PAIReID: Progressive Adaptation and Interactive Reflection for Person Re-Identification

Anonymous authors
Paper under double-blind review

Abstract

Interactive person re-identification (Re-ID) aims to retrieve a target pedestrian through multi-round interaction, where follow-up questions progressively enrich an initially incomplete description. Existing methods mainly improve the interaction process through supervised question learning, predefined questioning strategies, or LLM-based reasoning, while retrieval typically relies on a shared cross-modal alignment throughout the dialogue. We identify two overlooked challenges. First, as the query evolves from sparse attributes to increasingly fine-grained compositional cues, a fixed alignment may not be equally suitable for different query states. Second, a plausible follow-up question may still be redundant, visually unsupported, or uninformative at the current interaction stage. To address these issues, we propose PAIReID, which combines **P**rogressive **A**daptation with **I**nteractive **R**eflection. On the retrieval side, PAIReID learns lightweight low-rank alignment anchors from progressive query states and dynamically composes them according to the current query semantics, without using explicit interaction-round identities. On the interaction side, a State-Aware Reflective Questioner tracks appearance coverage and selectively refines unsuitable questions without task-specific fine-tuning. Experiments on Interactive-PEDES show that PAIReID achieves 74.08% Rank-1 after five interaction rounds with a zero-shot Questioner and a BRI of 0.485, while evaluations on RSTPReid, CUHK-PEDES, and ICFG-PEDES further demonstrate its applicability to different Text-ReID Retrievers.

1 Introduction

Text-based person re-identification (Re-ID) aims to retrieve images of a target person from a gallery based on a natural-language description, enabling person retrieval when query images are unavailable Li et al. (2017); Zheng et al. (2020); Jiang & Ye (2023); Qin et al. (2024). In practical scenarios, however, an initial description is often sparse and captures only a limited set of appearance cues Cai et al. (2021); Levy et al. (2023); Bai et al. (2025); Lu et al. (2025); Bai et al. (2026). Targeted follow-up questions can elicit additional visual details that are difficult to provide at once. Interactive person Re-ID therefore extends conventional text-based retrieval with multi-round question answering, progressively incorporating the acquired information into the retrieval query Bai et al. (2025); Qin et al. (2025); Xie et al. (2025); Lu et al. (2025); Bai et al. (2026). As the interaction proceeds, the query evolves from a coarse description to a richer representation that integrates fine-grained appearance cues, allowing the target to be distinguished from visually similar candidates.

This interactive retrieval paradigm has also been explored in general image retrieval, where a system iteratively asks follow-up questions to clarify the user's search intent and refines the textual query with the resulting answers Guo et al. (2018); Cai et al. (2021); Nie et al. (2021); Levy et al. (2023). Interactive person Re-ID adapts this paradigm to identity-oriented retrieval Bai et al. (2025); Qin et al. (2025); Lu et al. (2025); Bai et al. (2026), where the interaction focuses on progressively uncovering appearance cues of a specific person. A representative implementation, LLaVA-ReID Lu et al. (2025), organizes the interaction around three main components: a Retriever, a Questioner, and an Answerer. The Retriever matches the accumulated textual description against the person gallery and provides candidate images; the Questioner uses the current dialogue and visual candidates to generate a follow-up question; and the Answerer simulates the witness by supplying additional information about the target. The resulting answer is incorporated into the description for the next

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.

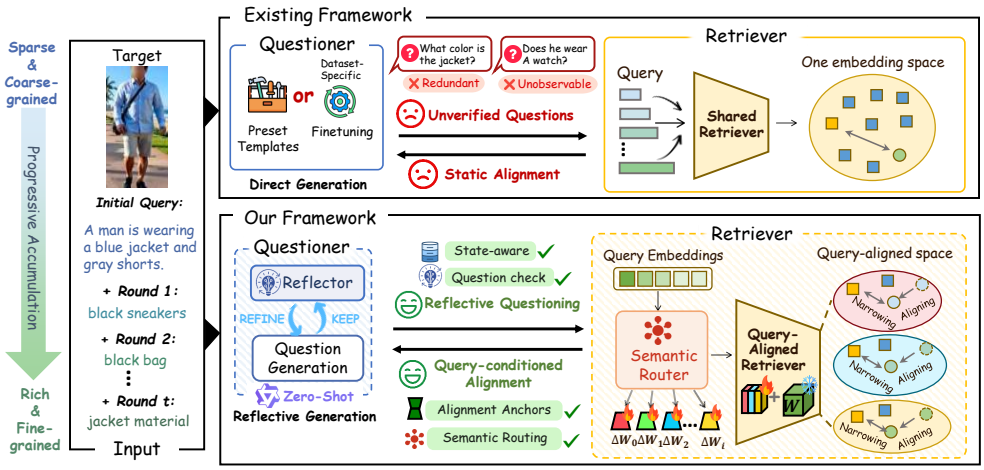

Figure 1: Motivation of PAIReID. Existing interactive person Re-ID methods progressively enrich the query, while retaining fixed retrieval alignment and lacking explicit state-aware question evaluation. PAIReID addresses both through query-conditioned adaptation and reflective questioning.

retrieval round. As this cycle repeats, both the textual evidence available to the Retriever and the information needs faced by the Questioner evolve with the interaction.

Existing interactive person Re-ID methods have explored several ways to make better use of progressively acquired appearance information, including dialogue-refined cross-modal alignment Bai et al. (2025), human-centered attribute interaction Qin et al. (2025), chat-driven textual enrichment Xie et al. (2025), and multi-turn dialogue modeling Bai et al. (2026). As shown in Figure 1, these approaches improve the representation or retention of accumulated cues, but the retrieval function itself is typically shared across interaction rounds Bai et al. (2025); Qin et al. (2025); Xie et al. (2025); Lu et al. (2025). This becomes restrictive as the semantic composition of the query changes over time. An initial description containing only a few attributes mainly supports coarse filtering, whereas later queries combine multiple fine-grained cues that must be considered jointly to distinguish visually similar candidates. For example, clothing color may narrow the initial search space, while subsequently acquired details such as bag type and footwear become important for resolving remaining ambiguity. Previous works typically retain the same cross-modal matching function throughout the interaction. As the query evolves from sparse attributes to richer compositional descriptions, such fixed alignment may become less suited to the changing combination of discriminative cues, limiting the effective use of newly acquired information Bai et al. (2025); Qin et al. (2025); Xie et al. (2025); Lu et al. (2025). This raises the first question: **I) How can cross-modal alignment adapt to the changing composition of appearance cues as the query is progressively enriched?**

