RAG-ReID: Reliable Query Augmentation via Self-Supervised Alignment for Person Re-Identification

Anonymous ACL submission

Abstract

Text-based Person Re-Identification (TB-ReID) often fails when a single-shot query underspecifies user intent. While recent Multimodal Large Language Models (MLLMs) offer new avenues for query refinement, standard generation paradigms are prone to severe hallucinations and redundancy, leading to unreliable retrieval contexts. To address this without relying on costly human-in-the-loop dialogues, we propose RAG-ReID, an automated closed-loop framework that ensures reliable query augmentation via self-supervised alignment. Our decoupled pipeline first employs an MLLM Reranker to purify noisy initial retrieval results into a clean Top-N visual context. Second, rather than relying on unconstrained MLLM generation, we introduce a lightweight, token-level reward model trained via Direct Preference Optimization (DPO) on self-mined multi-view preference pairs. This plug-and-play guidance steers a frozen MLLM generator during decoding to synthesize highly discriminative and factually grounded augmented queries, effectively mitigating visual distractors with negligible computational overhead. By transforming ambiguous initial matches into reliable cues, RAG-ReID achieves robust iterative refinement entirely offline. Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid demonstrate that our self-supervised alignment significantly suppresses hallucinations and achieves state-of-the-art performance.

1 Introduction

Text-based Person Re-Identification (TB-ReID) (Cao et al., 2024; Ding et al., 2021; Wang et al., 2022a; Chen et al., 2022; Zhu et al., 2021; Wang et al., 2020; Gao et al., 2024) retrieves target persons from large galleries using natural language descriptions. It is highly practical for surveillance (Bukhari et al., 2023) and crowd management (Galiyawala and Raval, 2021) where query images are unavailable. Unlike image-to-image ReID (Zhang et al., 2024b; Zuo et al., 2023, 2024a; Wang et al., 2025), TB-ReID faces a severe modality gap. Most prior works improve static models via attribute modeling (Zuo et al., 2024b; Yan et al., 2023b) or domain pre-training (Tan et al., 2024; Yang et al., 2023). However, they assume the initial query perfectly encodes user intent. In reality, queries are often concise, vague, or ambiguous. **As illustrated in the top half of Fig. 1, this standard single-shot paradigm yields noisy retrieval contexts and brittle localization.** While recent chat-based retrieval methods attempt to clarify intent, they rely heavily on continuous human-in-the-loop feedback. This is computationally expensive and impractical for automated surveillance, exposing a significant bottleneck in the standard one-pass retrieval paradigm.

To break this bottleneck without relying on costly human interaction, we propose RAG-ReID: an **automated closed-loop** Retrieval–Augmentation–Generation–Re-retrieval pipeline **(see the bottom half of Fig. 1)**. The core idea is to transform static querying into a self-optimizing process that leverages gallery feedback for **reliable query augmentation**. To achieve this robustly, we decouple the retrieval analysis into two dedicated modules:

First, an *MLLM Reranker* performs context purification. Given the noisy initial retrieval (Top-K), this lightweight module executes fast relevance filtering, removing identity-irrelevant candidates to produce a clean Top-N context.

Second, an *MLLM Generator* distills features from this purified context to synthesize an augmented query. However, recent multimodal generation frameworks face a critical issue here: unconstrained MLLM generators often suffer from severe hallucinations. If hard negatives bypass the reranker, the generator might mistakenly reinforce

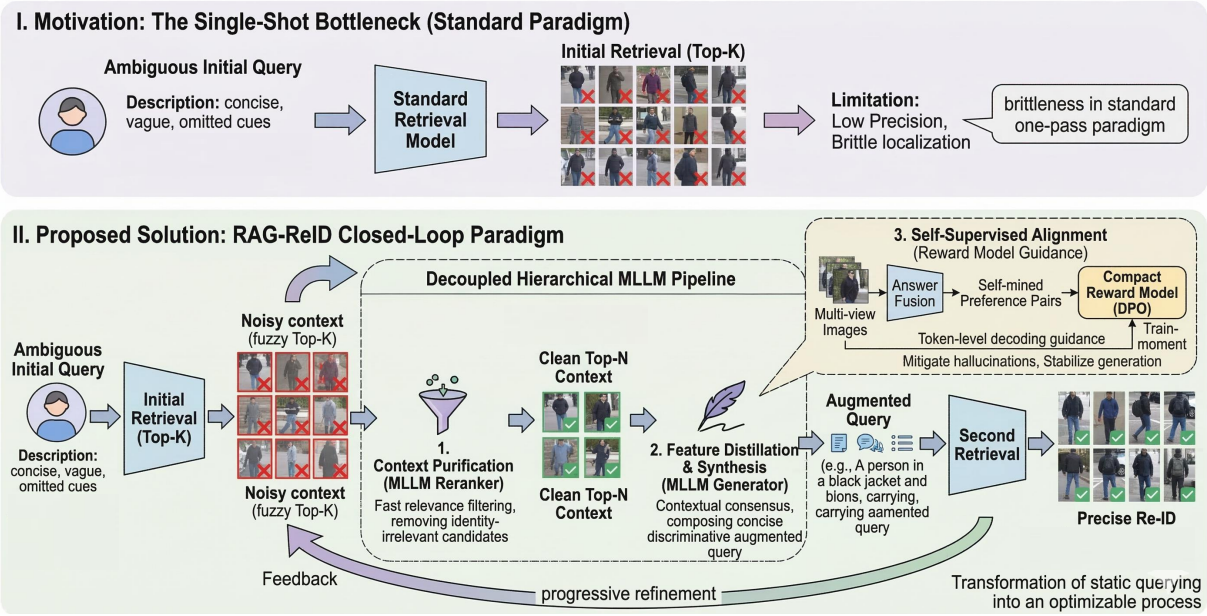

Figure 1: The proposed RAG-ReID framework recasts text-based person re-identification from a brittle single-shot retrieval paradigm into a dynamic and optimizable closed-loop process.

visual distractors.

