Cross-modal Fuzzy Alignment Network for Text-Aerial Person Retrieval and A Large-scale Benchmark

Supplementary Material

7. CoT-based Caption Generation Framework

To generate high-quality, fine-grained person captions that are closely aligned with visual information, we propose a staged description generation framework based on the Chain-of-Thought (CoT) paradigm. This framework decomposes the complex process of visual understanding and text generation into three tightly connected reasoning stages, thereby enhancing the accuracy, controllability, and interpretability of the descriptions. In the following, we provide a detailed overview of the implementation of each stage.

7.1. Human Attribute Parsing

For all collected pedestrian images, we begin by performing automated data filtering to eliminate low-quality samples that may introduce noise into the subsequent description-generation pipeline. The multimodal vision–language model (VLM) is able to automatically detect and discard problematic images, such as those that are severely blurred, poorly illuminated, extremely low in resolution, or heavily occluded based on predefined cleanliness criteria. This ensures that all retained images meet the minimum requirements for recognizability and reliable attribute extraction. After data cleaning, each remaining image is processed using the Figure 6 attribute-parsing prompt specifically designed for this step. The prompt provides explicit, structured instructions that guide the VLM to conduct fine-grained analysis of the person, covering multiple dimensions such as body parts, clothing categories, appearance attributes, and carried objects. Benefiting from the rigorously defined output format in the prompt, the model produces a standardized JSON representation that enumerates all detected attributes. Each attribute is accompanied by essential elements, including the attribute name, visual evidence from the image, and a visibility tag (e.g., visible, partially visible, or not visible). This stage enables high-quality and low-noise pedestrian attribute extraction without manual annotation, providing a precise and structured semantic foundation for the subsequent steps in the CoT reasoning process, such as semantic abstraction, sentence refinement, and final description synthesis.

7.2. Initial Caption Generation

After obtaining the structured set of pedestrian attributes, we proceed to generate an initial natural-language description of the person. In this stage, the multimodal large language model is instructed to act as a witness and is guided using the prompt template shown in Figure 7, enabling it to compose a first-impression description based on the attribute JSON file derived from the previous attribute-parsing process. Specifically, the model treats all non-empty fields in the JSON as observable evidence, under the constraints defined in the prompt, integrates these discrete attributes into a coherent, natural, and human-like descriptive sentence. The prompt explicitly requires the model to follow several principles during generation. First, each non-empty attribute must appear exactly once in the sentence to avoid omission or redundancy. Secondly, the ordering of attributes may be rearranged to ensure logical linguistic flow. Thirdly, list-like outputs or direct concatenation of attribute strings are prohibited, and the description must instead be phrased as a fluent witness-style narrative. Finally, any attribute left empty in the JSON must not appear in the generated sentence, ensuring that no hallucinated details are introduced. Through this "witness-perspective" formulation, the model transforms structured attribute elements into natural-language text, completing the conversion from discrete attribute units to an initial person description.

7.3. Caption Review and Correction

After obtaining the initial natural-language description, we introduce a vision-guided review and correction mechanism to further enhance the textual accuracy and visual consistency of the generated content. Specifically, we feed both the initial description and the original person image into another vision-language model, using the prompt template shown in Figure 8 to instruct the model to act as a visual auditor that examines and corrects the description in a fine-grained manner. In this stage, the model performs visual verification on each semantic component of the initial description, identifying parts that are inconsistent with the image, ambiguously expressed, or missing key details. The prompt enforces strict correction rules, all decisions must rely solely on clearly observable visual evidence from the image; any mismatched content must be directly replaced, removed, or completed within the sentence. The overall fluency and coherence of the description must be preserved; and no hallucinated or unverifiable details may be introduced. To ensure interpretability, the model outputs a correction trace alongside the refined description, documenting each modification action (such as retention, replacement, deletion, or addition) and its corresponding visual justification. Through this vision-guided, reflective correction process, the system effectively eliminates potential hallucina-

Table 6. Comparison results on the RSTPReid, CUHK-PEDES, and ICFG-PEDES datasets.

