Listening to the Retriever: Perturbation-Sensitive Question Selection for Interactive Person Retrieval

Yang Bai[1]*, **Yunhui Shao**[1]*, **Shuai You**[1]†, **Bin Yang**[1], **Jie Qiao**[1], **Min Cao**[2], **Mang Ye**[1]†
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, China
[2]School of Computer Science and Technology, Soochow University, Suzhou, China
{ybai.ai,youshuai_ai,yemang}@whu.edu.cn
https://github.com/Flame-Chasers/SCOUT

Abstract

Interactive person retrieval extends text-to-image person retrieval by allowing follow-up questions when the initial description is incomplete. The key challenge lies in selecting a follow-up question that can effectively differentiate the current retrieval candidates. Existing methods guide question selection with retriever-external signals, which may not directly reflect the discriminative value of different questions for the current candidate set. To address this, we propose SCOUT, an interactive person retrieval framework that derives questions from the retriever's own response. Specifically, we introduce a Perturbation-Sensitive Question Selection strategy that selects the next question based on Retrieval Perturbation Sensitivity (RPS), a test-time measure of perturbation-induced ranking change. Additionally, SCOUT is training-free and can be directly applied to off-the-shelf text-to-image person retrievers, without the dialogue-data curation or task-specific retraining required by existing interactive methods. Extensive experiments demonstrate that SCOUT yields consistent improvements over baseline methods, validating this RPS-based strategy for interactive person retrieval.

1 Introduction

Text-to-image person retrieval aims to identify a target person from a large surveillance gallery using a natural-language description [1, 2, 3, 4, 5]. Existing works, such as IRRA [6] and RDE [7], have established strong single-shot baselines through effective cross-modal alignment, while data-centric methods further scale performance with large MLLM-generated annotations [8, 9, 10, 11, 12]. However, these methods implicitly assume that the user can provide a single, sufficiently discriminative sentence before retrieval. This assumption is typically fragile in real-world scenarios: the initial description may contain only a subset of the target person's observable characteristics, while visually similar candidates may differ in fine-grained attributes that are not specified in the query. To address this limitation, interactive person retrieval [13, 14, 15, 16] allows the system to ask follow-up questions and progressively refine the query. Nevertheless, interaction alone does not solve the problem: the central challenge lies in selecting follow-up questions that reveal the missing information most useful for resolving ambiguity among the current retrieval candidates.

Current interactive person retrieval methods select questions using different signals, as illustrated in Figure 1. ChatPR [13] constructs LLM-based multi-round dialogues to progressively supplement missing details into the initial description. ICL [15] anchors interaction to top-ranked candidate images and applies MLLM-based VQA with a predefined attribute questionnaire. LLaVA-ReID

*Equal contribution.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

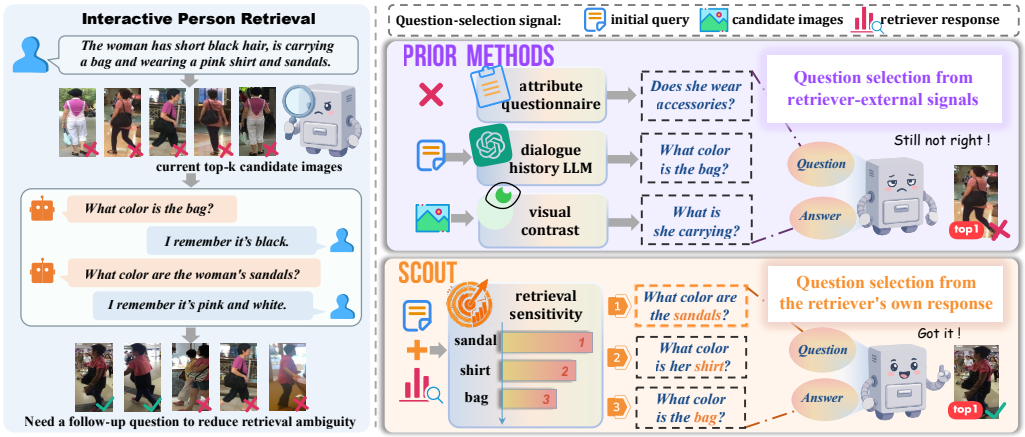


Figure 1: Interactive person retrieval requires choosing the next question to reduce ambiguity in the current rank list. Prior methods rely on retriever-external signals, while SCOUT uses the retriever's response to candidate attributes to select the next question by Retrieval Perturbation Sensitivity.

[14] generates targeted questions from visual and textual contexts, using a look-forward strategy to prioritize informative questions as training supervision. Although these methods can generate informative follow-up questions, their selection signals are derived from retriever-external sources rather than from the current retriever's own response. Consequently, a question that appears informative in general may not be the one most effective for resolving the ambiguity faced by the current retriever. This raises a central problem: **I) how can the system select a follow-up question that effectively reduces ambiguity among the current retrieval candidates?**

Identifying the most useful follow-up question requires estimating question utility under the current retrieval state. Existing interactive retrieval works select clarifying questions based on expected utility or informativeness [17, 14, 18]. Neural EVPI [17] makes this idea concrete by ranking clarification questions according to the expected value of their answers. Recent interactive person retrieval follows a similar motivation: the look-forward strategy in LLaVA-ReID [14] uses post-answer improvement of the target person's ranking score as training supervision. However, these utility-aware methods typically learn a questioner or question-ranking model from offline utility signals, which may not fully reflect the current retriever's behavior at test time. More importantly, the effect of a candidate question cannot be directly observed before it is asked. This raises the second problem: **II) how can the current retriever itself provide a reliable test-time signal for selecting the next question?**

To address the above problems, we use the current retriever itself as the reference for determining which missing information is more relevant to its current retrieval behavior for resolving ambiguity in the current retrieval state. For the ***first problem***, we propose SCOUT, an interactive person retrieval framework that shifts question selection from retriever-external signals to the retriever's own response to candidate attributes, using this response to identify the missing detail most relevant to its current retrieval ambiguity. For the ***second problem***, we introduce a Perturbation-Sensitive Question Selection strategy, which systematically constructs controlled semantic perturbations associated with candidate questions and compares the retriever's scoring behavior before and after each perturbation. We quantify the resulting change as Retrieval Perturbation Sensitivity (RPS), a retriever-specific test-time signal for selecting the next follow-up question. Additionally, SCOUT requires no additional training and builds on off-the-shelf text-to-image person retrievers, avoiding the dialogue-data curation and retraining overhead of existing interactive methods.

Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid with diverse retriever baselines show that SCOUT yields consistent improvements over both single-shot retrievers and interactive baselines. Ablation studies further show that the gains come from measuring Retrieval Perturbation Sensitivity and using it to guide Perturbation-Sensitive Question Selection. Theoretical analysis further provides formal justification for RPS by linking perturbation sensitivity to changes in the retrieval distribution and information gain. Together, these results support RPS as a principled criterion for selecting follow-up questions according to the current retriever's response. This allows SCOUT to adapt question selection to the current retrieval state without additional training.

Overall, the primary contributions of this work can be summarized as follows:

- ❶ ***SCOUT: An Interactive Person Retrieval Framework via Retriever Response.*** A framework that shifts question selection from retriever-external signals to the retriever's own response, plug-and-play with existing retrievers without dialogue-data curation or retraining.
- ❷ ***Perturbation-Sensitive Question Selection.*** A question selection strategy that probes semantic perturbations and measures Retrieval Perturbation Sensitivity (RPS), turning perturbation-induced ranking changes into a direct signal for question selection.
- ❸ ***Empirical Validation and Theoretical Analysis.*** Consistent gains over single-shot and interactive baselines across standard benchmarks, supported by theoretical analysis and ablation studies confirming the advantage of perturbation-sensitive question selection.

2 Related Work

2.1 Text-to-Image Person Retrieval

Text-to-image person retrieval (TIReID) aims to retrieve the target person from an image gallery using a natural-language description, with the central challenge in accurately matching person appearance with fine-grained textual descriptions [3, 5, 6, 19, 20, 21]. Existing methods mainly improve TIReID from two perspectives: architecture design and training objectives. The first line develops fine-grained cross-modal alignment mechanisms to associate visual regions, body parts, or attributes with textual phrases [22, 23, 24, 25, 26]. The second line introduces auxiliary objectives and stronger supervision, including masked language modeling, contrastive learning, relation modeling, attribute-aware learning, and noisy-correspondence learning [6, 7, 27, 28, 29]. More recent work continues to extend these directions with stronger vision-language priors [30, 31, 32] and richer data resources [11, 33, 34, 35, 36]. Overall, these advances have substantially improved the effectiveness of TIReID and enabled increasingly accurate retrieval from fine-grained textual descriptions.

Despite these advances, standard TIReID pipelines assume that the user provides one sufficiently complete description before retrieval [13, 14, 37, 38]. This assumption is often unrealistic in scenarios where users describe a target person from memory: test-time descriptions can be partial, uncertain, or revised after feedback. Although recent fine-grained alignment and MLLM-based caption augmentation methods enrich static image-text matching [34, 39, 9, 11], they focus on improving the retriever or its training data rather than actively eliciting missing details during retrieval. This motivates interactive person retrieval, where the system actively asks follow-up questions to refine the initial description over multiple rounds.

2.2 Interactive Retrieval and Query Refinement

Interactive retrieval progressively refines user intent through feedback, clarification, or multi-turn dialogue [40, 41, 42, 43, 44, 45]. Existing studies have explored interpretable feedback based on relative attributes or natural-language clarification for iteratively refining retrieval queries [46, 47, 48]. ChatIR [49] prompts an LLM over dialogue history, while PlugIR [43] generates questions from textual descriptions of candidate clusters. Conversational search studies how to rewrite dialogue history into retrieval-compatible queries [50, 51, 52], while clarifying-question methods explicitly request missing intent information in open-domain retrieval [53, 54]. Recent studies have also explored broader model adaptation problems, including concept control and generalization [55, 56, 57]. Most existing methods focus on dialogue modeling, intent clarification, or query reformulation, while the retriever itself is rarely used to guide question selection.

Within the field of person retrieval, ChatPR [13] introduces multi-round dialogue to progressively enrich incomplete initial descriptions. ICL [15] further grounds the interaction in top-ranked candidate images and uses an attribute questionnaire for VQA-based refinement. LLaVA-ReID [14] generates targeted questions through a look-forward strategy to obtain informative questions as training supervision. While these methods progressively introduce richer dialogue and visual context into person retrieval, their question-selection signals still come from retriever-external sources, such as fixed questionnaires or offline look-forward supervision. In contrast, we introduce SCOUT to derive the next question from the current retriever's test-time response to controlled attribute perturbations, thereby adapting question selection to the retriever's current state.