To solve this reliability bottleneck, we introduce a lightweight **self-supervised alignment** scheme. We automatically mine preference pairs via multi-view image augmentations and train a compact reward model using Direct Preference Optimization (DPO). During decoding, this model provides plug-and-play, token-level guidance. By steering the frozen MLLM generator away from hallucinations with negligible overhead, it ensures the augmented query reliably describes the true target rather than visual distractors. This work makes the following key contributions:

- We introduce RAG-ReID, a framework that overcomes the single-shot bottleneck by transforming TB-ReID into an **automated, closed-loop** process, eliminating the need for costly human-in-the-loop dialogues.
- We propose a decoupled hierarchical MLLM pipeline that separates context purification (Reranker) from feature distillation (Generator), enabling robust consensus over multiple candidates.
- We present a novel self-supervised alignment module. By utilizing a DPO-trained reward model for token-level decoding guidance, it effectively mitigates hallucinations and stabilizes query generation without modifying the backbone MLLM.

Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid demonstrate that RAG-ReID substantially improves ranking accuracy and robustness, particularly under ambiguous descriptions, validating the effectiveness of our reliable closed-loop design.

2 Related Work

2.1 Text-based Person Re-Identification

Text-based Person Re-Identification (TB-ReID) retrieves target persons using natural language descriptions (Zhu et al., 2021; Bai et al., 2023; Jiang and Ye, 2023; Qin et al., 2024; Tan et al., 2024). Prior works largely focused on improving cross-modal alignment via modality-specific encoders (Wang et al., 2020; Wu et al., 2021), vision–language pre-training (Radford et al., 2021; Li et al., 2022), or large-scale synthetic data generation (Yang et al., 2023; Tan et al., 2024). However, these static, single-round retrieval paradigms assume that initial queries fully encode user intent. In practice, ambiguous or concise queries create a significant accuracy bottleneck. While recent interactive paradigms (Xie et al., 2025) attempt to overcome this by clarifying intent via chat-based retrieval, they require continuous, labor-intensive human engagement. This reliance on human-in-the-loop feedback makes them highly impractical for automated, large-scale surveillance scenarios. In contrast, RAG-ReID recasts TB-ReID as an **auto-**

mated, closed-loop process that leverages gallery feedback to progressively refine queries entirely offline, eliminating the need for costly human intervention.

2.2 Retrieval-Augmented Generation and MLLMs

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and Multimodal Large Language Models (MLLMs) (Chen et al., 2024; Asai et al., 2024) have shown immense promise in automated query refinement. For instance, a closely related concurrent work, ICL (Qin et al., 2025), employs MLLMs to filter candidates and generate descriptive text for interactive query refinement. However, its reliance on the unconstrained generation capabilities of base MLLMs makes it highly susceptible to hallucinations and noise amplification. To address these critical reliability limitations, we introduce a decoupled hierarchical architecture where an MLLM Reranker first purifies noisy Top-K candidates, and a frozen MLLM Generator then performs consensus-driven synthesis. Crucially, rather than relying on unconstrained generation, we guide our generator using a lightweight, DPO-trained token-level reward model. This plug-and-play self-supervised alignment strictly anchors the augmented query to the purified visual context, effectively mitigating hallucinations and transforming static TB-ReID into a robust, highly reliable closed-loop system.

3 Methodology

In this section, we present **RAG-ReID**, a novel framework that transforms **Text-based Person Re-Identification (TB-ReID)** from a static one-shot retrieval task into a dynamic, closed-loop refinement process. To address the bottleneck of ambiguous or incomplete user queries, our approach establishes an iterative pipeline that progressively enhances query specificity by leveraging gallery feedback. The core of our design is a **Hierarchical MLLM Pipeline** that decouples the critical analysis phase into two specialized stages, thereby avoiding the task interference and suboptimal performance that arise when a single model attempts to simultaneously filter candidates and distill features. Specifically, we introduce (1) an **MLLM Reranker** dedicated to *context purification*—identifying truly relevant candidates from the initial Top-K retrieval set to produce a clean Top-N context; and (2) a **Preference-Guided MLLM Generator** specialized in *feature distillation*—performing cross-candidate consensus reasoning over the purified Top-N to synthesize a refined, discriminative query for re-retrieval. This separation of concerns allows each module to focus on its core competency, collectively converting imperfect initial matches into precise augmentation signals.

We structure the presentation as follows: Section 3.1 formalizes the problem and provides a complete pipeline overview. Section 3.2 details our **MLLM Reranker** for context purification. Section 3.3 introduces our **Preference-Guided MLLM Generator**, where the preference alignment mechanism acts as a quality assurance system to mitigate hallucinations and redundancy during query synthesis.