Method	Ref	RSTPReid					CUHK-PEDES					ICFG-PEDES				
		R-1	R-5	R-10	mAP	RSum	R-1	R-5	R-10	mAP	RSum	R-1	R-5	R-10	mAP	RSum
UniPT [35]	ICCV23	51.85	74.85	82.85	-	209.55	68.50	84.67	90.38	-	243.55	60.09	76.19	82.46	-	218.74
IRRA [14]	CVPR23	60.20	81.30	88.20	47.17	229.70	73.38	89.93	93.71	66.13	257.02	63.46	80.25	85.82	38.06	229.53
MALS [49]	MM23	61.90	80.60	89.30	48.08	231.80	74.05	89.48	93.64	66.57	257.17	64.37	80.75	86.12	38.85	231.24
CFAM [55]	CVPR24	62.45	83.55	91.10	49.50	237.10	75.60	90.53	94.36	67.27	260.49	65.38	81.17	86.35	39.42	232.90
IRLT [26]	AAAI24	61.49	82.26	89.23	-	232.98	74.46	90.19	94.01	-	258.66	64.72	81.35	86.31	-	232.38
Propot [48]	MM24	61.87	83.63	89.70	47.82	235.20	74.89	89.90	94.17	67.12	258.96	65.12	81.57	86.97	42.93	233.66
SAP-SAM [41]	MM24	62.85	82.65	89.85	-	235.35	75.05	89.93	93.73	-	258.71	63.97	80.84	86.17	-	230.98
NAM [39]	CVPR24	68.50	87.15	92.10	53.02	247.75	76.82	91.16	94.46	69.55	262.44	67.05	82.16	87.33	41.51	236.54
LERF [51]	PR25	46.75	71.30	81.60	-	199.65	65.84	84.24	90.22	-	240.30	57.23	76.64	83.11	-	216.98
HCA [4]	MM25	69.70	86.95	92.45	53.72	249.10	77.44	91.23	94.61	68.92	263.28	66.95	82.34	87.06	40.13	236.35
HAM [15]	CVPR25	71.69	87.85	**93.30**	55.19	252.84	**77.71**	**91.42**	94.57	**69.68**	**263.70**	68.25	83.30	**88.15**	42.30	239.70
WoRA [38]	WWW25	66.85	86.45	91.10	52.49	244.40	76.38	89.72	93.49	67.22	259.59	68.35	83.10	87.53	**42.60**	238.98
Ours	-	**72.15**	**87.95**	93.20	**55.88**	**253.30**	77.52	91.26	**94.90**	69.33	263.68	**68.59**	**83.41**	88.13	42.11	**240.13**

tions or inaccuracies in the initial description, ensuring that the final text is semantically precise, visually reliable, and faithfully aligned with the original image.

8. Evaluation Metric

Rank-k: Rank-k(k=1,5,10) measures the probability that at least one correct matching image appears within the top-k retrieved results for a given textual query. Specifically, Rank-1 indicates the probability of the correct image being the top-1 result, Rank-5 within the top-5 results, and Rank-10 within the top-10 results. Higher Rank-k values reflect better retrieval accuracy at different candidate thresholds.

mAP: mAP evaluates the overall ranking quality of the retrieved images by computing the average precision for each query and then averaging over all queries. It reflects both the correctness and ranking order of all relevant images in the retrieval list. A higher mAP indicates more precise and ordered retrieval.

RSum: RSum is defined as the sum of Rank-1, Rank-5, and Rank-10 scores. It provides a single metric summarizing the overall retrieval performance across different Rank thresholds. A higher RSum implies consistently strong performance across multiple top-k retrieval levels.

9. Generalization Analysis

To further assess the generalization capability of the proposed method, we conduct comprehensive evaluations on three widely used text–image person retrieval benchmarks, RSTPReid [6], CUHK-PEDES [24] and ICFG-PEDES [54]. As summarized in the Table 6, our approach achieves 72.15% Rank-1, 55.88% mAP, and an RSum of 253.30% on RSTPReid, outperforming all existing state-of-the-art methods across the board. On CUHK-PEDES, the method attains a Rank-10 accuracy of 94.90%, demonstrating competitive performance comparable to the most advanced models. Moreover, on ICFG-PEDES, we obtain 68.59% Rank-1 and an RSum of 240.13%, establishing a new SoTA.

You are an expert assistant specializing in fine-grained pedestrian attribute recognition. Given an image of a person, your task is to extract detailed, fine-grained pedestrian attributes following the strict schema defined below. For every attribute, you must output a structured triple: "attribute", "evidence", and "visibility".
General Guidelines：
1. Output Structure: Attribute – Evidence – Visibility
For each attribute, you must output:
attribute: A flat string describing the semantic attribute (e.g., "male", "wearing a white T-shirt"). If not visible → set to "".
evidence: A short, precise phrase describing what visual evidence supports this attribute. If the attribute is not visible → set to "".
visibility: Must be one of: "visible" – clearly observable, "partially_visible" – partly blocked / ambiguous, "not_visible" – no visual evidence
2. Flat Strings Only: All values MUST be flat strings. No lists, no objects, no sentence-level descriptions.
3. Strict Non-Guessing Rule: Do not infer or guess any attribute. If the attribute cannot be clearly determined → set "attribute": "" and "visibility": "not_visible".
4. Maximum Detail When Visible: For clothing/hair/accessories/items: Describe colors, styles, textures, patterns, materials, layers.
Use phrases like: "black short hair", "wearing a white short-sleeved T-shirt with yellow geometric patterns".
5. Evidence Must Be Explicit: For visible attributes, evidence should directly point to the visual cue, e.g.: "short hair visible above ears", "white T-shirt with printed yellow triangles".
6. No full sentences: Use descriptive fragments only.
Output Format (JSON)： You must output exactly this JSON format. Do not add, remove, rename, or restructure any field.

```
{
 "gender": { "attribute": "", "evidence": "", "visibility": "" },
 "age": { "attribute": "", "evidence": "", "visibility": "" },
 "height": { "attribute": "", "evidence": "", "visibility": "" },
 "body_shape": { "attribute": "", "evidence": "", "visibility": "" },
 "skin_color": { "attribute": "", "evidence": "", "visibility": "" },
 "hair": { "attribute": "", "evidence": "", "visibility": "" },
 "upper_clothing": { "attribute": "", "evidence": "", "visibility": "" },
 "lower_clothing": { "attribute": "", "evidence": "", "visibility": "" },
 "socks": { "attribute": "", "evidence": "", "visibility": "" },
 "shoes": { "attribute": "", "evidence": "", "visibility": "" },
 "accessories": { "attribute": "", "evidence": "", "visibility": "" },
 "bag": { "attribute": "", "evidence": "", "visibility": "" },
 "other_items": { "attribute": "", "evidence": "", "visibility": "" },
 "action_status": { "attribute": "", "evidence": "", "visibility": "" }
}
```

