Iterative Self-Correction for Text-Driven Person Re-Identification with Large Vision-Language Models

Guijin Luo[1,†], **Zequn Xie**[1,†], **Sihang Cai**[1], **Chuxin Wang**[1], **Zhou Zhao**[1], **Yixuan Tang**[2,*]
[1]Zhejiang University [2]National University of Singapore

Abstract

Person Re-Identification (ReID) has long struggled with the semantic gap between low-level visual features and high-level identity concepts. While Vision-Language Models (VLMs) offer promising semantic understanding, existing methods typically adopt a static "one-pass" paradigm, converting images to text once for retrieval. This approach suffers from two critical flaws: Information Bottleneck, where converting rich visuals into text causes detail loss, and Open-Loop Failure, where initial hallucinations propagate without recourse. To address this, we propose Auto-ReID, a novel framework that reformulates ReID as an iterative "Think-and-Refine" process. We first introduce a Hierarchical Progressive Tuning strategy to transform a generic VLM into a specialized Re-ID expert. During inference, we deploy a closed-loop architecture comprising a Reasoner for structured attribute extraction, a Hybrid Retriever that anchors dynamic semantic queries with stable visual features to prevent drift, and a Corrector that deconstructs and verifies candidates to iteratively optimize the search. Extensive experiments on ReID datasets demonstrate that our method significantly outperforms state-of-the-art approaches, particularly in complex occlusion scenarios. Code can be found at: https://github.com/GridNexus/Auto-ReID.

1 Introduction

Person ReIDentification (ReID) aims to match the same individual across non-overlapping cameras and is central to large-scale surveillance and urban safety systems (Yuan et al., 2020; Fu et al., 2022; Niu et al., 2025a). The dominant paradigm in ReID is to learn discriminative visual embeddings using CNNs or Vision Transformers (ViTs), and perform retrieval via nearest-neighbor search (Filax and Ortmeier, 2021; Fu et al., 2025). Despite remarkable progress, purely vision-based pipelines still face persistent failure modes in the wild, including severe occlusion, background clutter, viewpoint change, and hard negatives (e.g., two pedestrians wearing nearly identical outfits but differing in subtle identity cues such as hair style, shoes, or carried items).

Recently, Vision-Language Models (VLMs) (Wang et al., 2024a; Yu et al., 2025a,b)have demonstrated strong multi-modal understanding and instruction following (He et al., 2024, 2025), inspiring a new direction for ReID: converting the query image into a textual description and performing text-guided retrieval. Representative works such as ChatReID (Niu et al., 2025b) show that a VLM can be adapted to pedestrian domains and achieve competitive results. However, we argue that existing VLM-based ReID methods remain largely *open-loop*: they generate a description or representation once and proceed to retrieval without any mechanism to revise the query when the initial interpretation is wrong.

This open-loop design is fundamentally misaligned with how humans search for a person. In practice, human analysts perform an iterative verification process: they form an initial hypothesis (e.g., “a man in a black jacket”), inspect top candidates, notice contradictions (e.g., “the target has a backpack but retrieved candidates do not”), and then refine the search criteria. This behavior resembles the well-known dual-process theory: fast, intuitive *System-1* perception followed by slow, analytical *System-2* correction. In ReID, System-1 corresponds to a single-pass captioning or embedding extraction, while System-2 corresponds to verifying retrieved candidates and updating the search intent.

However, enabling such a closed-loop process is non-trivial for three reasons. First, image-to-text conversion inevitably introduces an information bottleneck: fine details critical to identity (logos, textures, shoe shapes) may be omitted. Second,

[†]Equal contribution. [*]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 6292–6301

VLMs can hallucinate attributes under challenging conditions (illumination, occlusion), leading to semantic drift if the system over-trusts generated text. Third, ReID is an open-set fine-grained task: the notion of "similarity" depends on the query context, and the system must adaptively emphasize identity-preserving cues rather than superficial clothing cues.

To address these challenges, we propose Auto-ReID, a closed-loop framework that reformulates ReID as an iterative think-and-refine process. We further propose a Hierarchical Progressive Tuning (HPT) strategy too empower a general-purpose VLM with the specialized capabilities required by these modules. HPT gradually adapts the VLM through a curriculum of tasks, from fine-grained attribute recognition to pairwise verification and corrective instruction generation, transforming it into a ReID expert.

In summary, our contributions are threefold:

- We propose **Auto-ReID**, the first fully autonomous closed-loop framework for VLM-based person ReID that performs iterative self-correction without human intervention.
- We introduce a novel **Hybrid Retriever** with a visual anchor, which stabilizes retrieval by combining dynamic semantic queries with static visual features, effectively preventing semantic drift.
- We design a **Hierarchical Progressive Tuning** strategy to adapt general VLMs into ReID specialists capable of fine-grained attribute reasoning and corrective feedback generation.

