modal prototypical contrastive learning for weakly supervised text-based person retrieval, 2025. URL https://arxiv.org/abs/2401.10011.

Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. Hierarchical gumbel attention network for text-based person search. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 3441–3449, 2020a.

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23, May 2020b. ISSN 1551-6865. doi: 10.1145/3383184. URL http://dx.doi.org/10.1145/3383184.

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, 2021. URL https://arxiv.org/abs/2109.05534.

APPENDIX OVERVIEW

To enhance the readability and navigation of this supplementary material, we outline the structural organization of the appendix as follows:

A Notation Summary 15
B Prompt Template for SATA 16
C Experimental Setup Details 16
C.1 Datasets 16
C.2 Protocol 17
C.3 Implementation Details 17
D Detailed Mechanism of Style-Agnostic Text Alignment (SATA) 17
E Annotation Cost and Computational Overhead 18
E.1 Offline Generation Cost vs. Manual Annotation 18
E.2 Training and Inference Complexity of STaG 19
F Pseudo-Text Reliability and Visual Robustness 19
F.1 Pseudo-Text Quality and Generation Robustness 19
F.2 Robustness to Image Corruptions 20
G Necessity of Semantic Alignment and Graph Topology 21
G.1 SATA: Quantifying Style Invariance 21
G.2 SATA: Valid Semantics vs. Unstructured Noise 21
G.3 HGMA: KNN Topology vs. Random Neighbors 22
G.4 Contribution Beyond Plain Pseudo-Text Adaptation 23
H Additional Evaluations 23
H.1 Comparative Analysis with Human-Annotation-Accessible Baselines 23
H.2 Generalization Across Diverse Source Domains 24
H.3 Architectural Robustness 24
H.4 Cross-Dataset Generalization of CAE 25
I Impacts and Limitations 26
I.1 Limitations 26
I.2 Impact Statement 26
J Visualization Analysis 26
J.1 Visualization of Image Activation Maps 27
J.2 Visualization of Attribute-Level Attention in Pseudo-Texts 28
J.3 Visualization of Retrieved HGMA Neighbors 28
J.4 Visualization of Retrieval Results 28

A NOTATION SUMMARY

In Table 5, we present a comprehensive summary of the key notations and variables used in the formulation of our proposed STaG framework.

Table 5: Summary of notations used in this paper, categorized by method components.

Symbol	Description	Reference
D^s, D^t	Source and target domain datasets	Sec. 3
v_i, t_i	ℓ_2-normalized image and text embeddings	Eq. 1
$\mathcal{L}_{itc}$	Symmetric image-text contrastive (InfoNCE) objective	Eq. 3
$\hat{T}^s, \hat{T}^t$	Generator-synthesized pseudo-descriptions	Sec. 3.2
$\hat{\mathcal{D}}^s, \hat{\mathcal{D}}^t$	Augmented datasets with generated pairs	Sec. 3.2
$\mathcal{L}_{\text{SATA}}$	Style-Agnostic Text Alignment optimization objective	Sec. 3.2
$\mathcal{M}_v^s, \mathcal{M}_t^s, \mathcal{M}_v^t$	FIFO Semantic Anchor Memories (SAMs)	Sec. 3.3
$\mathcal{N}_v^t$	Retrieved Top-K_g intra-domain neighbor set	Eq. 6
$\text{Agg}(\cdot)$	Attention-based neighbor aggregation function	Eq. 7
$w_{t,n}^{(x)}$	Normalized attention weight for neighbor n	Eq. 7
$\hat{v}_t$	Fused target feature incorporating heterogeneous cues	Eq. 8
g_t	Confidence-based residual gate for feature fusion	Sec. 3.3
$\tilde{v}_t$	Final corrected target image embedding	Sec. 3.3
p_k	Matching distribution over Top-K_e source texts	Eq. 10
$\mathcal{L}_{\text{CAE}}$	Confidence-Aware Entropy regularizer	Eq. 10
$\mathcal{L}$	Overall training objective for STaG	Eq. 11

B Prompt Template for SATA

Using the same generator and prompt encourages $\hat{T}^s$ and $\hat{T}^t$ to share similar generation styles, which facilitates learning style-invariant yet identity-discriminative representations rather than overfitting to domain-specific textual artifacts. To encourage visual grounding and ensure descriptive fidelity, our prompt explicitly instructs the generator to describe solely observable attributes. Furthermore, we enforce identical decoding strategies for all inferences—specifically fixing the temperature τ, nucleus sampling parameter top-p, and maximum sequence length L. The specific instruction used for generation is presented below:

System Instruction:
```
Write a description about the appearance of the person in the
image.  Do not imagine any contents that are not in the image.
```

C Experimental Setup Details

C.1 Datasets

Our experiments involve three open benchmark datasets.

CUHK-PEDES (Li et al., 2017) is the most commonly-used dataset in TBPS. It consists of $40,206$ images and $80,440$ texts from $13,003$ identities in total, which are split into $34,054$ images and $68,126$ texts from $11,003$ identities in the training set, $3,078$ images and $6,158$ texts from $1,000$ identities in the validation set, and $3,074$ images and $6,156$ texts from $1,000$ identities in the test set. The average length of all texts is 23.

ICFG-PEDES (Ding et al., 2021) contains $54,522$ images from $4,102$ identities. Each of the images is described by one text. The dataset is split into $34,674$ images from $3,102$ identities in the training set, and $19,848$ images from $1,000$ identities in the test set. There are 37 words per text on average.

RSTPReid (Zhu et al., 2021) consists of $20,505$ images of $4,101$ identities. Each identity has 5 corresponding images captured from different cameras. Each image is annotated with 2 textual descriptions, and each description is no shorter than 23 words. There are $3,701/200/200$ identities utilized for training/validation/testing, respectively.

C.2 PROTOCOL

We adopt the widely-used Rank@K (R@K for short, K=1, 5, 10) metric to evaluate the performance of the proposed framework. Specifically, given a query text, we rank all the test images via the similarity with the query text, and the search is deemed to be successful if top-K images contain any corresponding identity. R@K is the percentage of successful searches. In addition, we also adopt the mean average precision (mAP) as a complementary metric. Rank@K reflects the accuracy of the first few retrieval results, while mAP emphasizes the comprehensive performance of the method.

C.3 IMPLEMENTATION DETAILS

During training, all images are resized to 224×224, and each text input is truncated to at most 77 tokens. We optimize the retrieval model using AdamW with an initial learning rate of 1×10^{-5} and a cosine learning-rate scheduler, and train for 5 epochs with a batch size of 128. We adopt InternVL2.5-8B (Chen et al., 2025) as the generator. The hyperparameters are set as follows. We set the numbers of neighbors $K_g = 5$ and the number of entropy candidates $K_e = 10$ (in Section 3.3). The temperature parameters in Eq. 10 is set to $\tau_e = 0.2$. The capacity of SAMs in Section 3.3 is set to 2^{14} . All experiments are conducted on 4 NVIDIA L40 GPUs.

D DETAILED MECHANISM OF STYLE-AGNOSTIC TEXT ALIGNMENT (SATA)

To provide a deeper mechanistic understanding of how SATA effectively bridges the cross-modal domain gap, Fig. 4 visually dissects the text generation and alignment pipeline. Specifically, we elucidate the module's operation across three critical dimensions:

❶ The Stylistic Divergence and Shared Artifacts: As depicted in the text boxes of Fig. 4, human-written descriptions in the source domain (T^s) typically exhibit flexible phrasing and focus directly on discriminative attributes, such as "The woman is wearing a red and yellow shirt...". In contrast, texts generated by the MLLM, denoted as $\hat{T}^s$ and $\hat{T}^t$, suffer from rigid, repetitive stylistic artifacts. These include introductory boilerplate like "The person in the image is..." and structured enumerations like "She is wearing..." or "She has...". This severe stylistic divergence is precisely why relying solely on target-domain pseudo-texts for cross-modal alignment yields sub-optimal performance, a vulnerability empirically validated by the limited gains of variant M_2 in our ablation study shown in Table 3. Overcoming this fundamental bottleneck is the core motivation behind SATA. In Fig. 4, we highlight these uninformative, MLLM-specific syntactic templates in red, while the actual identity-discriminative semantic concepts (clothing colors and accessories), are highlighted in green.

