BEYOND VERBALIZED FEEDBACK: MINING INTENT-ALIGNED VISUAL CLUES FOR INTERACTIVE TEXT-TO-IMAGE PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

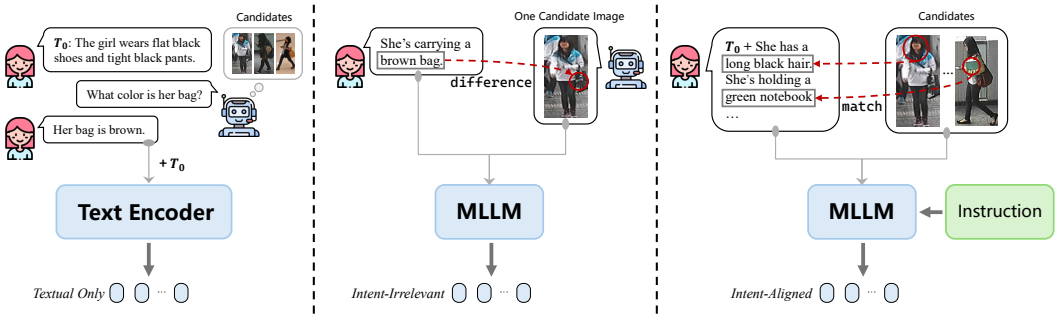
Text-to-image person re-identification (TIReID) aims to retrieve a target pedestrian from a large-scale image gallery using a natural-language description. Since a single description rarely captures the fine-grained details that distinguish visually similar pedestrians, interactive TIReID progressively completes the initial description through multi-round user feedback. However, existing methods rely on user feedback, and retrieved candidates are rarely used as a direct source of visual clues. As a result, subtle variations in color tone, texture, and clothing style, which are difficult to convey through textual feedback, can be easily overlooked. To address this issue, we propose Visual Clue-guided Interactive Retrieval (VCIR), a plug-and-play framework that mines intent-aligned visual clues from retrieved candidates. Specifically, we first design a Visual Clue Extraction (VCE) module that guides a Multi-modal Large Language Model (MLLM) to contrast these candidates against the user intent and extract discriminative clues. Then, we introduce a learnable retrieval token that distills these clues into a continuous representation, bypassing the information loss caused by verbalizing them into discrete text. We further embed VCE into a multi-round interactive retrieval framework, where the extracted clues directly serve as the retrieval query at each round and progressively steer retrieval toward the target pedestrian. VCIR introduces only **32.9M** trainable parameters and easily integrates with existing interaction paradigms. Extensive experiments on Interactive-PEDES and MInterPEDES demonstrate that VCIR achieves state-of-the-art retrieval performance. The code will be released.

1 INTRODUCTION

Text-to-image person re-identification (TIReID) (Li et al., 2017b; Zhang & Lu, 2018; Ding et al., 2021; Jiang & Ye, 2023; Yan et al., 2023; Qin et al., 2024) aims to retrieve images of a target pedestrian from a large-scale gallery given a short natural-language description. It plays an important role in a broad set of applications, such as public security (Ye et al., 2021; Zheng et al., 2016), social media analysis (Jiang et al., 2025a), and intelligent video surveillance (Wang, 2013; Galiyawala & Raval, 2021). However, a fundamental challenge of TIReID lies in the query itself. A single-sentence description can hardly characterize the target completely (Lu et al., 2025; Wang et al., 2020), and its inherently coarse granularity is insufficient to support the discrimination of subtle differences between pedestrians of similar appearance.

To address this issue, recent studies propose interactive TIReID, which iteratively refines the textual query through multi-round interaction with the user. Existing methods fall into two categories. The first category employs a Multi-modal Large Language Model (MLLM) to raise questions that maximally distinguish the remaining candidates, thereby narrowing the search space round by round, and then combines the user's answers with the initial description into a textual query (Lu et al., 2025), as illustrated in Figure 1(a). The second category directly presents a candidate image to the user, who then describes how its appearance differs from the target (Bai et al., 2026), as shown in Figure 1(b). Both categories treat the user as the sole source of new information, while the retrieved images merely shape the questions or comparisons presented to the user. The candidates themselves are rarely treated as a direct source of fine-grained visual clues, particularly those that are difficult to express in language.

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



(a) Chat-based Interaction (b) Multi-Modal Interaction (c) Our VCIR Method

Figure 1: (a) Chat-based interaction refines the initial description through rounds of question answering. The retrieved candidates only guide question generation, and the query remains textual. (b) Multi-modal interaction presents a candidate image for the user to comment on, and encodes the feedback via an MLLM with the image as context. If the candidate is a non-target image, encoding it in full introduces target-irrelevant content. (c) VCIR employs the MLLM as a clue extractor, condensing the intent-aligned clues of candidates into a single retrieval token. The candidates thus serve as a source of clues rather than mere context, without modifying the retrieval pipeline.

Consequently, a natural question arises: *can fine-grained clues be extracted directly from the retrieved candidates?* The candidates attain a high rank because they share attributes with the target and thus encode visual clues beyond the coverage of textual description. As a non-target sample, however, it is only partially aligned with the user intent. Therefore, the central difficulty lies in determining which clues genuinely reflect that intent and in aggregating them into a fine-grained embedding.

Herein, we propose **VCIR** (**V**isual **C**lue-guided **I**nteractive **R**etrieval), a framework that leverages the strong vision-language understanding of MLLMs to extract visual clues from the candidates of each round and directly retrieve the target pedestrian with them. Specifically, each interaction round of VCIR consists of three steps. First, for clue extraction, the **VCE** (**V**isual **C**lue **E**xtraction) module guides an MLLM that takes the top-K candidate images from the previous round, the user feedback, and a carefully designed instruction as input. Constrained by the known textual description, it identifies the fine-grained visual content that conforms to the user intent. Second, to form the query, the MLLM encodes these clues together with the textual description into a learnable retrieval token, which directly serves as the retrieval query. Finally, this query is matched against the gallery. The retrieved candidates are returned to the user and also serve as the source of clue extraction in the next round. This process is agnostic to the form of user feedback, so both chat-based (Lu et al., 2025; Bai et al., 2025b) and multi-modal (Bai et al., 2026) interactions can be plugged in directly. More importantly, the clues extracted by the MLLM need not be discretized into language. Instead, they enter retrieval directly as continuous representations.

Nevertheless, applying an MLLM to retrieval is not straightforward. As a general-purpose generative model, its output is oriented toward natural language generation rather than cross-modal matching, leaving a substantial representation gap from the discriminative embedding required for retrieval. To bridge this gap, we introduce two lightweight components into the MLLM. The first is clue prompt shared across interaction rounds. It learns which appearance details distinguish pedestrians, while the instruction conveys the user intent of each round. The second is a learnable retrieval token inserted into the LLM backbone, which continues absorbing information from both the visual and textual sides during the forward pass, and whose last-layer hidden state is taken as the clue representation for the current round. An additional advantage of this design is that the representation ability required for retrieval is concentrated in a small number of newly introduced parameters.