3.1 Problem and Symbol Statement

Suppose that we have the query text $q \in \mathcal{Q}$, a pedestrian image $v \in \mathcal{V}$, where $\mathcal{Q}$ and $\mathcal{V}$ are the text query set and image gallery. The purpose of TB-ReID is to utilize the text query to retrieve the ideal images from the gallery, thus achieving person searching by identities of retrieved images. To achieve this, existing methods usually train an offline cross-modal model $f_{\text{cross}} = \{f, g\}$ to measure the similarities between text queries and pedestrian images. For a text q and an image v, the similarity to measure the matching degree can be represented as $S_{q,v} \equiv \text{Sim}(f(q), g(v))$, where $\text{Sim}(*)$ is the similarity function, f and g are the text and image encoders, respectively. Then, we can utilize q to search relevant candidate images $\hat{v}$ from $\mathcal{V}$ as follows:

$$\hat{\mathcal{V}}(q) = \{\hat{v}_k\}_{k=1}^{K} = \text{Top-}K_{v \in \mathcal{V}}(S_{q,v}), \quad (1)$$

where $\hat{\mathcal{V}}(q)$ is the candidate set for q and K is the number of candidate images. In addition, our RAG-ReID also involves MLLMs to conduct interactions and training augmentation, which we denote as $\mathcal{M}$, and the prompt template function for $\mathcal{M}$ is denoted by the symbol $\mathcal{T}$. Based on the above definitions, we will elaborate on RAG-ReID in the following sections.

3.2 MLLM-Driven Candidate Reranking

The initial Top-K retrieval set $\hat{\mathcal{V}}(q)$ obtained via Eq. 1 often contains mixed identities that partially match the query q due to shared superficial attributes (e.g., similar clothing colors). To extract a

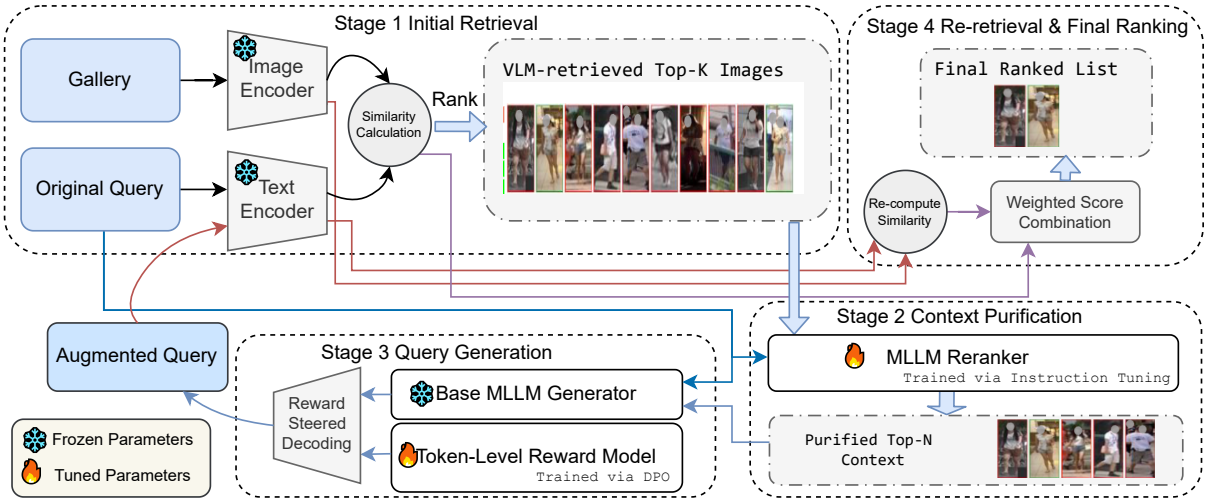

Figure 2: **Overview of the RAG-ReID framework. Stage 1:** Vision-language encoders retrieve a noisy Top-K candidate set from the gallery. **Stage 2:** An MLLM Reranker filters identity-irrelevant images to produce a clean Top-N context. **Stage 3:** An MLLM Generator analyzes the purified context to compose an augmented query; a lightweight Token-Level Reward Model (trained with DPO) provides plug-and-play decoding guidance to reduce hallucinations without fine-tuning the base MLLM generator. **Stage 4:** The augmented query drives re-retrieval, and scores from both rounds are fused to produce the final ranking.

clean, high-confidence subset for subsequent query augmentation, we introduce an **MLLM Reranker** that performs fine-grained text-image relevance assessment, filtering out identity-irrelevant candidates while retaining true matches.

3.2.1 Binary Relevance Filtering via MLLM Judgment

Relevance Scoring Mechanism. Rather than relying solely on the cross-modal encoder's similarity scores, we leverage the MLLM $\mathcal{M}$'s natural language reasoning ability to make explicit relevance judgments. For each candidate image $\hat{v}_k$ in $\hat{\mathcal{V}}(q)$, we query the MLLM with a binary verification prompt:

$$a^q_{\hat{v}_k} = \mathcal{M}(\mathcal{T}_{\text{filter}}(q, \hat{v}_k)), \quad k \in \{1, \ldots, K\}, \quad (2)$$

where $\mathcal{T}_{\text{filter}}$ asks: "Does the provided text description accurately match the person shown in this image?", and $a^q_{\hat{v}_k} \in \{\texttt{Yes}, \texttt{No}\}$ is the MLLM's binary response. This process enables the model to identify subtle mismatches that similarity-based ranking might miss.