2 Related Work

2.1 Traditional Person ReID

Person ReIDentification (ReID) is a fundamental task in computer vision, which aims to match the same individual across different camera views based on visual features. Recent studies in person ReID carefully designed settings and developed models to tackle every specific scenario. Standard person ReID (Zheng et al., 2017; Tan et al., 2021; Ning et al., 2020; Yuan et al., 2020), which aims to match individuals across cameras based on visual features. These methods distinguish pedestrian identities based on body posture and appearance. Cloth-changing ReID (CC ReID) (Qian et al., 2020a; Bansal et al., 2022; Hong et al., 2021; Guo et al., 2023) is a more challenging variant where individuals change their clothing between camera views. It assists the model in extracting non-clothing information for identity determination. CSSC (Wang et al., 2024b) introduces a framework that leverages abundant semantics within pedestrian images to extract identity features. However, different settings within person ReID focus on distinct visual features, making it difficult to effectively integrate these settings into a single model. Consequently, we intend to develop a versatile 'one-for-all' framework to interactively ask the machine to help with the person retrieval task.

2.2 VLM-driven Person ReID

Vision-language models (VLMs) (Yang et al., 2023; Liu et al., 2024) have shown strong multimodal understanding, inspiring their use in person ReID. Recent works explore two main directions: text-to-image ReID, where methods like MLLM-ReID (Yang and Zhang, 2024) and CHAT (Xie et al., 2025) generate textual descriptions for retrieval, and instruction-based ReID, where Instruct-ReID (He et al., 2024) unifies multiple settings via task-specific instructions. However, these approaches operate in a *static, one-pass* manner, lacking mechanisms to correct initial errors. ChatReID (Niu et al., 2025b) introduces interactivity, allowing human feedback to refine results, but this *human-in-the-loop* design limits scalability. A common limitation across all existing VLM-based ReID methods is the absence of an *autonomous, closed-loop correction* mechanism. They cannot self-verify and iteratively refine retrievals without external input, leaving them vulnerable to information loss and error propagation. In contrast, we propose Auto-ReID, the first framework to implement a fully autonomous iterative self-correction loop for VLM-based ReID, closing this critical gap.

3 Methodology

The core philosophy of Auto-ReID is to transition from a static "encode-and-match" paradigm to a dynamic "reason-and-refine" loop. Our framework, depicted in Figure 1, is architected as a closed-loop system where the output of one cycle informs and improves the next. We first detail the specialization of the VLM through Hierarchical Progressive Tuning (3.1), then describe the iterative inference pipeline (3.2).