Unlike prior methods that rely solely on the user for new information, VCIR directly mines it from the retrieved candidates. Since the clues are carried by a single retrieval token, adapting the MLLM to retrieval only requires training 32.9M parameters rather than fine-tuning the full backbone. This yields a plug-and-play interactive retrieval framework and significantly reduces the training cost. Extensive experiments justify the effectiveness of our design and show that VCIR achieves state-of-the-art performance on the recent Interactive-PEDES and MInterPEDES benchmarks.

2 Related Work

Text-to-Image ReID. TIReID aims to retrieve images of a specific pedestrian according to a natural-language description. Existing TIReID methods mainly address two key problems: how to design effective cross-modal feature alignment strategies and how to enlarge the scale and diversity of training data to improve model generalization.

For cross-modal feature alignment, early methods project global text and image features into a shared embedding space (Li et al., 2017b; Zhang & Lu, 2018; Zheng et al., 2020). However, pedestrians often share similar clothing and poses, while global features miss the word-region correspondences needed to distinguish them. Later studies pursue part-level alignment through explicit or implicit methods. Explicit alignment methods rely on external tools to extract local cues, such as attributes (Wang et al., 2020), poses (Jing et al., 2020), body parts (Ding et al., 2021), noun phrases (Niu et al., 2020), or color-aware regions (Wu et al., 2021). In contrast, implicit alignment methods discard external tools and instead promote fine-grained image-text matching through cross-modal attention (Li et al., 2017a), masked language modeling (Yang et al., 2023), multi-granularity interaction (Gao et al., 2021), and token-level relation reasoning (Jiang & Ye, 2023). Recently, both strategies have adapted pre-trained models such as CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021), exploiting their already aligned cross-modal representations to further model fine-grained correspondence (Yan et al., 2023; Bai et al., 2023; Cao et al., 2024; Qin et al., 2024). Despite their success within TIReID benchmarks, these models generalize poorly to unseen data due to the limited scale of human-annotated training data.

Another line of work generates diverse textual descriptions for pedestrian images to construct large-scale image-text pairs for TIReID pre-training (Zou et al., 2026). Some studies produce pseudo descriptions with predefined templates (Zuo et al., 2024; Shao et al., 2023). More recently, researchers have begun to use MLLMs to generate higher-quality descriptions. For instance, Tan et al. (2024) heuristically design 46 description templates with ChatGPT to enrich the instructions given to an MLLM, while HAM (Jiang et al., 2025b) goes a step further by learning the linguistic styles of different human annotators, thereby achieving multi-style and high-quality annotation. Despite these advancements, these methods still follow a "one-shot" querying paradigm, often yielding incomplete or ambiguous descriptions, which results in suboptimal retrieval performance.

Interactive Cross-Modal Retrieval. Interactive cross-modal retrieval encourages users to provide additional information about the target through interactions with the retrieval system and has advanced substantially in image (Cai et al., 2021; Guo et al., 2018) and video retrieval (Madasu et al., 2022). Existing approaches mainly adopt either chat-based or multimodal interaction.

In chat-based interaction, the system leads the retrieval process by asking users for target details or clarification (Zhu et al., 2024; He et al., 2025; Bai et al., 2025b). For example, Levy et al. (2023) use an LLM to generate questions from dialog history, while Lu et al. (2025) train a multi-image questioner to identify differences among candidates from visual and textual context and generate more discriminative questions. In contrast, multimodal interaction is user-led, allowing users to directly highlight differences between the target and candidate images. Bai et al. (2026) introduce this paradigm to interactive TIReID by enabling visual feedback in later retrieval rounds.

However, these methods ultimately aim to construct a more complete textual description, implicitly assuming that sufficient feedback enables text alone to identify the target. This assumption is restrictive because TIReID often depends on fine-grained appearance details that are difficult to fully express in language. Consequently, textual feedback alone cannot bridge the modality gap. To overcome this limitation, we propose VCIR, which uses an MLLM to extract fine-grained visual clues from ranked candidates and then converts these clues into features directly usable for retrieval.

3 Method

The overview of our framework is illustrated in Figure 2. Section 3.1 defines the problem setup. Section 3.2 introduces **VCE**, which extracts visual clues consistent with user intent from the current candidates and aggregates them into a continuous representation that refines the textual query. Section 3.3 presents the entire **VCIR** framework, which progressively approaches the target through multi-round interaction.

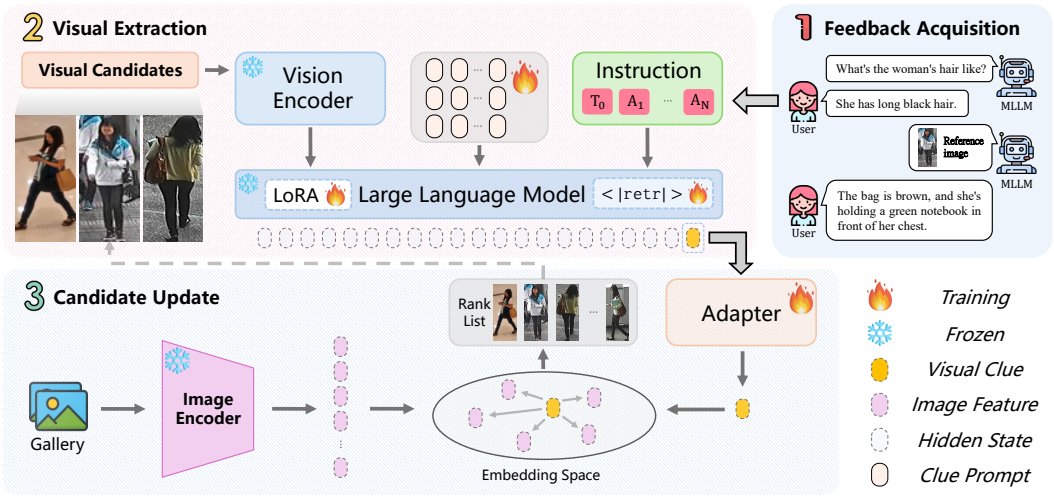


Figure 2: Overview of our VCIR framework. We extract fine-grained visual clues from the candidate images of each round and inject them into the query. (1) The user provides feedback on the target, either by answering a question or by commenting on the candidate images. (2) Conditioned on the clue prompts, the accumulated feedback and the instruction, the MLLM inspects the current candidates and aggregates the intent-relevant clues onto a retrieval token. (3) The extracted visual clues and the textual intent from the user are jointly encoded into a single query embedding that directly reranks the entire gallery and produces refined candidates for the next round.

3.1 Problem Setup

Unlike traditional "one-shot" TIReID, interactive TIReID is an iterative retrieval task that allows a dialog between the user and the system. The user holds a fine-grained memory of the target image I and supplies additional information about it in each round. The dialog C during retrieval can generally be formulated as:

$$C = \{T_0, (R_1, A_1), (R_2, A_2), \ldots, (R_N, A_N)\}, \tag{1}$$

where N is the round budget and T_0 denotes the coarse-grained textual description initially provided by the user for I, which is usually incomplete. The system issues a request R_n in either of two forms, i.e., asking a question or showing an image, and the user provides an answer A_n according to the memory of I.

Despite the answer A_i accumulated as the interaction proceeds, language conveys attributes rather than the pixel-level details that distinguish visually similar identities, so T_0 and A_i remain insufficient to discriminate between them.