Supervised Fine-Tuning. Generic MLLMs are pre-trained on natural images, not ReID pedestrian photos, so relevance judgments can be unreliable. We address this with parameter-efficient fine-tuning using LoRA (Hu et al., 2021), adapting $\mathcal{M}$ for person search while preserving base capabilities. We minimize cross-entropy loss over binary responses:

$$\mathcal{L}_{\text{adapt}} = -\sum_{(x,y)\in\mathcal{D}_{\text{adapt}}} \sum_{t=1}^{|y|} \log p_{\mathcal{M}}(y_t \mid x, y_{<t}), \quad (3)$$

where x is an input prompt and y is the target response. The adaptation dataset $\mathcal{D}_{\text{adapt}} = \mathcal{D}^+ \cup \mathcal{D}^-$ includes:

- **Positive**: $\mathcal{D}^+$ consists of prompts $\{\mathcal{T}_{\text{filter}}(q_i, v_i^+)\}$, pairing training queries q_i with their ground-truth images v_i^+, expecting `Yes`.
- **Negative**: $\mathcal{D}^-$ draws Top-100 non-matching images (different IDs) for each query q_i via f_{cross} to build hard negatives $\{\mathcal{T}_{\text{filter}}(q_i, v_i^-)\}$, expecting `No`. This hard-negative mining promotes truly discriminative judgments.

3.2.2 Iterative Purification

Confidence-Based Filtering Strategy. Querying the MLLM for all K candidates is costly. We propose a **stratified filtering strategy**: prioritize high-confidence cases and reserve MLLM for hard cases.

Round 1 (High-Confidence Filtering): Only consider candidates where:

- The Top-1 similarity exceeds a threshold: $S_{q,\hat{v}_1} > \xi$, where ξ is set based on similarity statistics.

- The MLLM judges the Top-1 as relevant: $a^q_{\hat{v}_1} = \texttt{Yes}$.

If both hold, proceed directly to attribute extraction (Section 3.2.3).

Round $k > 1$ (Low-Confidence Recovery): For queries without a Top-1 match (above), sequentially evaluate $\hat{v}_k$ if $S_{q,\hat{v}_1} \leq \xi$, stopping at the first Yes or after Top-K. This two-tier strategy prevents excess computation on easy or impossible cases.

Formalized Purification Logic. Let $\mathcal{V}_{\text{filtered}}(q)$ denote all candidates approved by the MLLM:

$$\mathcal{V}_{\text{filtered}}(q) = \{\hat{v}_k \in \hat{\mathcal{V}}(q) : S_{q,k} > \xi \text{ and } a_k = \texttt{Yes}\}. \tag{4}$$

In practice, for each query q, $\mathcal{V}_{\text{filtered}}(q)$ includes all candidates receiving both high similarity and a Yes judgment. The resulting $|\mathcal{V}_{\text{filtered}}(q)| = N_q \geq 1$ is variable.

This adaptive mechanism efficiently reduces MLLM invocations while maintaining filtering accuracy.

3.2.3 Attribute Extraction and Query Enhancement

Once a purified set $\bar{\mathcal{V}}(q) = \{\bar{v}_i\}_{i=1}^{N_q}$ ($N_q \geq 1$, adaptively selected) is obtained from the above, we extract fine-grained attributes to augment the original query q.

Structured Attribute Querying. For each anchor $\bar{v}_i \in \bar{\mathcal{V}}(q)$, the MLLM is queried with N_{attr} structured questions $\{c_j\}_{j=1}^{N_{\text{attr}}}$:

$$r_{\bar{v}_i} = \mathcal{M}(\mathcal{T}_{\text{extract}}(\{c_1, \ldots, c_{N_{\text{attr}}}\}, \bar{v}_i)), \tag{5}$$

where $\mathcal{T}_{\text{extract}}$ is a multi-question prompt and $r_{\bar{v}_i}$ contains the responses for gender, hair, upper/lower body garments, footwear, and accessories.

Attribute Fusion and Sequence Extension. Direct concatenation of $r_{\bar{v}_i}$ and q may yield redundancy or exceed encoder limits (e.g., CLIP's 77 tokens). Following Long-CLIP (Zhang et al., 2024a) and FineCLIP (Asokan et al., 2025), we expand CLIP's token limit to 248 via positional embedding interpolation.

The MLLM then synthesizes an augmented query:

$$q' = \mathcal{M}(\mathcal{T}_{\text{merge}}(q, \{r_{\bar{v}_i}\}_{i=1}^{N_q})), \tag{6}$$

where $\mathcal{T}_{\text{merge}}$ prompts a comprehensive description combining the original query and all extracted attributes.

3.2.4 Re-Ranking via Score Fusion

Given the augmented query q', re-ranking is performed by weighted similarity fusion:

$$\hat{S}_{q,v} = \lambda \cdot S_{q,v} + (1 - \lambda) \cdot S_{q',v}, \tag{7}$$

where $S_{q,v}$ comes from Eq. 1, $S_{q',v} = \text{Sim}(f(q'), g(v))$, and $\lambda \in [0, 1]$ (default $\lambda = 0.5$).

3.2.5 Complete Pipeline

Computational Complexity. MLLM inference dominates cost. The framework supports flexible trade-offs: stronger MLLMs yield better filtering at higher cost, while early stopping via confidence ensures efficiency.