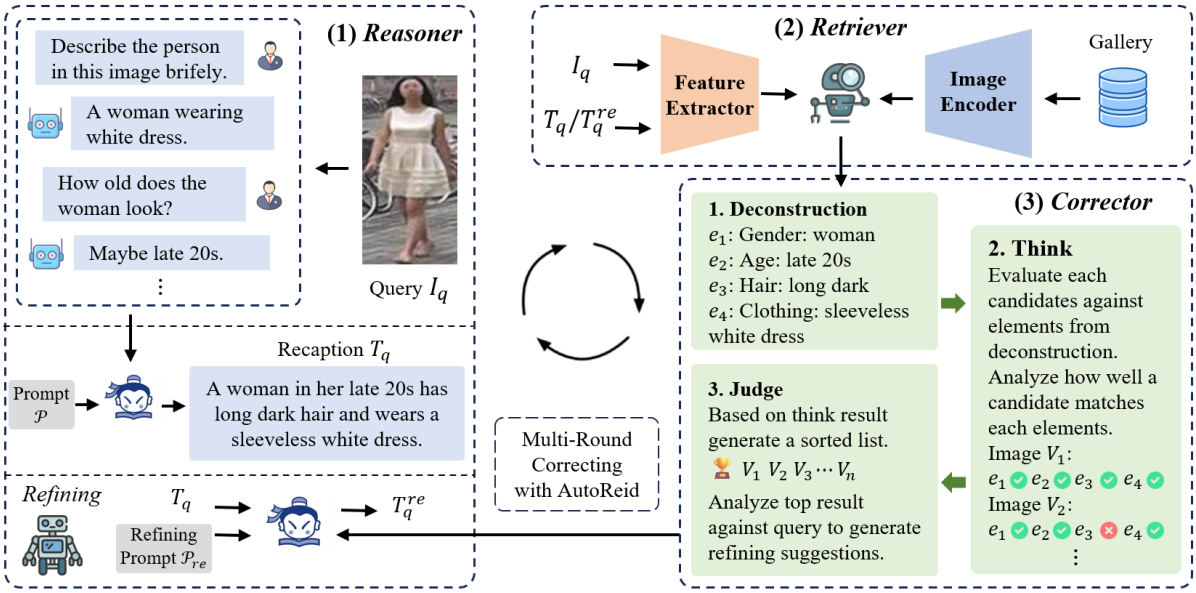

Figure 1: **Illustration of our iterative self-correcting pipeline.** Unlike traditional one-pass methods, our framework introduces a feedback loop. The *Reasoner* first perceives the query image I_q. The *Retriever* then fetches potential matches based on hybrid visual-semantic cues. Crucially, the *Corrector* scrutinizes these results by deconstructing attributes $(e_1, e_2, \ldots)$ and verifying them against the retrieved images $(V_1, V_2, \ldots)$. If mismatches are detected (e.g., wrong clothing color), the Corrector feeds negative constraints back to the Reasoner to refine the search query T_q^{re}, progressively narrowing down the search space.

3.1 Hierarchical Progressive Tuning of the VLM

A general-purpose VLM lacks the precise attribute sensitivity and verification logic required for fine-grained ReID. To bridge this gap, we propose a two-stage tuning strategy that incrementally builds ReID expertise. We keep the base VLM weights frozen and apply parameter-efficient Low-Rank Adaptation (LoRA) to the attention layers.

Stage 1: Fine-Grained Attribute Alignment. The objective is to align the VLM's visual encoder and text decoder with the granular semantics of pedestrian appearance. We construct the data exclusively from the official training splits of Market-1501 and MSMT17, and generate structured descriptions only for training images. Instead of free-form captions, we use a fixed prompt template that elicits a comprehensive, attribute-oriented analysis:

> *Describe the person in the image. Include: gender, estimated age group, hairstyle (length/color), upper clothing (type/color/pattern), lower clothing (type/color), footwear, and any carried items or notable accessories.*

Training the VLM to generate such structured descriptions forces it to attend to and verbally articulate details that are often critical for identity discrimination, thereby reducing coarse-grained hallucinations.

Stage 2: Identity Verification and Feedback Generation. This stage teaches the VLM the logical operations required by the Corrector module. As illustrate in figure 2, we formulate multi auxiliary tasks presented to the model via multi-turn instruction tuning:

1. **Pairwise Verification:** Given two pedestrian images, the model must predict "Yes" or "No" to the question "Are these the same person?" and provide a concise reasoning based on attribute consistency or discrepancy.

2. **Attribute-Specific Q&A:** Given an image and a specific attribute question (e.g., "Is the person carrying a backpack?"), the model learns to answer accurately. This directly supports the attribute verification step in the Corrector.

3. **Corrective Feedback Synthesis:** This is the key task for enabling the refinement loop. The model is presented with a query image, its current textual description, and a set of hard negative images that are visually similar but belong to different identities. The model is

Training Tasks	Input Prompt	Response
Attribute Annotations Matching	 *<Image> Extract and list the visible attributes (e.g., gender, clothing, accessories) for the pedestrian in image.*	*The attributes are:*
Attribute Difference Mining	 *<Image 1><Image 2> Compare these two pedestrians. List their shared features and identify what makes each appearance distinct.*	*Shared: Unique to A: Unique to B:*
Image to image matching	 *<Image 1><Image 2> Verify if the person depicted in these two images is the same individual.*	*Yes/No*
Image to images retrieval	 *<Query Image> Find all occurrences of this specific pedestrian within the provided gallery:<Image>*	*Images <X> and <Y> cont ain the target pedestrian.*
Image to texts retrieval	 *<Image> Identify the most accurate description for this person from the options below:<Caption 1> <Caption 2> <Caption 3>*	*Option 2 describes the image best.*
Text to image matching	 *<Image> Assess whether this image corresponds to the description: <Caption>.*	*Yes/No*
Text to images retrieval	 *From the candidate images <Image>, pick the one that aligns with this description: <Caption>.*	*Image <X> matches the description.*

Figure 2: Multi-task Design Diagram for joint training in the stage 2. Concretely, we conduct seven distinct matching and retrieval tasks between text and image modalities, encouraging VLMS to acquire an initial capability for fine-grained image retrieval based on images, textual descriptions, and pedestrian attributes.

trained to generate a natural language instruction that would help retrieve the true match, such as "*The target has a red logo on the left chest; exclude candidates without this detail*" or "*Focus on the distinctive white stripes on the shoes*".

This progressive curriculum transforms the VLM into a specialized agent capable of detailed perception, logical verification, and actionable feedback generation—the three pillars of the Auto-ReID loop.