❷ Source Domain: Calibrating the Semantic Filter via $\mathcal{L}_{SATA}^{Src}$. The left portion of Fig. 4 illustrates the core calibration phase. In the source domain, a single image anchor I^s is paired with both its original human annotation T^s and its MLLM-generated counterpart $\hat{T}^s$. When optimizing the contrastive objective $\mathcal{L}_{SATA}^{Src}$, the Retrieval Model is mathematically forced to pull these two stylistically distinct texts toward the identical visual point in the shared embedding space. To minimize this loss, the text encoder has no choice but to act as a semantic filter: it must actively discard the differing surface-level phrasing highlighted in red, and isolate the shared underlying visual attributes highlighted in green. This process explicitly decouples the style-invariant semantic features from the machine-generated noise.

❸ Target Domain: Robust Cross-Modal Adaptation via $\mathcal{L}_{SATA}^{tgt}$. The crucial bridge between the source and target domains is denoted by the "Similar Structure" arrow in Fig. 4(a). The pseudo-descriptions for both the source $\hat{T}^s$ and target $\hat{T}^t$ domains are synthesized utilizing the identical frozen MLLM and prompt template, they inherently share the exact same rigid syntactic structures, referred to as the red artifacts. Consequently, when the target objective $\mathcal{L}_{SATA}^{tgt}$ aligns the unlabeled target images I^t with $\hat{T}^t$, the text encoder has already been pre-calibrated by the source domain to ignore these specific formatting artifacts. Instead of overfitting to the MLLM's phrasing, the model robustly extracts the pure identity semantics, successfully mitigating the cross-domain shift and establishing highly reliable image-text alignment in the human-annotation-free target domain.

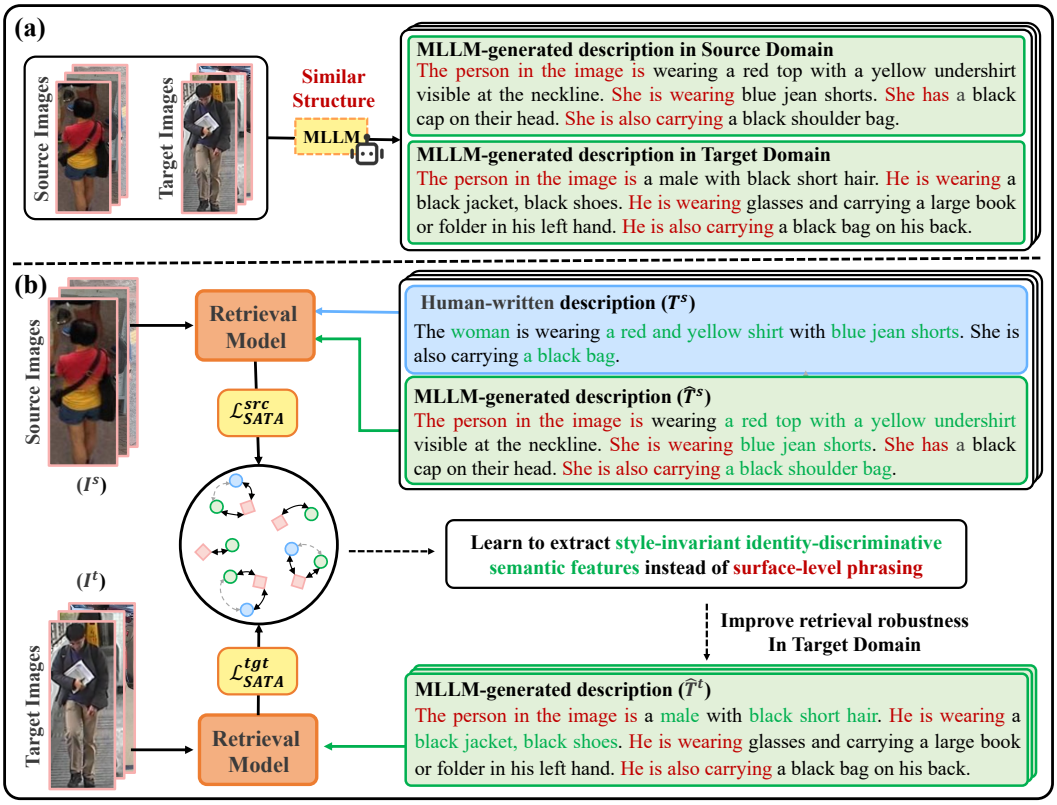


Figure 4: **Detailed mechanism of the Style-Agnostic Text Alignment (SATA) module.** (a) Given the identical MLLM generator and prompt, the synthesized pseudo-descriptions in both domains exhibit highly similar, rigid syntactic structures (highlighted in red), which severely diverge from the flexible phrasing of human-written queries. (b) illustrates the fundamental stylistic divergence between human-written queries and MLLM-generated pseudo-descriptions. By utilizing a shared MLLM generator, machine-generated artifacts remain structurally consistent across domains. SATA forces the retrieval model to align both text modalities to a shared visual anchor in the source domain ($\mathcal{L}_{SATA}^{Src}$), compelling the network to discard rigid syntactic templates (highlighted in red) and extract pure identity-discriminative semantics (highlighted in green). This calibrated capability seamlessly generalizes to the target domain ($\mathcal{L}_{SATA}^{tgt}$), ensuring robust adaptation.

E Annotation Cost and Computational Overhead

As highlighted in our motivation (Sec. 1), the primary bottleneck for real-world UDA-TBPR deployment is the prohibitive cost of acquiring target-domain textual annotations. To demonstrate that introducing an MLLM is a genuine cost-reduction paradigm, rather than simply replacing human labor costs with expensive computational overhead, we provide a comprehensive analysis of the offline preparation expenses and the training complexity of our framework.

E.1 Offline Generation Cost vs. Manual Annotation

❶ Economic Cost and Efficiency. Manual annotation for TBPR is notoriously labor-intensive and expensive. On typical crowdsourcing platforms (e.g., Amazon Mechanical Turk), accurately annotating a single pedestrian image requires 1 to 2 minutes. At a standard minimum wage of \$6.00/hour, the manual annotation cost equates to approximately \$0.10 per image. In contrast, our offline generation using InternVL2.5-8B on four NVIDIA L40 GPUs processes images at approximately 0.11 seconds per image. Under a system-level compute price of approximately \$4.00 per four-GPU hour, this corresponds to about \$0.00012 per image, yielding roughly an $800\times$ reduction compared with the estimated \$0.10 manual annotation cost.

❷ **Scalability for Large-Scale Deployment.** Beyond direct monetary costs, MLLM generation scales deterministically and is exempt from human-centric bottlenecks such as annotator fatigue, subjective inconsistency, and logistical coordination overhead.

For instance, manually annotating a new target domain of 100,000 surveillance images would cost roughly $10,000 and require months to complete. Our framework accomplishes the identical task in approximately 3 hours for about $12 in compute costs. The estimated energy footprint for this process (based on the 300W TDP of an L40 GPU) is around 3.67 kWh, which is a negligible expenditure given the magnitude of the data processed.