On the questioning side, existing methods obtain follow-up information through predefined attribute questions Qin et al. (2025), prompting pretrained language models Levy et al. (2023), or supervised question generation conditioned on dialogue and visual context Lu et al. (2025). Some approaches further filter generated questions to reduce redundancy and improve contextual relevance Lee et al. (2024). However, a question that is plausible in isolation is not necessarily worth asking at the current interaction state. It may repeat an established fact, target a visually indiscernible detail, or concern an attribute that offers little additional information for resolving the remaining ambiguity. Conversely, an attribute mentioned earlier may still be only partially explored: knowing that a person carries a bag, for instance, leaves its color and type unresolved. Determining the utility of a candidate question therefore requires jointly considering what is already known, what remains unresolved, and what can be reliably inferred from the available visual evidence. Otherwise, redundant or low-value questions may waste the limited interaction budget and contribute little useful information for resolving the remaining retrieval ambiguity. This raises the second question: **II) How can we determine whether a generated question is still worth asking at the current interaction state?**

To address these two questions, we propose **PAIReID** (Figure 2), an interactive person Re-ID framework consisting of a **Progressive Query-Aligned Retriever** and a **State-Aware Reflective Ques-**

tioner. The two components address complementary interaction dynamics: the Retriever adapts *how* the evolving query is matched to the gallery, while the Questioner reasons about *what* unresolved information should be acquired next. To accommodate the evolving composition of appearance cues, the Progressive Query-Aligned Retriever learns lightweight low-rank adaptations from queries at different stages of information accumulation and composes them according to the semantic content of the current query. Unlike using a fixed cross-modal alignment function throughout the interaction, this design enables query-dependent reconfiguration of image–text matching while retaining a shared frozen backbone. To determine whether a generated question remains useful at the current interaction state, the State-Aware Reflective Questioner maintains an attribute-level state describing established information and unresolved attribute coverage. Given a candidate question, it jointly considers the state and the available visual evidence to identify potential defects, such as redundancy, insufficient remaining information, or unreliable visual support. A suitable proposal is retained, while for a defective one, the Reflector identifies an alternative unresolved factor and the Generator reformulates the question accordingly. This reflective questioning process requires no task-specific fine-tuning of the questioner. Together, the two components form a progressive interaction loop in which newly acquired evidence updates both the retrieval query and the state used to guide subsequent questioning. Our main contributions are summarized as follows:

- **Problem Insight — Evolving-Query Alignment:** We identify a previously underexplored issue in interactive person Re-ID. As dialogue progressively enriches the query with fine-grained appearance cues, its semantic composition changes, while existing pipelines typically retain a shared cross-modal matching function throughout the interaction, which may limit their ability to effectively exploit newly acquired information.
- **Adaptive Retrieval — Query-Conditioned Alignment Reconfiguration:** We propose a Progressive Query-Aligned Retriever that learns lightweight low-rank adaptations from different stages of information accumulation and composes them according to the semantics of the current query. This allows the Retriever to better exploit newly acquired fine-grained appearance cues under different query compositions, facilitating more discriminative matching among visually similar candidates without relying on explicit interaction-round identities.
- **Reflective Questioning — State-Aware Question Evaluation:** We introduce a State-Aware Reflective Questioner that tracks established and unresolved appearance information and evaluates each generated question using both the interaction state and visual evidence, retaining suitable proposals while selectively revising those with identifiable defects without task-specific fine-tuning.
- **Empirical Validation — Interactive Person Re-ID:** Extensive experiments on Interactive-PEDES and three Text-ReID benchmarks demonstrate the effectiveness and transferability of PAIReID, while ablation studies further validate the proposed retrieval adaptation and reflective questioning.

2 Related Works

2.1 Text-Based Person Re-Identification

Text-based person Re-ID aims to retrieve pedestrian images from a gallery using natural-language descriptions. Early studies mainly learned global cross-modal representations between pedestrian images and textual descriptions Zheng et al. (2020), while subsequent methods introduced attribute-level or multi-granularity modeling to capture finer visual–textual correspondences Wang et al. (2020); Shao et al. (2022). With the development of vision-language pretraining, recent studies increasingly build on pretrained models such as CLIP Radford et al. (2021) and improve identity-sensitive image–text alignment through stronger cross-modal objectives and semantic modeling Jiang & Ye (2023); Qin et al. (2024); Bai et al. (2023a); Cao et al. (2024); Jiang et al. (2025). More recent works further exploit large vision-language models to obtain richer person descriptions or transferable semantic supervision Tan et al. (2024); Bai et al. (2023b). Together, these methods substantially improve cross-modal matching for detailed textual descriptions. However, most existing Text-ReID methods assume a given textual query and perform one-shot retrieval, whereas interactive person Re-ID progressively updates the query as additional appearance information is acquired through multi-round interaction Lu et al. (2025). Our work focuses on cross-modal alignment under this evolving-query setting.

2.2 Interactive Image Retrieval and Person Re-Identification

Interactive image retrieval extends conventional one-shot retrieval by allowing a system to iteratively refine an incomplete query through user interaction. Early approaches explored dialogue-based retrieval and active query enrichment by requesting additional or clarifying information from users Guo et al. (2018); Cai et al. (2021). Recent methods further leverage pretrained language models to generate follow-up questions and incorporate the resulting feedback into subsequent retrieval Levy et al. (2023); Liang & Albanie (2023); Lee et al. (2024).

This interactive paradigm has recently been introduced into person retrieval, where interaction is used to progressively uncover appearance cues that are absent or ambiguous in the initial description. Representative methods explore different interaction strategies, including predefined attribute-oriented questions Qin et al. (2025), dialogue-based query refinement Bai et al. (2025), and supervised multimodal question generation conditioned on dialogue history and retrieved visual candidates Lu et al. (2025). Other studies improve the interaction process by filtering candidate questions to reduce redundancy and improve contextual relevance Lee et al. (2024), or by explicitly modeling accumulated multi-turn dialogue information Bai et al. (2026).