3.2 Iterative Self-Correction Inference Loop

Formally, given a query image I_q and a gallery $\mathcal{G} = \{I_g^i\}_{i=1}^N$, Auto-ReID iteratively refines a textual query $T_q^{(t)}$ over $t = 0, \ldots, T_{max}$ steps to improve ranking. The visual representation of I_q, denoted $\mathbf{v}_q = f_{vis}(I_q)$, remains fixed as an anchor. The loop consists of three interconnected modules.

3.2.1 Reasoner: Structured Semantic Perception

At iteration t, the Reasoner's role is to generate or refine the textual query $T_q^{(t)}$. For $t = 0$, it performs an initial analysis of I_q using the structured prompt from Stage 1 tuning:

$$T_q^{(0)} = \mathcal{M}_{VLM}(I_q, \mathcal{P}_{struct}), \tag{1}$$

where $\mathcal{M}_{VLM}$ is our tuned model. For $t > 0$, it integrates the feedback $\mathcal{F}^{(t-1)}$ from the previous Corrector step:

$$T_q^{(t)} = \mathcal{M}_{VLM}(I_q, T_q^{(t-1)}, \mathcal{F}^{(t-1)}). \tag{2}$$

This generates a refined description, e.g., evolving from "a man in a black jacket" to "a man in a black jacket *with a white logo on the sleeve and no backpack*".

3.2.2 Hybrid Retriever: Anchored Semantic Search

To retrieve candidates, we employ a dual-encoder scheme that balances semantic guidance with visual fidelity. A text encoder f_{txt} (from CLIP) embeds the textual query $T_q^{(t)}$ into $\mathbf{h}_q^{(t)}$. The similarity between the query and a gallery image I_g is a convex combination of visual and textual similarities:

$$S^{(t)}(I_q, I_g) = \alpha \cdot \frac{\mathbf{v}_q \cdot \mathbf{v}_g}{\|\mathbf{v}_q\|\|\mathbf{v}_g\|} + (1-\alpha) \cdot \frac{\mathbf{h}_q^{(t)} \cdot \mathbf{h}_g}{\|\mathbf{h}_q^{(t)}\|\|\mathbf{h}_g\|}, \tag{3}$$

where $\mathbf{v}_g = f_{vis}(I_g)$, $\mathbf{h}_g = f_{txt}(\phi(I_g))$, and $\phi(I_g)$ is a generic caption of I_g. The hyperparameter $\alpha \in (0, 1)$ controls the trust between modalities. The fixed visual anchor ($\mathbf{v}_q$) ensures the search remains grounded to the original appearance, preventing catastrophic drift due to erroneous text, while the dynamic text term ($\mathbf{h}_q^{(t)}$) allows the search focus to be iteratively refined. The top-K candidates, $\mathcal{C}^{(t)} = \{I_{c1}, ..., I_{cK}\}$, are passed to the Corrector.

3.2.3 Corrector: Deconstruction, Verification and Feedback

The Corrector mimics the "System 2" slow-thinking process. It takes the candidate set $\mathcal{C}^{(t)}$ and scrutinizes it to generate a refined query $T_q^{(t+1)}$. This module executes three logical steps:

Algorithm 1 Auto-ReID inference process.

Input: Query image I_q, gallery $\mathcal{G}$, max iterations $T_{\max}$, candidate size K, consistency threshold τ_{low}.

Parameter: Tuned VLM $\mathcal{M}_{ReID}$, visual encoder f_{vis}, text encoder f_{txt}, mixing coefficient α.

$\mathbf{v}_q \leftarrow f_{vis}(I_q)$ {Visual anchor}
$T_q \leftarrow \mathcal{M}_{ReID}(I_q, \mathcal{P}_{struct})$ {Initial description}
$\mathcal{C}_{prev} \leftarrow \emptyset$
for $t = 0$ to $T_{\max} - 1$ **do**
  // Hybrid retrieval
  for each $I_g \in \mathcal{G}$ **do**
    $S \leftarrow \alpha \cdot sim(\mathbf{v}_q, f_{vis}(I_g)) + (1 - \alpha) \cdot sim(f_{txt}(T_q), f_{txt}(\phi(I_g)))$
  end for
  $\mathcal{C} \leftarrow top{-}\text{K}by\text{S}$
  if $t = T_{\max} - 1$ **then**
    return $\mathcal{C}$
  end if
  // Correction
  $\mathcal{A} \leftarrow Parse(T_q)$
  $\mathcal{F} \leftarrow Verify(\mathcal{C}, \mathcal{A}, \tau_{low})$
  if $\mathcal{F} = \emptyset$ **and** $IoU(\mathcal{C}, \mathcal{C}_{prev}) > 0.9$ **then**
    return $\mathcal{C}$ {Early termination}
  end if
  // Reasoning
  $T_q \leftarrow \mathcal{M}_{ReID}(I_q, T_q, \mathcal{F})$
  $\mathcal{C}_{prev} \leftarrow \mathcal{C}$
end for
return $\mathcal{C}$

Semantic Deconstruction. To perform precise verification, the monolithic query $T_q^{(t)}$ is parsed into atomic attribute-value pairs $\mathcal{A}^{(t)} = \{(k_j, v_j)\}_{j=1}^{M}$. Example: $\mathcal{A}^{(t)} = \{Gender : Female, Hair : Long, Bag : None\}$.