3.3 Preference-Guided MLLM Generator

After purifying the initial retrieval into an adaptive set $\bar{\mathcal{V}}(q) = \{\bar{v}_1, \ldots, \bar{v}_N\}$, where N denotes all candidates that meet both the high similarity criterion and receive positive MLLM judgments, we synthesize an augmented query q' that captures shared discriminative characteristics among these images. Directly prompting the MLLM $\mathcal{M}$ to "describe common features" often results in **(i) hallucination** (inventing details not present) or **(ii) redundancy** (simply restating the original query). To address these issues, we introduce a **token-level reward model** that steers generation toward factually grounded and informative outputs, while keeping the parameters of $\mathcal{M}$ frozen.

3.3.1 Preference Data from Multi-View Observations

Our approach builds on a simple observation: different visual transformations of the same images highlight complementary discriminative details. Synthesizing descriptions across these perspectives constructs training signals that reward comprehensive, detail-rich outputs.

Preference Pair Generation. Given query q and purified candidates $\bar{\mathcal{V}}(q) = \{\bar{v}_1, \ldots, \bar{v}_N\}$, we construct training triplet $(q, \bar{\mathcal{V}}(q), y^+, y^-)$ where y^+ and y^- denote preferred and baseline responses:

1. **Baseline Response y^-:** Generate description from original images:

$$y^- = \mathcal{M}(\mathcal{T}_{\text{gen}}(q, \bar{\mathcal{V}}(q))). \tag{8}$$

2. **Multi-View Descriptions:** Apply K visual transformations $\{T_1, \ldots, T_K\}$ to obtain augmented views:

$$\tilde{y}_k = \mathcal{M}(\mathcal{T}_{\text{gen}}(q, \{T_k(\bar{v}_i)\}_{i=1}^{N})), \quad k = 1, \ldots, K, \tag{9}$$

where T_k includes cropping, color jittering, and flipping operations.

3. **Preferred Response y^+:** Synthesize all perspectives via consensus prompting:

$$y^+ = \mathcal{M}(\mathcal{T}_{\text{merge}}([\tilde{y}_1; \ldots; \tilde{y}_K])), \quad (10)$$

where $\mathcal{T}_{\text{merge}}$ instructs merging into a comprehensive description.

We model preference via Bradley-Terry:

$$p(y^+ \succ y^- \mid q, \bar{\mathcal{V}}) = \sigma\big(r(y^+) - r(y^-)\big), \quad (11)$$

where $r(y)$ denotes the reward for response y and $\sigma(\cdot)$ is the sigmoid function.

3.3.2 Token-Level Reward Model

We train a compact autoregressive reward scorer π_r (150M parameters) that scores each token during generation. Let $y = (y_1, \ldots, y_{|y|})$ denote a response and $y_{<t}$ its prefix up to position $t-1$. The total reward decomposes across timesteps:

$$r(y) = \sum_{t=1}^{|y|} \log \pi_r(y_t \mid q, \bar{\mathcal{V}}, y_{<t}), \quad (12)$$

enabling single-pass evaluation versus costly sequence-level rollouts.

Training Objective. We optimize π_r on preference dataset $\mathcal{D}_{\text{pref}} = \{(q_i, \bar{\mathcal{V}}(q_i), y_i^+, y_i^-)\}$ using DPO loss:

$$\mathcal{L}_{\text{DPO}}(\pi_r) = -\mathbb{E}_{(y^+,y^-)\sim\mathcal{D}_{\text{pref}}}\Big[\log \sigma\big(\beta(r(y^+) - r(y^-))\big)\Big] \text{with } (y^+, y^-) \sim \mathcal{D}_{\text{pref}}. \quad (13)$$

where $\sigma(\cdot)$ is the sigmoid function, $\beta = 0.1$ controls the preference margin, and $r(y)$ is defined in Eq. 12.

3.3.3 Reward-Guided Decoding

At inference, we blend the base generator $\mathcal{M}$ with the reward model π_r via multiplicative reweighting:

$$\pi_{\text{decode}}\big(y_t \mid q, \bar{\mathcal{V}}, y_{<t}\big) \propto \pi_{\mathcal{M}}\big(y_t \mid q, \bar{\mathcal{V}}, y_{<t}\big) \cdot \Big[\pi_r\big(y_t \mid q, \bar{\mathcal{V}}, y_{<t}\big)\Big]^{\lambda}, \quad (14)$$

where $\lambda \in [0.5, 2.0]$ controls the influence of reward guidance. This approach biases generation toward high-reward outputs without updating the parameters of $\mathcal{M}$. **Cross-Tokenizer Alignment.** If π_r and $\mathcal{M}$ use different tokenization schemes, we align their outputs using a top-k mapping ($k = 50$): we extract the top-k tokens from π_r, convert these to strings, re-tokenize with $\mathcal{M}$'s vocabulary, and allocate the associated probabilities accordingly. In practice, this affects token selection in fewer than 2% of cases. Algorithm 1 provides a summary of the full procedure.