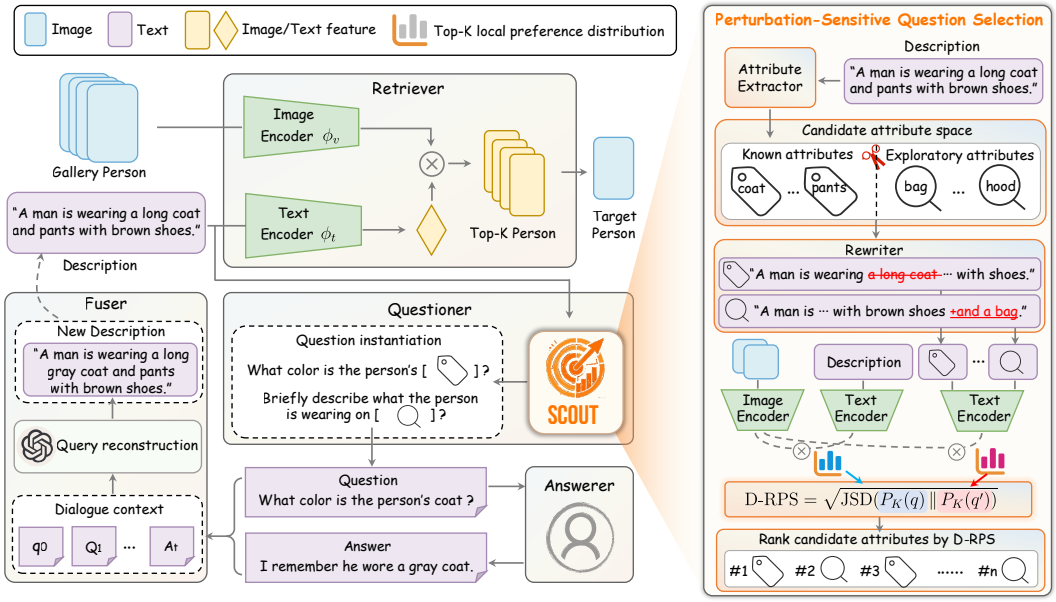


Figure 2: The overall framework of SCOUT. The left panel illustrates the multi-round interaction process, where the frozen retriever ranks the gallery, the Questioner selects a follow-up question, and the Fuser updates the query using the user's response. The right panel illustrates Perturbation-Sensitive Question Selection, which compares the retriever's local preference distributions before and after controlled query perturbations over a shared support, and selects the most sensitive attribute or region for question generation.

3 Method

We formulate interactive person retrieval as iterative refinement over a fixed gallery $\mathcal{G} = \{I_i\}_{i=1}^{N}$ containing the target person I_{gt}. Given an initial description q_0, the system retrieves with the current query q_{t-1} at round $t = 1, \ldots, T$, asks a follow-up question Q_t, receives an answer A_t, and updates the dialogue context $\mathcal{D}_t = \{q_0, (Q_1, A_1), \ldots, (Q_t, A_t)\}$. A Fuser reconstructs q_t from $\mathcal{D}_t$ for the next round. We propose SCOUT, a training-free framework that keeps the retriever $\mathcal{R}$ fixed and improves retrieval by refining the query. Its Questioner performs Perturbation-Sensitive Question Selection: it probes candidate attribute perturbations, computes their Retrieval Perturbation Sensitivity (RPS), and selects the next question according to this sensitivity. We present the SCOUT framework in Section 3.1, the selection strategy in Section 3.2, and RPS in Section 3.3.

3.1 SCOUT Framework

SCOUT forms a closed loop over four components (left of Figure 2): the Retriever ranks the gallery under the current query; the Questioner selects the next attribute or region to ask about; the Answerer provides the answer; and the Fuser reconstructs the query from the updated dialogue context. This loop repeats for T=5 rounds, after which the Retriever produces the final ranked output using q_T.

Retriever. The frozen dual-encoder retriever $\mathcal{R}$ encodes the current query q_{t-1} and all gallery images $\mathcal{G} = \{I_i\}_{i=1}^{N}$ via text encoder ϕ_t and image encoder ϕ_v, producing a ranked gallery list. Neither encoder is modified during interaction.

Questioner. Given the retrieval state and dialogue context $\mathcal{D}_{t-1}$ under the current query q_{t-1}, the Questioner uses a Perturbation-Sensitive Question Selection strategy that evaluates candidate attributes by estimating their Retrieval Perturbation Sensitivity (RPS) and outputs Q_t.

Answerer. The Answerer responds to Q_t with an answer A_t based on the target person. In real use, this role is played by the user; in our experiments, answers are simulated by a uniform VLM-based protocol applied to all methods under the same interaction setting [13, 14, 15].

Fuser. After receiving A_t, SCOUT appends the latest QA pair and updates the dialogue context as $\mathcal{D}_t = \{q_0, (Q_1, A_1), \ldots, (Q_t, A_t)\}$. The Fuser then reconstructs the refined query q_t in single-shot mode, conditioned on the updated context:

$$q_t = \mathcal{F}(\mathcal{D}_t). \tag{1}$$

Rather than appending the new answer to q_{t-1}, which can inflate query length and compound early errors, $\mathcal{F}$ reconstructs the description from the full dialogue context each round, keeping the output compatible with the frozen encoder. The prompt for $\mathcal{F}$ is in Appendix D.

3.2 Perturbation-Sensitive Question Selection

The Questioner is the core component that decides what to ask next. It performs Perturbation-Sensitive Question Selection (right of Figure 2): instead of relying on retriever-external signals, such as language priors, visual contrast, or an attribute questionnaire, it probes the frozen retriever with controlled query perturbations, following the broader idea of using input perturbations to examine model behavior [58, 59]. We write a general RPS function as $\Psi(q, q')$, which measures how much changing q into q' affects the retriever's current output state. This follows the interactive-retrieval principle of choosing feedback by expected utility or informativeness [40]. At round t, the Questioner constructs perturbation variants from the current query q_{t-1}, computes their RPS values using Ψ, and instantiates the most sensitive attribute or region as the next question Q_t.

Constructing Perturbed Query Variants. To evaluate the retrieval value of any attribute, we simulate what the retriever would see if that attribute were added to or removed from the current query. We first characterize the initial attribute coverage. An attribute extractor parses q_0 to identify mentioned attributes, initializing the candidate attribute space into *known attributes* (present in q_0) and *exploratory attributes* (absent from q_0). For a known attribute c, a rewriter removes the corresponding descriptive phrase from the current query q_{t-1}, yielding a *removal variant* $q_{t-1}^{(-c)}$. For an exploratory attribute p, the rewriter appends a standardized phrase, yielding an *insertion variant* $q_{t-1}^{(+p)}$. The rewriter preserves grammatical fluency and semantic coherence, reducing the chance that the measured sensitivity reflects rewriting artifacts rather than the intended attribute perturbation.

Computing Retrieval Perturbation Sensitivity. Let $\Psi(q, q')$ denote an RPS function that measures how strongly the perturbation from q to q' changes the retriever's output. Each variant is re-encoded by $\mathcal{R}$ and compared with the original query using Ψ:

$$u_{\text{remove}}(c;\, q_{t-1}) := \Psi(q_{t-1},\, q_{t-1}^{(-c)}), \quad u_{\text{insert}}(p;\, q_{t-1}) := \Psi(q_{t-1},\, q_{t-1}^{(+p)}). \tag{2}$$

A high u_{remove} suggests that the current ranking is sensitive to the known attribute c, so clarifying its missing modifiers may be useful. A high u_{insert} suggests that information about the exploratory attribute p could reshape the retriever's local preference. For exploratory attributes, the Questioner does not directly ask the single highest-RPS attribute, since many absent attributes may also be absent from the target. Instead, it aggregates their RPS values at the body-region level:

$$U_\Psi(r) = \frac{\sum_{p \in r} u_{\text{insert}}(p;\, q_{t-1})}{\sqrt{|r|}}, \tag{3}$$

where r denotes a body region and $|r|$ is the number of exploratory attributes there. The square-root normalization suppresses count-driven inflation while still rewarding regions with high sensitivity.

Question Instantiation. The Questioner instantiates Q_t according to the selected candidate type. For known attributes, it asks a color clarification question for the highest-RPS attribute whose color modifier is missing, since color is a prominent and readily interpretable cue in text-based person retrieval [60]. For exploratory attributes, it selects the region with the highest aggregated sensitivity $U_\Psi(r)$ and asks a descriptive question about that region. Attributes or regions already queried in previous rounds are masked to prevent repetition. When high-sensitivity attribute questions are exhausted while the fixed interaction budget remains, the Questioner asks a coarse open-ended question rather than spending the round on low-sensitivity attributes.

3.3 Retrieval Perturbation Sensitivity

The Questioner uses Perturbation-Sensitive Question Selection to select questions by comparing how candidate query perturbations affect the retriever. We formalize this as Retrieval Perturbation

Sensitivity (RPS), a family of functions mapping (q, q') to a scalar sensitivity value. We use Distributional RPS (D-RPS) as the default instantiation. For a query q, let $s_i(q) = \mathrm{sim}(\phi_t(q), \phi_v(I_i))$, and let $\mathrm{Top}_K(q)$ be the K highest-scoring gallery images; for q and q', the shared local support is

$$\mathcal{U}_K(q, q') = \mathrm{Top}_K(q) \cup \mathrm{Top}_K(q'). \tag{4}$$

3.3.1 Distributional RPS

D-RPS compares normalized local preference distributions over $\mathcal{U}_K(q, q')$. For any item in $\mathcal{U}_K$ absent from $\mathrm{Top}_K(q)$, we assign the K-th score under q as a conservative tail estimate; the same rule is applied to q'. We then apply z-normalization and softmax over $\mathcal{U}_K$ to obtain distributions $P_K(q)$ and $P_K(q')$. The default RPS function is defined as

$$\begin{aligned} \Psi(q, q') = \text{D-RPS}(q, q') &= \sqrt{\mathrm{JSD}(P_K(q) \,\|\, P_K(q'))}, \\ \mathrm{JSD}(P\|Q) = \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M), &\quad M = \frac{1}{2}(P+Q). \end{aligned} \tag{5}$$

D-RPS has three immediate properties: affine invariance (z-normalization removes dependence on positive affine score transformations); metric structure ($\sqrt{\mathrm{JSD}}$ satisfies the triangle inequality); and distribution faithfulness (D-RPS is computed before rank compression, retaining the local preference distribution).

