Supplementary Material: Tackling Alignment Ambiguity in Person Retrieval through Conversational Attribute Mining

Hao Zou[1] Runqing Zhang[1] Jin Ding[1] Xue Zhou[2,3,1*] Jianxiao Zou[3] Mingzhu Cai[3]
[1] University of Electronic Science and Technology of China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China

In this supplementary material, we provide additional results and details to further support our main paper. Specifically, we include:

- A. Algorithm Process
- B. Dataset Descriptions
- C. Question Templates
- D. Samples with original query and generated summary
- E. Examples of Retrieval
- F. Ablation study on the fusion weight

A. Algorithm Process

The training process of the proposed CECA (Conversation-Enhanced Cross-modal Alignment) framework begins by initializing the backbone encoders with the pretrained CLIP weights, while all conversation-related modules, the Conversation Attribute Mining (CAM) are randomly initialized. For each image–text pair in the training set, CECA first employs a multimodal large language model (MLLM) to conduct multi-round visual question answering based on a set of predefined question templates. The generated dialogue responses are summarized into a compact description that captures fine-grained pedestrian attributes. The original image and the summarized text are then encoded by the visual and textual encoders to obtain their respective token-level representations. Next, the image tokens and conversation tokens are fed into the Bidirectional Cross-Attention Mixer (BCM), where self-attention and cross-attention are alternately applied to refine both modalities. The refined tokens are fused through a gated residual connection to form the conversation-enhanced visual embedding. To achieve robust alignment, a Confidence-Aware Weighting Loss (CAWL) is computed, which dynamically adjusts the contribution of each sample based on the consistency between the image, summary, and text features. The total training objective combines the global and local Triplet Alignment Loss (TAL) terms with the CAWL term to jointly optimize cross-modal consistency.

B. Dataset Descriptions

We conduct experiments on three commonly used text-to-image person retrieval (TIPR) datasets, namely CUHK-PEDES, RSTPReid, and ICFG-PEDES, which are widely adopted benchmarks for evaluating cross-modal alignment and retrieval performance. A brief introduction of each dataset is given below.

CUHK-PEDES is the first large-scale benchmark dedicated to text-to-image person retrieval. It contains 40,206 images corresponding to 13,003 unique person identities, and each image is annotated with two textual descriptions, resulting in a total of 80,412 sentences. Following the official data split, we use 34,054 images for training, 3,078 for validation, and 3,074 for testing. This dataset provides diverse visual conditions and natural language descriptions, making it the most widely used standard for TIPR evaluation.

RSTPReid is another large-scale benchmark constructed from the MSMT17 dataset. It consists of 20,505 pedestrian images belonging to 4,101 identities, with each image paired with two textual descriptions. The official training, validation and test split follows 3,701 and 200 and 200 identities. Compared with CUHK-PEDES, RSTPReid includes richer illumination and pose variations, posing additional challenges for robust cross-modal alignment.

ICFG-PEDES is a challenging benchmark that emphasizes fine-grained semantic alignment between text and image. It contains 54,522 images of 4,102 unique identities, each associated with a single natural language description. The dataset is collected from MSMT17 and contains more noisy and diverse descriptions, reflecting complex real-world scenes. Following prior works, we use 3,102 identities for training and 1,000 identities for validation, as the dataset does not include an official test split.

C. Question Templates

To extract structured and fine-grained pedestrian attributes, we design several attribute-oriented question templates used during the multi-turn dialogue process. The question templates are listed as follows:

- **Q1:** What is the gender of the person in the image? Describe the hair of the person.
- **Q2:** What kind of clothes is the person wearing on the upper body? Is the person wearing a hat?
- **Q3:** What kind of pants or lower-body garment is the person wearing?

*Corresponding author: Xue Zhou (zhouxue@uestc.edu.cn).