Attribute Consistency Verification. The Corrector evaluates the consistency of the top-K candidates against these attributes. Using the discrimination capability learned in Stage 2, the VLM estimates the probability $P(v|I_{cj}, k)$ that candidate I_{cj} matches attribute value v. We calculate the **Attribute Consistency Score (ACS)** for the retrieved set:

$$ACS(k, v) = \frac{1}{K} \sum_{j=1}^{K} P(v|I_{cj}, k) \tag{4}$$

A low $ACS(k, v)$ indicates a conflict: the retrieved results diverge from the query description regarding attribute k. For instance, if the query states "No Bag" but the top results mostly show "Backpacks," a conflict is flagged.

Feedback Generation. Upon detecting a conflict, the Corrector generates a feedback instruction $\mathcal{I}_{feedback}$.

- **Negative Constraints:** If distractors share a common feature (e.g., backpacks), the instruction is: *"Exclude candidates with backpacks."*
- **Attribute Emphasis:** If a key visual trait (e.g., white dress) is missed, the instruction is: *"Prioritize candidates wearing white dresses."*

The refined text description is then generated by the Reasoner:

$$T_q^{(t+1)} = LLM(T_q^{(t)}, \mathcal{I}_{feedback}, I_q) \tag{5}$$

This updated prompt $T_q^{(t+1)}$ is fed back into the Hybrid Retriever for iteration $t + 1$. The loop terminates when $t = T_{max}$ (set to 3) or when the candidate set stabilizes (intersection over union of top-K results > threshold).

Table 1: Statistics of datasets used in the paper.

Dataset	Image	ID	Cam&View
MSMT17	126,441	4,101	15
Market-1501	32,668	1,501	6
CUHK03	13,164	1,467	2
Occluded-Duke	35,489	1,404	8
PRCC	33,698	221	-
LTCC	17,119	152	-

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the reasoning and robustness of Auto-ReID, we conduct experiments on both standard and challenging settings: (1) Standard ReID: Market-1501 (Zheng et al., 2015), MSMT17 (Wei et al., 2018) and CUHK03 (Li et al., 2014). (2) Occluded ReID: Occluded-Duke (Miao et al., 2019), where pedestrians are partially visible and global appearance cues are unreliable. (3) Cloth-changing ReID: PRCC (Yang et al., 2020) and LTCC (Qian et al., 2020b). Table 1 shows details of above datasets.

Evaluation Metrics. We follow the standard evaluation protocol in person ReID and report Rank-1 accuracy and mean Average Precision (mAP).

Table 2: Comparison with SOTA methods on standard(Market1501, MSMT17, CUHK03) and occluded(Occluded-Duke) person ReID datasets. **Bold** indicates the best performance. Our Auto-ReID consistently outperforms existing approaches, showing significant gains in occlusion scenarios where reasoning is crucial.

Methods	Market-1501		MSMT17		CUHK03		Occluded-Duke	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SAN (Jin et al., 2020)	88.0	96.1	-	-	76.4	80.1	-	-
TransReID (He et al., 2021)	86.8	94.4	61.0	81.8	-	-	55.6	64.2
PASS (Zhu et al., 2022)	93.0	96.8	71.8	88.2	-	-	57.1	65.8
HumanBench (Tang et al., 2023)	89.5	-	69.1	-	77.7	-	59.3	68.0
CLIP-ReID (Li et al., 2023)	90.5	95.4	75.8	89.7	-	-	60.3	67.2
PHA (Zhang et al., 2023)	90.2	96.1	68.9	86.1	83.0	84.5	66.5	70.4
IRM (He et al., 2024)	93.5	96.5	72.4	86.9	85.4	86.5	-	-
ChatReID (Niu et al., 2025b)	96.4	97.2	87.5	90.1	91.3	92.7	-	-
Auto-ReID (Ours)	**97.1**	**97.8**	**89.2**	**91.8**	**92.4**	**93.1**	**72.4**	**79.5**

Table 3: Comparison of Auto-ReID with SOTA methods across two cloth-changing ReID (CC ReID) datasets.