❸ **Zero Inference Overhead.** Crucially, the MLLM is employed exclusively as an offline data preparation engine. Once the pseudo-texts are generated and the model is trained, the MLLM is entirely discarded. During online deployment and inference, only the lightweight CLIP-style dual-encoder is active. Thus, STaG introduces zero architectural changes, latency delays, or deployment complexities to the final retrieval system. Trading a one-time, marginal offline computational cost for substantial cross-domain performance improvements represents a highly practical and favorable trade-off for real-world applications.

E.2 Training and Inference Complexity of STaG

We analyze both the asymptotic complexity and the measured computational overhead of STaG. Let B denote the mini-batch size, D the embedding dimension, and M the fixed capacity of each Semantic Anchor Memory.

❶ **SATA Training Complexity.** SATA introduces no additional trainable network module beyond the underlying retrieval backbone. Its main additional computation arises from image–text contrastive optimization over generated pairs. Computing a dense pairwise similarity matrix for a batch of B image and text embeddings of dimension D requires $\mathcal{O}(B^2D)$ operations. The pseudo-descriptions themselves are generated offline and remain fixed throughout retrieval-model training.

❷ **HGMA Training Complexity.** The Heterogeneous Graph Memory Alignment (HGMA) module introduces structural alignment. It maintains three fixed memory queues ($\mathcal{M}_v^s, \mathcal{M}_t^s, \mathcal{M}_v^t$), each with a capacity of M (set to 16,384) and a dimension of D (set to 512). The static memory complexity is $O(MD)$ for each memory bank, and remains independent of the overall dataset size. The computational bottleneck is nearest-neighbor retrieval across the memory banks. Computing similarities between a batch of B query features and M memory features of dimension D requires $O(BMD)$ operations. Since M is fixed, the memory and retrieval costs remain bounded with respect to the dataset size. Crucially, since M is a predefined constant, this similarity computation is strictly restricted to a local subset, completely decoupling the graph construction overhead from the overall target dataset scale.

❸ **Measured Training and Memory Overhead.** Under the ICFG-PEDES $\rightarrow$ CUHK-PEDES setting, a complete training run takes approximately 25 minutes on four NVIDIA L40 GPUs, with approximately 40 GB peak memory usage measured on a single L40 GPU. STaG introduces 1.05M additional trainable parameters, while the three Semantic Anchor Memories require approximately 96 MiB in FP32 of storage in total. For inference, we measure a single-query latency of 5.78 ms and a throughput of 373 queries/s with a batch size of 64.

❹ **Online Inference Complexity.** Most importantly, the added complexity of STaG is strictly confined to the training phase. Once the model is fully trained, the MLLM, the Semantic Anchor Memories (SAMs), and the graph aggregation operations are entirely discarded. During online deployment and inference, only the lightweight CLIP-style dual-encoder is active. Thus, STaG introduces zero architectural changes, latency delays, or deployment complexities to the final retrieval system.

F Pseudo-Text Reliability and Visual Robustness

F.1 Pseudo-Text Quality and Generation Robustness

SATA relies on pseudo-descriptions generated by an off-the-shelf MLLM. We therefore examine their reliability from three aspects: intrinsic quality, sensitivity to the generator choice, and robustness to hallucinated attributes.

Table 6: Robustness to different MLLM generators (ICFG-PEDES → CUHK-PEDES). Retrieval performance remains stable regardless of which MLLM is used for offline pseudo-text generation.

Generator Model	R@1	R@5	R@10	mAP
InternVL2.5-8B (Default)	44.23	65.09	73.08	38.64
InternVL2-8B	44.07	64.31	71.99	38.40
Qwen2-VL-7B-Instruct	43.84	64.11	71.15	37.94

Table 7: Robustness to hallucination noise (ICFG-PEDES → CUHK-PEDES). We randomly alter observable attributes in the target pseudo-texts to simulate varying degrees of MLLM hallucinations.

Noise Rate	R@1	R@5	R@10	mAP
0% (Original Pseudo-texts)	44.23	65.09	73.08	38.64
10%	43.87	64.55	72.63	38.28
20%	44.05	64.88	72.99	38.49
30%	43.13	62.85	71.23	37.67

❶ **Quality of Generated Pseudo-Texts.** We compare the pseudo-descriptions generated by InternVL2.5-8B with human annotations on CUHK-PEDES. They achieve a CLIP cosine similarity of 73.89, a ROUGE-L score of 24.79, and BLEU-1/2/4 scores of 28.01/17.23/6.68. The 97.39% unique-caption ratio further indicates that the generated descriptions retain substantial lexical diversity. These results support their use as target-domain cross-modal supervision.

❷ **Robustness to the Choice of MLLM Generator.** Table 6 compares InternVL2.5-8B with InternVL2-8B and Qwen2-VL-7B-Instruct on ICFG-PEDES → CUHK-PEDES. Rank@1 varies only from 43.84% to 44.23%, showing that STaG is relatively insensitive to the choice of generator among the evaluated MLLMs.

❸ **Robustness to Hallucination Noise.** To evaluate tolerance to factual errors, we randomly corrupt $r\%$ of the target pseudo-descriptions by replacing observable attributes, such as clothing colors, garment types, and hairstyles, with plausible but incorrect alternatives. Here, r denotes the proportion of corrupted descriptions rather than the token-level corruption ratio. As shown in Table 7, Rank@1 decreases only from 44.23% to 43.13% when 30% of the pseudo-descriptions are corrupted.

This robustness benefits from two complementary mechanisms. SATA anchors the representation space with clean source-domain annotations, while HGMA aggregates multiple source-side semantic references, reducing the influence of individual noisy pseudo-descriptions.

F.2 Robustness to Image Corruptions

We further evaluate the robustness of STaG to visual degradation at test time. Specifically, we apply five common corruptions (Gaussian blur, motion blur, low illumination with noise, partial occlusion, and JPEG compression) to the target-domain test images under ICFG-PEDES → CUHK-PEDES.

As shown in Table 8, STaG exhibits only a 0.19-point average Rank@1 drop under mild corruptions, compared with 1.27 points for the baseline. Under severe corruptions, both methods degrade, but STaG consistently remains above the baseline, achieving an average Rank@1 of 37.65 versus 28.87.

These results show that the adaptation learned by STaG remains effective when target-domain gallery images suffer from common test-time degradations. Since the corruptions are introduced only at evaluation time, this experiment specifically measures retrieval robustness to degraded visual inputs rather than the robustness of MLLM pseudo-text generation.

Table 8: **Robustness to target-domain test-image corruptions** on ICFG-PEDES → CUHK-PEDES. Δ denotes the R@1 change relative to clean test images.

Corruption	Parameters	STaG		Baseline	
		R@1	Δ	R@1	Δ
Original images	–	44.23	–	35.99	–
Mild corruption					
Gaussian blur	$\sigma = 1$	44.39	+0.16	35.77	-0.22
Motion blur	$k = 7$	44.57	+0.34	35.86	-0.13
Low light + noise	$\gamma = 1.5,\ \sigma_n \approx 0.008$	44.61	+0.38	34.70	-1.29
Occlusion	10% image area	43.30	-0.93	32.75	-3.24
JPEG compression	$q = 50$	43.32	-0.91	34.51	-1.48
Average	–	44.04	-0.19	34.72	-1.27
Severe corruption					
Gaussian blur	$\sigma = 3$	36.04	-8.19	28.11	-7.88
Motion blur	$k = 15$	41.07	-3.16	32.56	-3.43
Low light + noise	$\gamma = 2.5,\ \sigma_n \approx 0.024$	35.99	-8.24	25.77	-10.22
Occlusion	25% image area	35.52	-8.71	26.81	-9.18
JPEG compression	$q = 20$	39.61	-4.62	31.08	-4.91
Average	–	37.65	-6.58	28.87	-7.12

G Necessity of Semantic Alignment and Graph Topology

To further examine the roles of SATA and HGMA, we provide several complementary sanity checks and control experiments on the ICFG-PEDES → CUHK-PEDES setting. These analyses evaluate style alignment, semantic validity, neighborhood topology, and the contribution beyond plain pseudo-text adaptation.