These studies substantially improve how additional appearance information is acquired or maintained during interaction. However, existing methods mainly focus on question generation, dialogue aggregation, or candidate-question filtering, while the evolving *information state* of individual appearance factors remains less explicitly modeled. In particular, an attribute mentioned earlier may still contain unresolved details, whereas a seemingly relevant question can become redundant or visually unsupported as the interaction progresses. Our work therefore introduces state-aware reflection that explicitly tracks the coverage of appearance information and evaluates each proposed question jointly against the current interaction state and retrieved visual context before deciding whether it should be retained or revised.

3 Methodology

Formulation. Interactive person Re-ID aims to retrieve gallery images of a target pedestrian by progressively acquiring appearance information through multi-round dialogue. Let $\mathcal{G} = \{I_j\}_{j=1}^{N}$ denote an image gallery and c_0 the initial textual description of the target. After t interactions, the dialogue history is $\mathcal{D}_t = \{c_0, (q_1, a_1), \ldots, (q_t, a_t)\}$, where q_i and a_i denote the follow-up question and its corresponding answer at interaction i. We construct the accumulated textual query as

$$x_t = \mathrm{Concat}(c_0, a_1, \ldots, a_t),$$

with $x_0 = c_0$. The Retriever ranks gallery images according to their similarity to x_t, and the top-ranked candidates are subsequently used as visual context for generating the next follow-up question. Following the interactive retrieval protocol, target-side information used to simulate witness feedback is accessible only to the Answerer and is never exposed to the Questioner, which observes only the dialogue history and retrieved candidates.

Overview. PAIReID addresses two coupled problems in interactive person Re-ID: how an evolving query should be aligned with the gallery, and what unresolved appearance information should be acquired next. As shown in Figure 2, the Progressive Query-Aligned Retriever dynamically composes lightweight alignment anchors according to the current query semantics, while the State-Aware Reflective Questioner tracks appearance coverage and selectively refines unsuitable follow-up questions. Each new answer updates both the retrieval query and the interaction state, forming a closed progressive retrieval loop.

3.1 Progressive Query-Aligned Retriever

Alignment anchors for evolving queries. As interaction proceeds, the retrieval query changes from an initial sparse description to a composition of progressively acquired appearance cues. We therefore construct six lightweight alignment anchors, one for each progressive query state $k \in \{0, \ldots, 5\}$. Let g and f_0 denote the frozen image and text encoders of the base Retriever. For each state k, we independently optimize a LoRA adaptation in the text encoder while keeping the entire

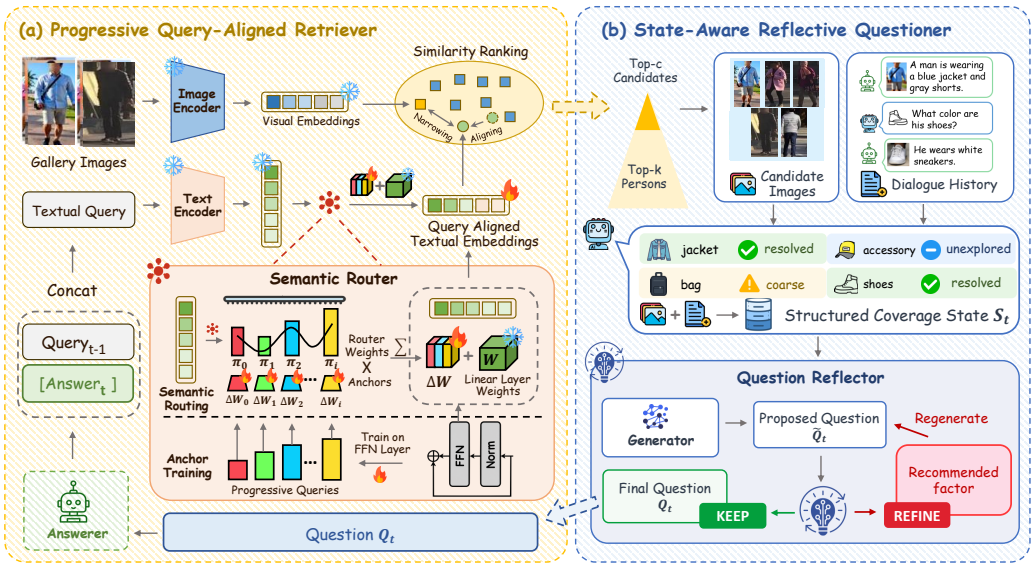

Figure 2: Overview of PAIReID. (a) The Progressive Query-Aligned Retriever dynamically composes alignment anchors according to the current query semantics. (b) The State-Aware Reflective Questioner tracks appearance coverage and selectively refines unsuitable follow-up questions.

base Retriever fixed. For an adapted linear layer ℓ,

$$\Delta W_{\ell,k} = \frac{\alpha}{r} B_{\ell,k} A_{\ell,k}. \tag{1}$$

All anchors therefore share the same frozen backbone and visual embedding space while capturing alignment behaviors associated with different progressive query distributions.

Semantic composition of alignment updates. The progressive query states used to construct the anchors provide a set of representative alignment behaviors, but interaction progress cannot be fully characterized by a discrete round index. Queries occurring at the same interaction round may contain different combinations and granularities of appearance information. We therefore determine the alignment configuration directly from the semantics of the current query rather than explicitly switching anchors according to interaction round.

Given the accumulated query x_t, we first obtain a routing representation from the frozen base text encoder with all alignment anchors disabled. A lightweight semantic Router then predicts a distribution over the six anchors:

$$e_t = f_0(x_t)_{\mathrm{EOT}}, \qquad \boldsymbol{\pi}_t = \mathrm{softmax}(R_\theta(e_t)), \tag{2}$$

where EOT denotes the end-of-text representation, R_θ is the Router, and $\boldsymbol{\pi}_t = \{\pi_{t,k}\}_{k=0}^{5}$ gives the query-dependent contribution of each alignment anchor. Since the base text encoder remains frozen, the routing representation reflects the current query without being affected by any individual alignment anchor.

The resulting query-conditioned update at layer ℓ is

$$\Delta W_\ell(x_t) = \sum_{k=0}^{5} \pi_{t,k} \Delta W_{\ell,k}, \qquad h'_\ell = W_\ell h_\ell + \Delta W_\ell(x_t) h_\ell, \tag{3}$$

where W_ℓ denotes the frozen base weight and h_ℓ the input representation of the corresponding linear transformation. The same query-level routing distribution $\boldsymbol{\pi}_t$ is shared across all adapted layers, yielding a consistent alignment configuration for the current query. We mix complete low-rank updates rather than their individual factors, thereby preserving a weighted combination of learned

anchor transformations. Because routing is conditioned on the current query representation, queries at the same interaction round may induce different alignment compositions, and no round index is provided to the Router.