METHODS	LTCC		PRCC	
	mAP	Rank-1	mAP	Rank-1
HACNN (Li et al., 2018)	26.7	60.2	-	21.8
RGA-SC (Zhang et al., 2020)	27.5	65.0	-	42.3
PCB (Sun et al., 2018)	30.6	65.1	38.7	41.8
IANet (Hou et al., 2019)	31.0	63.7	45.9	46.3
CAL (Gu et al., 2022)	40.8	74.2	-	-
TransReID (He et al., 2021)	-	-	44.2	-
IRM (He et al., 2024)	52.0	75.8	52.3	54.2
Auto-ReID(Ours)	**66.3**	**79.7**	**65.6**	**73.2**

Rank-1 reflects the probability that the correct identity is retrieved at the first position, while mAP evaluates the overall ranking quality and is more sensitive to hard negatives and long-tail retrieval behavior. Since our method aims to refine top-ranked candidates through iterative verification, improvements in mAP are especially indicative of better re-ranking quality.

Implementation Details. We implement Auto-ReID using PyTorch on 4 NVIDIA A100 GPUs. We utilize InternVL3.5-8B (Wang et al., 2025) as the VLM backbone for the Reasoner and Corrector. The visual encoder and text encoder are initialized with Siglip2 (Tschannen et al., 2025)(siglip2-base-patch16-224). For HPT, we apply LoRA (rank=16) to the VLM's attention modules, training for 3 epochs with a learning rate of 1e-5. Unless specified, we keep the VLM backbone frozen except the LoRA adapters to avoid overfitting and to keep training efficient. During inference, we set $\lambda_1 = 0.65$ and $\lambda_2 = 0.35$ to prioritize the visual anchor while allowing semantic guidance. We set $T_{\max} = 3$ and $K = 20$. Images are resized to 256×128.

Evaluation Strategy for VLM-based Iterative ReID. Directly evaluating VLMs in a naive "query + full gallery" manner is infeasible due to the context-length constraint and the computational cost of repeatedly prompting the VLM with thousands of images. We adopt an efficient two-stage evaluation strategy aligned with our framework: (1) Visual pre-filtering: we compute the similarity between the query and all gallery images using E_{vis} and keep a shortlist (top-N or those above a threshold). This step is fast and ensures that the candidate set retains high recall. (2) Closed-loop refinement on Top-K: our Corrector only inspects the Top-K candidates returned by the Hybrid Retriever. This "lazy check" design makes the per-query cost independent of the gallery size, enabling scalable evaluation while preserving the benefits of iterative reasoning.

4.2 Comparison with State-of-the-Art Methods

We compare Auto-ReID against a comprehensive set of state-of-the-art methods, including both traditional vision-based approaches and recent VLM-based techniques. Results are presented across multiple benchmarks in Tables 2 and 3.

4.2.1 Performance on Standard Benchmarks

Table 2 shows that Auto-ReID achieves state-of-the-art performance on all standard benchmarks. On the challenging MSMT17 dataset, Auto-ReID attains 89.2% mAP and 91.8% Rank-1, surpassing the previous best VLM-based method (ChatReID) by +1.7% mAP and +1.7% Rank-1. This improvement demonstrates the effectiveness of our iterative refinement approach in handling large-scale,

Table 4: Ablation study on MSMT17 demonstrating the contribution of each component. VA: Visual Anchor, HPT: Hierarchical Progressive Tuning, Loop: Iterative correction.

Settings	VA	Init Text	HPT	Loop	mAP / Rank-1
1. Baseline (Visual Only)	✓	-	-	-	72.4 / 86.9
2. Text-Only Loop	-	✓	✓	✓	66.5 / 79.2
3. Hybrid (Static)	✓	✓	✓	-	76.8 / 89.1
4. No-Tuning Loop	✓	✓	-	✓	74.2 / 87.5
5. **Full Model**	✓	✓	✓	✓	**89.2 / 91.8**

diverse environments where single-pass methods may misinterpret subtle identity cues.

On Market-1501, Auto-ReID achieves 97.1% mAP and 97.8% Rank-1, setting a new state-of-the-art despite the dataset's near-saturation. Moreover, on CUHK03, our method outperforms existing approaches with 92.4% mAP and 93.1% Rank-1. The consistent gains across all standard benchmarks validate that our closed-loop reasoning framework provides robust improvements beyond what can be achieved with static representations.

4.2.2 Performance on Occluded Scenarios

The advantages of iterative reasoning become particularly evident in occlusion scenarios. As shown in Table 2, Auto-ReID achieves 72.4% mAP on Occluded-Duke, representing substantial improvements of +12.1% over CLIP-ReID (60.3%) and +5.9% over PHA (66.5%). This significant gain highlights how our Corrector module effectively compensates for missing visual information: by systematically verifying visible attributes against candidate matches and refining the search based on consistent cues, Auto-ReID can overcome the limitations of incomplete visual data that plague traditional methods.