G.1 SATA: Quantifying Style Invariance.

To directly validate that SATA successfully bridges the stylistic domain gap between human-written queries and machine-generated descriptions, we evaluate the feature embeddings of both textual modalities on the source domain. As summarized in Table 9 and visually corroborated by Fig. 5, SATA aligns the representations across two critical levels:

❶ **Macroscopic Distribution Alignment:** As illustrated in Fig. 5, the baseline model exhibits two disjoint clusters for human and generated texts, indicating a severe stylistic gap. SATA elegantly maps these distinct styles into a unified semantic space, merging the initially separated distributions and reducing their Centroid L2 Distance by more than half ($0.2821 \rightarrow 0.1275$).

❷ **Microscopic Instance-Level Alignment:** At the instance level, SATA significantly boosts the Pairwise Cosine Mean between human queries and their corresponding pseudo-texts by $+0.1109$. The proportion of highly similar human–generated text pairs, measured by $P(\cos > 0.90)$, increases by an absolute 13.55%.

Together, these measurements provide evidence that SATA reduces the representation gap between human-written and generated descriptions while preserving their shared semantic content. Token-level attention visualization can be seen in Appendix J.

G.2 SATA: Valid Semantics vs. Unstructured Noise.

We first investigate whether the SATA module relies on meaningful visual-semantic grounding or if it merely acts as a generic text-based regularizer. To this end, we replace the valid generated pseudo-descriptions with randomly sampled noise. Concretely, we randomly select 30% of the target samples and, within each selected sample, replace the attribute tokens with words drawn uniformly at random from the vocabulary; the remaining 70% retain their original pseudo-texts. This matches

Table 9: Quantifying style invariance (Human vs. Generated Texts). We compare the feature embeddings of human annotations and MLLM-generated pseudo-texts on the source domain.

Metric	Baseline	Baseline + SATA (Ours)	Change
Centroid L2 Distance ↓	0.2821	0.1275	$\Delta = -0.1546$
Pair Cosine Mean ↑	0.7883	0.8992	$\Delta = +0.1109$
$P(\cos > 0.90)$ ↑	37.80%	51.35%	$\Delta = +13.55\%$

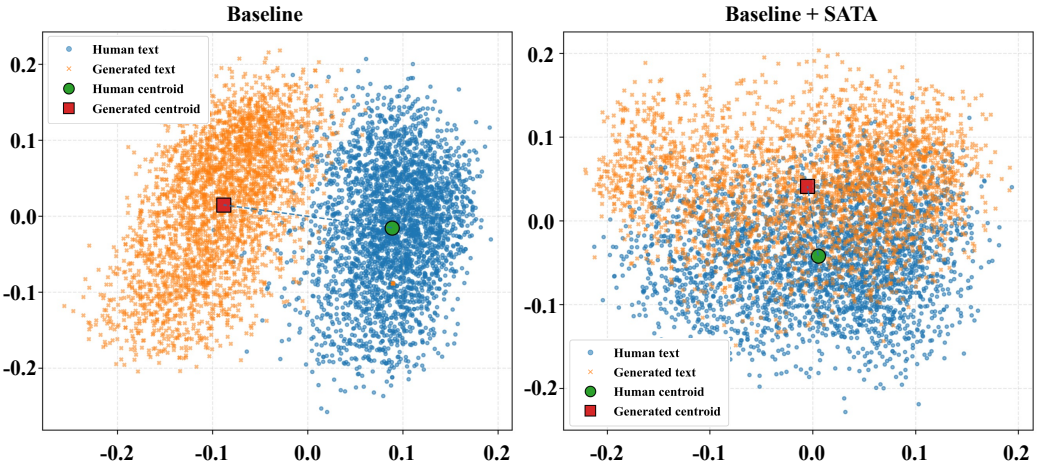


Figure 5: Visualization of feature distributions for human-written texts (blue dots) and MLLM-generated pseudo-texts (orange crosses). **Left:** The baseline model exhibits a distinct stylistic domain gap, with the two text modalities forming disjoint clusters. **Right:** By integrating SATA, the two distributions are pulled into a highly overlapping and unified semantic space, significantly reducing the distance between their centroids (green circle and red square).

the sample-level selection used in the hallucination experiment, but substitutes random tokens rather than plausible but incorrect ones.

As shown in Table 10, random token corruption yields 38.50% Rank@1, notably below the 43.13% obtained with structured hallucination noise at the same 30% sample corruption rate. This gap indicates that SATA benefits from meaningful visual semantic content rather than from the mere presence of an auxiliary text signal. At the same time, the smaller degradation under structured hallucinations suggests that the framework tolerates moderate attribute-level errors better than unstructured semantic corruption.

G.3 HGMA: KNN Topology vs. Random Neighbors.

We apply the analogous test to HGMA by asking whether its gains arise from the structure of the retrieved neighborhood or merely from the addition of memory-based features. We replace the K-Nearest Neighbor (K_g) retrieval mechanism in Eq. 6 with uniform random sampling from the semantic anchor memories, keeping memory capacity, attention aggregation, residual gating, and the CAE regularizer unchanged. Table 11 shows that random-neighbor aggregation yields only +0.54% Rank@1 over the SATA-only baseline (41.03% → 41.57%), whereas KNN-based HGMA yields +3.20% (41.03% → 44.23%). The gap indicates that HGMA benefits from similarity-based neighborhood construction rather than generic feature aggregation. Retrieving feature-similar neighbors provides more coherent cross-domain references than random sampling, without assuming any cross-domain identity correspondence.

Table 10: Sanity check on the SATA module (ICFG-PEDES → CUHK-PEDES). We compare the impact of introducing unstructured random noise versus structured MLLM hallucinations.

Method	R@1	R@5	R@10	mAP
Baseline (HGMA Only)	36.63	58.08	67.46	32.94
+ Random Noise Text	38.50	60.00	68.47	33.24
+ 30% Hallucination	43.13	62.85	71.23	37.67
+ SATA (Original)	**44.23**	**65.09**	**73.08**	**38.64**

Table 11: Sanity check on the HGMA module (ICFG-PEDES → CUHK-PEDES). We compare random neighbor sampling against our proposed KNN-based graph retrieval.

Method	R@1	R@5	R@10	mAP
Baseline (SATA Only)	41.03	59.71	67.46	36.11
+ Random Neighbor	41.57	60.29	67.88	36.40
+ HGMA (KNN-based)	**44.23**	**65.09**	**73.08**	**38.64**

G.4 Contribution Beyond Plain Pseudo-Text Adaptation

To isolate the contribution of STaG from that of MLLM-generated supervision itself, we additionally evaluate a plain pseudo-text baseline that directly adapts the retrieval model using the same target image–pseudo-text pairs, without SATA or HGMA.

As shown in Table 12, pseudo-text supervision alone improves Rank@1 from 35.99 to 38.85, confirming that the generated descriptions provide useful target-domain supervision. STaG further increases Rank@1 to 44.23 and mAP to 38.64, outperforming the plain pseudo-text baseline by 5.38 and 3.85 points, respectively. Together with the component ablations in the main paper, this result shows that the improvement of STaG extends beyond the benefit of raw pseudo-text supervision.

H Additional Evaluations

This section provides additional evaluations of STaG from several complementary perspectives, including comparisons across different supervision settings, generalization across source domains, architectural compatibility, cross-dataset behavior of CAE, and run-to-run stability. These experiments supplement the controlled comparisons reported in the main paper.