Theoretical properties. We analyze D-RPS from three perspectives relevant to question selection: retrieval-event probability change, local sensitivity geometry, and an information-theoretic view of question value. The following three theorems formalize these properties and justify D-RPS as the default RPS function. Full proofs are in Appendix B.

Theorem 1 *(Event-level probability control).* For any $A \subseteq \mathcal{U}_K$:

$$|P_K(q)(A) - P_K(q')(A)| \le \sqrt{2} \cdot \text{D-RPS}(q, q'). \tag{6}$$

Proof sketch. $|P(A) - Q(A)| \le \mathrm{TV}(P, Q) \le \sqrt{2\,\mathrm{JSD}(P\|Q)}$ by probability-metric and Pinsker-type bounds [61]. Thus, a small D-RPS guarantees limited probability variation, whereas a large D-RPS permits substantial redistribution of probability mass over the candidate set.

Theorem 2 *(Fisher-geometric interpretation).* Let $p = \mathrm{softmax}(\ell)$ and $\ell + \eta$ a small perturbation:

$$\text{D-RPS}(q, q')^2 = \frac{1}{8}\eta^\top(\mathrm{Diag}(p) - pp^\top)\eta + o(\|\eta\|^2). \tag{7}$$

Proof sketch. First-order softmax expansion gives $p' = p + (\mathrm{Diag}(p) - pp^\top)\eta + o(\|\eta\|)$; substituting into the second-order JSD expansion yields the result. Since $\mathrm{Diag}(p) - pp^\top$ is the categorical Fisher information matrix [62], D-RPS measures the local Fisher-weighted change in preference along η.

Theorem 3 *(Information-gain connection).* Suppose attribute $A \in \mathcal{A}$ with prior $\pi(a)$ induces conditional distribution P_a over $\mathcal{U}_K$, and let $M = \sum_a \pi(a) P_a$. Then:

$$\Delta L = H(M) - \sum_a \pi(a) H(P_a) = \mathrm{JSD}_\pi(\{P_a\}) = I(A; X). \tag{8}$$

Under a binary uniform prior, $\text{D-RPS}(q, q')^2$ equals the information gain of querying the attribute, giving an information-theoretic interpretation of perturbation sensitivity.

3.3.2 Analysis of Alternative RPS Measures

To motivate D-RPS as SCOUT's default measure, we discuss two simpler ways to instantiate RPS by first compressing the retriever output. Set-RPS measures top-K membership change:

$$\text{Set-RPS}(q, q') = 1 - \frac{|\,\mathrm{Top}_K(q) \cap \mathrm{Top}_K(q')|}{K}. \tag{9}$$

It is equivalent to top-K overlap change, yielding only $K{+}1$ discrete values and ignoring score redistribution within an unchanged top-K set. Rank-RPS measures positional displacement over $\mathcal{U}_K$:

$$\text{Rank-RPS}(q, q') = \frac{1}{K(K+1)} \sum_{u \in \mathcal{U}_K} |\bar{r}_q(u) - \bar{r}_{q'}(u)|, \tag{10}$$

where $\bar{r}_q(u) = \mathrm{rank}_q(u)$ if $u \in \mathrm{Top}_K(q)$ and $K{+}1$ otherwise; $K(K{+}1)$ is the exact maximum displacement. Rank-RPS is richer than Set-RPS, but remains purely ordinal.

Both variants compress the score vector into ranks or top-K membership; by data processing, such compression can only discard information. Two examples illustrate failure modes.

Proposition 1 *(Within-top redistribution is invisible).* $K{=}3$, $P{=}(1{-}3\epsilon, 2\epsilon, \epsilon)$, $Q{=}(0.5, 0.3, 0.2)$, same ranking, $\epsilon \to 0$: $\text{Set-RPS} = \text{Rank-RPS} = 0$, yet $\mathrm{JSD}(P\|Q) \to \frac{3}{4}\ln\frac{4}{3} > 0$.

Proposition 2 *(Rank permutations exaggerate negligible changes).* $\ell{=}(1, \epsilon, 0)$, $\ell'{=}(1, 0, \epsilon)$: rankings differ so $\text{Rank-RPS} > 0$, while $\mathrm{JSD}(\mathrm{softmax}(\ell)\|\mathrm{softmax}(\ell')) = O(\epsilon^2) \to 0$.

Thus, Set-RPS and Rank-RPS can miss large distributional shifts, while Rank-RPS can also overstate negligible score changes. D-RPS avoids both by operating on the normalized local preference distribution before rank compression. Empirical validation is in Section 4.2.

4 Experiment

We conduct experiments on three standard TIReID benchmarks—CUHK-PEDES [3], ICFG-PEDES [63], and RSTPReid [64]—with three retrievers, IRRA [6], RDE [7], and HAM [11], and two interactive baselines, ICL [15] and LLaVA-ReID [14]. Implementation details, evaluation metrics, and the unified five-round evaluation protocol are provided in Appendix A.

4.1 Main Results

Table 1 and Figure 3 summarize the overall performance. Multi-round interaction consistently improves over the corresponding single-shot backbone, showing that targeted follow-up questions provide information missing from the initial description. The effect is strong enough to change the relative ordering of retrievers: on CUHK-PEDES, interactive IRRA variants can surpass stronger single-shot backbones, indicating that effective interaction can compensate for weaker initial retrieval.

Among interactive methods, SCOUT achieves the strongest and most consistent performance, obtaining the best Rank-1 in 8 of 9 backbone–dataset settings. It consistently improves over the corresponding single-shot retrievers and also delivers clear gains in mAP, indicating improvements to the overall gallery ranking. For example, SCOUT improves Rank-1 by up to 12.15% on CUHK-PEDES. The only Rank-1 exception occurs on RSTPReid with IRRA, where SCOUT is slightly below LLaVA-ReID, while remaining clearly superior with RDE and HAM.

The round-wise curves in Figure 3 further reveal how different interaction strategies accumulate evidence. ICL exhibits a delayed-gain pattern, with little first-round improvement because its anchor-based procedure primarily updates high-confidence top-1 cases before harder queries benefit from later rounds. LLaVA-ReID and SCOUT improve earlier, as both can ask discriminative questions without waiting for anchor confirmation. SCOUT shows the strongest cumulative improvement on CUHK-PEDES and ICFG-PEDES, and remains within a narrow margin of LLaVA-ReID on RSTPReid. This suggests that perturbation-sensitive selection provides a stable way to accumulate retrieval-relevant evidence across rounds, rather than merely producing a one-shot correction.

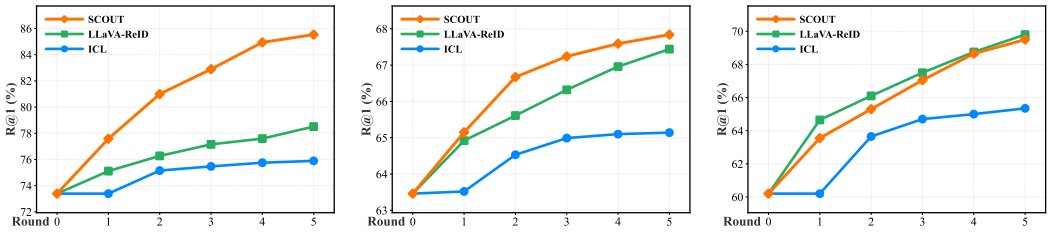
4.2 Ablation Study

We conduct a series of analyses on CUHK-PEDES with IRRA to evaluate the effectiveness and robustness of SCOUT, including the question-selection strategy, the RPS instantiation, the support size K, the realized utility of selected questions, and the robustness to missing feedback.

Question selection strategy. Table 2 isolates the question-selection strategy while holding all other components fixed. Random selection (+4.07) shows that interaction alone is not sufficient without a useful priority signal. Look-forward, which transplants LLaVA-ReID's Selector and Questioner [14], reaches only +4.77, suggesting that utility learned offline from visual candidates may not fully match the current retriever state. Fixed Questionnaire, using ICL's 15-question list [15], improves to +7.87, but may spend rounds on attributes already specified in the current query or weakly related to the current retrieval state. History-Based LLM, following the dialogue-history-driven design of ChatPR [13], further improves to +9.66 by adding redundancy awareness, yet still lacks a direct measure of retrieval impact. SCOUT achieves the best result on every metric (+12.15 Rank-1),

Table 1: Retrieval performance on CUHK-PEDES, ICFG-PEDES, and RSTPReid under a five-round interaction protocol. "+" denotes applying an interactive method on top of the retrieval backbone.

Methods	Venue	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Text-to-image person retrieval methods													
APTM[5]	ACM MM'23	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
RaSa[28]	IJCAI'23	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	66.90	86.50	91.35	52.31
TBPS-CLIP[65]	AAAI'24	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	62.10	81.90	87.75	48.00
CFAM[34]	CVPR'24	75.60	90.53	94.36	67.27	65.38	81.17	86.35	39.42	62.45	83.55	91.10	49.50
Bi-IRRA[66]	TPAMI'25	79.43	92.59	95.68	70.51	70.36	83.86	88.47	43.28	72.50	88.15	92.45	57.32
Interactive person retrieval methods													
IRRA[6]	CVPR'23	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
+ICL[15]	CVPR'25	75.89	89.44	93.62	67.28	65.14	80.11	85.63	38.46	65.35	81.35	88.10	48.13
+LLaVA-ReID[14]	ICML'25	78.51	92.43	95.67	70.61	67.44	82.91	87.69	40.60	69.85	88.10	92.55	54.92
+SCOUT	Ours	85.53	95.14	97.25	75.66	67.84	83.61	88.31	40.54	69.50	87.75	92.35	53.35
RDE[7]	CVPR'24	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
+ICL[15]	CVPR'25	77.91	90.27	94.14	69.13	69.02	82.45	87.36	41.21	70.55	85.95	91.65	53.68
+LLaVA-ReID[14]	ICML'25	79.39	93.02	95.96	71.70	69.98	84.32	88.60	42.12	72.00	88.95	93.55	56.06
+SCOUT	Ours	86.84	95.39	97.45	76.51	71.36	85.40	89.61	42.43	76.75	90.35	94.20	58.52
HAM[11]	CVPR'25	77.99	91.34	95.03	69.72	69.95	83.88	88.39	42.72	72.50	87.70	91.95	55.47
+ICL[15]	CVPR'25	79.11	91.55	94.74	70.98	70.95	84.28	88.68	44.01	73.90	86.85	91.60	55.99
+LLaVA-ReID[14]	ICML'25	80.95	92.53	95.52	73.34	71.38	84.04	88.43	45.90	75.30	88.20	93.40	60.16
+SCOUT	Ours	**88.94**	**96.43**	**97.87**	**79.11**	**75.46**	**87.74**	**91.49**	**46.39**	**83.75**	**93.30**	**96.00**	**63.20**



(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 3: Round-wise Rank-1 curves of IRRA across three benchmarks and three interactive methods.

indicating that its advantage comes from Perturbation-Sensitive Question Selection rather than random choice, visually grounded offline utility, attribute-questionnaire coverage, or dialogue history.