Consequently, we propose the **VCIR** framework, which exploits the strong vision-language understanding ability of MLLMs to mine the visual information that matches the user intent. Nevertheless, directly employing an MLLM is impractical, since it autoregressively produces discrete textual tokens, whereas retrieval requires a continuous representation that is aligned with the gallery (Cui et al., 2026; Yang et al., 2026). To address this, **VCIR** adopts a generate-to-represent strategy. It first applies **VCE**, where the MLLM keeps aggregating information from the candidate images and user feedback. Then, instead of decoding a token into text, we take its last-layer hidden state and project it into the image embedding space of CLIP, so that the aggregated clues serve directly as the retrieval query in the form of a continuous and intent-aligned representation.

3.2 Visual Clue Extraction

Visual **C**lue **E**xtraction (**VCE**) consists of three components in addition to the MLLM, namely a retrieval token that carries the clues, an intent-aware instruction that specifies which clues to keep, and clue prompts that encode retrieval-specific priors.

Retrieval Token. We extend the vocabulary of MLLM with a special token: $\mathcal{V}' = \mathcal{V} \cup <|\texttt{retr}|>$. The embedding of this token is randomly initialized and learned during training. We append this token to the input sequence so that it can attend to all preceding visual and textual tokens. The resulting hidden state captures instruction-relevant visual clues and text-confirmed attributes. We use it as the retrieval representation.

Instruction Design. We design an intent-aware instruction P for VCE. Given the current top-K candidate images, the initial description T_0, and the feedback F_n accumulated through n rounds, P specifies three constraints for integrating these inputs into the hidden state of $<|\texttt{retr}|>$. A simplified version of P is shown below, and the complete instruction is provided in Appendix F.

> *"We are performing an interactive text-to-image person retrieval task. The target person is described by the "$\{T_0\}$" and "$\{F_n\}$". The input images are the current top-K candidates. These images may not be of the target person.*
> *1. Preserve confirmed target attributes.*
> *2. Emphasize user feedback.*
> *3. Ignore appearance details in the images that differ from the target person.*
> *Finally, encode the representation into the $<|\texttt{retr}|>$ token for retrieval."*

Here, F_n denotes the concatenation of all user feedback collected up to the n-th round.

The three rules address three failure modes in interactive retrieval, respectively. Rule 1 prevents the query from drifting across rounds, i.e., discarding previously confirmed discriminative cues while attending to new information. Rule 2 highlights the feedbacks of the current round, which are usually given in response to the confusion exposed by the previous retrieval. Rule 3 targets a risk specific to this task: in most rounds, the input candidates are not the target, and their appearance is a mixture of target attributes and details of another identity. Absorbing it indiscriminately would pull the query towards the current erroneous candidates. We thus explicitly state in the instruction that the candidates may not be the target, so that they are treated as a reference rather than a description.

Clue Prompt. A discrete instruction alone is insufficient to bridge the gap between general-purpose instruction following and the retrieval objective, since it cannot enumerate which fine-grained cues are discriminative for pedestrian identity. We therefore introduce a set of clue prompt embeddings $E_p \in \mathbb{R}^{L \times d}$ that are shared across all rounds. L is the prompt length. E_p is optimized during training and encodes such retrieval-specific priors implicitly.

3.3 Visual Clue-guided Interactive Retrieval

In this section, we present the multi-round retrieval pipeline of VCIR. Rather than serving merely as a ranking result, the candidates of each round are fed back as the clue source for VCE in the next round, so that the retrieval and VCE reinforce each other. As illustrated in Figure 2, each round consists of three steps, i.e., **Feedback**, **Extraction**, and **Retrieve**.

Initialization. Let f_v and f_t denote the image and text encoders of the base TIReID model, respectively. Given an image gallery $\mathcal{G} = \{v_1, \ldots, v_M\}$, we extract its visual features $\{f_v(v_j)\}_{j=1}^{M}$. At the beginning, only the coarse-grained description T_0 is available. We set $F_0 = \varnothing$ and encode T_0 into the initial query $q_0 = f_t(T_0)$. We then rank all gallery images in descending order of $\mathrm{Sim}\,(q_0, f_v(v_j))$, where $\mathrm{Sim}(\cdot)$ denotes the cosine similarity, and obtain the top-K candidate list:

$$\hat{\mathbf{V}}_0 = \left(\hat{v}_0^{(1)}, \hat{v}_0^{(2)}, \ldots, \hat{v}_0^{(K)}\right). \tag{2}$$

In general, $\hat{\mathbf{V}}_n$ denotes the top-K candidate list retrieved at the n-th round.

Feedback Acquisition. At the n-th round, the user provides a single sentence A_n about the target, and we update the accumulated feedback as:

$$F_n = F_{n-1} \oplus A_n, \tag{3}$$

where $\oplus$ denotes concatenation. VCIR is agnostic to how A_n is obtained, and can therefore be directly plugged into both the chat-based and the multi-modal interaction paradigms. Regardless of what is displayed to the user, $\hat{\mathbf{V}}_{n-1}$ always serves as the visual content of VCE.

Visual Extraction. The query of the n-th round is produced entirely by VCE, which reads the textual intent and the previous candidates jointly and extracts from $\hat{\mathbf{V}}_{n-1}$ the fine-grained appearance details that language can hardly depict. Specifically, following the instruction design in Section 3.2, we construct the round-specific instruction $P(T_0, F_n)$ from the initial description T_0 and the accumulated feedback F_n, so that every attribute stated by the user so far is conveyed to VCE. VCE then takes the clue prompt, the previous round's top-K candidates, and $P(T_0, F_n)$ as input:

$$X_n = \left[E_p \; ; \; \hat{\mathbf{V}}_{n-1} \; ; \; P(T_0, F_n) \right]. \tag{4}$$

VCE encodes X_n and outputs the clue representation of the n-th round:

$$c_n = \mathrm{VCE}(X_n), \tag{5}$$

where $\mathrm{VCE}(\cdot)$ returns the last-layer hidden state at the position of <|retr|>. Since c_n lies in the representation space of the LLM backbone, we map it into the image embedding space of CLIP with a lightweight linear adapter and directly use the result as the query:

$$q_n = \mathrm{Adapter}(c_n) \in \mathbb{R}^{d_e}, \tag{6}$$

where d_e is the dimension of the CLIP embedding space. Residing in the same space as $\{f_v(v_j)\}_{j=1}^{M}$, q_n can be matched against the gallery without any text-side encoding.

Candidates Update. We retrieve the gallery with q_n and obtain the new top-K list $\hat{\mathbf{V}}_n$ in the same way as Eq. 2, which serves both as the retrieval result of the n-th round and as the visual content for VCE at the $(n{+}1)$-th round. The interaction terminates once the target is ranked first or the maximum number of rounds N is reached.

Training Objective. To encourage the target to surface in as few rounds as possible, we supervise every round rather than only the final state. Within an interaction trajectory, both losses are computed between q_n and the visual feature of its target image. The overall objective is formulated as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{L}_{\text{itc}}^{(n)} + \mathcal{L}_{\text{sdm}}^{(n)} \right), \tag{7}$$

where $\mathcal{L}_{\text{itc}}$ (Cao et al., 2024) is the image-text contrastive loss and $\mathcal{L}_{\text{sdm}}$ (Jiang & Ye, 2023) is the similarity distribution matching loss that alleviates the false negatives caused by multiple gallery images sharing one identity.