Let $f_{\boldsymbol{\pi}_t}$ denote the text encoder equipped with the resulting mixed alignment updates. We compute the normalized textual and visual representations as

$$z_t = \frac{f_{\boldsymbol{\pi}_t}(x_t)_{\mathrm{EOT}}}{\|f_{\boldsymbol{\pi}_t}(x_t)_{\mathrm{EOT}}\|_2}, \qquad v_j = \frac{g(I_j)}{\|g(I_j)\|_2}, \tag{4}$$

and rank gallery image I_j according to

$$s_t(j) = z_t^\top v_j, \tag{5}$$

Inference therefore involves two text-side forward passes. The first obtains the semantic routing representation from the frozen base encoder, while the second encodes the query using the resulting mixture of alignment updates. Gallery representations can be reused throughout interaction because the image branch remains unchanged. Thus, the proposed composition operates within the text encoder at the level of low-rank parameter updates, rather than averaging features produced by independently executed experts.

Learning query-dependent composition. Anchor construction and semantic routing are optimized in two stages. In the first stage, the entire base Retriever is frozen and the six alignment anchors are independently optimized using the corresponding progressive query states. After anchor construction, we freeze both the base Retriever and all six anchor parameters and optimize only the Router R_θ.

For each training instance, one query is uniformly sampled from its six progressive states $\{x_0, \ldots, x_5\}$. The corresponding state index determines only which textual query is sampled and is never supplied to the Router as either input or supervision. The Router is optimized using the same retrieval objectives as the base Retriever:

$$\mathcal{L}_{\mathrm{ret}} = \mathcal{L}_{\mathrm{SDM}} + \lambda_{\mathrm{ID}}\mathcal{L}_{\mathrm{ID}} + \lambda_{\mathrm{MLM}}\mathcal{L}_{\mathrm{MLM}}, \tag{6}$$

where similarity distribution matching encourages image–text similarity distributions to respect identity correspondence, identity classification promotes identity-discriminative representations, and masked language modeling provides token-level cross-modal supervision. For masked language modeling, the masked query reuses the routing distribution inferred from its unmasked counterpart, preventing artificial token corruption from changing the alignment configuration.

No auxiliary state-classification objective or routing-specific balancing regularization is introduced. The Router therefore learns query-dependent anchor composition solely through the retrieval supervision in Eq. (6).

3.2 State-Aware Reflective Questioner

Structured coverage state. A dialogue history alone does not explicitly indicate whether an appearance factor has been resolved, only coarsely specified, or remains unexplored. We therefore convert the accumulated interaction history into a structured coverage state that records, for each appearance factor, its observed facts, whether it has been queried, the granularity of the available information, and an ordinal estimate of its remaining acquisition value.

In our person Re-ID setting, we instantiate the appearance space using nine semantically stable groups covering clothing, hair, footwear, carried items, accessories, posture, and body build. State construction separates semantic fact extraction from persistent state management: given the dialogue history D_t, a frozen language model extracts attribute-level facts together with their reliability, while deterministic operations validate, merge, and update the structured records. An indiscernible answer does not establish a new appearance fact, but the corresponding factor remains marked as queried to avoid repeatedly spending the interaction budget on the same observation.

The acquisition value is derived from the current coverage state. Resolved or previously queried factors are down-weighted, partially specified factors remain eligible for refinement, and unexplored visually stable factors receive higher preference. This ordinal value serves only as a deterministic acquisition signal rather than a numerical estimate of information gain, allowing the same factor to evolve from unexplored to partially specified and eventually resolved.

Candidate-aware question proposal. The coverage state indicates which appearance factors remain unresolved, but an unresolved factor may not be reliably observable in the current retrieval context. We therefore complement symbolic coverage with candidate-conditioned visual support. Given the retrieval scores from Sec. 3.1, we collect the current top-ranked candidates as

$$\mathcal{C}_t = \mathrm{TopK}_{I_j \in \mathcal{G}}\, s_t(j). \tag{7}$$

For appearance factors whose visibility can be unreliable in retrieved images, we apply a conservative support check over $\mathcal{C}_t$. Let $\mathrm{Visible}(I, s)$ indicate whether factor s is sufficiently observable in candidate image I. We define

$$b_t(s) = \mathbb{I}\left[\sum_{I \in \mathcal{C}_t} \mathrm{Visible}(I, s) \geq m\right]. \tag{8}$$

A factor without sufficient candidate support is temporarily suppressed from acquisition. Importantly, retrieved candidates are used only to assess whether a question is visually supportable and never to establish appearance facts about the unknown target. Conditioned on the resulting coverage state and candidate support, the Generator proposes a single follow-up question targeting one eligible unresolved factor, while avoiding redundant or overly specific queries.

State-conditioned reflective verification. Constraining the proposal space reduces obviously unsuitable questions, but generation remains an open-ended process and may still produce a question that is inconsistent with the current interaction state. We therefore introduce a separate verification stage before committing an interaction round. Given the proposed question $\widetilde{q}_{t+1}$, the current coverage state, and the retrieved candidates $\mathcal{C}_t$, a frozen multimodal Reflector determines whether the proposal should be retained or redirected.

The verification considers whether the proposal contributes unresolved information, remains non-redundant with the accumulated state, is visually supportable under the current candidates, and uses the limited interaction budget at an appropriate level of specificity. A suitable proposal is returned unchanged with a KEEP decision; otherwise, the Reflector returns REFINE and identifies an alternative unresolved factor s^*_{t+1}.

When refinement is required, the Reflector does not directly rewrite the question. Instead, the Generator reformulates it under the same state and generation constraints, conditioned on the recommended factor. A deterministic consistency check ensures that the revised question follows the intended direction, while a canonical group-level formulation is used only as a fallback. Formally,

$$q_{t+1} = \begin{cases} \widetilde{q}_{t+1}, & d_{t+1} = \mathrm{KEEP}, \\ \mathcal{G}\left(s^*_{t+1}, \mathcal{S}_t, \mathcal{C}_t\right), & d_{t+1} = \mathrm{REFINE}, \end{cases} \tag{9}$$

where $d_{t+1} \in \{\mathrm{KEEP}, \mathrm{REFINE}\}$ denotes the reflection decision, $\mathcal{S}_t$ the current coverage state, and $\mathcal{G}$ the constrained question generator. This separation decouples diagnosis from question realization: the Reflector determines whether and where revision is required, while the Generator determines how the final question is expressed.