Sensitivity to K. Figure 4 shows how retrieval performance varies with the local support size K. Three regimes are visible: at K=1, the support is too narrow to capture stable local preference changes, making the sensitivity estimate less reliable; from K=2 to 8, performance rises rapidly as the support becomes large enough to form a meaningful distribution; from $K \geq 12$, performance plateaus with variation below 0.12 Rank-1 across $K \in \{12, 16, 32, 48\}$. We adopt K=12 as the default: smaller values underestimate sensitivity due to insufficient support, while larger values increase computation without improving signal quality. The plateau confirms that D-RPS is robust to the precise choice of K once a moderate support size is reached.

RPS instantiation. Figure 5 compares three RPS instantiations within SCOUT at K=12: Set-RPS, Rank-RPS, and D-RPS. D-RPS achieves the highest Rank-1 (85.53) and mAP (75.66), while the set- and rank-level alternatives fall short, validating the distributional perspective developed in Section 3.3. Figure 5(b–d) explains this gap using three diagnostics: score diversity measures how many candidate attributes receive distinguishable scores, rank stability measures whether attribute rankings remain consistent across K, and selection clarity measures whether the top-RPS candidate is uniquely determined. D-RPS saturates near 100% across all three dimensions well before the alternatives, while Set-RPS fails to distinguish candidates on 32–47% of queries and Rank-RPS never consistently reaches $\tau_B > 0.9$ in the K=1–100 range. Consistent with the analysis in Section 3.3, set- and

Table 2: Analysis of question selection strategy

Question Selection Strategy	R@1	R@5	R@10	mAP
baseline	73.38	89.93	93.71	66.13
Random	77.45	91.68	94.95	68.87
Look-forward(LLaVA-ReID)	78.15	91.47	94.72	70.18
Fixed Questionnaire(ICL)	81.25	94.00	96.90	73.06
History-Based LLM(ChatPR)	83.04	94.13	96.77	73.38
Perturbation-Sensitive(Ours)	**85.53**	**95.14**	**97.25**	**75.66**

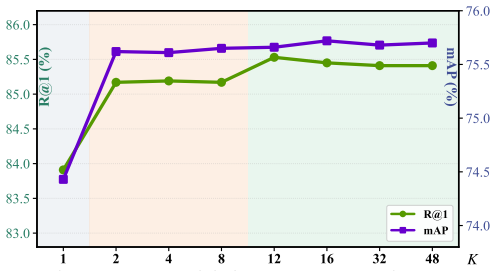
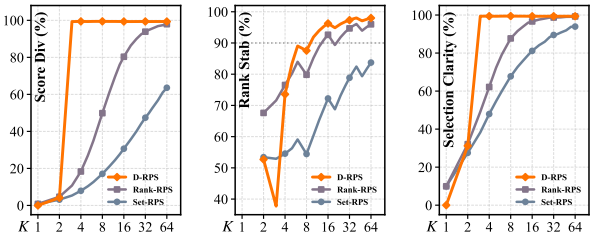


Figure 4: Sensitivity to support size K.

RPS Type	Div↑	Stab↑	Clar↑	R@1	mAP
Set-RPS	24.43	65.26	76.10	84.99	75.41
Rank-RPS	69.42	88.81	93.99	85.07	75.44
D-RPS	**99.40**	**93.95**	**99.40**	**85.53**	**75.66**

(a) Comparison of RPS instantiations.



(b) Score diversity (c) Rank stability (d) Selection Clarity

Figure 5: Analysis of RPS instantiations on CUHK-PEDES with IRRA.

Table 3: Analysis of realized question utility

Strategy	Avg. ΔRR ↑	Improve ↑	Unchanged ↓	Regress ↓
Random	0.09	42.7%	43.7%	13.6%
D-RPS	**0.14**	**68.0%**	**22.0%**	**10.0%**
Gain	**+0.05**	**+25.3%**	**-21.7%**	**-3.6%**

Table 4: Robustness to missing feedback

Available Answers	R@1	mAP	Gain (R@1 / mAP)
3	82.20	73.10	+8.82 / +6.97
4	83.02	74.08	+9.64 / +7.95
5	**85.53**	**75.66**	+12.15 / +9.53

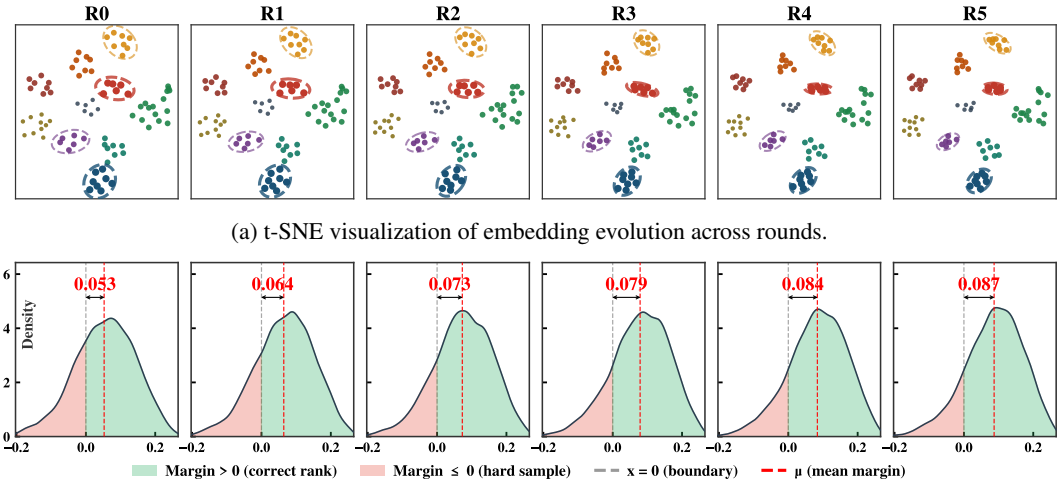
rank-level compression discards distributional information that D-RPS retains, and the gap is large enough to translate into retrieval performance.

Question utility analysis. To evaluate the utility of RPS-selected questions, we compare them with randomly selected feasible questions on 100 challenging CUHK-PEDES queries. After obtaining and incorporating the corresponding answer, we measure the change in reciprocal rank as $\Delta\mathrm{RR} = 1/r_{\mathrm{after}} - 1/r_{\mathrm{before}}$, where a positive value indicates that the target moves toward the top of the ranking. We also report the proportions of queries whose target rank improves, remains unchanged, or regresses. As shown in Table 3, RPS achieves a higher average $\Delta\mathrm{RR}$ than random selection (0.14 vs. 0.09) and increases the proportion of rank-improving queries from 42.7% to 68.0%. Meanwhile, unchanged and regressed cases are both reduced. These results support RPS as an effective pre-answer criterion for prioritizing follow-up questions.

Robustness to missing feedback. To reflect practical interactions where feedback may be unavailable in some rounds, we evaluate SCOUT under missing-feedback settings. Starting from the five-round setting, we drop one or two answers and leave the query unchanged at the corresponding rounds. As shown in Table 4, performance decreases gradually as feedback becomes less complete. Even with only three available answers, SCOUT still maintains a clear improvement over the initial retrieval result. This indicates that the method can tolerate occasional missing responses and does not require feedback to be available at every interaction round.

4.3 Visualization

Figure 6a shows t-SNE projections [67] of query text embeddings for ten identities across rounds R_0 to R_5, sharing a single coordinate space. At R_0, queries for the same identity are scattered; by R_5, they converge into compact clusters, with within-cluster cosine similarity rising from 0.471 to 0.555 (+18.0%), while the retriever encoder remains frozen throughout. Figure 6b tracks the per-query retrieval margin $S^{\max}_{\mathrm{pos}}(q) - S^{\max}_{\mathrm{neg}}(q)$ across rounds: the distribution shifts steadily rightward, with the mean increasing monotonically from μ=0.053 at R_0 to μ=0.087 at R_5. Together, the two



(a) t-SNE visualization of embedding evolution across rounds.

(b) Retrieval margin distribution across rounds.

Figure 6: Representation-level refinement on CUHK-PEDES with IRRA. SCOUT makes same-identity query embeddings more compact and steadily increases retrieval margins over rounds.

views suggest that SCOUT improves retrieval by progressively refining the query representation: same-identity queries become more coherent in embedding space while simultaneously becoming less confusable with hard negatives, rather than merely changing isolated retrieval scores.

5 Conclusion

In this paper, we introduce SCOUT, a training-free framework for interactive person retrieval that selects follow-up questions by directly probing the current retriever. At its core, SCOUT introduces Retrieval Perturbation Sensitivity (RPS) to characterize how candidate attributes affect the retriever's current preference over gallery images, turning this response into a criterion for deciding what information to request next. The distributional formulation of RPS further provides a principled view of this sensitivity by relating changes in the retrieval distribution to candidate discrimination and information gain. SCOUT requires neither an additional question-selection model nor retriever modification, while consistently improving existing text-to-image person retrieval models in interactive settings. Our experimental results further suggest that a retriever's own response can guide interactive information acquisition for resolving incomplete user queries.

6 Limitations

We introduce SCOUT to guide follow-up question selection in interactive person retrieval by leveraging the current retriever's response and state. In this work, the candidate question space is restricted to a predefined set of appearance-related attributes, which provides a controlled and comparable basis for perturbation-based question selection but does not fully cover more open-ended concepts. A broader question space could further improve the flexibility of SCOUT in handling diverse user descriptions and interaction needs. Future work will explore dynamically constructed question spaces while preserving reliable and comparable sensitivity estimation.

7 Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant (2026YFE0202100), the National Natural Science Foundation of China under Grants (T2541022, 62476188, 62361166629 and 62501428), the Major Project of Science and Technology Innovation of Hubei Province under Grant (2025BEA002), the Hubei Provincial Natural Science Foundation of China under Grant (2025AFB219), and the China Postdoctoral Science Foundation under Fellowship (2026T190427). The numerical calculations were supported by the supercomputing system at the Supercomputing Center of Wuhan University.