H.1 Comparative Analysis with Human-Annotation-Accessible Baselines

We include fully supervised and text-accessible UDA-TBPR results in Tables 13 to 15 as references across different target-domain supervision settings. Fully supervised configurations use paired target-domain image–text annotations, whereas the text-accessible UDA-TBPR configurations use human-written target-domain descriptions during adaptation. In contrast, STaG uses labeled source-domain data and unlabeled target-domain images, with target-domain textual supervision provided by MLLM-generated pseudo-descriptions rather than human-written annotations. STaG therefore operates under a more restrictive target-domain annotation-access requirement. Comparisons with these reference groups contextualize retrieval performance across supervision settings, rather than establish methodological superiority under identical supervision.

Despite this restriction, STaG remains competitive in several evaluated transfer settings. For example, on CUHK-PEDES → RSTPReid, RaSa+STaG achieves 60.75% Rank@1, compared with 57.44% for GCKD and 52.15% for ReCLIP in the text-accessible UDA-TBPR group of Table 13. These cross-setting observations complement the component ablations in Sec. 4.2 and the comparison with plain pseudo-text adaptation in Appendix G.4, which assess the contributions of STaG within the human-annotation-free setting.

Table 12: **Comparison with plain pseudo-text adaptation** on ICFG-PEDES → CUHK-PEDES.

Method	R@1	R@5	R@10	mAP
Baseline	35.99	56.94	66.23	32.07
Plain pseudo-texts	38.85	59.07	67.28	34.79
STaG	**44.23**	**65.09**	**73.08**	**38.64**

Table 13: Comparisons with state-of-the-art methods using **CUHK-PEDES** as the source domain and **ICFG-PEDES / RSTPReid** as target domains. We report results under three target-domain supervision settings: (1) **Fully Supervised**: uses ground-truth paired image–text annotations from the target domain, serving as a reference under full target-domain supervision; (2) **Text-Accessible UDA-TBPR**: uses human-written target-domain descriptions during adaptation without requiring paired target image–text annotations; and (3) **Human-Annotation-Free UDA-TBPR (Ours)**: uses unlabeled target images and MLLM-generated pseudo-descriptions, without accessing human-written target-domain descriptions during training or adaptation. Results across different supervision settings are reported as cross-setting references rather than strictly controlled comparisons. "Backbone+STaG" further evaluates the compatibility of STaG with different retrieval architectures under the human-annotation-free setting. Controlled analyses under matched human-annotation-free conditions are provided in Sec. 4.2 and Appendix G.4. * indicates results reproduced by us due to the unavailability of official source code.

Setting	Method	Ref.	CUHK-PEDES → ICFG-PEDES				CUHK-PEDES → RSTPReid			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Fully Supervised	IRRA (Jiang & Ye, 2023)	CVPR'23	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
	RaSa (Bai et al., 2023a)	IJCAI'23	65.28	80.40	85.12	41.29	66.90	86.50	91.35	52.31
	TBPS-CLIP (Cao et al., 2023)	AAAI'24	65.05	80.34	85.47	39.83	61.60	81.20	87.70	47.71
	RDE (Qin et al., 2024)	CVPR'24	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
	PTMI (Lu et al., 2025)	TIFS'25	66.54	81.50	86.50	46.05	63.35	81.95	89.10	52.46
	GAHR (Qi et al., 2025)	TIFS'25	68.69	82.83	87.40	42.10	68.85	86.50	91.10	53.60
	ICL (Qin et al., 2025)	CVPR'25	69.22	83.49	88.06	42.34	70.00	86.60	91.70	54.16
UDA-TBPR	POUF (Tanwisuth et al., 2023)	ICML'23	41.08	61.13	69.02	19.06	43.04	64.10	71.09	21.05
	ReCLIP (Hu et al., 2023)	WACV'24	48.17	69.08	75.22	26.04	52.15	76.20	82.16	37.10
	GCKD* (Luo et al., 2025)	AAAI'25	51.16	71.45	78.50	31.34	57.44	78.50	84.70	40.12
Human-Annotation-Free UDA-TBPR	RaSa+STaG	Ours	50.45	67.24	73.55	24.23	60.75	82.25	87.15	45.88

H.2 Generalization Across Diverse Source Domains

We further evaluate STaG with different datasets serving as the source domain. With CUHK-PEDES as the source, STaG reaches 50.45% Rank@1 on ICFG-PEDES and 60.74% on RSTPReid. When ICFG-PEDES is used as the source, STaG obtains 57.44% Rank@1 on CUHK-PEDES and 55.72% on RSTPReid. The consistent gains across these transfer directions indicate that the effectiveness of STaG is not limited to a single source–target configuration. At the same time, the varying absolute performance across datasets reflects the different levels of cross-domain difficulty.

H.3 Architectural Robustness

Table 15 completes our comprehensive evaluation matrix by utilizing RSTPReid as the source domain, successfully adapting it to both CUHK-PEDES and ICFG-PEDES. Since RSTPReid is significantly smaller in scale and possesses distinct imaging characteristics compared to the other two datasets, it imposes an extreme visual domain shift.

This table serves two crucial purposes. ❶ it verifies the architectural agnosticism of STaG. By integrating our framework with three structurally distinct baselines (RaSa, IRRA, and TBPS-CLIP), we observe universal and consistent performance surges, proving STaG is a highly compatible "plug-and-play" solution. ❷ it highlights STaG's remarkable rescue effect under severe domain degradation. When transferring from RSTPReid to CUHK-PEDES, the domain gap is so drastic that the strong baseline IRRA collapses to a mere 32.65% R@1. However, with the integration of STaG, the performance experiences a massive rebound, yielding a remarkable absolute improvement of

Table 14: Comparisons with state-of-the-art methods using **ICFG-PEDES** as the source domain and **CUHK-PEDES / RSTPReid** as target domains. We follow the same target-domain supervision settings and cross-setting reference protocol defined in Table 13.

Setting	Method	Ref.	ICFG-PEDES → CUHK-PEDES				ICFG-PEDES → RSTPReid			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Fully Supervised	IRRA (Jiang & Ye, 2023)	CVPR'23	73.38	89.93	93.71	66.13	60.20	81.30	88.20	47.17
	RaSa (Bai et al., 2023a)	IJCAI'23	76.51	90.29	94.25	69.38	66.90	86.50	91.35	52.31
	TBPS-CLIP (Cao et al., 2023)	AAAI'24	73.54	88.19	92.35	65.38	61.60	81.20	87.70	47.71
	RDE (Qin et al., 2024)	CVPR'24	75.94	90.14	94.12	67.56	65.35	83.95	89.90	50.88
	PTMI (Lu et al., 2025)	TIFS'25	76.02	89.93	94.14	70.85	63.35	81.95	89.10	52.46
	GAHR (Qi et al., 2025)	TIFS'25	76.64	90.46	94.10	66.81	68.85	86.50	91.10	53.60
	ICL (Qin et al., 2025)	CVPR'25	78.18	91.63	94.83	69.58	70.00	86.60	91.70	54.16
UDA-TBPR	POUF (Tanwisuth et al., 2023)	ICML'23	20.50	39.39	49.03	18.48	36.60	61.80	73.15	28.28
	ReCLIP (Hu et al., 2023)	WACV'24	33.48	55.25	64.26	29.19	50.35	73.25	81.20	42.75
	GCKD (Luo et al., 2025)	AAAI'25	52.70	72.76	80.15	45.97	59.95	79.05	85.95	49.68
Human-Annotation-Free UDA-TBPR	RaSa + STaG	Ours	57.44	76.13	83.25	50.97	55.70	76.95	84.15	46.88

Table 15: Comparisons with state-of-the-art methods using **RSTPReid** as the source domain and **CUHK-PEDES / ICFG-PEDES** as target domains. We follow the same target-domain supervision settings and cross-setting reference protocol defined in Table 13. Within the human-annotation-free setting, "Backbone+STaG" evaluates the compatibility of STaG with different retrieval architectures, and Δ denotes the absolute performance change relative to the corresponding backbone baseline. * indicates results reproduced by us due to the unavailability of official source code.