Algorithm 1 Reward-Guided Query Augmentation

Require: Query q, purified images $\bar{\mathcal{V}}(q)$, MLLM $\mathcal{M}$, transformations $\{T_k\}_{k=1}^K$, guidance weight λ
Ensure: Enhanced query q'

```
// Phase A: Construct Preference Data
y⁻ ← M(T_gen(q, V̄))
for k = 1 to K do
    ỹ_k ← M(T_gen(q, {T_k(v̄_i)}_{i=1}^N))
end for
y⁺ ← M(T_merge([ỹ_1; ...; ỹ_K]))
D_pref ← D_pref ∪ {(q, V̄, y⁺, y⁻)}
// Phase B: Train Reward Model
Optimize π_r via L_DPO on D_pref
// Phase C: Guided Decoding
y ← ∅
for t = 1 to T_max do
    π_decode(y_t) ∝ π_M(y_t | q, V̄, y_<t) · [π_r(y_t | q, V̄, y_<t)]^λ
    y_t ← arg max π_decode(y_t | q, V̄, y_<t)
    y ← y ⊕ y_t
    if y_t = [EOS] then
        break
    end if
end for
return q' = y
```

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. We evaluate on three benchmarks: CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021). All datasets follow their official data splits. Further details are provided in the supplementary materials. **Evaluation Protocols.** We report Rank-K ($K \in \{1, 5, 10\}$) as the main metric, and also include mean Average Precision (mAP) following (Jiang and Ye, 2023; Qin et al., 2024).

Table 1: Performance comparisons on the CUHK-PEDES dataset. The best results are in **bold**, and the second best are underlined.

Methods	ImageEnc.	TextEnc.	R-1	R-5	R-10	mAP
ViTAA (Wang et al., 2020)	RN50	LSTM	54.92	75.18	82.90	51.60
Lapscore (Wu et al., 2021)	RN50	BERT	63.40	-	87.80	-
LBUL (Wang et al., 2022b)	RN50	BERT	64.04	82.66	87.22	-
IVT (Shu et al., 2022)	ViT-Base	BERT	65.59	83.11	89.21	-
CFine (Yan et al., 2023b)	CLIP-ViT	BERT	69.57	85.93	91.15	-
IRRA (Jiang and Ye, 2023)	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13
PBSL (Shen et al., 2023)	RN50	BERT	65.32	83.81	89.26	-
BEAT(Ma et al., 2023)	RN101	BERT	65.61	83.45	89.54	-
LCR^{2}S (Yan et al., 2023a)	RN50	TextCNN	67.36	84.19	89.62	59.24
APTM (Yang et al., 2023)	Swin-Transformer	Bert	76.53	90.04	94.15	66.91
UniPT (Shao et al., 2023)	CLIP-ViT	CLIP-Xformer	68.50	84.67	-	-
RaSa (Bai et al., 2023)	ALBEFF	ALBEFF	76.51	90.29	94.25	69.38
TBPS (Cao et al., 2024)	CLIP-ViT	CLIP-Xformer	73.54	88.19	92.35	65.38
DP (Song et al., 2024)	CLIP-ViT	CLIP-Xformer	75.66	90.59	94.07	66.58
IRRA+IRLT (Liu et al., 2024)	CLIP-ViT	CLIP-Xformer	74.46	90.19	94.01	-
UASA (Zhao et al., 2024)	CLIP-ViT	CLIP-Xformer	74.25	89.83	93.58	66.15
CFAM (Zuo et al., 2024b)	CLIP-ViT	CLIP-Xformer	75.60	90.53	-	67.27
MGRL (Lv et al., 2024)	CLIP-ViT	CLIP-Xformer	73.91	90.68	-	67.28
DM-Adapter (Liu et al., 2025)	CLIP-ViT	CLIP-Xformer	72.17	88.74	92.85	64.33
OCDL (Li et al., 2025)	CLIP-ViT	CLIP-Xformer	75.10	89.43	-	68.18
HAM(1.0M) (Jiang et al., 2025)	CLIP-ViT	CLIP-Xformer	77.71	91.42	94.57	69.68
Baseline (RDE)	C-ViT	C-Xfmr	75.94	90.14	94.12	67.56
Ours: RAG-ReID	**Reranker** (M_R)	**Generator** (M_G)	**R-1**	**R-5**	**R-10**	**mAP**
RAG-ReID	Qwen3-VL-8B+SFT	Qwen3-VL-8B	78.91	91.10	**95.19**	**70.68**
RAG-ReID	Qwen3-VL-8B	Qwen3-VL-8B+DPO	77.81	91.24	94.79	69.46
RAG-ReID-full	Qwen3-VL-8B+SFT	Qwen3-VL-8B+DPO	**79.00**	**91.31**	94.82	**70.68**

4.2 Implementation Details

We use RDE (Qin et al., 2024) with CLIP-ViT-B/16 (Radford et al., 2021) as the base retrieval model, extending CLIP's token limit from 77 to 248 by interpolating positional embeddings (learning rate 5×10^{-5}). For the MLLM reranker, Qwen2.5-VL-7B-Instruct (Wang et al., 2024) is fine-tuned via LoRA (Hu et al., 2021) for binary relevance filtering (see Section 3.2.1), using 2 epochs, batch size 4 per device, gradient accumulation steps 16, learning rate 5×10^{-5}, LoRA rank $r = 8$, and scaling factor $\alpha = 16$. In the generator stage, we freeze the large Qwen3-VL-8B-Instruct and Qwen3-VL-2B-Instruct models during inference and train only a lightweight autoregressive reward model using DPO (Rafailov et al., 2024) for token-level guidance (see Section 3.3.2). All experiments are performed on 8×NVIDIA A6000 GPUs (40GB each).