References

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2872–2893, 2022.

[2] Haitao Shi, Meng Liu, Yongqi Li, Xianjing Han, Yangyuan He, and Liqiang Nie. A survey on text-based person search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2026.

[3] 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*, pages 5187–5196, 2017.

[4] Kai Niu, Yanyi Liu, Yuzhou Long, Yan Huang, Liang Wang, and Yanning Zhang. An overview of text-based person search: Recent advances and future directions. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 7803–7819, 2024.

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

[6] 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*, pages 2787–2797, 2023.

[7] 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*, pages 27197–27206, 2024.

[8] 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*, pages 757–767, 2023.

[9] Wentao Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of MLLMs for transferable text-to-image person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–32, 2024.

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

[11] 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*, pages 9220–9230, 2025.

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

[13] 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*, pages 3952–3962, 2025.

[14] 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 *Proceedings of the 42nd International Conference on Machine Learning*, pages 40868–40887, 2025.

[15] 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*, pages 14390–14399, 2025.

[16] 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.

[17] Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2737–2746, 2018.

[18] Ziliang Zhao, Zhicheng Dou, and Yujia Zhou. Generating intent-aware clarifying questions in conversational information retrieval systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3384–3394, 2024.

[19] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 686–701, 2018.

[20] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. ViTAA: Visual-textual attributes alignment in person search by natural language. In *Proceedings of the European Conference on Computer Vision*, pages 402–420, 2020.

[21] Xiao Wang, Bing Wang, Bin Yang, Cuiqun Chen, Xin Xu, and Mang Ye. Blurring modal boundaries: A unified survey from single-to multi-modal person re-ldentification. *International Journal of Computer Vision*, page 473, 2026.

[22] Chenyang Gao, Guanyu Cai, Xingyu Jiang, Feng Zheng, Jun Zhang, Yanan Gong, Feng Lin, Xing Sun, and Xiang Bai. Conditional feature learning based transformer for text-based person search. *IEEE Transactions on Image Processing*, pages 6097–6108, 2022.

[23] 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*, pages 5566–5574, 2022.

[24] Ding Jiang and Mang Ye. Transformer network for cross-modal text-to-image person re-identification. *Journal of Image and Graphics*, pages 1384–1395, 2023.

[25] Lei Tan, Weihao Li, Pingyang Dai, Jie Chen, Liujuan Cao, and Rongrong Ji. Knowing where to focus: Attention-guided alignment for text-based person search. *International Journal of Computer Vision*, page 72, 2026.

[26] Hao Pan, Jian Dai, Yuan Sun, Zhenwen Ren, and Xingfeng Li. Difference-aware decision learning for multimodal image fusion. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 95745–95762, 2026.

[27] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2617–2625, 2020.

[28] 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*, pages 555–563, 2023.

[29] Yuan Sun, Kaiming Liu, Yongxiang Li, Zhenwen Ren, Jian Dai, and Dezhong Peng. Distribution consistency guided hashing for cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5623–5632, 2024.

[30] 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*, pages 8748–8763, 2021.

[31] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C H Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, pages 9694–9705, 2021.

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900, 2022.

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

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

[35] Xudong Wang, Lei Tan, Pingyang Dai, Liujuan Cao, and Rongrong Ji. Gpt-reid: Learning fine-grained representation with gpt for text-based person retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3585–3594, 2025.

[36] Fangyi Liu, Mang Ye, and Bo Du. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, page 28, 2024.

[37] Chen He, Shen Li, Zheng Wang, Hua Chen, Fumin Shen, and Xing Xu. Chatting with interactive memory for text-based person retrieval. *Multimedia Systems*, 2025.

[38] Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24245–24254, 2025.

[39] Yuhao Wang, Min Yang, and Rui Cao. Fine-grained semantic alignment with transferred person-sam for text-based person retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5432–5441, 2024.

[40] 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*, pages 678–688, 2018.

[41] 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*, pages 1835–1844, 2021.

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

[43] 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*, pages 791–809, 2024.

[44] Zhiyu Cao, Peifeng Li, and Qiaoming Zhu. ICR: Iterative clarification and rewriting for conversational search. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9821–9835, 2025.

[45] Yanglin Feng, Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, and Peng Hu. Interactive cross-modal learning for text-3d scene retrieval. In *Advances in Neural Information Processing Systems*, pages 97972–97992, 2025.

[46] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59403–59420, 2024.

[47] Yanzhe Chen, Zhiwen Yang, Jinglin Xu, and Yuxin Peng. Mai: A multi-turn aggregation-iteration model for composed image retrieval. In *International Conference on Learning Representations*, pages 18960–18979, 2025.

[48] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021.

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

[50] Shi Yu, Jiahua Liu, Jingtao Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1933–1936, 2020.

[51] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems*, pages 1–29, 2021.

[52] Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, 2022.

[53] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 475–484, 2019.

[54] Sedrick Scott Keh, Justin T. Chiu, and Daniel Fried. Asking more informative questions for grounded retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4429–4442, 2024.

[55] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8496–8504, 2025.

[56] Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, et al. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, pages 27550–27571, 2026.

[57] Xingfeng Li, Hao Pan, Honglin Yuan, Yuan Sun, Xujian Zhao, Jiaqi Lin, and Zhenwen Ren. Anti-degradation lifelong multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8749–8759, 2026.

[58] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2950–2958, 2019.

[59] Negar Arabzadeh, Radin Hamidi Rad, Maryam Khodabakhsh, and Ebrahim Bagheri. Noisy perturbations for estimating query difficulty in dense retrievers. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3722–3727, 2023.

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

[61] James Melbourne, Saurav Talukdar, Shreyas Bhaban, Mokshay Madiman, and Murti V Salapaka. The differential entropy of mixtures: New bounds and applications. *IEEE Transactions on Information Theory*, pages 2123–2146, 2022.

[62] Alessandro Manenti and Cesare Alippi. Beyond softmax: A natural parameterization for categorical random variables. In *Forty-third International Conference on Machine Learning*, 2026.

[63] 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.

[64] Aichun Zhu, Zhe 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*, pages 209–217, 2021.

[65] 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*, pages 465–473, 2024.

[66] Min Cao, Xuyang Zhou, Ding Jiang, Bo Du, Mang Ye, and Min Zhang. Multilingual text-to-image person retrieval via bidirectional relation reasoning and aligning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1961–1977, 2025.

[67] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, pages 2579–2605, 2008.

Appendix Overview

This supplementary document is organized as follows:

- Sec. A provides experimental details, including datasets, evaluation metrics, implementation settings, interactive baseline adaptation, and runtime analysis.
- Sec. B contains proofs and mathematical details for the propositions and theorems in the main paper.
- Sec. C discusses the broader impact and potential societal implications of interactive person retrieval.
- Sec. D lists the prompts used in SCOUT's LLM/VLM components.
- Sec. E shows qualitative interaction examples and retrieval trajectories.

A Experimental Details

A.1 Datasets

We evaluate on three standard TIReID benchmarks spanning different gallery scales and retrieval conditions.

CUHK-PEDES [3] contains 40,206 images of 13,003 identities, each annotated with two natural-language descriptions averaging 24 words. The test split comprises 3,074 gallery images from 1,000 identities and 6,156 query descriptions. The descriptions follow a free-form narrative style and include substantial cross-camera viewpoint variation.

ICFG-PEDES [63] contains 54,522 images of 4,102 identities. The test split contains 19,848 image-text pairs from 1,000 identities. The captions contain about 37 words on average and follow a more structured attribute-list style, making fine-grained discrimination challenging in the large gallery.

RSTPReid [64] contains 20,505 images of 4,101 identities, with 5 images per identity and 2 descriptions per image. The test split comprises 1,000 gallery images from 200 identities and 2,000 queries averaging 27 words. It includes challenging visual conditions such as occlusion, motion blur, and low resolution.

A.2 Evaluation Metrics

Recall@k. Recall@k measures the percentage of queries for which the ground-truth person appears in the retrieved top-k images. We report $k \in \{1, 5, 10\}$.

Mean Average Precision (mAP). mAP measures overall retrieval quality by averaging precision over all rank positions and all queries, computed over the full gallery.

A.3 Implementation Details

All interactive methods are evaluated under the same five-round protocol ($T{=}5$), with one question asked in each round. Answers are generated from visual evidence according to each method's original interaction design. For SCOUT, a VLM answerer observes the target image and responds to the selected question. For LLaVA-ReID, because the standard benchmarks do not provide fine-grained textual annotations, the answerer uses the ground-truth image as visual input, consistent with the authors' benchmark evaluation practice. For ICL, VQA is applied to candidate images to anchor and refine retrieval. This keeps all compared methods grounded in visual evidence rather than privileged textual annotations.

All retrievers are kept frozen throughout interaction. We use publicly released checkpoints for IRRA [6] and RDE [7]. Since HAM [11] does not release pretrained checkpoints, we train it following the official implementation and obtain performance consistent with the original paper. Because SCOUT reconstructs the full description at each round, query length remains within each retriever's original maximum token limit without positional embedding interpolation.

For SCOUT, the local support size is fixed to $K{=}12$ throughout, as motivated by the sensitivity analysis in the main paper. Already queried attributes are masked in later rounds to prevent repetition. The attribute extractor, rewriter, fuser, and answer simulator all use Gemini-2.5-Flash; their prompts are listed in Appendix D. All experiments are run on a single NVIDIA RTX 3090 GPU (24 GB) with PyTorch 2.10.0+cu128.

A.4 Interactive Baseline Adaptation

We adapt interactive baselines only to provide a unified evaluation interface across the three standard benchmarks, retrieval backbones, and five interaction rounds. Following the official implementations whenever available, we keep the core question-selection mechanism of each method unchanged and only make the minimal changes

needed for dataset loading, feature extraction, and visual answer generation under our evaluation protocol. For methods requiring visual answers, the same visual-evidence setting is used as described in Sec. A.3. We verify the adapted pipelines against reported settings before extending them to additional retrieval backbones.

A.5 Runtime Analysis

Under the unified five-round protocol, SCOUT takes approximately 2 hours on CUHK-PEDES, 6 hours on ICFG-PEDES, and 1 hour on RSTPReid. These numbers include retriever re-encoding, LLM API calls, and VLM answer simulation.

SCOUT remains practical for full-benchmark evaluation: even on ICFG-PEDES, the largest gallery among the three benchmarks, a complete five-round run finishes within a few hours on a single GPU. The main overhead comes from scoring candidate perturbations with the frozen retriever and calling LLM/VLM components for text processing and answer simulation. Since SCOUT does not train a questioner or modify the retriever, its cost is limited to inference-time query perturbation and re-encoding.