After the final question is issued, the Answerer returns a_{t+1}. The newly acquired information updates both the accumulated query x_{t+1} and the structured coverage state. The Progressive Query-Aligned Retriever then recomputes its alignment composition from the updated query and produces a new candidate set for the next interaction. In this way, each interaction jointly updates what information is available for retrieval and what information remains worth acquiring, closing the loop between progressive retrieval adaptation and state-aware questioning.

Table 1: Comparison with interactive retrieval methods on Interactive-PEDES. Best and second-best results are highlighted in **bold** and underline, respectively.

Method	Questioner Type	Round 3				Round 5				BRI ↓
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
SimIRV Liang & Albanie (2023)	Heuristic	50.45	73.39	82.49	41.00	61.27	82.00	88.36	46.53	1.024
ChatIR Levy et al. (2023)	Few-shot	57.85	78.69	86.08	44.92	63.86	83.81	89.84	48.05	0.935
PlugIR Lee et al. (2024)	Few-shot	60.34	81.34	87.89	<u>47.28</u>	65.44	85.33	91.01	49.89	0.849
LLaVA-ReID Lu et al. (2025)	Zero-shot	46.91	70.64	79.43	33.76	52.00	75.34	83.30	37.23	0.973
LLaVA-ReID Lu et al. (2025)	Finetuned	<u>63.96</u>	<u>84.70</u>	<u>91.27</u>	**48.45**	<u>73.20</u>	<u>90.62</u>	**95.95**	**53.31**	<u>0.719</u>
PAIReID (Ours)	**Zero-shot**	**66.93**	**87.18**	**92.68**	47.04	**74.08**	**91.48**	<u>95.39</u>	<u>50.85</u>	**0.485**

4 EXPERIMENTS

Datasets. We conduct experiments on four person retrieval benchmarks. Interactive-PEDES Lu et al. (2025) serves as the primary benchmark for interactive person Re-ID. It contains 54,749 pedestrian images from 13,051 identities collected from CUHK-PEDES Li et al. (2017) and ICFG-PEDES Ding et al. (2021), with each image paired with an initial coarse-grained description and multi-round dialogue annotations. We further evaluate PAIReID on three widely used Text-ReID benchmarks, CUHK-PEDES, ICFG-PEDES, and RSTPReid Zhu et al. (2021), following their official data partitions. These datasets complement Interactive-PEDES by evaluating transferability to conventional Text-ReID settings.

Evaluation Protocols. We define retrieval using only the initial description as Round 0. At each subsequent round, the system generates one follow-up question, obtains an answer, and incorporates the newly acquired information into the accumulated query. Unless otherwise specified, we conduct five interaction rounds. Following prior work, we report R@1, R@5, R@10, and mAP for retrieval performance. We additionally report BRI↓ as a trajectory-level interaction metric, where a lower value indicates more effective refinement across successive rounds.

Implementation Details. We use IRRA with a CLIP ViT-B/16 backbone as the base Retriever and construct six LoRA alignment anchors for progressive query states from Round 0 to Round 5, with LoRA applied to the text-side feed-forward projections while the base Retriever remains frozen. After independently training the anchors, we freeze them together with the base Retriever and optimize a lightweight two-layer semantic Router that performs dense softmax composition over all six anchors without round-index supervision. We use Qwen-2.5-VL as the zero-shot Questioner and retain the top four retrieved candidates for question generation and reflection. Further implementation details are provided in Appendix A.1.

4.1 COMPARISON WITH STATE-OF-THE-ART METHODS

Performance on the dialogue-based benchmark. As shown in Table 1, PAIReID achieves strong performance on Interactive-PEDES with a zero-shot Questioner. At Round 3, our method reaches 66.93% R@1, 87.18% R@5, and 92.68% R@10, outperforming the finetuned LLaVA-ReID by 2.97%, 2.48%, and 1.41%, respectively. After five interaction rounds, PAIReID further achieves 74.08% R@1 and 91.48% R@5, surpassing LLaVA-ReID by 0.88% and 0.86%. Meanwhile, PAIReID obtains the best BRI of 0.485, compared with 0.719 for LLaVA-ReID, demonstrating effective retrieval refinement throughout the interaction without task-specific Questioner finetuning.

Performance on traditional Text-to-Image Re-ID benchmarks. As shown in Table 2, PAIReID also exhibits strong applicability on RSTPReid, CUHK-PEDES, and ICFG-PEDES. With RDE as the caption Retriever, our method achieves 75.30% R@1 and 56.90% mAP on RSTPReid, outperforming LLaVA-ReID by 3.30% and 0.84%, respectively. On CUHK-PEDES, PAIReID reaches 80.86% R@1 and 93.11% R@5, while matching the best R@10 of 95.96%. On ICFG-PEDES, our method achieves 69.24% R@1, 83.65% R@5, and 88.38% R@10, ranking second on all three recall metrics. Similar trends are observed with IRRA, showing that PAIReID can be effectively integrated with different underlying Text-ReID Retrievers.

Table 2: Performance comparison on three conventional text-based person Re-ID benchmarks. Best and second-best results in each column are highlighted in **bold** and underline, respectively.

Method	RSTPReid				CUHK-PEDES				ICFG-PEDES			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
CFiNe Yan et al. (2023)	50.55	72.50	81.60	–	69.57	85.93	91.15	–	60.83	76.55	82.42	–
RaSa Bai et al. (2023a)	66.90	86.50	91.35	52.31	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29
APTM Yang et al. (2023)	67.50	85.70	91.45	52.56	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22
AUL Li et al. (2024)	71.65	87.55	92.05	–	77.23	90.43	94.41	–	69.16	83.32	88.37	–
IRRA Jiang & Ye (2023)	60.20	81.30	88.20	47.19	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06
+ LLaVA-ReID	69.85	88.10	92.55	54.92	78.51	92.43	95.67	70.61	67.44	82.91	87.69	40.60
+ **PAIReID (Ours)**	72.85	89.45	**94.50**	54.97	78.85	92.59	95.94	69.98	65.79	81.68	86.85	38.11
RDE Qin et al. (2024)	65.35	83.95	89.90	50.88	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06
+ LLaVA-ReID	72.00	88.95	93.55	56.06	79.39	93.02	**95.96**	**71.70**	**69.98**	**84.32**	**88.60**	**42.12**
+ **PAIReID (Ours)**	**75.30**	**90.35**	94.40	**56.90**	**80.86**	**93.11**	**95.96**	71.15	69.24	83.65	88.38	40.18

Table 3: Ablation Study on Routing Strategy.