Figure 6. Prompt used for human attribute parsing, implemented with the Qwen-7B model.

You are a witness model responsible for converting structured pedestrian attributes into a natural-language description while providing a transparent reasoning trace. You will receive a JSON object containing pedestrian attributes extracted from an image. Each key corresponds to an attribute category, and each value is either: a non-empty flat string describing the visible attribute, or an empty string if the attribute is not visible.
Your task consists of two outputs:
Output Part 1 — Initial Caption: Generate one coherent, natural, and human-like sentence describing the pedestrian using every non-empty attribute exactly once.
Rules:
1. You must use every non-empty attribute value exactly once.
2. You may reorder attributes freely to produce a fluent, grammatical sentence.
3. Do not include empty attributes.
4. The sentence must sound natural and cohesive, not like a list.
5. Do not mention attribute names (e.g., "gender", "upper clothing"); only describe the person.

Output Part 2 — Reasoning Trace
Provide a structured explanation describing how each non-empty attribute contributed to the final sentence, including:
how the attribute value was rewritten or integrated into the sentence, where it appears conceptually in the sentence order,
how it logically connects with other attributes.
Format the trace as a JSON object:

```
"trace_gen": {
  "gender": "how this attribute was used",
  "upper_clothing": "how this attribute was used",
  ...
}
```

Output Format:
Return the final answer using the exact structure below:

```
{
 "Initial Caption": "your generated sentence here",
 "Reasoning Trace": {
   "attr1": "explanation",
   "attr2": "explanation",
   ...
  }
}
```

Figure 7. Prompt used for initial caption generation, implemented with the Qwen-7B model.

Your task is to review and correct the pedestrian description generated in Stage-2, based on the original image. You will receive: Original Image of a pedestrian and Initial Caption.
Rules:
1. Correct any semantic mismatches, omissions, or inaccuracies.
2. Allowed actions per attribute:
   "retain" → keep the phrase as is
   "replace" → modify the phrase to match the image
   "delete" → remove if incorrect or misleading
   "add" → insert missing information if clearly visible
3. Do not invent details that cannot be visually confirmed.
4. Maintain the description as one coherent sentence.

Correction Guidelines: For every non-empty attribute used in Initial Caption, record:
"attribute" → the attribute name
"action" → one of "retain", "replace", "delete", "add"
"visual_evidence" → brief description of image region or visual cue that justifies the action

Output Format:
Return exactly the following JSON:

```
{
 "Refined Caption": "your corrected sentence here",
 "Trace Correction": {
   "attribute1": {
     "action": "retain|replace|delete|add",
     "visual_evidence": "short description of supporting region"
   },
   "attribute2": {
     ...
   }
 }
}
```

Refined Caption must be one complete, coherent sentence. Do not output anything outside this JSON.

Figure 8. Prompt used for caption review and correction, implemented with the Chatgpt-4o model.

The strong performance across these text–image benchmarks indicates that our method effectively captures robust cross-modal semantic associations even beyond aerial scenarios. This further demonstrates that the proposed approach not only excels in the challenging text–aerial person retrieval task but also maintains strong generalization and robustness in conventional text–image retrieval settings.

Figure 9. Visualization of Top-10 retrieval results for selected queries from the AERI-PEDES dataset. For each query, the first row shows the baseline results and the second row shows the results of our method. Green bounding boxes denote correct retrievals, while red bounding boxes indicate incorrect ones.

10. Qualitative Analysis

As shown in the Figure 9, we present the Top-10 retrieval results for several queries from the AERI-PEDES dataset, where the first row of each result set corresponds to the baseline method and the second row corresponds to our method. For the first query, the baseline method fails to retrieve any correct samples, while our method successfully retrieves four accurate instances, demonstrating stronger semantic discrimination ability under challenging viewpoint conditions. In the second query, our method retrieves an additional correct instance whose lower body is occluded due to the camera angle. This improvement is attributed to our proposed fuzzy token alignment module, which enables robust text–image alignment even when only upper-body cues are visible in the aerial image. In the third query, our method retrieves two more correct results than the baseline. Both of these images are captured from near-vertical viewpoints, where the available visual evidence is extremely limited. Nevertheless, the fuzzy alignment mechanism effectively focuses on reliable cues, leading to more accurate cross-modal matching. These qualitative results clearly indicate that our method is more effective at bridging the semantic and viewpoint gap between textual descriptions and aerial-view person images, significantly improving the robustness and accuracy of cross-modal alignment.