4.2.3 Performance on Cloth-changing Scenarios

Table 3 presents results on cloth-changing ReID benchmarks, where individuals change clothing between camera views. Auto-ReID demonstrates strong performance in this challenging setting, achieving 66.3% mAP on LTCC and 65.6% mAP on PRCC. These results represent improvements of +14.3% and +13.3% over the previous best method (IRM) on LTCC and PRCC, respectively.

The success in cloth-changing scenarios underscores a key advantage of our semantic reasoning approach: by decomposing identity into atomic attributes and focusing verification on non-clothing characteristics (e.g., facial features, body shape, hairstyle), Auto-ReID maintains robustness against appearance changes that would confuse purely visual matching methods.

4.3 Ablation Study

To validate the contribution of each component in Auto-ReID, we conduct comprehensive ablation studies on MSMT17, with results summarized in Table 4.

Visual Anchor: Comparing rows 1 and 2 reveals the critical importance of the visual anchor. When relying solely on textual descriptions (Text-Only Loop), performance drops to 66.5% mAP due to semantic drift—errors in the initial text generation propagate without correction. The visual anchor stabilizes retrieval by grounding it in robust visual features, as evidenced by the 76.8% mAP achieved by the static hybrid model (row 3).

Hierarchical Progressive Tuning: The necessity of specialized tuning is demonstrated by comparing rows 4 and 5. Without HPT, the generic VLM lacks the fine-grained discrimination capability needed for effective correction, resulting in suboptimal performance (74.2% mAP). Our two-stage tuning strategy teaches the model to recognize subtle pedestrian attributes and generate precise feedback, enabling the full system to achieve 89.2% mAP.

Iterative Correction Loop: The value of iterative reasoning is confirmed by comparing rows 3 and 5. The static hybrid model performs reasonably well at 76.8% mAP, but introducing the correction loop provides an additional +12.4% improvement. This substantial gain demonstrates that the "System 2" slow-thinking process—verifying candidates, identifying conflicts, and refining the search—is crucial for resolving challenging cases where initial retrievals contain hard negatives.

These ablation results confirm that all three components work synergistically: the visual anchor provides stability, HPT enables precise reasoning, and the iterative loop applies this reasoning to con-

tinuously improve retrieval quality.

4.4 Hyperparameter Analysis

4.4.1 Impact of Iteration Depth

We analyze the sensitivity to iteration depth ($T_{\max}$) in Table 5. The static hybrid retrieval ($t = 0$) achieves 85.8% mAP, serving as a strong baseline. The first correction iteration ($t = 1$) provides the largest gain (+2.3% mAP), as it addresses the most obvious semantic mismatches. Subsequent iterations yield diminishing returns, with performance saturating at $t = 3$ (89.2% mAP). Beyond three iterations, no significant improvement is observed, but computational cost increases linearly. Therefore, we select $T_{\max} = 3$ as the optimal trade-off between performance and efficiency.

4.4.2 Modality Balancing

The mixing coefficient α controls the balance between visual and textual modalities in hybrid retrieval. We find that $\alpha = 0.65$ (visual weight) provides optimal results across all datasets. When $\alpha > 0.9$, the system essentially reverts to visual-only retrieval, losing the semantic flexibility needed for refinement. Conversely, when $\alpha < 0.4$, the system becomes overly dependent on potentially noisy text descriptions, leading to semantic drift. The 65:35 visual-to-text ratio allows the visual anchor to stabilize retrieval while leaving sufficient room for semantic guidance to refine results.

4.4.3 Candidate Set Size for Verification

The parameter K (number of candidates verified by the Corrector) balances verification thoroughness against computational cost. We evaluate $K \in \{5, 10, 20, 30, 50\}$ and find that $K = 20$ provides the best trade-off: smaller values ($K \leq 10$) may miss important patterns in the retrieved results, while larger values ($K \geq 30$) increase VLM inference time without improving performance. At $K = 20$, the Corrector captures sufficient context to identify systematic retrieval errors while maintaining reasonable computational overhead.

Table 5: Hyperparameter analysis on MSMT17. For each parameter, the optimal value is **bold**.

Parameter	Value	mAP	Rank-1
Iteration Depth (T)	0	85.8	89.5
	1	88.1	90.8
	2	89.0	91.5
	3	**89.2**	**91.8**
	4	89.2	91.8
Modality Mixing (α)	0.4	85.2	89.0
	0.5	86.3	89.5
	0.6	88.1	90.6
	0.65	**89.2**	**91.8**
	0.7	88.8	91.2
	0.8	87.5	90.5
Candidate Size (K)	5	87.5	90.2
	10	87.5	90.8
	20	**89.2**	**91.8**
	30	89.2	91.8
	50	89.2	91.7

5 Conclusion

This paper presents Auto-ReID, the first autonomous closed-loop framework for iterative self-correction in VLM-based person ReID. Auto-ReID addresses two key limitations of existing methods: the information bottleneck of image-to-text conversion and the error propagation in open-loop retrieval. Our framework integrates a Reasoner for semantic description, a Hybrid Retriever that combines visual and textual cues to prevent drift, and a Corrector that autonomously verifies and refines the search. Coupled with a Hierarchical Progressive Tuning strategy, Auto-ReID achieves consistent improvements across standard, occluded, and cloth-changing benchmarks, with notable gains in challenging occlusion scenarios. This work establishes a new paradigm of autonomous reasoning for person retrieval.