B Proofs and Mathematical Details

This appendix provides complete proofs for the propositions and theorems in Section 3.3. We use the notation of the main paper throughout. For query q and perturbed variant q', let $\mathcal{U}_K(q, q') = \mathrm{Top}_K(q) \cup \mathrm{Top}_K(q')$ denote the local support and let $P_K(q)$, $P_K(q')$ be the corresponding local preference distributions, so that $\text{D-RPS}(q, q') = \sqrt{\mathrm{JSD}(P_K(q) \| P_K(q'))}$.

B.1 Auxiliary Definitions

For distributions P, Q on a common finite support $\mathcal{U}$:

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}\,\mathrm{KL}(P\|M) + \frac{1}{2}\,\mathrm{KL}(Q\|M), \quad M = \frac{1}{2}(P+Q),$$

$$\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1.$$

By a Pinsker-type bound for Jensen–Shannon divergence:

$$\mathrm{TV}(P,Q) \le \sqrt{2\ \mathrm{JSD}(P\|Q)}.$$

B.2 Proof of Proposition 1

Statement. Let $K{=}3$, $P{=}(1{-}3\epsilon, 2\epsilon, \epsilon)$, $Q{=}(0.5, 0.3, 0.2)$, same ranking, $\epsilon \to 0$. Then the corresponding Set-RPS and Rank-RPS values are zero, yet $\mathrm{JSD}(P\|Q) > 0$.

Proof. For sufficiently small $\epsilon > 0$, both P and Q satisfy the strict order $1 \succ 2 \succ 3$. Since Set-RPS depends only on top-K membership and Rank-RPS only on rank displacement, both are zero when the ranking is unchanged. As $\epsilon \to 0$, $P \to (1,0,0)$ while $Q = (0.5, 0.3, 0.2)$, so $\mathrm{JSD}(P\|Q) \to \mathrm{JSD}((1,0,0)\|(0.5,0.3,0.2)) = \frac{3}{4}\ln\frac{4}{3} > 0$. The set- and rank-level signals vanish while the distributional shift remains substantial.

B.3 Proof of Proposition 2

Statement. Let $\ell{=}(1, \epsilon, 0)$, $\ell'{=}(1, 0, \epsilon)$, $\epsilon \to 0$. Rankings differ, so $\text{Rank-RPS}(q, q') > 0$, yet $\mathrm{JSD}(\mathrm{softmax}(\ell)\|\,\mathrm{softmax}(\ell')) = O(\epsilon^2) \to 0$.

Proof. For any $\epsilon > 0$, ℓ and ℓ' differ only in the ordering of their second and third entries, and therefore induce different rank permutations, yielding $\text{Rank-RPS}(q, q') > 0$.

As $\epsilon \to 0$, both logits converge to the same vector $(1, 0, 0)$. By continuity of the softmax function,

$$\mathrm{softmax}(\ell),\ \mathrm{softmax}(\ell') \to p_0 := \mathrm{softmax}(1,0,0) = \left(\frac{e}{e+2}, \frac{1}{e+2}, \frac{1}{e+2}\right).$$

Moreover, since softmax is smooth,

$$\mathrm{softmax}(\ell) = p_0 + O(\epsilon), \qquad \mathrm{softmax}(\ell') = p_0 + O(\epsilon),$$

and hence

$$\mathrm{softmax}(\ell) - \mathrm{softmax}(\ell') = O(\epsilon).$$

Since JSD is second-order around equality,

$$\mathrm{JSD}(\mathrm{softmax}(\ell)\|\,\mathrm{softmax}(\ell')) = O(\epsilon^2) \to 0.$$

Thus, Rank-RPS remains nonzero even when the underlying distributions become arbitrarily close.

B.4 Proof of Theorem 1

Statement. For any $A \subseteq \mathcal{U}_K(q, q')$:

$$|P_K(q)(A) - P_K(q')(A)| \le \sqrt{2}\ \text{D-RPS}(q, q').$$

Proof. For any distributions P, Q and event A:

$$|P(A) - Q(A)| \le \text{TV}(P, Q) \le \sqrt{2\ \text{JSD}(P\|Q)}.$$

Substituting $P = P_K(q)$ and $Q = P_K(q')$ and using $\text{D-RPS}(q, q') = \sqrt{\text{JSD}(P_K(q)\|P_K(q'))}$ gives:

$$|P_K(q)(A) - P_K(q')(A)| \le \sqrt{2}\ \text{D-RPS}(q, q'). \qquad \square$$

B.5 Proof of Theorem 2

Statement. Let $p = \text{softmax}(\ell)$ and $\ell + \eta$ a small perturbation. Then:

$$\text{D-RPS}(q, q')^2 = \frac{1}{8}\,\eta^\top \big(\text{Diag}(p) - pp^\top\big)\eta + o(\|\eta\|^2).$$

Proof. The Jacobian of the softmax map at ℓ is $J(\ell) = \text{Diag}(p) - pp^\top$, giving $p' = p + J(\ell)\eta + o(\|\eta\|)$. The second-order expansion of JSD around $p' = p$ is:

$$\text{JSD}(p\|p') = \frac{1}{8}(p' - p)^\top\ \text{Diag}(1/p)(p' - p) + o(\|p' - p\|^2).$$

Since $p' - p = J(\ell)\eta + o(\|\eta\|)$, we compute:

$$(p' - p)^\top\ \text{Diag}(1/p)(p' - p) = \eta^\top J(\ell)^\top\ \text{Diag}(1/p)\ J(\ell)\,\eta + o(\|\eta\|^2).$$

Using $J(\ell) = \text{Diag}(p) - pp^\top$ and the identity $J(\ell)^\top\ \text{Diag}(1/p) J(\ell) = J(\ell)$:

$$\eta^\top J(\ell)^\top\ \text{Diag}(1/p)\ J(\ell)\,\eta = \eta^\top J(\ell)\,\eta = \eta^\top(\text{Diag}(p) - pp^\top)\eta.$$

Substituting back gives $\text{JSD}(p\|p') = \frac{1}{8}\eta^\top(\text{Diag}(p) - pp^\top)\eta + o(\|\eta\|^2)$. Since $\text{D-RPS}^2 = \text{JSD}$, the stated expansion follows.

B.6 Proof of Theorem 3

Statement. Suppose attribute $A \in \mathcal{A}$ with prior $\pi(a)$ induces conditional distribution P_a over $\mathcal{U}_K$, and let $M = \sum_a \pi(a)P_a$. Then:

$$\Delta L = H(M) - \sum_a \pi(a)H(P_a) = \text{JSD}_\pi(\{P_a\}) = I(A; X).$$

Under a binary uniform prior, $\text{D-RPS}(q, q')^2$ equals the information gain of querying the corresponding attribute.

Proof. Without querying A, the optimal predictor uses the mixture M with expected log-loss $H(M)$. After observing $A{=}a$, the predictor uses P_a, giving expected log-loss $\sum_a \pi(a)H(P_a)$. The reduction is:

$$\Delta L = H(M) - \sum_a \pi(a)H(P_a).$$

By the entropy decomposition of weighted JSD:

$$\text{JSD}_\pi(\{P_a\}) = H\Big(\sum_a \pi(a)P_a\Big) - \sum_a \pi(a)H(P_a) = \Delta L.$$

The mutual information identity $I(A; X) = H(X) - H(X|A)$ with $H(X) = H(M)$ and $H(X|A) = \sum_a \pi(a)H(P_a)$ gives $I(A; X) = \Delta L$. Under a binary uniform prior, JSD_π reduces to the two-distribution JSD, so $\text{D-RPS}(q, q')^2 = \text{JSD}(P_K(q)\|P_K(q'))$ equals the information gain.

C Broader Impact

Interactive person retrieval can make human-in-the-loop search more flexible by allowing incomplete descriptions to be progressively refined through follow-up questions. However, person retrieval is privacy-sensitive and may raise concerns regarding surveillance, consent, and fairness in practical deployment. Our experiments use public research benchmarks and do not collect new personal data. SCOUT is intended for research on interactive retrieval and should not be interpreted as endorsing unauthorized or privacy-invasive use.

D Prompts

SCOUT relies on four prompt-based components, all implemented with Gemini-2.5-Flash. The **attribute extractor** receives the initial description q_0 and returns a structured list of mentioned attributes, partitioning the attribute space into known and exploratory sets (prompt in Table 5). The **rewriter** takes the current query and a target attribute, producing either a removal variant $q^{(-c)}$ or an insertion variant $q^{(+p)}$ while keeping the output fluent and encodable by the frozen retriever (prompt in Table 6). The **answerer** receives the target image and the current question, then generates a natural-language answer as if from a user (prompt in Table 7). The **fuser** receives the dialogue context $\mathcal{D}_t$ and reconstructs a refined query q_t from scratch in a single LLM pass, avoiding length inflation and error compounding from naive concatenation (prompt in Table 8).

Table 5: Prompt for attribute extraction.

You are an information extractor for person appearance attributes.

We extract ONLY 6 categories:

- **upper**: tops/jackets/coats/shirts/hoodies/sweaters/dresses etc.
- **lower**: pants/jeans/shorts/skirts etc.
- **footwear**: shoes/sneakers/boots/sandals etc.
- **accessory**: bags/backpacks/purses/hats/glasses/belts/ties/watches AND carried/held objects/props like umbrella/phone/bottle/cup/jug/book/suitcase/shopping bag etc.
- **hair**: hair color/length/style; ponytail/bangs etc.
- **body**: canonical items only (gender, age, height, build, facial_hair, race, ethnicity, skin_tone, etc.)

INPUT SENTENCE:
"{sentence}"

TASK:
Return ALL appearance-related items and traits as STRUCTURED objects that are robust to different phrasings. Do NOT rely on a single "COLOR ITEM" surface form. Correctly attach modifiers even if expressed as predicates.

OUTPUT:
A JSON array. Each element MUST follow this schema:

```
[{
   "category": "upper" | "lower" | "footwear" | "accessory" | "hair" | "body",
   "item": string | null,
   "modifiers": {
      "color": string | null,
      "pattern": string | null,
      "material": string | null,
      "style": string | null,
      "placement": string | null
   },
   "evidence_phrases": [string],
}]
```

CRITICAL RULES:

1) Focus ONLY on appearance: clothing, footwear, accessories, hair, body traits. Ignore actions/locations/backgrounds.
2) Extract ITEMS FIRST. One concrete item noun = one object (e.g., pants, shoes, belt, tie, backpack, glasses, jacket).
3) Attach modifiers to the correct item even if expressed differently:
   - "black pants" → item=pants, color=black
   - "pants are black" → item=pants, color=black

3b) Placement modifiers: If described with a wearing/carrying placement (e.g., "draped over...", "on his shoulder"), record that phrase in `modifiers.placement`.