Routing Strategy	R1	mAP
Base	69.73	47.54
Uniform	73.36	50.26
Round-indexed	71.67	48.88
Semantic	**74.08**	**50.85**

Table 4: Ablation Study on State-Aware Reflective Questioner.

Coverage	Reflection	R1	mAP	BRI ↓
×	×	51.47	37.71	0.673
✓	×	63.52	44.98	0.586
✓	✓	**74.08**	**50.85**	**0.485**

4.2 Ablation Study

Effect of Progressive Query-Aligned Retriever. As shown in Table 3, introducing progressive alignment anchors consistently improves the shared base Retriever. Uniformly combining the anchors increases R@1 from 69.73% to 73.36% and mAP from 47.54% to 50.26%, indicating that complementary adaptations learned from different query states are beneficial for evolving retrieval queries. In contrast, directly selecting the alignment according to interaction rounds achieves 71.67% R@1 and 48.88% mAP. Our semantic routing strategy obtains the best performance, reaching 74.08% R@1 and 50.85% mAP. The improvement over round-indexed routing suggests that query semantics provide a more effective signal for composing alignment adaptations than explicit interaction-round identities.

Effect of State-Aware Reflective Questioner. As shown in Table 4, we evaluate the two components of the Questioner. Without coverage modeling or reflection, the model achieves 51.47% R@1 and 37.71% mAP at Round 5. Introducing structured coverage improves R@1 to 63.52% and mAP to 44.98%, while reducing BRI from 0.673 to 0.586. Adding reflective verification further improves performance to 74.08% R@1 and 50.85% mAP, with the lowest BRI of 0.485. These results confirm that coverage tracking guides information acquisition, while reflection further refines unsuitable questions, providing complementary benefits throughout multi-round interaction.

5 Conclusion

In this work, we present PAIReID, a progressive adaptation and reflective interaction framework for interactive person re-identification. We identify two coupled challenges in multi-round interaction: the evolving semantic composition of retrieval queries and the changing demand for follow-up information. To address them, PAIReID combines a Progressive Query-Aligned Retriever that dynamically composes lightweight alignment adaptations according to the current query semantics with a State-Aware Reflective Questioner that tracks unresolved appearance information and verifies candidate questions before each interaction. Experiments on Interactive-PEDES demonstrate strong progressive retrieval performance, achieving 74.08% Rank-1 after five interactions with a zero-shot Questioner and a BRI of 0.485. Results on RSTPReid, CUHK-PEDES, and ICFG-PEDES further show that PAIReID can be effectively integrated with different text-based person retrieval systems.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and writing refinement. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies person retrieval using existing research benchmarks and does not introduce new human-subject data collection. Person re-identification technologies can support legitimate retrieval applications, but they may also raise privacy, surveillance, and potential misuse concerns when deployed in real-world environments. We therefore restrict our claims to research evaluation and emphasize that practical deployment should comply with applicable privacy regulations, access-control requirements, and appropriate oversight. The proposed method should not be interpreted as removing the broader ethical risks associated with person identification and retrieval systems.

REPRODUCIBILITY STATEMENT

We describe the task formulation, model architecture, training objectives, and inference procedure in Sec. 3, and provide the datasets, evaluation protocol, and main experimental settings in Sec. 4. Additional implementation details, prompts, and analysis are provided in the Appendix. All reported comparisons follow the official dataset partitions and evaluation metrics of the corresponding benchmarks. We will provide the implementation and configuration files necessary to reproduce the reported results as supplementary material.

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*, pp. 555–563, 2023a.

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

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3952–3962, 2025.

Yang Bai, Tingfeng Wang, Bin Yang, Min Cao, Jinqiao Wang, and Mang Ye. Interactive person retrieval via multi-turn multimodal conversation. In *Forty-third International Conference on Machine Learning*, 2026.

Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: Active detail enriching for cross-modal retrieval with partial query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1815–1824, 2021.

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

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

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2787–2797, 2023.

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

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 791–809, 2024.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems*, pp. 61437–61449, 2023.

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 AAAI Conference on Artificial Intelligence*, pp. 3172–3180, 2024.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5187–5196, 2017.

Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11091–11101, 2023.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. In *Forty-second International Conference on Machine Learning*, 2025.

Liqiang Nie, Fangkai Jiao, Wenjie Wang, Yinglong Wang, and Qi Tian. Conversational image search. *IEEE Transactions on Image Processing*, pp. 7732–7743, 2021.

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

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14390–14399, 2025.

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. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, pp. 5566–5574, 2022.

Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 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*, pp. 17127–17137, 2024.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Proceedings of the 16th European Conference on Computer Vision*, pp. 402–420, 2020.

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

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, pp. 6032–6046, 2023.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 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*, pp. 4492–4501, 2023.

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*, pp. 1–23, 2020.

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

A Appendix

A.1 Implementation Details

Retriever and Alignment Anchors. We use IRRA with a CLIP ViT-B/16 backbone as the base Retriever. We construct six alignment anchors corresponding to the progressive query states from Round 0 to Round 5. Specifically, the Round-0 anchor is trained using the initial query, while the Round-t anchor ($t > 0$) is trained using the initial query augmented with the first t dialogue answers.

Each anchor is parameterized by LoRA with rank $r = 8$, scaling factor $\alpha = 16$, and dropout rate 0.05. LoRA modules are inserted into the two feed-forward projections, `c_fc` and `c_proj`, of all 12 text-Transformer blocks, resulting in 24 adapted linear layers. The entire base Retriever is frozen, and only the LoRA parameters are optimized. Each anchor contains 491,520 trainable parameters.