Setting	Method	Ref.	RSTPReid→CUHK-PEDES				RSTPReid→ICFG-PEDES			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Fully Supervised	IRRA (Jiang & Ye, 2023)	CVPR'23	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06
	RaSa (Bai et al., 2023a)	IJCAI'23	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29
	TBPS-CLIP (Cao et al., 2023)	AAAI'24	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83
	RDE (Qin et al., 2024)	CVPR'24	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06
	PTMI (Lu et al., 2025)	TIFS'25	76.02	89.93	94.14	70.85	66.54	81.50	86.50	46.05
	GAHR (Qi et al., 2025)	TIFS'25	76.64	90.46	94.10	66.81	68.69	82.83	87.40	42.10
	ICL (Qin et al., 2025)	CVPR'25	78.18	91.63	94.83	69.58	69.22	83.49	88.06	42.34
UDA-TBPR	POUF (Tanwisuth et al., 2023)	ICML'23	17.84	35.62	46.72	16.30	21.27	37.87	46.66	11.04
	ReCLIP (Hu et al., 2023)	WACV'24	21.04	40.29	50.28	18.92	27.18	42.51	50.28	17.83
	GCKD (Luo et al., 2025)	AAAI'25	51.01	70.29	77.29	43.71	46.40	61.17	66.97	26.06
Human-Annotation-Free	IVT (Shu et al., 2022)	ECCVW'22	16.73	33.65	43.01	14.42	19.43	35.22	43.83	13.52
	POUF (Tanwisuth et al., 2023)	ICML'23	16.41	33.26	43.29	15.10	18.33	34.80	44.09	11.31
	CFine (Yan et al., 2022)	TIP'23	20.79	38.76	48.93	15.77	24.83	39.49	46.92	17.04
	ReCLIP (Hu et al., 2023)	WACV'24	20.36	38.57	48.63	20.15	26.32	39.93	47.78	16.30
	GCKD* (Luo et al., 2025)	AAAI'25	39.15	58.76	67.30	34.92	42.23	58.29	65.24	25.03
	RaSa (Bai et al., 2023a)	IJCAI'23	42.85	61.97	69.79	35.64	41.30	56.18	62.36	22.39
	+ STaG	Ours	54.97	74.43	81.69	48.00	45.10	62.16	69.58	22.73
	Improvement Δ		**+12.12**	**+12.46**	**+11.90**	**+12.36**	**+3.80**	**+5.98**	**+7.22**	**+0.34**
	IRRA (Jiang & Ye, 2023)	CVPR'23	32.65	55.20	65.38	30.17	32.35	49.68	57.74	20.57
	+ STaG	Ours	46.44	67.54	75.88	41.01	40.64	59.23	67.15	20.92
	Improvement Δ		**+13.79**	**+12.34**	**+10.50**	**+10.84**	**+8.29**	**+9.55**	**+9.41**	**+0.35**
	TBPS-CLIP (Cao et al., 2023)	AAAI'24	32.09	53.57	63.28	29.40	37.04	52.65	60.13	23.23
	+ STaG	Ours	40.14	59.58	68.03	35.27	44.14	61.17	68.13	24.19
	Improvement Δ		**+8.05**	**+6.01**	**+4.75**	**+5.87**	**+7.10**	**+8.52**	**+8.00**	**+0.96**

+13.79% in R@1 and +10.84% in mAP. This demonstrates that when direct cross-domain inference suffers substantial performance degradation, STaG can effectively improve the representation space and retrieval performance.

H.4 Cross-Dataset Generalization of CAE

To examine whether the distribution sharpening introduced by CAE causes over-specialization to the adaptation domain, we directly evaluate the ICFG-PEDES → CUHK-PEDES adapted model on RSTPReid without further adaptation.

Table 16: **Cross-dataset generalization of CAE.** The STaG variants adapted on ICFG-PEDES → CUHK-PEDES are directly evaluated on RSTPReid without further adaptation. Baseline denotes the source-only model trained on ICFG-PEDES.

Model	R@1	R@5	R@10	mAP
Baseline	44.50	67.00	76.85	33.45
STaG w/o $\mathcal{L}_{\text{CAE}}$	50.15	72.85	80.75	37.86
STaG w/ $\mathcal{L}_{\text{CAE}}$	**50.55**	**73.50**	**81.45**	**37.92**

As shown in Table 16, adding CAE consistently improves all retrieval metrics, increasing Rank@1 from 50.15 to 50.55. This result provides evidence that CAE does not impair cross-dataset transferability in this setting, while slightly improving the generalization of the adapted representation.

I Impacts and Limitations

I.1 Limitations

While our proposed STaG framework significantly reduces manual annotation costs and bridges the domain gap in the Human-Annotation-Free UDA-TBPR setting, it still exhibits a few limitations that warrant future investigation:

❶ Although STaG demonstrates strong robustness against typical environmental shifts, extreme physical scenarios (such as profoundly blurred surveillance footage, severe occlusion, or low-light nighttime environments) inherently lack extractable visual cues. Under such boundary conditions, fine-grained attributes become exceedingly difficult to capture, which disrupts the explicit cross-modal alignment in the target domain and consequently degrades the final retrieval performance.

❷ Our framework is primarily designed to effectively mitigate structural domain gaps across standard RGB environments. However, if the target domain involves a drastically different imaging modality (*e.g.*, infrared or thermal imagery) that diverges fundamentally from the RGB source, the cross-domain graph must contend with an entirely new level of modality discrepancy.

Addressing both of these challenges presents highly promising directions for future research.

I.2 Impact Statement

The proposed STaG framework significantly advances the practical deployment of Text-Based Person Retrieval (TBPR) by eliminating the need for costly target-domain text annotations, thereby facilitating rapid, scalable adaptation to new surveillance environments. While this enhances public safety through the efficient location of missing persons or suspects, it also necessitates rigorous ethical oversight regarding privacy and the potential for unauthorized tracking. By introducing Style-Agnostic Text Alignment (SATA) and Heterogeneous Graph Memory Alignment (HGMA), our work mitigates the risks of stylistic bias and visual domain shift; however, we emphasize that such systems should serve as assistive tools for human operators rather than autonomous decision-makers. To ensure positive societal outcomes, future research must prioritize demographic fairness auditing and the integration of privacy-preserving training protocols to safeguard individual rights while maintaining high-fidelity cross-modal retrieval.

J Visualization Analysis

To complement the quantitative evaluations, we provide qualitative analyses from four perspectives: spatial cross-modal grounding, attention over generated pseudo-texts, neighborhood retrieval in HGMA, and final retrieval behavior. These visualizations provide an intuitive view of how SATA and HGMA affect the learned representations under domain shift.

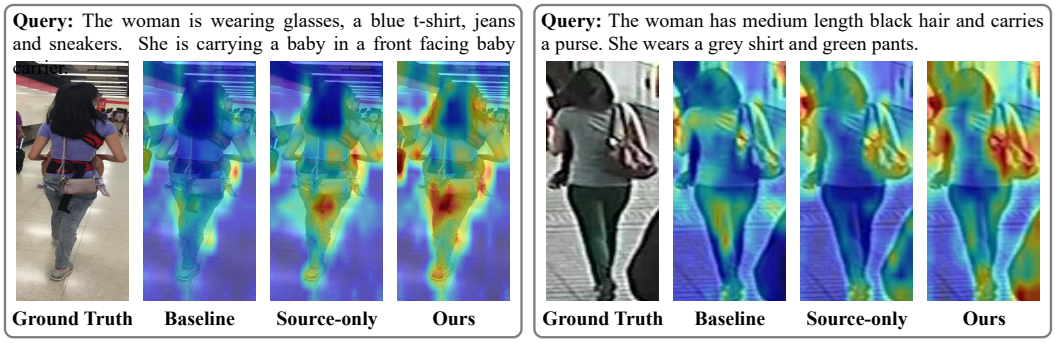


Figure 6: **Visualization of activation maps and token-level importance.** We compare three configurations on representative positive image–text pairs for the transfer ICFG-PEDES → CUHK-PEDES: **Baseline** denotes TBPS-CLIP (Cao et al., 2023) trained on the source domain and directly evaluated on the target domain without any adaptation; **Pseudo-Text Adaptation** denotes TBPS-CLIP directly fine-tuned with target-domain image–pseudo-text pairs without SATA or HGMA; **Ours** denotes the full STaG framework. The **Ground Truth** column shows the original target-domain image for reference. Warmer colors indicate stronger spatial activation.