3c) Held/Carried objects rule: Any noun phrase after "carrying/holding/with/in hand/under arm" should be extracted as an accessory item. Split multiple carried objects into multiple objects.

3d) If a sentence indicates someone is carrying/holding "something/an object" that is not a specific thing, still identify it as an accessory object.

4) Coordination sharing a modifier MUST be split:
   - "shoes and belt are black" → TWO objects: shoes(color=black) + belt(color=black)

5) If an item has NO modifiers, still extract it with modifiers all null.

6) `evidence_phrases`: Use short phrases copied from the sentence. Include at least the item word and, if present, the phrase that expresses its color/pattern.

IMPORTANT: Return ONLY valid JSON. No explanations. No markdown.

Table 6: Prompt for query rewriting.

You are a careful editor for person appearance descriptions used for retrieval.

GOAL:
COMPLETELY REMOVE the TARGET ATTRIBUTE (item + all its modifiers) from the sentence. Keep all other attributes UNCHANGED. Fix grammar at the deletion point so the sentence reads naturally.

INPUT SENTENCE:
"{sentence}"

TARGET ATTRIBUTE TO DELETE:

- **category**: {category} (one of upper/lower/footwear/accessory)
- **item**: {item} (core noun, e.g., "shirt", "pants", "shoes", "belt", "bag")
- **modifiers**: color: {color}, pattern: {pattern}, material: {material}, style: {style}, placement: {placement}
- **evidence_phrases** (may appear in the sentence): {evidence_phrases}

STRICT RULES:

1) DELETE the entire mention of the target item AND all its modifiers from the sentence.
2) Do NOT add new details or change other attributes.
3) You MAY adjust a small local region around the deletion point to fix grammar:
   - edit nearby articles (a/the), conjunctions (and/or), punctuation, and short connectors.
   - you MAY split one clause into two short clauses if needed.
4) If the target item shares a clause with other attributes (e.g., "X and Y are COLOR"), restructure only minimally to preserve non-target info.

Table 7: Prompt for answer simulation.

You are a user looking at a photo and answering a question about a specific person in it.

{context_block}

IMPORTANT — HANDLING UNCLEAR IMAGES:
Some images may be blurry, small, or partially obscured. When visual details are ambiguous or hard to make out:

- Prefer "not sure" over guessing.
- Do NOT confidently deny something just because you cannot clearly see it. "Cannot see it" $\neq$ "It is not there."
- If you can see a vague shape or color but are not confident, describe it briefly and add "but not sure" (e.g., "looks dark, maybe black, but not sure").
- If a body part or item is NOT VISIBLE in the image (e.g., legs/feet are cropped out, or the person is seen from behind and the front is hidden), you MUST say "not sure". Do NOT guess details for parts you cannot see at all.

RULES:

1) Answer only the asked question, briefly and naturally (1-10 words preferred).
2) For yes/no or confirmation questions, start with "Yes" or "No", then add brief details if needed.
3) For color questions, state the dominant/primary color you see. If multiple colors are clearly visible, mention the main colors briefly. Do NOT give uncertain alternatives like "Blue or black".
4) For type questions, use common general terms (e.g., "shoes" not "Nike running shoes").
5) If you truly cannot see or determine something, say "not sure".
6) If the question refers to a specific item (e.g., "the backpack") but you see a different item instead (e.g., a handbag), clarify what you actually see rather than answering about the wrong item.
7) Be honest about what you see, even if it seems to contradict the description.
8) Focus ONLY on the described person, ignore other people and background.
9) Focus on what the question is actually asking about (the attribute or object), NOT on subject words like "woman", "man", or "person". These are just references to identify WHO is being asked about, not WHAT is being asked. For example, if asked "What is the woman holding?", answer about the held object, not about whether the person is a woman.

Question: {question}

Answer:

Table 8: Prompt for the Fuser.

You are a careful editor for person appearance descriptions.

TASK:
You are given an ORIGINAL DESCRIPTION and ALL the Q&A evidence collected across multiple dialog rounds. Produce a single, refined description that integrates confirmed information from the Q&A pairs into the original description while keeping the description faithful and concise.

QUESTION TYPES (for your reference):

- **[SUB]** — Asks about a specific attribute, such as color, of an item already in the description.
- **[ADD]** — Asks about a body region. The answer may re-describe items already in the original using different words, and/or mention new items.
- **[GENERAL]** — Open-ended question. Treat same as **[ADD]**.

PRINCIPLE — VOCABULARY CONSISTENCY:
The ORIGINAL DESCRIPTION provides the base wording for the person description. Q&A answers may use more generic or slightly different terms for the same visible item (e.g., "coat" → "jacket", "sneakers" → "shoes", "trousers" → "pants"). To keep the refined description consistent with the original, preserve the original clothing and accessory nouns unless the Q&A evidence clearly indicates a genuinely new item or an explicit correction.

ORIGINAL DESCRIPTION:
"{initial_query}"

ALL COLLECTED Q&A EVIDENCE:
{qa_text}

RULES:

1) Do not replace clothing/accessory nouns from the ORIGINAL with different words unless the Q&A evidence clearly indicates a correction.
2) You may add new modifiers, such as color, length, sleeve-style, or pattern, to existing nouns when supported by Q&A evidence. For example, "coat" → "long dark coat" is allowed.
3) You may add genuinely new items for body regions not already covered in the ORIGINAL.
4) Start from the ORIGINAL DESCRIPTION. Preserve its general structure, wording, and style when possible.
5) UNCERTAINTY ("not sure", "unclear", "can't tell", etc.) → do not modify that attribute.
6) CONFIRMATION — If evidence confirms information already in the ORIGINAL, keep it as is.
7) When later answers contradict earlier answers, use the latest answer if it is clear and confident.
8) COLOR CONSISTENCY — If the ORIGINAL already has a color for an item, keep it unless the Q&A evidence clearly and confidently corrects that color. Add color only to previously uncolored items when supported.
9) Output ONE line ending with a period. Keep concise — similar length to original ($\pm$ 10 words).
10) MEMORY DENIAL → treat as uncertainty, skip.
11) For **[ADD]** questions: the answer may redescribe existing items with different vocabulary. Extract only new modifiers or genuinely new items. Avoid noun substitutions.
12) Do NOT add information not supported by the original description or Q&A evidence.
13) Partial "or" answers — ignore the ambiguous "or" part, but keep any definite information around it.

E Qualitative Interaction Examples and Retrieval Trajectories

This section provides qualitative examples of how SCOUT selects follow-up questions and how the ranked list evolves during interaction. The three visualizations cover complementary cases: clarifying known but underspecified attributes, exploring missing attributes absent from the initial description, and completing a mixed multi-round interaction trajectory.

Figure 7 illustrates known-attribute clarification. The initial query mentions jeans, sneakers, and a backpack, but their colors are unspecified. SCOUT ranks these mentioned attributes by D-RPS and asks color-clarification questions for them. In this example, the attribute with the highest D-RPS also yields the largest rank improvement among the shown candidates, indicating that higher perturbation sensitivity can correspond to a more retrieval-effective clarification for known attributes.

Figure 8 illustrates exploratory attribute discovery. The initial query omits several potentially useful cues, so SCOUT ranks unmentioned candidate attributes by D-RPS and asks the corresponding exploratory questions. Among the shown candidates, the attribute with the highest D-RPS also yields the largest rank improvement, suggesting that perturbation sensitivity can help identify missing cues with high retrieval impact.

Figure 9 shows a complete interaction trajectory. At each round, SCOUT scores both known and exploratory candidates by D-RPS, asks the highest-sensitivity remaining question, receives the user's answer, and progressively moves the target upward in the ranked list.

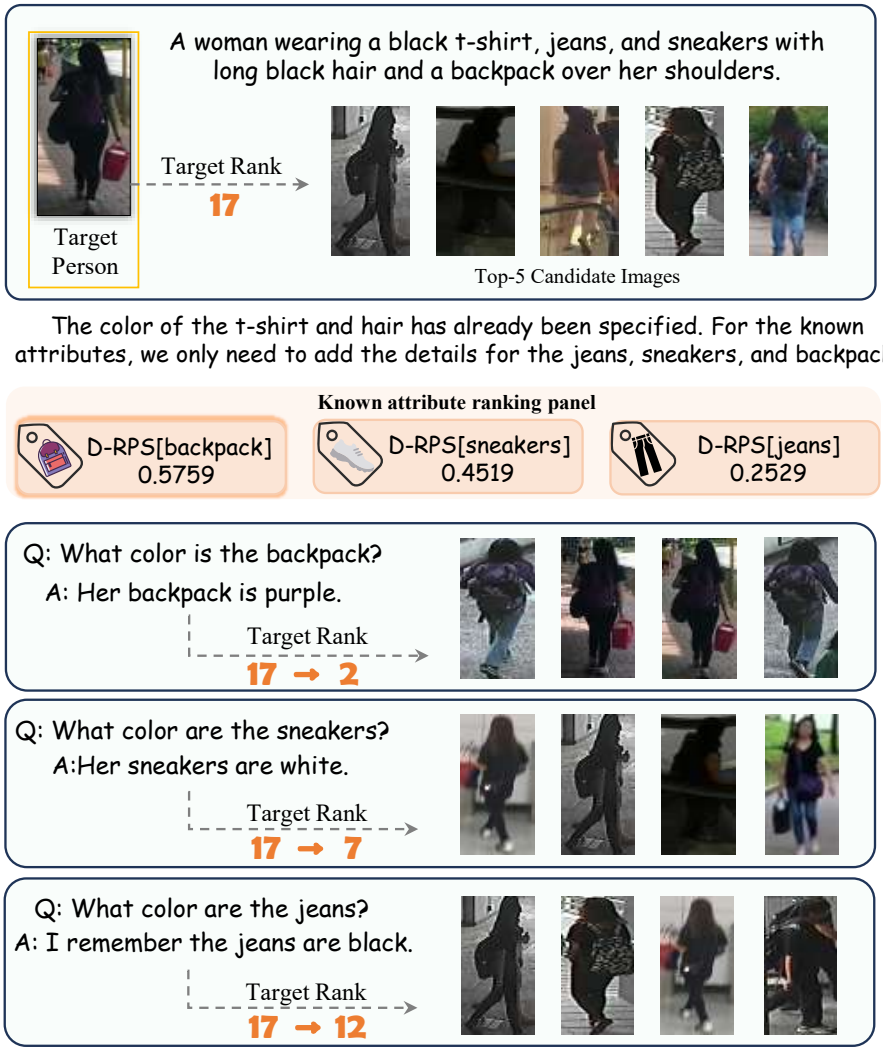


Figure 7: Known-attribute clarification example. SCOUT ranks mentioned but underspecified attributes by D-RPS and evaluates their corresponding color-clarification questions.