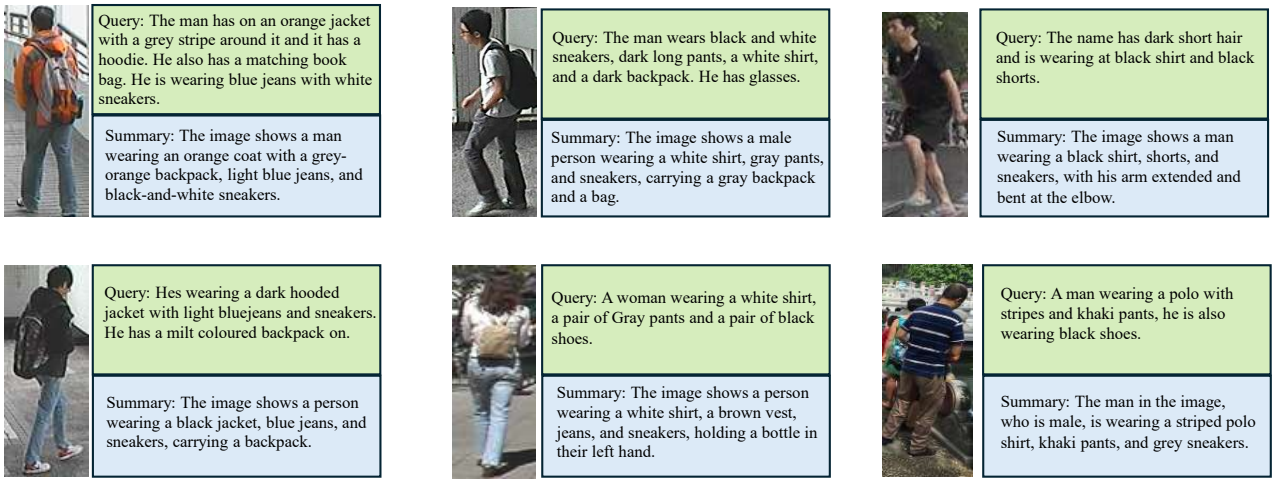

Figure 1. Examples of image, original query and generated summary in CUHK-PEDES

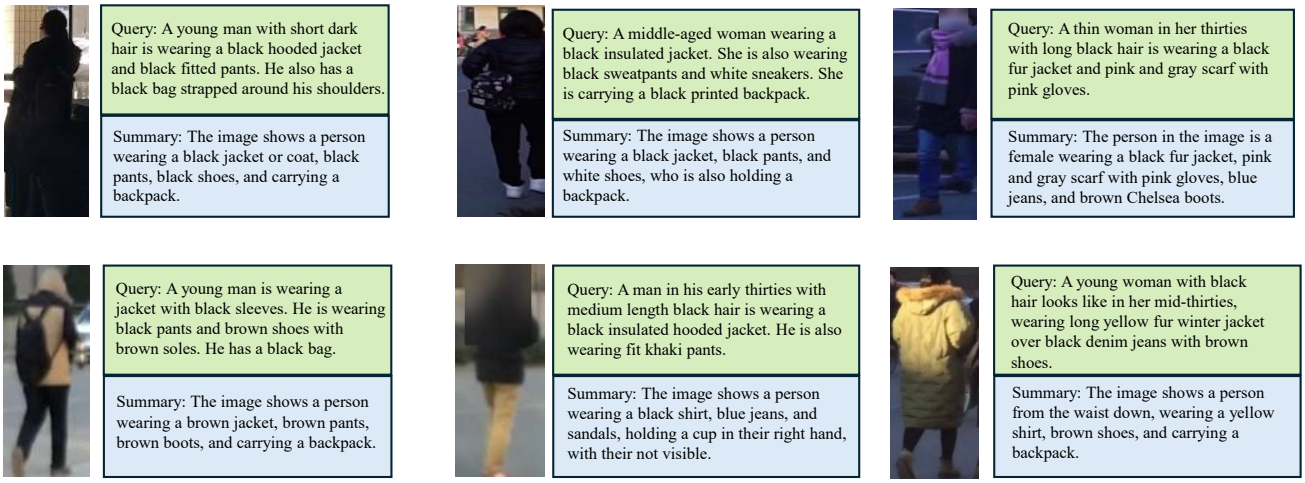

Figure 2. Examples of image, original query and generated summary in ICFG-PEDES

- **Q4:** What kind of shoes is the person wearing?
- **Q5:** Is the person holding anything in hand?

The MLLM is ultimately required to summarize the answers to all questions together with the visual content to produce a coherent description of the person in the image.

Require: Based on the image and the answers to the previous questions, please provide a summary description of the person.

These templates serve as the basis for guiding the MLLM to produce interpretable and fine-grained attribute descriptions.

D. Samples with original query and generated summary

To better illustrate the capability of the proposed CECA framework in extracting fine-grained semantic information, we provide qualitative samples from the CUHK-PEDES, RSTPReid, and ICFG-PEDES datasets. For each dataset, we randomly select six examples, each consisting of the original pedestrian image, its corresponding textual query, and the generated summary produced by the Conversation Attribute Mining (CAM) module. These summaries are obtained through multi-round visual dialogues between the multimodal large language model (MLLM) and the visual encoder, which capture detailed and discriminative attributes such as clothing color, appearance patterns, and carried objects. Unlike retrieval examples, these samples are intended solely to visualize the semantic refinement process, showing how the generated summaries enrich textual semantics and reduce ambiguity in cross-modal representation learning.