4 Experiments

4.1 Datasets and Metrics

We evaluate VCIR on two interactive person retrieval benchmarks: Interactive-PEDES (Lu et al., 2025) and MInterPEDES (Bai et al., 2026), following their official protocols. The former uses chat-based interaction, while the latter adds visual feedback to candidate images. We report the results after the third and fifth rounds on Interactive-PEDES, and after the last round on MInterPEDES. To validate the generalizability of our method, we also extend the interactive evaluation to three standard TIReID benchmarks, i.e., CUHK-PEDES (Li et al., 2017b), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021). Due to space limitations, detailed descriptions of the datasets are provided in Appendix A, and the results on TIReID benchmarks are reported in Appendix C.

Following previous work (Jiang & Ye, 2023; Cao et al., 2024; Qin et al., 2024), we report Rank@1/5/10, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) (Ye et al., 2021) on all benchmarks. mINP measures the cost of retrieving the hardest positive and thus reflects whether all images of the target are ranked high. For the interactive setting, we further report the Best log Rank Integral (BRI) (Lee et al., 2024). BRI accounts for the whole interaction process rather than only the final round, and thus rewards methods that localize the target in earlier rounds. A lower BRI indicates better interaction quality.

4.2 Implementation Details

We freeze both the language and vision backbones of Qwen3-VL-Instruct-8B (Bai et al., 2025a) and optimize only the clue prompt, <|retr|> token, adapter, and LoRA (Hu et al., 2021) parameters inserted into the query and value projections of the attention layers. The LoRA rank and scaling factor are set to $r = 64$ and $\alpha = 128$, respectively, with a dropout rate of 0.05. The clue prompt length is set to $L = 32$. In total, only approximately 32.9M parameters are trainable. We adopt the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.01. The learning rates are set to 2×10^{-4} for the clue prompt and the <|retr|> token, and 4×10^{-5} for the LoRA modules. Training is conducted with batch size of 64 for 5 epochs on $8\times$ NVIDIA GeForce RTX 3090 GPUs.

As for the retrieval backbone, we adopt an IRRA model (Jiang & Ye, 2023) pre-trained on HAM-PEDES (Jiang et al., 2025b) and keep it frozen during training. Following Long-CLIP (Zhang et al., 2024), we extend the context length of its text encoder from 77 to 192 by retaining the first 20 positional embeddings and linearly interpolating the remaining 57 to fill the other 172 positions. The causal attention mask is rebuilt accordingly, and the stretched embeddings are not fine-tuned.

4.3 Ablation Studies

Table 1: Ablation studies on Interactive-PEDES. The best results are highlighted in **bold.**

Method	R@1	mAP	mINP
(1) HAM	59.38	45.52	24.71
(2) Baseline	77.17	60.06	36.29
(3) VCIR (random images)	77.57	59.91	35.88
(4) VCIR (w/o <|retr|>)	78.71	67.72	48.17
(5) VCIR (w/o clue prompt)	79.62	68.95	49.56
(6) VCIR	**80.08**	**70.00**	**51.14**

In Table 1, we perform ablation studies to investigate the effectiveness of each proposed component in VCIR. All the experiments are conducted with five rounds of interaction. Exp. (1) reports the performance of our basic TIReID model (Jiang et al., 2025b). In this setting, the query of each round is the accumulated textual description alone. Exp. (2) reports our baseline. It shares the same architecture as VCIR, except that no candidate image is fed to the MLLM and the image-related part is removed from the instruction P, with the remaining content unchanged. Exp. (3) to Exp. (5) modify one design of the full model at a time.

We first study whether VCIR indeed learns to extract effective visual clues. Compared with Exp. (1), our baseline in Exp. (2) promotes R@1 and mAP by 17.79% and 14.54%, showing that the MLLM branch itself already refines the textual query. However, replacing the candidates with random gallery images in Exp. (3) yields almost identical results to Exp. (2), i.e., 77.57% versus 77.17% in R@1. In comparison, the full model in Exp. (6) further promotes R@1, mAP and mINP by 2.91%, 9.94% and 14.85% over Exp. (2). This indicates that the model does not blindly absorb whatever image is provided, and that its improvement comes from the clues extracted from the retrieved candidates.

Then, we study the two designs that convert the MLLM into a retrieval-oriented model. Replacing our <|retr|> token with the last-layer hidden state of the End-Of-Sequence (EOS) token in Exp. (4) reduces R@1, mAP and mINP by 1.37%, 2.28% and 2.97%. This is probably because the EOS token is optimized towards language generation rather than cross-modal matching. Removing the clue prompt in Exp. (5) also brings a consistent degradation, mainly on mAP and mINP by 1.05% and 1.58%. For both designs, the degradation on mINP is larger than that on mAP, suggesting that they mainly help retrieve the hard positives of the target. The above experiments justify the effectiveness of each of our proposed components.

Finally, we study the optimal value for the candidate number K. As expected, the performance improves as K increases, since a larger K introduces more available visual clues. For space reasons, more parameter studies are provided in Appendix D.

4.4 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 2: Comparison to state-of-the-art methods on Interactive-PEDES. A lower BRI indicates better performance. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Round 3				Round 5				BRI ↓
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
SimIRV (Liang & Albanie, 2023)	50.45	73.39	82.49	41.00	61.27	82.00	88.36	46.53	1.024
ChatIR (Levy et al., 2023)	57.85	78.69	86.08	44.92	63.86	83.81	89.84	48.05	0.935
PlugIR (Lee et al., 2024)	60.34	81.34	87.89	47.28	65.44	85.33	91.01	49.89	0.849
LLaVA-ReID (Lu et al., 2025)	63.96	84.70	91.27	48.45	73.20	**90.62**	**95.95**	53.31	0.719
VCIR (Ours)	**75.06**	**85.66**	**91.31**	**66.13**	**80.08**	88.43	93.54	**70.00**	**0.467**

Results on Interactive-PEDES. We compare VCIR with four state-of-the-art interactive retrieval methods, and report the results at Round 3 and Round 5. As shown in Table 2, VCIR consistently achieves the highest R@1 and mAP scores at both rounds as well as the lowest BRI, leading to the best overall performance. In particular, compared to the second-best LLaVA-ReID, VCIR increases the R@1 accuracy by 11.10% and 6.88% and the mAP by 17.68% and 16.69% at Round 3 and 5, respectively, and reduces BRI from 0.719 to 0.467. Notably, with only three rounds of interaction, VCIR already surpasses the Round-5 mAP of all competing methods by 12.82% to 19.60%. Its Round-3 R@1 of 75.06% also exceeds the best Round-5 R@1 among competing methods, which is 73.20% by LLaVA-ReID. Although LLaVA-ReID obtains higher R@5 and R@10 at Round 5, VCIR is more effective at promoting the target to the very top of the ranking list, since the injected visual clues mainly disambiguate candidates that are already similar to the target. These results indicate that VCIR locates the target person more accurately with markedly fewer interaction rounds.