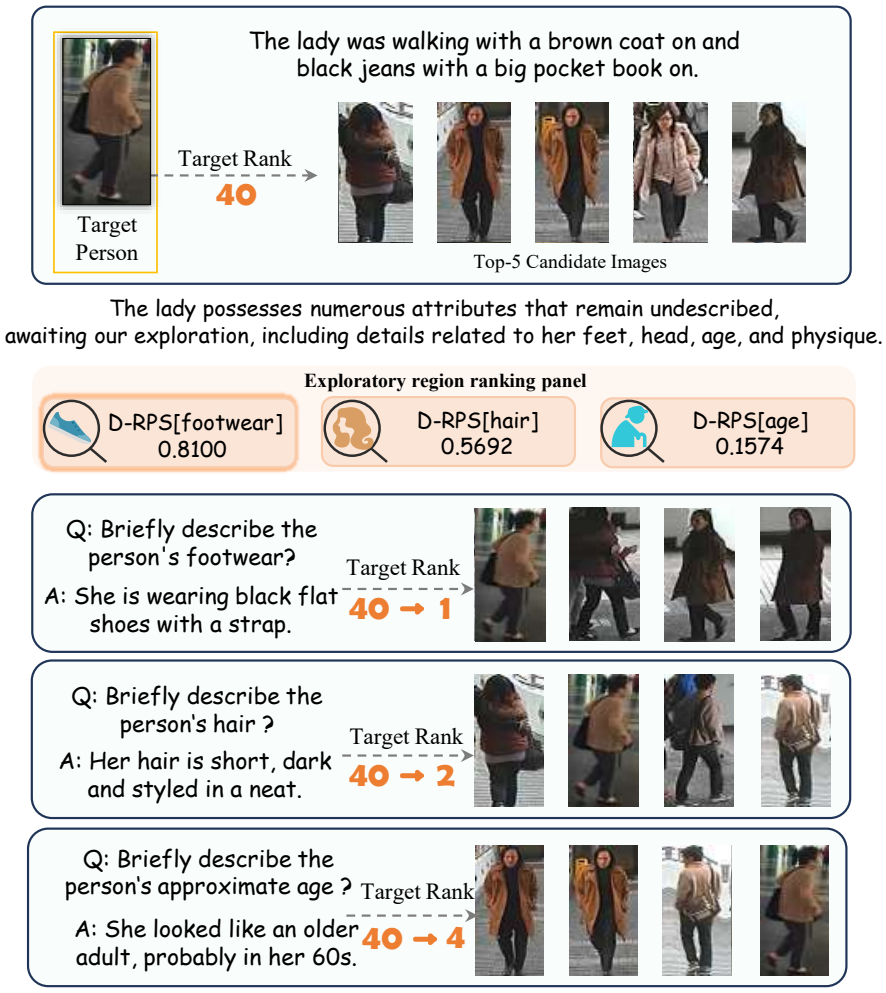


Figure 8: Exploratory-attribute discovery example. SCOUT ranks unmentioned candidate attributes by D-RPS and evaluates their corresponding exploratory questions.

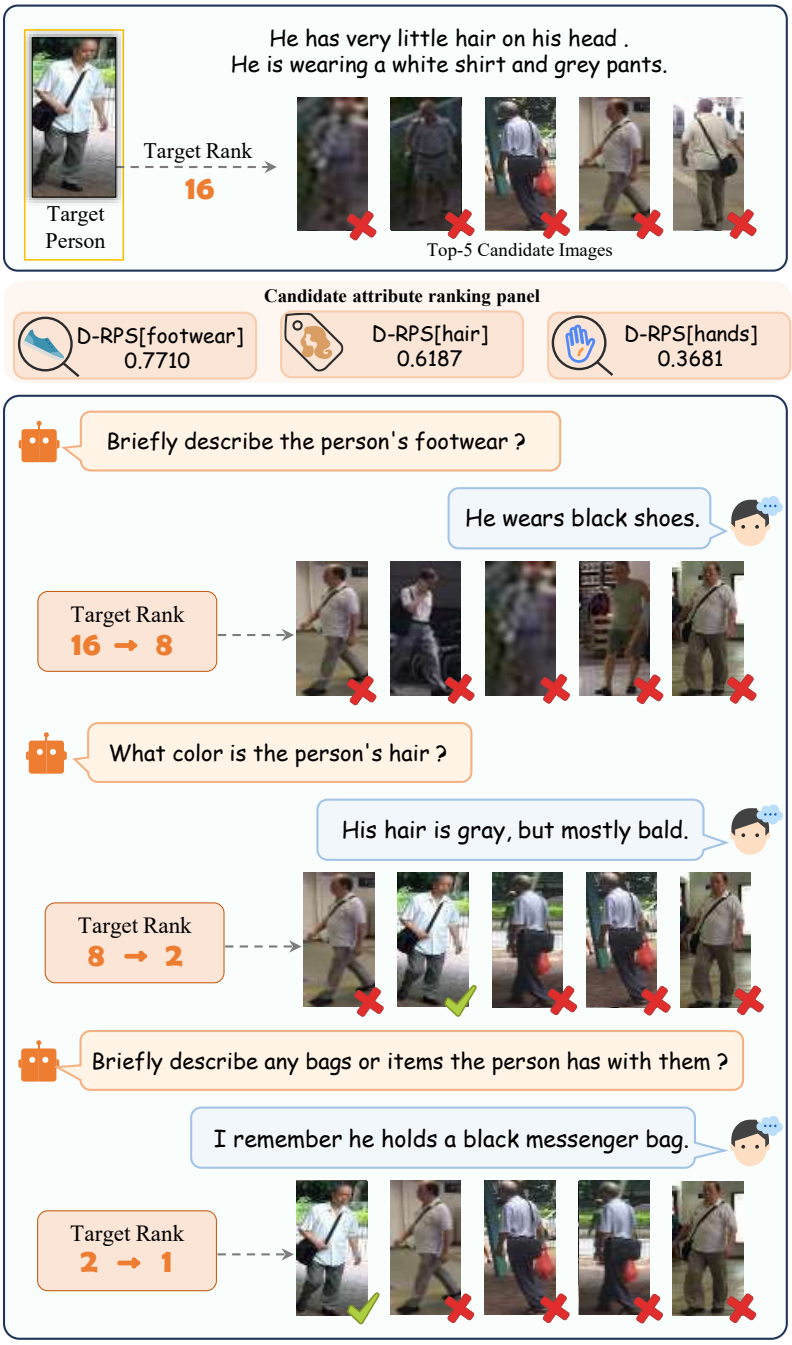


Figure 9: Complete multi-round interaction example. SCOUT scores known and exploratory candidates by D-RPS and asks questions in descending sensitivity order across successive rounds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main claims, including the SCOUT framework, Perturbation-Sensitive Question Selection, Retrieval Perturbation Sensitivity (RPS), and the empirical evaluation across standard TIReID benchmarks under an interactive protocol. These claims are supported by the method description, theoretical analysis, and experiments in Sections 3–4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 6, particularly the restriction to a predefined appearance-related question space and the need for more flexible dynamically constructed question spaces.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3.3 defines the RPS family, including D-RPS as the default instantiation and Set-RPS/Rank-RPS as alternatives, and provides propositions illustrating the failure modes of set- and rank-level signals. The main theoretical properties of D-RPS are stated with their assumptions in Section 3.3, with complete proofs and auxiliary definitions provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the SCOUT framework, RPS computation, and the unified five-round evaluation setting in the main paper. Appendix A provides the experimental details needed for reproduction, including datasets, metrics, implementation settings, interactive baseline adaptation, and runtime analysis. The prompts used by the LLM/VLM components are listed in Appendix D. We also provide an anonymized code package with reproduction instructions for the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit an anonymized code package for review, including the implementation of SCOUT and scripts/configurations needed to reproduce the main results. The experiments use public TIReID benchmarks and publicly available retrievers or reproducible retriever implementations, with dataset and implementation details described in Appendices A.1 and A.3.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Dataset statistics, test splits, evaluation metrics, the unified five-round evaluation setting, retriever/backbone usage, SCOUT configuration, interactive baseline adaptation, and runtime information are provided in Appendix A. The main experimental results are summarized in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following standard TIReID evaluation practice, we report fixed-split retrieval metrics, including R@1, R@5, R@10, and mAP, on established benchmarks, as described in Appendix A.2. We do not report error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A.3 reports the computing environment, including the GPU type, memory, and software version. Appendix A.5 reports the approximate SCOUT runtime for complete five-round evaluation runs on each benchmark.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses public benchmarks and public retriever implementations or checkpoints, does not collect new personal data, and preserves anonymity in the submission. We have reviewed and followed the NeurIPS Code of Ethics throughout the research and submission process.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader societal impact of interactive person retrieval is discussed in Appendix C, including improved human-in-the-loop search as well as privacy, surveillance, consent, and fairness concerns. We also clarify the intended research scope of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release a new person-image dataset, a scraped dataset, a pretrained LLM/VLM, or a deployment-ready surveillance system. The released asset is an anonymized research code package for reproducing experiments on public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original creators of the datasets, retrievers, and interactive baseline methods used in the paper, with dataset, implementation, and baseline details provided in Appendices A.1, A.3, and A.4. We use public benchmarks, public implementations, and released checkpoints according to their public terms, and include license or usage information in the anonymized code package where applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce an anonymized SCOUT code package for review, including implementation, scripts, configurations, and instructions for reproducing the main experiments. The paper does not introduce a new dataset or pretrained model; implementation details and prompts are documented in Appendices A and D.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject studies. User answers in the experiments are simulated under the unified VLM-based evaluation setting described in Appendix A.3.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject studies, and no new data are collected from human participants. Therefore, IRB approval or equivalent human-subject review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This paper uses LLM/VLM components in the research pipeline, including SCOUT's attribute extractor, rewriter, fuser, and answer simulator. The usage, evaluation setting, and prompts are described in Appendices A.3 and D. All usage follows the terms specified by the respective model providers and the NeurIPS LLM usage policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.