The anchors are trained independently for 30 epochs using Adam with a learning rate of 1×10^{-5}, weight decay 4×10^{-5}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. We adopt a cosine learning-rate schedule with five epochs of linear warm-up, initialized at 0.1 times the base learning rate. The per-device batch size is 32, and no gradient accumulation is used. The Round-0 checkpoint is selected according to initial-query Rank-1, whereas the Round-1–5 checkpoints are selected according to fine-grained-description Rank-1.

Semantic Router. After constructing the anchors, we freeze both the base Retriever and all six LoRA anchors and optimize only the semantic Router. For a query x_t, we first obtain the EOT representation from the frozen base text encoder and normalize it as

$$e_t = \frac{f_0(x_t)_{\mathrm{EOT}}}{\|f_0(x_t)_{\mathrm{EOT}}\|_2} \in \mathbb{R}^{512}. \tag{10}$$

The Router is implemented as a two-layer MLP consisting of LayerNorm, a $512 \to 256$ linear layer, GELU activation, dropout with rate 0.1, and a $256 \to 6$ output layer. It contains 133,894 trainable parameters. The final projection is initialized to zero, such that routing starts from a uniform distribution over the six anchors.

The routing probabilities are computed as

$$\boldsymbol{\pi}_t = \mathrm{softmax}\left(\frac{R_\theta(e_t)}{\tau}\right), \qquad \tau = 1. \tag{11}$$

All six anchors participate in the composition; no hard expert selection or top-K routing is used. The same query-level routing distribution is shared across all 24 adapted layers. At layer ℓ, the composed LoRA update is

$$\Delta W_\ell^{(t)} = \sum_{k=0}^{5} \pi_{t,k} \frac{\alpha}{r} B_{\ell,k} A_{\ell,k}. \tag{12}$$

The complete low-rank updates are therefore mixed after each expert update is formed, avoiding cross-expert terms that would arise from independently mixing the low-rank factors.

The Router is trained for 10 epochs using AdamW with a learning rate of 1×10^{-4}, weight decay 1×10^{-4}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. We use the same cosine schedule with five epochs of linear warm-up. The per-device batch size is 32, and training is distributed over two GPUs, resulting in a global batch size of 64. No gradient accumulation is used.

During training, each sample uniformly draws one query state from Rounds 0–5. The corresponding round index is retained only for diagnostic statistics and is never provided to the Router as either input or supervision. The Router is optimized using the original retrieval objectives, including similarity distribution matching, identity classification, and masked language modeling. No auxiliary round-classification, entropy regularization, expert load-balancing, or routing-supervision objective is introduced. The best Router checkpoint is selected according to fine-grained-description Rank-1.

Numerical Precision. The frozen CLIP-based Retriever is represented in FP16, while the trainable LoRA parameters, semantic Router, and LoRA expert-update computations are maintained in FP32 for numerical stability.

Candidate Configuration. At each interaction round, we retain the top $K_{\text{ret}} = 4$ retrieved candidate images. The same four candidates are provided to the question-generation, reflection, and visual-feedback modules, i.e., $K_{\text{vis}} = 4$. Questions concerning carried items or accessories are permitted only when the corresponding attribute is judged sufficiently visible in at least three of the four candidates. Candidate images are used exclusively to assess whether an attribute is visually supportable for questioning and never to establish appearance facts about the target person.

Hardware and Training Cost. All experiments were conducted on NVIDIA GeForce RTX 4090 GPUs. Training an individual alignment anchor required approximately 68–71 minutes. The semantic Router was trained on two RTX 4090 GPUs and required approximately 45 minutes.

A.2 Coverage-State Construction

The State-Aware Reflective Questioner maintains an explicit coverage state rather than relying on the language model to implicitly infer the dialogue status from the complete interaction history. We organize pedestrian appearance information into nine semantically stable groups: *outerwear*, *upper clothing*, *lower clothing*, *head/hair*, *footwear*, *carried items*, *accessories*, *posture*, and *body build*.

For each appearance group s, the coverage state is represented as

$$S_t(s) = \{\texttt{asked}, \texttt{fact}, \texttt{fact_type}, \texttt{question_value}\}, \tag{13}$$

where `asked` indicates whether the factor has already been queried, `fact` stores the currently supported appearance information, `fact_type` characterizes its granularity as *none*, *coarse-grained*, or *fine-grained*, and `question_value` provides an ordinal acquisition preference.

Importantly, semantic fact extraction and persistent state management are separated. A frozen language model is used only to extract visual facts explicitly supported by the initial description and dialogue answers, together with their source turns and reliability status. Persistent state variables, including whether a factor has already been queried, the granularity of the available information, and its acquisition status, are subsequently validated and updated by deterministic operations. Therefore, the language model does not directly control persistent state transitions.

A single observation may update multiple appearance groups when it provides information relevant to more than one semantic factor. Explicitly unavailable or indiscernible information is distinguished from positively supported facts rather than being treated as a missing positive observation. In addition, visual details appearing only in the wording of a generated question are never incorporated as target facts; only information supplied by the Answerer is allowed to update the coverage memory.

The ordinal acquisition value is used only as a control signal for subsequent question generation and should not be interpreted as a learned estimate of numerical information gain. Resolved or previously queried factors are suppressed from repeated acquisition, partially specified factors may remain eligible for further refinement, and unexplored visually stable factors receive higher acquisition preference. Consequently, the mechanism models the progression of information coverage rather than imposing a fixed attribute sequence.

A.3 Candidate-Aware Visual Support

The retrieved candidates provide contextual support for question selection rather than evidence about the unknown target identity. Candidate-conditioned visual checking is applied only to visibility-sensitive appearance groups, specifically *carried items* and *accessories*, for which unreliable visibility can otherwise encourage speculative questions.

Given the current candidate set $\mathcal{C}_t$, the visual-feedback module independently determines for each candidate whether the relevant attribute region is sufficiently visible for reliable inspection. Occluded, blurred, excessively small, cropped-out, or otherwise uncertain regions are treated as unsupported. In our implementation, a visibility-sensitive factor becomes eligible for questioning only when sufficient visual support is observed in at least three candidate images.

This mechanism is deliberately conservative. Candidate observations never create or modify target appearance facts in the coverage state; they only control whether a visibility-sensitive factor is temporarily eligible for questioning. As the retrieval result changes across interaction rounds, an attribute suppressed at one stage may therefore become eligible again under a later candidate set.

Table 5: Roles of model-based components and deterministic safeguards in the State-Aware Reflective Questioner.