Table 3: Comparison to state-of-the-art methods on MInterPEDES. A lower BRI indicates better performance. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Chat-Based				Multi-Modal				BRI↓
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
IRRA (Jiang & Ye, 2023)	48.26	67.96	75.65	44.64	–	–	–	–	1.571
TBPS-CLIP (Cao et al., 2024)	50.21	69.27	76.43	45.44	–	–	–	–	1.557
RDE (Qin et al., 2024)	52.77	70.61	77.47	47.98	–	–	–	–	1.515
SPRC (Bai et al., 2024)	67.63	83.47	89.62	58.95	83.13	96.23	98.54	72.81	0.551
ICL (Qin et al., 2025)	73.56	87.68	93.14	64.59	88.73	95.56	97.38	77.53	0.468
LLaVA-ReID (Lu et al., 2025)	71.65	86.12	91.78	62.75	87.01	94.29	96.60	75.94	0.669
MNEMO (Bai et al., 2026)	78.09	**92.27**	**95.48**	69.18	92.26	**98.86**	**99.63**	80.31	0.495
VCIR (InternVL-2.5)	80.50	92.13	95.02	73.98	**93.64**	98.41	99.24	84.43	0.464
VCIR (Qwen3-VL)	**80.97**	92.26	95.06	**74.32**	93.43	98.41	99.27	**84.48**	**0.461**

Results on MInterPEDES. Table 3 compares VCIR with both traditional and interactive TIReID methods under the chat-based and multi-modal settings. We instantiate VCIR with two backbones, InternVL-2.5 and Qwen3-VL, where InternVL-2.5 is the same backbone used by MNEMO (Bai et al., 2026). With either backbone, VCIR achieves higher R@1 and mAP and a lower BRI than all compared methods in both settings. The BRI values of the compared methods are reproduced by us. With the same backbone, VCIR surpasses the previous state-of-the-art MNEMO by 2.41% and 4.80% in R@1 and mAP under the chat-based setting, and by 1.38% and 4.12% under the multi-modal setting. These gains therefore stem from the proposed design rather than a stronger MLLM. With Qwen3-VL, VCIR further achieves the best mAP in both settings, the best chat-based R@1, and the lowest BRI of 0.461. The slight drop in R@5 and R@10, less than 0.5%, probably results from a trade-off. VCIR places its most confident matches at the top, which benefits R@1 but allows confidently mismatched images to crowd the target out of the top-5 or top-10. Notably, the mAP gains consistently exceed the R@1 gains in both settings and with both backbones. Since each identity corresponds to multiple gallery images, the larger mAP gains suggest that VCIR captures identity-level fine-grained appearance rather than matching a single image by chance. These results indicate that converting the candidates of each round into exploitable visual clues is essential for interactive person retrieval.

4.5 Robustness to Initial Descriptions

In practice, a user starts from a coarse impression of the target, so we investigate whether VCIR remains effective when the initial description is vague or even erroneous. We perturb only the initial description T_0 by fixed rules and run five rounds of interaction for every perturbation, covering ***Vague*** (all color words removed, the second half of the sentence truncated, and the fully uninformative sentence "A person.") and ***Erroneous*** (all colors replaced, gender reversed, and the identity substituted by another one). Examples of the perturbed descriptions are provided in Appendix E. The question-answer pairs are generated following the interaction protocol of LLaVA-ReID (Lu et al., 2025). For comparison, we adopt the baseline in Exp. (2) of Section 4.3, which shares the same architecture as VCIR but feeds no candidate image to the MLLM, so that it refines the query from the dialogue alone. Round 0 denotes retrieval by HAM with only the perturbed initial description, which is shared by both models. Since the user feedback is generated from the clean initial description and kept unchanged, it never explicitly corrects the corrupted description. The erroneous prior therefore persists throughout the dialogue and can only be overcome by the model itself.

As shown in Table 4, VCIR achieves the highest mAP under all six perturbations and the highest R@1 under five of them, and recovers more from the same feedback when the starting point is partially degraded. In particular, its ΔR@1 is $+31.48$ on clean descriptions, but reaches $+41.86$, $+41.08$ and $+48.48$ when the color words are removed, the second half of the sentence is truncated, and the gender is reversed. Notably, with the gender reversed, VCIR reaches 77.61% R@1 and 68.42% mAP, surpassing the baseline fed with a *clean* description by 0.44% and 8.36%. These results demonstrate the strong robustness of VCIR to vague and erroneous initial descriptions.

Table 4: Robustness analysis on Interactive-PEDES. The best results are highlighted in **bold**.

Initial description	HAM (Round 0)		Baseline (Round 5)			VCIR (Round 5)		
	R@1	mAP	R@1	mAP	ΔR@1	R@1	mAP	ΔR@1
Clean	48.60	39.11	77.17	60.06	+28.57	**80.08**	**70.00**	**+31.48**
Vague								
Color words removed	17.89	15.96	58.02	44.47	+40.13	**59.75**	**52.57**	**+41.86**
Second half truncated	6.99	8.63	47.82	38.23	+40.83	**48.07**	**43.36**	**+41.08**
Fully uninformative	0.08	0.23	**33.01**	26.74	**+32.93**	32.88	**29.30**	+32.80
Erroneous								
All colors replaced	5.03	5.61	27.66	23.51	+22.63	**33.58**	**30.45**	**+28.55**
Gender reversed	29.13	24.66	75.74	58.20	+46.61	**77.61**	**68.42**	**+48.48**
Identity substituted	4.29	7.98	29.69	28.52	+25.40	**33.30**	**34.46**	**+29.01**

5 Conclusion and Limitations

In this paper, we identify a key limitation of existing interactive text-to-image person re-identification methods: they rely on verbalized user feedback while overlooking discriminative visual clues in retrieved candidates. To address this, we propose VCIR, a plug-and-play framework that turns top-ranked candidates into intent-aligned retrieval cues. Its Visual Clue Extraction module uses a clue prompt to guide a frozen MLLM in comparing candidates with the user's intent, while a learnable retrieval token captures the resulting clues in a continuous representation without requiring their verbalization. Used as the query at each interaction round, this representation progressively refines retrieval toward the target person. Experiments on Interactive-PEDES and MInterPEDES show that VCIR improves R@1 and mAP over competing methods, with only 32.9M trainable parameters. Ablation and robustness results further support the value of candidate-derived clues under vague or erroneous initial descriptions.

Limitations. VCIR currently uses a fixed number of retrieved candidates across interaction rounds, while the amount of useful visual context may vary with query difficulty and retrieval confidence. We plan to explore adaptive candidate selection strategies that dynamically balance retrieval effectiveness and inference efficiency in future work.

Ethics Statement

Our work builds on person re-identification datasets (e.g., CUHK-PEDES, ICFG-PEDES, RSTPReid) that contain images of pedestrians collected from public surveillance-style cameras. These datasets are widely used in the ReID research community and are released under their respective academic licenses. We use them strictly for research purposes and do not attempt to re-identify or de-anonymize any individual beyond the identity labels already provided by the dataset creators. We recognize that person retrieval technology can be misused for unauthorized surveillance or tracking, and we emphasize that our method is intended for retrieval-accuracy research rather than deployment in real-world monitoring systems.