4.3 Comparison with State-of-the-Arts

We evaluate RAG-ReID against state-of-the-art (SOTA) methods, including recent strong baselines like RDE (Qin et al., 2024) and CFAM (Zuo et al., 2024b), across three widely used datasets. Results are summarized in Table 1, Table 2, and Table 3.

Results on CUHK-PEDES & Component Analysis. Table 1 demonstrates our method's superiority and the contribution of each module. Compared to the strong RDE baseline (75.94% Rank-1), adding only the **SFT-trained MLLM Reranker** boosts Rank-1 to 78.91%, proving the necessity of context purification. Conversely, applying only the **DPO-guided MLLM Generator** improves Rank-1 to 77.81%. Our full **RAG-ReID** pipeline, combining both components, achieves a new SOTA Rank-1 of **79.00%** (+3.06% over RDE). This confirms that decoupled purification and reward-guided generation are highly complementary.

Results on ICFG-PEDES. As shown in Table 2, this dataset is notoriously challenging due to heavy visual noise and occlusions. Despite this, RAG-ReID achieves **69.79%** Rank-1, outperforming the RDE baseline (67.68%). This highlights the robustness of our framework: the SFT-Reranker successfully filters noisy candidates, allowing the DPO-Generator to synthesize reliable queries even under poor data conditions.

Results on RSTPReid. Table 3 evaluates performance on a smaller, high-quality dataset. Here, RAG-ReID dominates with **73.00%** Rank-1 and

55.30% mAP. This massive **7.65%** Rank-1 improvement over RDE validates the high precision of our closed-loop refinement when initialized with clean visual data.

In summary, RAG-ReID consistently achieves significant performance gains across diverse benchmarks, proving its effectiveness in both highly noisy and high-quality retrieval scenarios.

Table 2: Performance comparisons on the ICFG-PEDES dataset. The best results are in **bold**.

Methods	R-1	R-5	R-10	mAP	mINP
DualPath (Zheng et al., 2020)	38.99	59.44	68.41	-	-
CMPM/C (Zhang and Lu, 2018)	43.51	65.44	74.26	-	-
ViTAA (Wang et al., 2020)	50.98	68.79	75.78	-	-
SSAN (Ding et al., 2021)	54.23	72.63	79.53	-	-
IVT (Shu et al., 2022)	56.04	73.60	80.22	-	-
ISANet (Yan et al., 2022b)	57.73	75.42	81.72	-	-
CFine (Yan et al., 2022a)	60.83	76.55	82.42	-	-
IRRA (Jiang and Ye, 2023)	63.46	80.25	85.82	38.06	7.93
BiLMa (Fujii and Tarashima, 2023)	63.83	80.15	85.74	38.26	-
PBSL (Shen et al., 2023)	57.84	75.46	82.15	-	-
BEAT(Ma et al., 2023)	58.25	75.92	81.96	-	-
LCR^{2}S (Yan et al., 2023a)	57.93	76.08	82.40	38.21	-
DCEL (Li et al., 2023)	64.88	81.34	86.72	-	-
UniPT (Shao et al., 2023)	60.09	76.19	-	-	-
RaSa (Bai et al., 2023)	65.28	80.40	85.12	41.29	-
TBPS (Cao et al., 2024)	65.05	80.34	85.47	39.83	-
CFAM (Zuo et al., 2024b)	65.38	81.17	-	39.42	-
MGRL (Lv et al., 2024)	67.28	63.87	82.34	-	-
OCDL (Li et al., 2025)	64.53	80.23	-	40.76	-
Baseline (RDE)	67.68	82.47	87.36	40.06	7.87
RAG-ReID	**69.79**	**83.30**	**87.94**	**42.60**	**9.29**

Table 3: Performance comparisons on the RSTPReid dataset. The best results are in **bold**.

Methods	R-1	R-5	R-10	mAP	mINP
DSSL (Zhu et al., 2021)	39.05	62.60	73.95	-	-
SSAN (Ding et al., 2021)	43.50	67.80	77.15	-	-
LBUL (Wang et al., 2022b)	45.55	68.20	77.85	-	-
IVT (Shu et al., 2022)	46.70	70.00	78.80	-	-
CFine (Yan et al., 2022a)	50.55	72.50	81.60	-	-
IRRA (Jiang and Ye, 2023)	60.20	81.30	88.20	47.17	25.28
BiLMA (Fujii and Tarashima, 2023)	61.20	81.50	88.80	48.51	-
PBSL (Shen et al., 2023)	47.80	71.40	79.90	-	-
BEAT(Ma et al., 2023)	48.10	73.10	81.30	-	-
LCR^{2}S (Yan et al., 2023a)	54.95	76.65	84.70	40.92	-
DCEL (Li et al., 2023)	61.35	83.95	90.45	-	-
RaSa (Bai et al., 2023)	66.90	86.50	91.35	52.31	-
TBPS (Cao et al., 2024)	61.95	83.55	88.75	48.26	-
CFAM (Zuo et al., 2024b)	62.45	83.55	-	49.50	-
OCDL (Li et al., 2025)	61.60	82.35	-	49.77	-
Baseline (RDE)	65.35	83.95	89.90	50.88	28.08
RAG-ReID	**73.00**	**87.50**	**92.00**	**55.30**	**31.24**