Component	Main Input	Output	Deterministic Control
Coverage Extractor	Dialogue history	Supported facts	State validation/update
Visual Feedback	Retrieved candidates	Visibility judgments	Eligibility threshold
Generator	Coverage + candidates	One question	Valid-factor constraints
Reflector	State + candidates + question	KEEP/REFINE + factor	Forced refinement checks
Answerer	Target memory + question	Witness answer	Information isolation

A.4 Question Generation and Reflective Verification

Given the current coverage state and retrieved visual context, the Generator produces exactly one follow-up question targeting an eligible unresolved appearance factor. The generation process favors broad group-level descriptions rather than narrow hypotheses concerning individual colors, logos, or other micro-details. Questions about identity, face, age, background context, generic "other" attributes, or direct comparisons between retrieved candidates are excluded.

Although these constraints reduce obviously unsuitable proposals, question generation remains open-ended. We therefore apply a separate state-conditioned Reflector before committing an interaction round. The Reflector jointly examines the proposed question, the current coverage state, and the retrieved candidates. A proposal is redirected when it targets already queried or low-value information, combines multiple unrelated appearance factors, is excessively narrow, or lacks sufficient visual support for a visibility-sensitive factor.

The Reflector outputs a binary decision

$$d_t \in \{\texttt{KEEP}, \texttt{REFINE}\}. \tag{14}$$

For `KEEP`, the original proposal is issued unchanged. For `REFINE`, the Reflector additionally identifies one alternative unresolved appearance factor together with a short semantic focus. The Reflector does not directly rewrite the question. Instead, the Generator produces a new question conditioned on the recommended factor.

After regeneration, a deterministic consistency check verifies that the new question follows the redirected appearance factor. If this constraint is not satisfied, a canonical group-level formulation is used only as a fallback. This separation keeps the Reflector responsible for diagnosis and redirection, while leaving linguistic realization to the zero-shot Generator.

A.5 Deterministic Safeguards

PAIReID combines open-ended language-model reasoning with deterministic state-management safeguards. Specifically, the language model is responsible for semantic interpretation and extraction of supported visual facts, whereas persistent fields such as querying status, fact granularity, and acquisition value are maintained programmatically.

Previously queried or sufficiently resolved factors are prevented from being repeatedly selected. Candidate-aware eligibility is also verified after the visual-feedback stage using the predefined minimum visibility requirement. Questions targeting already queried or low-value factors are routed to refinement regardless of the initial generation output. After refinement, the resulting question is further checked against the appearance factor recommended by the Reflector before being issued to the Answerer.

These safeguards prevent transient language-model outputs from directly modifying the persistent interaction state and make multi-round state transitions reproducible given the same model outputs. In this way, semantic reasoning is delegated to pretrained models while the interaction policy remains explicitly controlled and auditable.

A.6 Witness Simulation and Information Isolation

The Answerer simulates a witness with access to target-specific information that is unavailable to both the Questioner and Retriever. Given the final question, the Answerer is instructed to return

only information concerning the requested appearance factor and to preserve the factual content, uncertainty, and negation contained in the underlying target memory.

The Answerer is explicitly prohibited from inferring unobserved attributes or introducing unrelated appearance information. Missing information is also distinguished from an explicitly supported absence. The resulting answer, rather than the hidden witness memory itself, is the only new target-specific information incorporated into the accumulated retrieval query and structured coverage state.

This information separation follows the standard interactive retrieval setting: the Answerer acts as the simulated witness, while both retrieval and question generation operate exclusively on information that has already been revealed through interaction.

A.7 Prompt Templates

The State-Aware Reflective Questioner consists of five prompt-driven operations: visual-fact extraction, candidate-aware visibility assessment, coverage-aware question generation, state-aware reflective verification, and targeted question refinement. The prompts constrain the output format and the semantic role of each module, while persistent interaction state and validity checks are handled separately by the deterministic mechanisms described above. The complete prompt templates and configuration files used in the experiments are provided in the supplementary implementation for reproducibility.

A.8 Detailed Routing Analysis

Figure 3 visualizes the round-averaged semantic routing behavior over the six alignment anchors. At Round 0, the Router strongly favors the initial-query anchor, whose mean routing weight is 0.81. As additional appearance information is accumulated, its contribution gradually decreases to 0.32 at Round 5, while the remaining anchors receive increasing weights.

Importantly, multiple anchors remain jointly active rather than forming a hard one-to-one correspondence between interaction rounds and alignment anchors. This observation is consistent with the routing ablation in Table 3, where semantic routing outperforms both uniform composition and explicit round-indexed selection. Together, these results indicate that the Router learns a soft composition of progressive alignment behaviors as the semantic content of the query evolves.

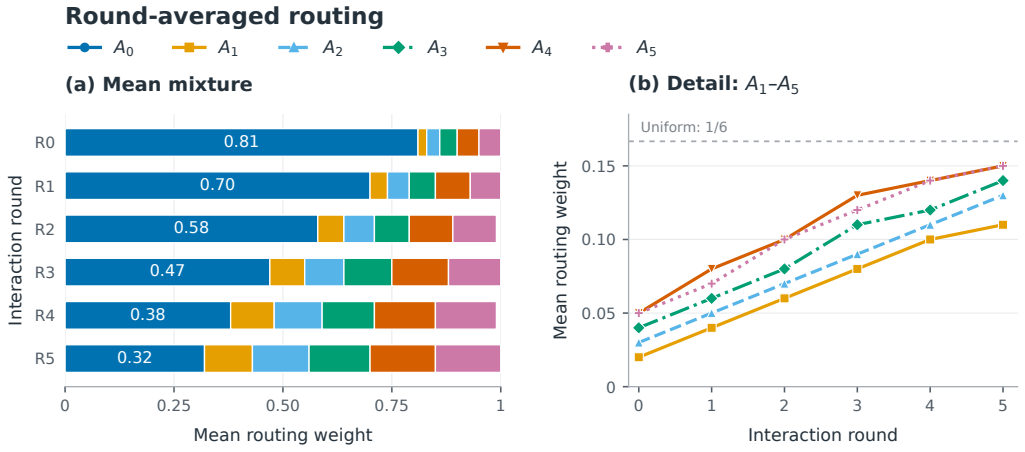

Figure 3: Detailed semantic-routing behavior across interaction rounds. The routing distribution gradually becomes less concentrated on the initial-query anchor as additional appearance information is accumulated.