Reproducibility Statement

We provide full implementation details to facilitate reproducibility. The architecture of the retrieval token mechanism, training objectives, and hyperparameters are described in Section 4.2 and Appendix B. Dataset statistics are detailed in Appendix A. We report results averaged over 5 runs with different random seeds where applicable. Our code and trained model checkpoints will be released upon acceptance.

AI Use Statement

We used AI to help polish the writing (grammar, phrasing, and clarity) of this manuscript and to assist with boilerplate code during implementation. All research ideas, experimental design, analysis, and conclusions are our own. The authors reviewed and take full responsibility for all content, including any AI-assisted text and code, in this paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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. *arXiv preprint arXiv:2305.13653*, 2023.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong GOH Mong, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 30531–30548, 2024.

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

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

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

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

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Mishra, Taipeng Tian, Qi Guo, et al. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations*, volume 2026, pp. 2690–2709, 2026.

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.

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

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

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018.

Chen He, Shenshen Li, Zheng Wang, Hua Chen, Fumin Shen, and Xing Xu. Chatting with interactive memory for text-based person retrieval. *Multimedia Systems*, 31(1):31, 2025.

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

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

Fanzhi Jiang, Su Yang, Mark W. Jones, and Liumei Zhang. From attributes to natural language: A survey and foresight on text-based person re-identification. *Information Fusion*, 118:102879, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9220–9230. IEEE, 2025b.

Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11189–11196, 2020.

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

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36:61437–61449, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021.

Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *2017 IEEE International Conference on Computer Vision*, pp. 1908–1917. IEEE, 2017a.

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

Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 11057–11067. IEEE, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025.

Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 356–365, 2022.

Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27187–27196. IEEE, 2024.

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

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

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 11140–11150. IEEE, 2023.

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

Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *European Conference on Computer Vision*, pp. 402–420. Springer, 2020.

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

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023.

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

Tianyu Yang, ChenWei He, Xiangzhao Hao, Tianyue Wang, Jiarui Guo, Haiyun Guo, Leigang Qu, Jinqiao Wang, and Tat-Seng Chua. Recall: Recalibrating capability degradation for mllm-based composed image retrieval. *arXiv preprint arXiv:2602.01639*, 2026.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024.

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

Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

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

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

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 978–987, 2024.

Hao Zou, Runqing Zhang, Jin Ding, Xue Zhou, Jianxiao Zou, and Mingzhu Cai. Tackling alignment ambiguity in person retrieval through conversational attribute mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9689–9698, 2026.

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. *Advances in Neural Information Processing Systems*, 37:45666–45702, 2024.

A Datasets Description

CUHK-PEDES (Li et al., 2017b) is the first dataset for the text-to-image ReID task. It contains a total of 40,206 images from 13,003 identities. Each image was annotated by two different annotators. We follow its official data splitting protocol, where the training set comprises 34,054 images from 11,003 identities, the validation set contains 3,078 images from 1,000 identities, and the testing set contains 3,074 images from 1,000 identities.

ICFG-PEDES (Ding et al., 2021) consists of 54,522 images from 4,102 identities. Each image has one caption. Its training set comprises 34,674 image-text pairs from 3,102 identities, while the testing set includes the remaining 19,848 image-text pairs from 1,000 identities.

RSTPReid (Zhu et al., 2021) contains 20,505 images from 4,104 identities across 15 different cameras. Each identity has 5 images from different cameras, and each image is paired with 2 captions. According to the official data splitting protocol, the training set includes image-text pairs from 3,701 identities, while both the validation and testing sets contain 200 identities.

Interactive-PEDES (Lu et al., 2025) is the first Chat-based interactive dataset in the TIReID, built upon CUHK-PEDES (Li et al., 2017b) and ICFG-PEDES (Ding et al., 2021). It contains a total of 54,749 images from 13,051 identities. Each image is annotated with an initial description and a multi-round dialogue. The training set comprises 47,376 images from 11,543 identities, and the testing set contains 7,373 images from 1,508 identities.

MInterPEDES (Bai et al., 2026) is the first multimodal dataset in the interactive TIReID. It contains a total of 37,128 images from 12,003 identities, with 74,256 annotations in total. It follows the data splitting protocol of CUHK-PEDES (Li et al., 2017b), where the training set comprises 34,054 images from 11,003 identities, and the testing set contains 3,074 images from 1,000 identities.

B Implementation Details

We build VCIR upon Qwen3-VL-8B-Instruct (Bai et al., 2025a). LoRA adapters (Hu et al., 2021) are inserted into the query and value projections of every self-attention layer of the language backbone. As the retriever we adopt a frozen IRRA model built on a CLIP ViT-B/16 visual encoder and a CLIP-Xformer text encoder. It is pretrained for 30 epochs on the large-scale synthetic TIReID dataset HAM-PEDES (Jiang et al., 2025b) and takes 384×128 images as input. We also extend the CLIP text context from 77 to 192 tokens. The number of candidates is set to $K = 3$. The clue prompt length is set to $L = 32$. We study the effect of the candidate number K in Appendix D.

Table 5: Training configuration of VCIR.

Hyperparameter	Value
LoRA rank r	64
LoRA scaling α	128
LoRA learning rate	4×10^{-5}
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	1×10^{-2}
Temperature τ	2×10^{-2}
Batch size	64
Training epochs	5
Precision	bfloat16
Max sequence length	4096
GPUs	8× RTX 3090

Evaluation protocol. We adopt four metrics. R@1, R@5 and R@10 are the proportions of queries whose top-1, top-5 and top-10 results contain an image of the target identity, respectively. mAP is

the mean over queries of the average precision computed over all images of the target identity. mINP is the ratio between the number of relevant images and the rank of the last retrieved relevant image, and thus reflects the cost of finding the hardest match. BRI (Lee et al., 2024) summarizes a whole session and rewards dialogues that push the target to the top within few rounds.

C Generalization to Text-to-Image Person Re-identification

In this experiment, we evaluate the generalization ability of VCIR on traditional text-to-image person re-identification (TIReID) using three widely used benchmarks, *i.e.*, CUHK-PEDES (Li et al., 2017b), ICFG-PEDES (Ding et al., 2021) and RSTPReid (Zhu et al., 2021). All the compared methods adopt CLIP as their backbone. We simulate the user with an MLLM, which takes all the captions of the target identity as its memory. In each round, the simulated user provides one new piece of information about the target based on the dialogue history. With its interaction module trained on Interactive-PEDES, VCIR is evaluated after three rounds of interaction on each benchmark.

As shown in Table 6, VCIR achieves R@1 of 85.59%, 69.79% and 76.10% on the three benchmarks, with the highest R@1 and mAP among all compared methods. This result demonstrates that the interaction ability generalizes well across datasets. We note that the advantage is the smallest on ICFG-PEDES, which is constructed as a fine-grained benchmark with identity-centric and detailed annotations (Ding et al., 2021). Such queries are already discriminative enough to localize the target, leaving a narrower margin for the visual clues extracted by VCE.