Limitations

Auto-ReID has several limitations. The iterative loop increases computational cost, which may affect real-time deployment. The method relies on a robust visual anchor; severe image degradation can reduce its effectiveness. Additionally, fundamental initial perception errors are difficult to correct within the loop. Future work will focus on efficiency optimization, enhancing anchor robustness, and incorporating error-detection mechanisms. Extending the framework to video Re-ID and leveraging external knowledge are promising directions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U24A20326) and State Grid Zhejiang Electric Power Cooperation Technology Project (No.B311DS240012).

References

Vaibhav Bansal, Gian Luca Foresti, and Niki Martinel. 2022. Cloth-changing person re-identification with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 602–610.

Marco Filax and Frank Ortmeier. 2021. On the influence of viewpoint change for metric learning. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–4. IEEE.

Lihua Fu, Yubin Du, Yu Ding, Dan Wang, Hanxu Jiang, and Haitao Zhang. 2022. Domain adaptive learning with multi-granularity features for unsupervised person re-identification. *Chinese Journal of Electronics*, 31(1):116–128.

Teng Fu, Haiyang Yu, Ke Niu, Bin Li, and Xiangyang Xue. 2025. Foundation model driven appearance extraction for robust multiple object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3031–3039.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. 2022. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069.

Peini Guo, Hong Liu, Jianbing Wu, Guoquan Wang, and Tao Wang. 2023. Semantic-aware consistency network for cloth-changing person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8730–8739.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, and 1 others. 2024. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531.

Weizhen He, Yiheng Deng, Yunfeng Yan, Feng Zhu, Yizhou Wang, Lei Bai, Qingsong Xie, Rui Zhao, Donglian Qi, Wanli Ouyang, and 1 others. 2025. Instruct-reid++: Towards universal purpose instruction-guided person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. 2021. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522.

Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9317–9326.

Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. 2020. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180.

Siyuan Li, Li Sun, and Qingli Li. 2023. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*.

Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551.

Xin Ning, Ke Gong, Weijun Li, Liping Zhang, Xiao Bai, and Shengwei Tian. 2020. Feature refinement and filter network for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3391–3402.

Ke Niu, Haiyang Yu, Xuelin Qian, Teng Fu, Bin Li, and Xiangyang Xue. 2025a. Synthesizing efficient data with diffusion models for person re-identification pre-training. *Machine Learning*, 114(3):1–25.

Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. 2025b. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. *ArXiv*, abs/2502.19958.

Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. 2020a. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*.

Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. 2020b. Long-term cloth-changing person re-identification. *arXiv preprint arXiv:2005.12633*.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496.

Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. 2021. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):160–171.

Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, and 1 others. 2023. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21970–21982.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qizao Wang, Xuelin Qian, Bin Li, Lifeng Chen, Yanwei Fu, and Xiangyang Xue. 2024b. Content and salient semantics collaboration for cloth-changing person re-identification. *arXiv preprint arXiv:2405.16597*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88.

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.

Qize Yang, Ancong Wu, and Wei-Shi Zheng. 2020. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence (DOI 10.1109/TPAMI.2019.2960509)*.

Shan Yang and Yongfei Zhang. 2024. Mllmreid: Multimodal large language model-based person re-identification. *arXiv preprint arXiv:2401.13201*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Haiyang Yu, Jinghui Lu, Yanjie Wang, Yang Li, Han Wang, Can Huang, and Bin Li. 2025a. Eve: Towards end-to-end video subtitle extraction with vision-language models. *arXiv preprint arXiv:2503.04058*.

Haiyang Yu, Siyang Yi, Ke Niu, Minghan Zhuo, and Bin Li. 2025b. Umit: Unifying medical imaging tasks via vision-language models. *arXiv preprint arXiv:2503.15892*.

Ye Yuan, Wuyang Chen, Yang Yang, and Zhangyang Wang. 2020. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *CVPR Workshops*, pages 354–355.

Guiwei Zhang, Zhang Yongfei, Zhang Tianyu, and Shiliang Pu Bo Li1. 2023. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14133–14142.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124.

Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *CVPR*, pages 1367–1376.

Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. 2022. Pass: Part-aware self-supervised pre-training for person re-identification. In *European conference on computer vision*, pages 198–214. Springer.