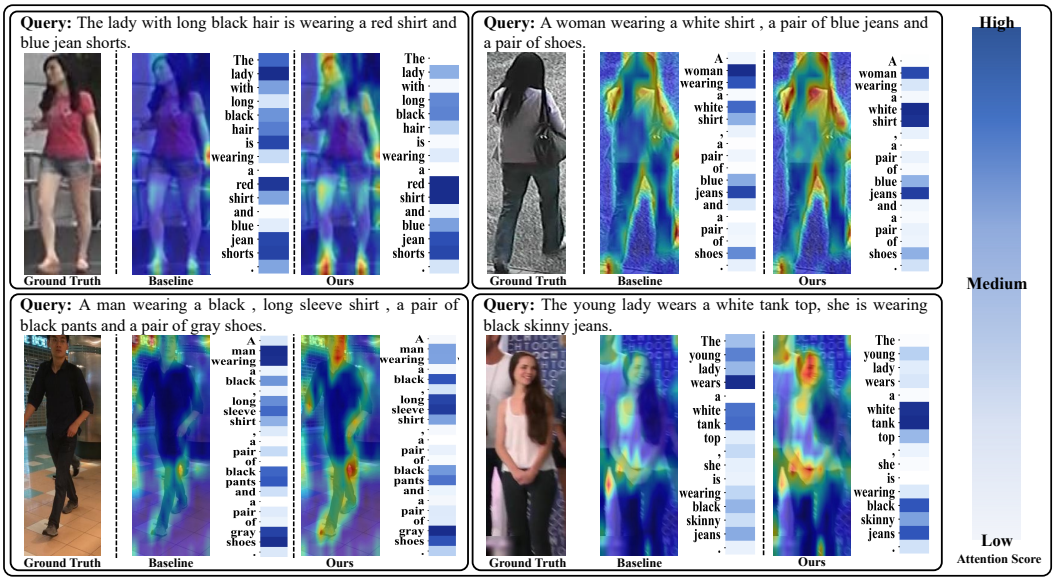


Figure 7: **Visualization of spatial activation and token-level importance.** We compare TBPS-CLIP (Cao et al., 2023) (Baseline) with STaG (Ours) on representative positive image-text pairs under the human-annotation-free UDA-TBPR setting, using ICFG-PEDES as the source domain and evaluating on target datasets. Warmer colors indicate stronger spatial activation on the image, and the bar plots show the normalized contribution of each word to the matching score.

J.1 Visualization of Image Activation Maps

Figs. 6 and 7 visualize the spatial and textual attention learned under domain shift. Compared with the baseline, STaG consistently produces more concentrated responses on query-relevant person regions, such as upper-body clothing, pants, and shoes, while reducing attention to less relevant visual regions. A similar tendency is observed on the text side, where attribute-bearing words related to colors, garments, and accessories contribute more strongly to the matching score.

The consistency between spatial localization and textual importance suggests that STaG improves not only global feature alignment, but also the correspondence between fine-grained visual evidence and descriptive attributes. These observations provide qualitative support for the improved cross-modal grounding achieved by STaG under domain shift.

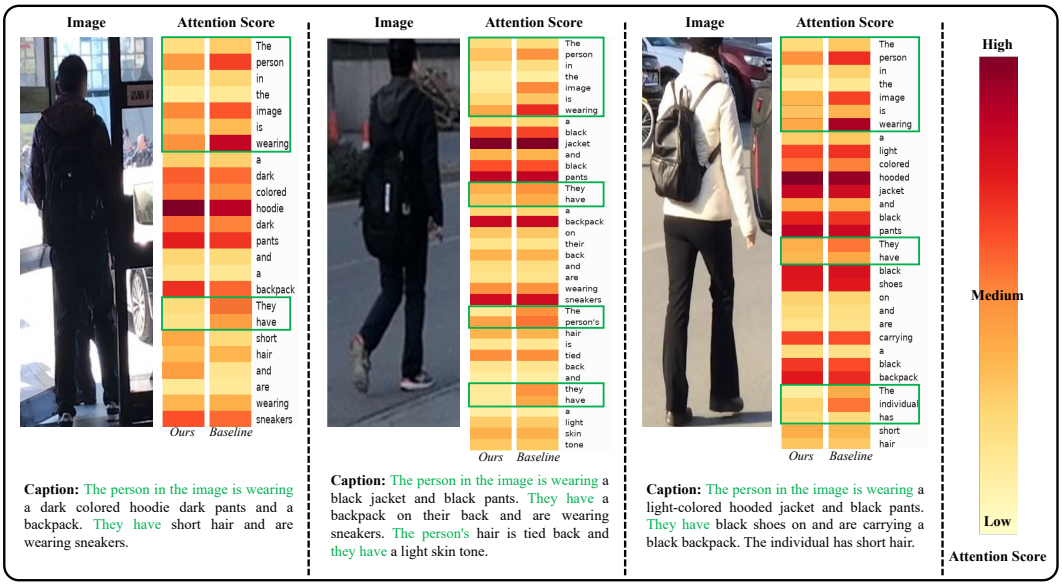


Figure 8: **Token-level attention on generated pseudo-texts.** We compare TBPS-CLIP (Cao et al., 2023) (Baseline) with STaG (Ours) on representative image–pseudo-text pairs under the human-annotation-free UDA-TBPR setting (ICFG-PEDES as source). Bar plots show the normalized contribution of each token to the matching score. Green boxes mark MLLM-specific stylistic templates, which carry no identity-discriminative information. The baseline distributes attention diffusely across these templates, whereas STaG concentrates attention on attribute-bearing tokens.

J.2 Visualization of Attribute-Level Attention in Pseudo-Texts

To complement the quantitative style-invariance analysis in Sec. 4.2.2, Fig. 8 visualizes token-level attention over generated pseudo-texts. The baseline assigns noticeable attention to recurring MLLM-specific templates such as "The person in the image is wearing...", whereas STaG shifts attention toward attribute-bearing words describing colors, garments, and accessories.

These template expressions occur frequently across generated descriptions but provide limited instance-specific visual information. Reducing their influence therefore helps prevent the model from relying excessively on generator-specific writing patterns. Meanwhile, the increased attention to observable attributes is consistent with the motivation of SATA: using shared visual anchors to encourage representations that preserve semantic content while reducing sensitivity to textual style. This visualization provides instance-level evidence complementary to the distribution-level analysis in Appendix G.1.

J.3 Visualization of Retrieved HGMA Neighbors

Fig. 9 visualizes representative neighbors retrieved from the source-image, source-text, and target-image memories. Although the cross-domain samples do not share identities with the target query, the retrieved anchors exhibit similar fine-grained appearance attributes, such as clothing color, garment type, and carried objects.

The three memories provide complementary cues: source-image neighbors offer cross-domain visual references, source-text neighbors introduce text-side semantic information, and target-image neighbors preserve the local structure of the target domain. Their semantic consistency illustrates how HGMA combines heterogeneous but feature-similar evidence rather than relying on assumed cross-domain identity correspondence. Together with the random-neighbor and memory-source ablations, this visualization qualitatively supports the neighborhood construction adopted by HGMA.