Table 6: Comparison with TIReID methods on three standard benchmarks

Method	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
IRRA (Jiang & Ye, 2023)	73.38	89.93	93.71	66.10	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
RDE (Qin et al., 2024)	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
TBPS-CLIP (Cao et al., 2024)	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	61.95	83.55	88.75	48.26
VCIR (Ours)	85.59	94.22	96.34	79.14	69.79	81.83	86.30	45.19	76.10	89.05	92.45	61.67

D Parameter Study

In this experiment, we investigate the optimal value of K. As K determines the number of candidate images observed at each round, it governs both the retrieval accuracy and the inference cost. We vary K in $\{1, 3, 5\}$ on Interactive-PEDES and MInterPEDES, and report the retrieval performance together with the evaluation time on the whole testing set.

Figure 3 shows that $K = 3$ offers the best trade-off between accuracy and evaluation time on Interactive-PEDES. Compared with $K = 1$, $K = 3$ improves R@1 by 5.4%, while further enlarging K to 5 brings only 0.9% additional R@1 at the cost of a longer evaluation time. Figure 4 shows that MInterPEDES is less sensitive to K in terms of R@1, where R@1 varies within 2% across all tested values. We therefore adopt $K = 3$ on both datasets as a unified setting.

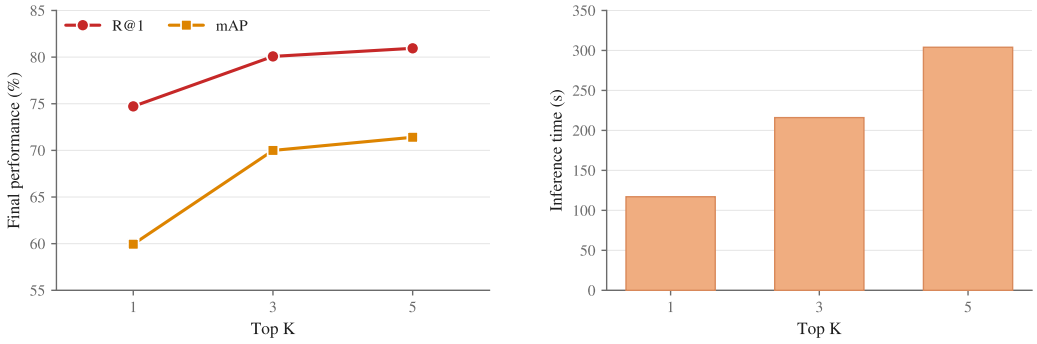


Figure 3: Parameter study on the top-K value on Interactive-PEDES

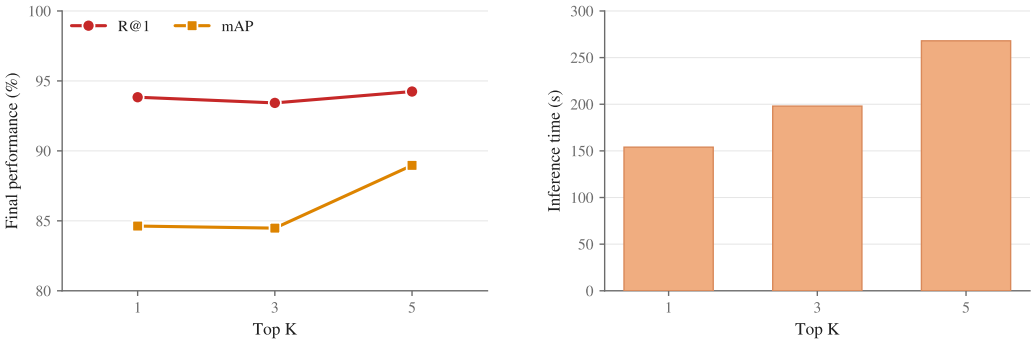


Figure 4: Parameter study on the top-K value on MInterPEDES

E Perturbed Initial Descriptions

In Section 4.5, we study whether VCIR remains effective when the initial description T_0 is vague or erroneous. For each test example in Interactive-PEDES, we construct six perturbed versions of T_0 using fixed rules. The question-answer pairs remain unchanged across these versions. Table 7 shows the example of each rule together with its clean description. The vague perturbations remove details while remaining consistent with the target. The erroneous perturbations introduce incorrect attributes or replace the description. Since the feedback is generated from the clean description, it does not explicitly identify the perturbation in T_0.

Table 7: Examples of the perturbed initial descriptions on Interactive-PEDES.

Initial description	Clean description	Perturbed description
	Vague	
Color words removed	*He is wearing a light blue striped polo shirt, black shorts, and purple shoes.*	*He is wearing a striped polo shirt, shorts, and shoes.*
Second half truncated	*A man is wearing a black jacket, dark pants, and a graphic t-shirt.*	*A man is wearing a black jacket.*
Fully uninformative	*The man is wearing a gray coat, blue jeans, and a dark backpack.*	*A person.*
	Erroneous	
All colors replaced	*I saw a man wearing a red jacket and carrying a black bag.*	*I saw a man wearing a green jacket and carrying a white bag.*
Gender reversed	*A man is wearing a bright red puffer jacket, blue jeans, and white sneakers.*	*A woman is wearing a bright red puffer jacket, blue jeans, and white sneakers.*
Identity substituted	*He is wearing a dark coat and light blue jeans.*	*I saw a man walking away, dressed in a long coat and casual pants.*

F Complete Instruction for VCE

Table 8 presents the complete instruction P used in VCE. At each round n, the placeholders $\{T_0\}$ and $\{F_n\}$ are filled with the initial description and the accumulated feedback, respectively. The three rules correspond to the three failure modes discussed in Section 3.2.

Table 8: Complete intent-aligned instruction P for VCE.

Instruction:

We are performing an interactive text-to-image person retrieval task in which the user will provide additional information about the target person at each round to guide the next retrieval.

The target person is described by the initial description and user feedbacks provided so far. The K input images are the current top-K candidates, ordered from highest to lowest retrieval score. These images may not include the target person.

Now, the initial description is "{$\boldsymbol{T_0}$}", and the user feedbacks is "{$\boldsymbol{F_n}$}".

Use the above information and the input images to build a complete representation of the target person, following these rules:

1. Preserve confirmed target attributes.

2. Emphasize user feedbacks.

3. Ignore appearance details in the candidate images that differ from the target person.

Finally, encode this representation into the <|retr|> token as the query for retrieval.

G QUALITATIVE RESULTS

As illustrated in Figure 5 and Figure 6, VCIR progressively refines the retrieval result on Interactive-PEDES and MInterPEDES as the dialogue proceeds.