4.4 Ablation Study

To dissect the individual contributions of each component within RAG-ReID, we conduct comprehensive ablation experiments on the CUHK-PEDES dataset. As reported in Table 1, upgrading the baseline RDE model to a more powerful MLLM backbone (Qwen3-VL-30B-A3B-Instruct) yields consistent improvements across all metrics, underscoring the necessity of high-capacity reasoning for retrieval augmentation. For the query augmentation stage, integrating DPO-guided reward modeling (Qwen3-VL-8B + DPO) significantly enhances retrieval precision by providing fine-grained, token-level steering during decoding. Simultaneously, applying supervised fine-tuning (SFT) to the MLLM Reranker (Qwen3-VL-8B + SFT) strengthens domain alignment for binary relevance, leading to superior Rank-1 and mAP scores through effective context purification. Notably, the full configuration combining SFT-based reranking with DPO-guided generation achieves the state-of-the-art performance, confirming that context purification and discriminative query synthesis are highly complementary within our hierarchical pipeline.

5 Conclusion

We have introduced RAG-ReID, a new framework for text-based person re-identification that rethinks the conventional single-shot, single-anchor paradigm. Unlike prior approaches that assume a single, static query or anchor image is sufficient, our method treats TB-ReID as a closed-loop process—actively leveraging gallery feedback to refine ambiguous or under-specified queries in a principled, multi-stage manner. Our hierarchical pipeline decouples context purification and query synthesis into two dedicated modules: an MLLM Reranker that filters identity-irrelevant images to derive a clean, high-confidence Top-N context, and an MLLM Generator that performs contextual consensus to synthesize a discriminative, augmented query for a second, more precise retrieval round. By further introducing a lightweight preference-aligned reward model, we ensure that our augmented queries are not only informative but also factually reliable, eliminating hallucination and redundancy in a plug-and-play manner. Extensive experiments validate that RAG-ReID robustly improves both ranking accuracy and generalization over competitive baselines, especially under real-world ambiguities where user queries are incomplete or vague. This closed-loop, decoupled paradigm demonstrates a clear path forward for practical, robust TB-ReID in complex and open-world settings—and we anticipate its ideas can inspire broader advances in retrieval-augmented generative modeling.

Limitations

While RAG-ReID achieves state-of-the-art performance by transforming text-based person re-identification into a robust closed-loop process, it presents a few limitations that offer avenues for future research. First, although our token-level reward model is lightweight , the reliance on large multimodal models for both the Reranker and Generator stages inherently introduces higher computational costs and inference latency compared to traditional single-shot retrieval paradigms. This could restrict its immediate deployment in strict real-time surveillance scenarios. Second, our decoupled pipeline heavily depends on the initial Top-K retrieval set to purify a visual context; if the base retriever completely fails to fetch any true positive images initially due to extreme query ambiguity, the framework lacks the necessary visual cues to synthesize an accurate augmented query. Finally, while we expanded the CLIP token limit to 248 via positional embedding interpolation to accommodate rich, fine-grained attributes, exceptionally complex generated queries could theoretically still exceed this capacity, potentially diluting attention weights within the frozen dual-encoder architecture.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. *arXiv preprint arXiv:2504.01916*.

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023. Rasa: relation and sensitivity aware representation learning for text-based person search. In *IJCAI*, pages 555–563.

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

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 465–473.

Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. 2022. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

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

Takuro Fujii and Shuhei Tarashima. 2023. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2786–2790.

Hiren Galiyawala and Mehul S Raval. 2021. Person retrieval in surveillance using textual query: a review. *Multimedia Tools and Applications*, 80(18):27343–27383.

Daming Gao, Yang Bai, Min Cao, Hao Dou, Mang Ye, and Min Zhang. 2024. Semi-supervised text-based person search. *arXiv preprint arXiv:2404.18106*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haiwen Li, Delong Liu, Fei Su, and Zhicheng Zhao. 2025. Object-centric discriminative learning for text-based person retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. 2023. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *CVPR*, pages 1970–1979.

Yating Liu, Zimo Liu, Xiangyuan Lan, Wenming Yang, Yaowei Li, and Qingmin Liao. 2025. Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5703–5711.

Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14052–14060.

Tianle Lv, Shuang Li, Jiaxu Leng, and Xinbo Gao. 2024. Mgrl: Mutual-guidance representation learning for text-to-image person retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2895–2899. IEEE.

Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. 2023. Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4157–4168.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. 2025. Human-centered interactive learning via mllms for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14390–14399.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184.

Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. 2023. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931.

Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer.

Zifan Song, Guosheng Hu, and Cairong Zhao. 2024. Diverse person: Customize your own dataset for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4943–4951.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. 2025. Demo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 402–420. Springer.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022a. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022b. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1984–1992.

Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. 2021. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633.

Zequn Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025. Chat-driven text generation and interaction for person retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5259–5270.

Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. 2023a. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2022a. Clip-driven fine-grained text-image person re-identification. *arXiv:2210.10276*.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2023b. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*.

Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022b. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365*.

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

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. 2024b. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17117–17126.

Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701.

Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. 2024. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7534–7542.

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

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217.

Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. 2024a. Cross-video identity correlating for person re-identification pre-training. *arXiv preprint arXiv:2409.18569*.

Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. 2023. Plip: Language-image pre-training for person representation learning. *Preprint*, arXiv:2305.08386.

Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. 2024b. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019.