J.4 Visualization of Retrieval Results

Fig. 10 presents representative top-5 retrieval results under two cross-domain transfer settings. Compared with the baseline, STaG places more correct matches at higher ranks and reduces errors caused

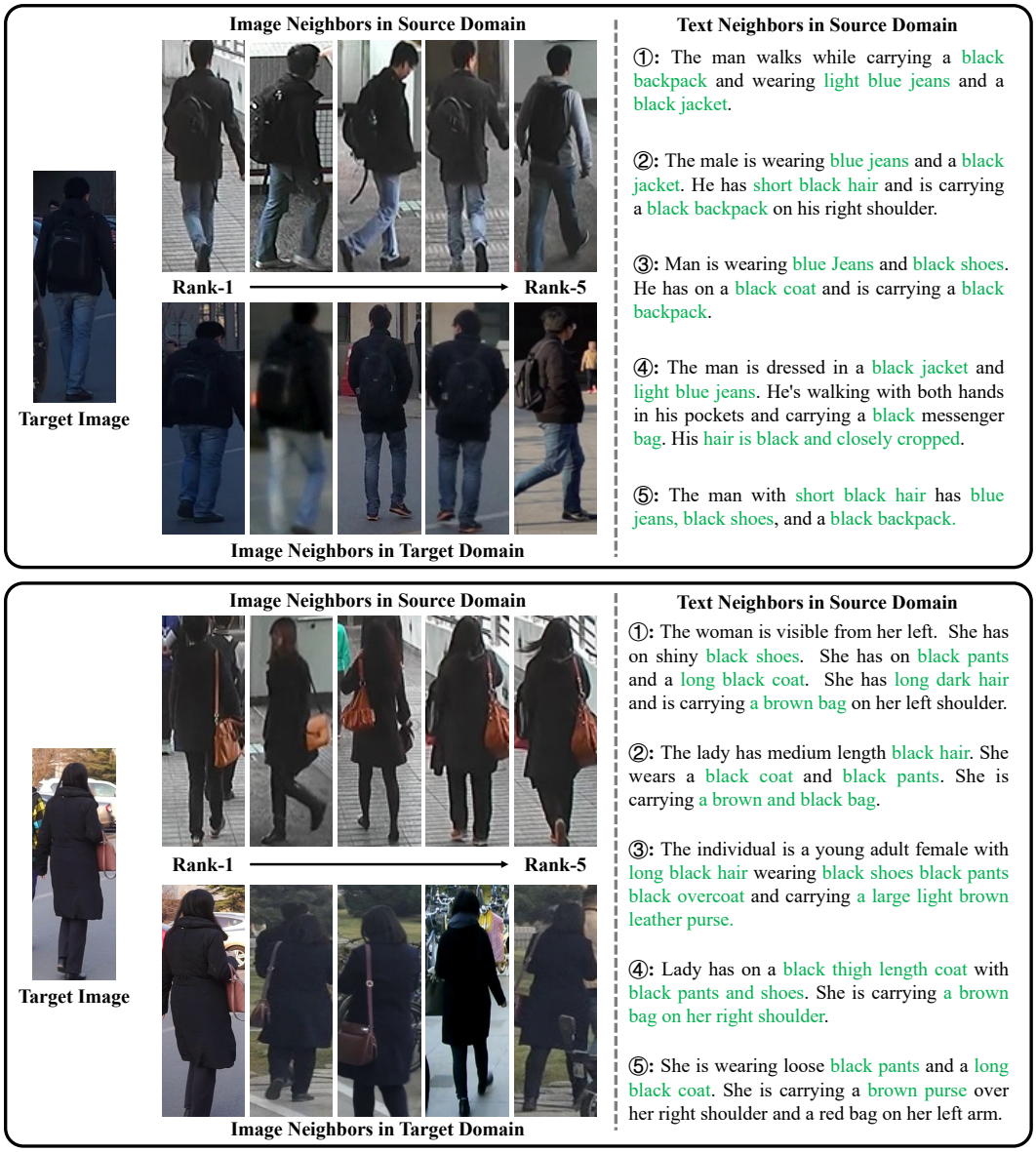


Figure 9: **Visualization of HGMA-retrieved neighbors.** Given a target query, HGMA retrieves feature-similar anchors from the source-image, source-text, and target-image memories. Target-domain annotations, if displayed, are used only for post-hoc visualization and are not accessed during adaptation.

by visually similar distractors. The baseline often retrieves incorrect candidates that share a dominant coarse attribute with the query, such as clothing color, while differing in other fine-grained cues such as garment type, footwear, or accessories.

In contrast, STaG better preserves combinations of multiple descriptive attributes and remains more stable under variations in illumination, viewpoint, and background. These examples suggest that the quantitative improvements in Rank@K and mAP are accompanied by more discriminative ranking behavior rather than merely changes in overall similarity scores.

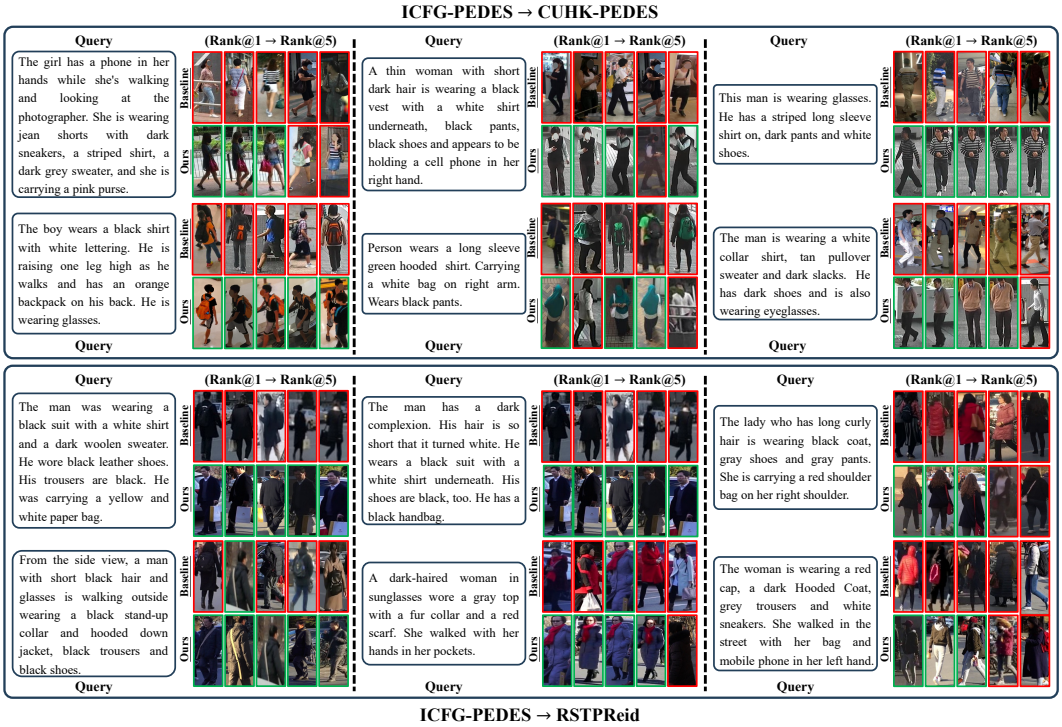


Figure 10: **Qualitative retrieval comparison (Rank@1→Rank@5).** We show the top-5 retrieved images of TBPS-CLIP (Cao et al., 2023) (Baseline) and STaG (Ours) for representative text queries under the human-annotation-free UDA-TBPR setting. Green and red boxes denote correct and incorrect matches, respectively.