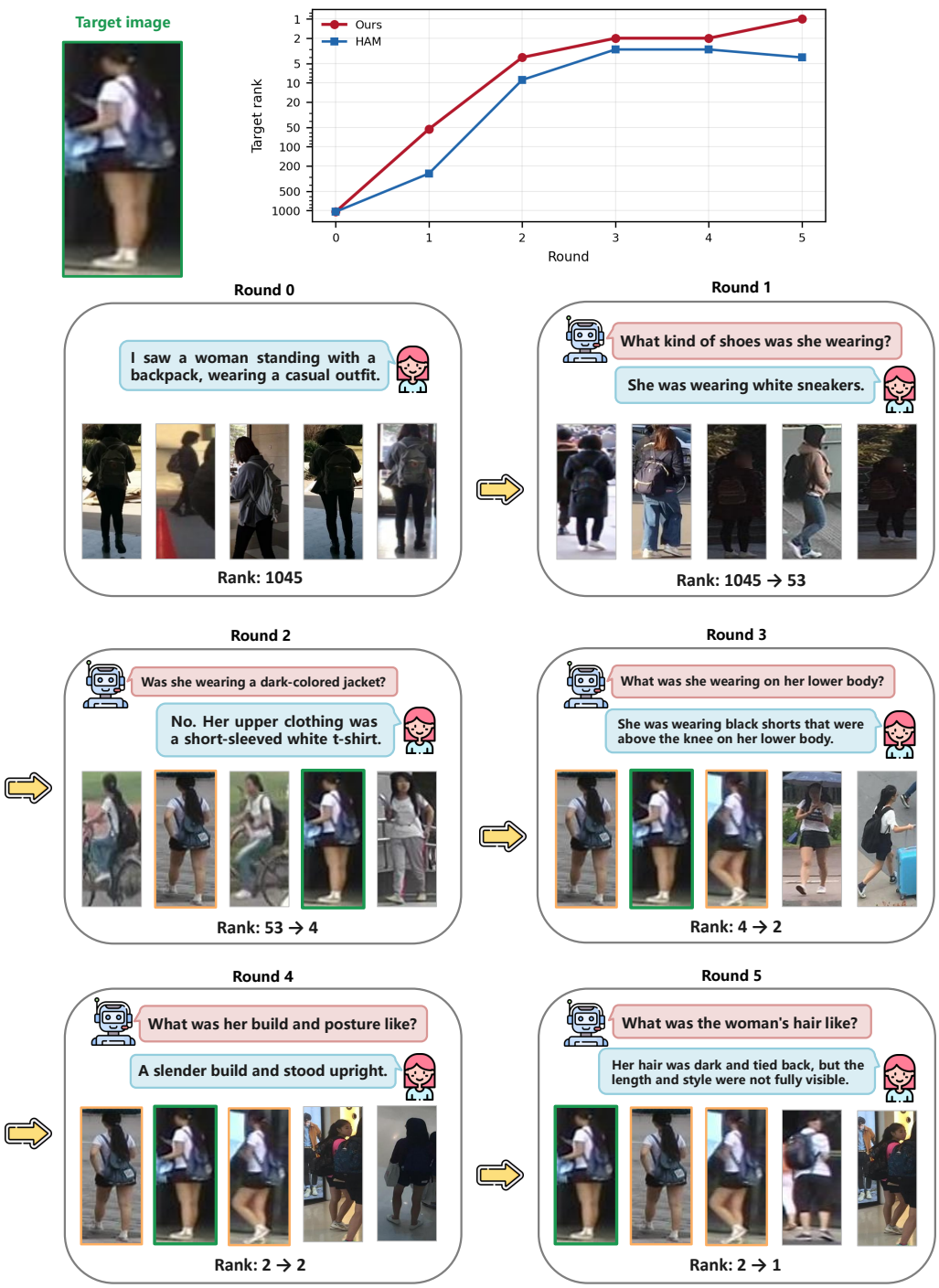


Figure 5: Qualitative result on Interactive-PEDES.

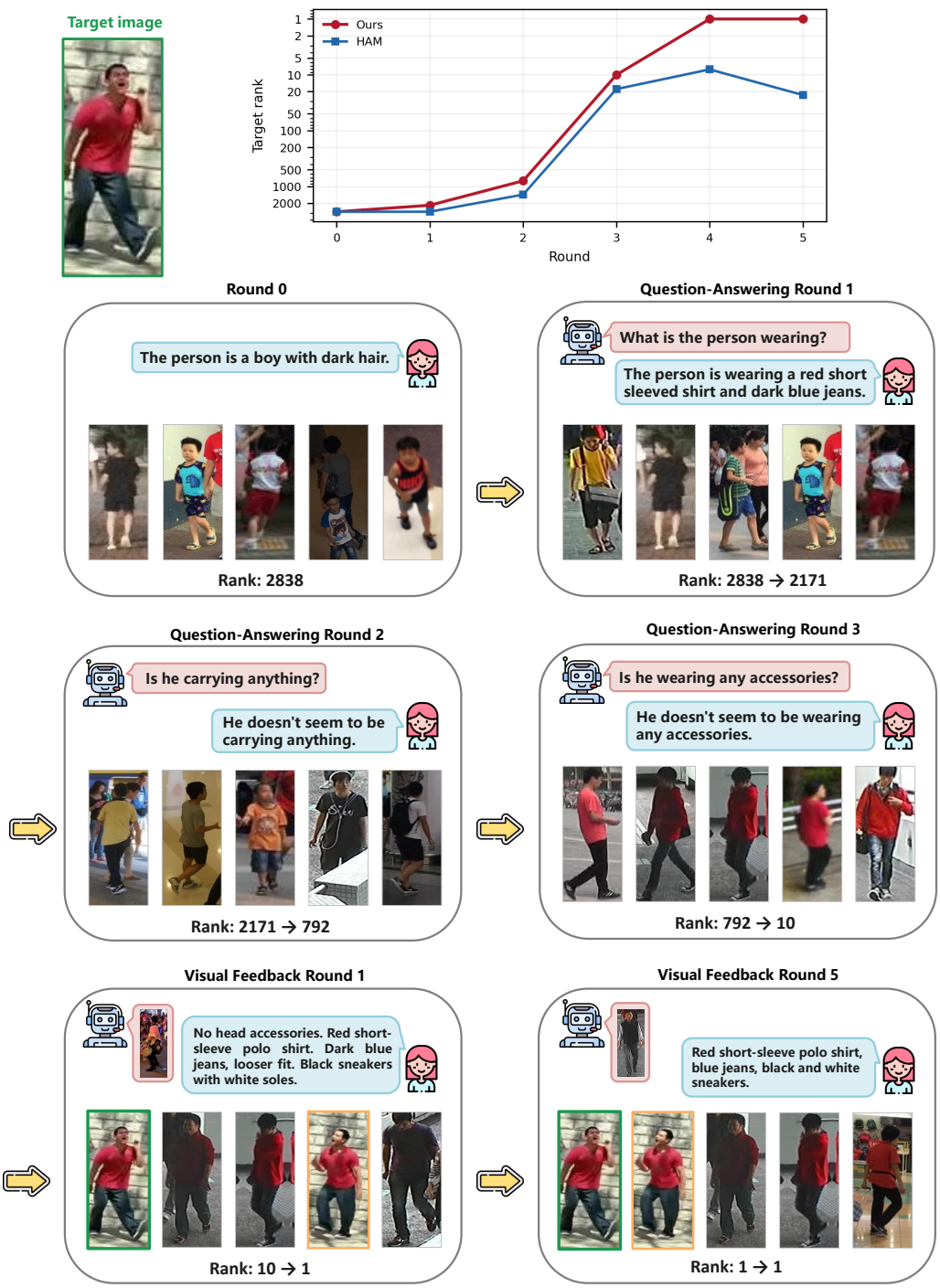


Figure 6: Qualitative result on MInterPEDES.