E. Examples of Retrieval

We provide qualitative retrieval examples comparing our CECA framework with the IRRA baseline. Each example contains two rows: the top row shows the retrieval results produced by IRRA, while the bottom row presents the results obtained by our method. A total of four representative query cases are displayed. These examples clearly demonstrate that CECA retrieves more accurate and semantically aligned pedestrian images, especially in scenarios involving

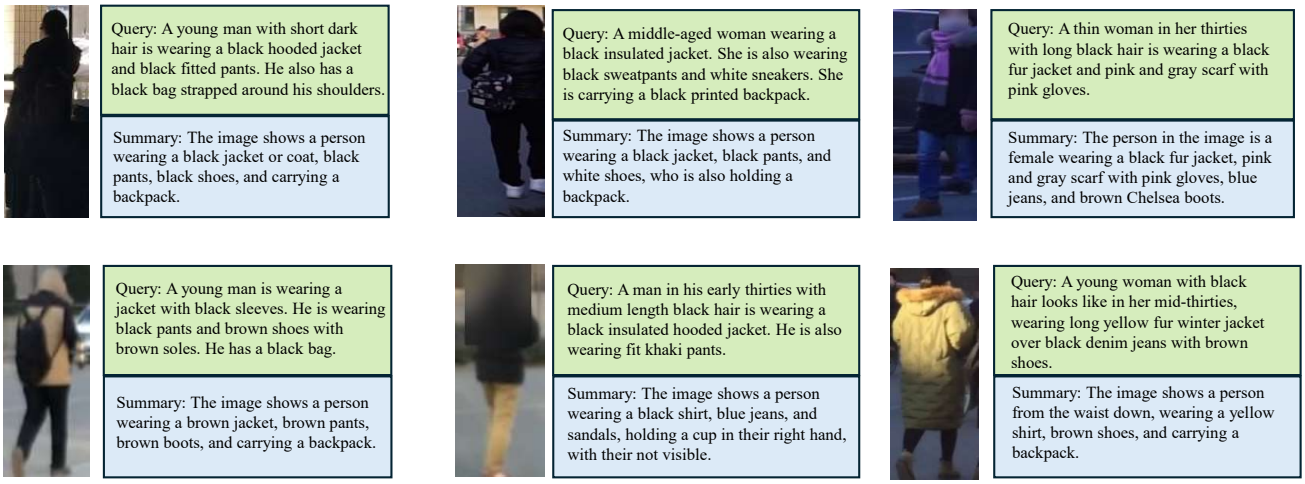

Figure 3. Examples of image, original query and generated summary in RSTPReid

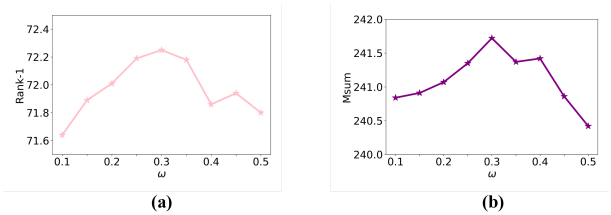

Figure 4. Parameter sensitivity analysis of the fusion weight ω on the ICFG dataset.

visually similar distractor identities.

As shown in the examples, CECA generally achieves better Rank-1 performance by retrieving the correct identity at the first position more frequently than IRRA. However, we also observe that in the subsequent positions of the rank list, the ground-truth image may appear farther down compared to IRRA. This phenomenon provides insight into why our method obtains strong Rank-1 accuracy but relatively lower mAP.

The underlying cause is related to the interaction between different views of the same identity and the MLLM-generated summaries. When images of the same person are captured from significantly different viewpoints, the MLLM may produce summaries that emphasize view-specific attributes (e.g., frontal vs. back view clothing cues). Such variations inadvertently amplify the intra-identity view discrepancy, causing some same-ID images to be ranked lower in the list despite correct semantic alignment at the top positions. Consequently, while CECA excels at identifying the most relevant match, the expanded view variance introduced by dialogue-generated summaries leads to a decline in overall mAP.

F. Ablation study on the fusion weight

We conduct a parameter analysis on the fusion weight ω for combining the global similarity and the refined similarity during retrieval. The experiments are performed on the ICFG dataset, and the results are illustrated in Figure 4.

As shown in the Rank-1 curve, the performance generally increases as ω grows from 0.1 to 0.3, reaching the highest Rank-1 accuracy at $\omega = 0.3$. Beyond this point, the accuracy begins to decline. A similar trend is observed in the Msum metric: the score steadily rises and also peaks around $\omega = 0.3$, followed by a noticeable drop when ω exceeds 0.4.

Query1:
A man with short black hair is wearing a black down jacket, a crimson shirt and dark blue jeans, pulling a trolley case.

Query2:
This female walker is wearing a long grey coat with fur collar and some white writing on it. She also wears a hat with zebra pattern and black pants.

Figure 5. Examples of Retrieval in RSTPReid

Query3:
The man is wearing a yellow screen printed tee shirt, camouflage cargo pants, and blue sneakers with white soles. He carries a black backpack with a water bottle in the left side pocket.

Query4:
Asian person wearing green short sleeve shirt, blue jeans, open toes sandals carrying a brown shoulder bag.

Figure 6. Examples of Retrieval in CUHK-PEDES