Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification

Supplementary Material

This supplementary material includes five sections. Section A illustrates the instruction we use in Qwen2.5-7B to remove discriminative words in pedestrian descriptions. Section B conducts ablation study on the value of some hyper-parameters in our method. The SDM loss [3] used in ReID model training is introduced in Section C. In Section D, we provide more details about the ablation experiment in Section 4.3, including the extraction of the descriptive templates in the experiment (3) in Table 1 and the detailed setting of DBSCAN [2] clustering in the experiment (7). Section E shows our MLLM-generated captions of the pedestrian images, which demonstrates the effectiveness of our method in simulating various human descriptive styles.

A. Instruction of Qwen2.5

In HAM, we remove information that reflects pedestrian identity from the textual descriptions for style feature extraction. To achieve this, we employ Qwen2.5-7B [8] to replace the words in a caption that describe pedestrian attributes, *e.g.*, clothing category, color, age, hairstyle with vague ones that are generic across identities. We design the instruction for Qwen2.5 and also provide some in-context examples to better align LLM's output with our expectations. The complete instruction for Qwen2.5 is shown in Figure A.

B. Hyper-Parameters Tuning

Following Section 4.3, we employ only the training set of CUHK-PEDES [4] to train the LLaVA1.6 [5] by our HAM method and then annotate 0.1 million images with one caption per image for a quick ablation study. Then we explore the optimal values of the following three hyper-parameters, *i.e.*, the cluster number K_1 of KMeans [6] and K_2 of UPS, the token number M of each style prompt and the number Q of samples contained in each UPS cluster.

Number of Clusters. In this experiment, we explore the optimal number of clusters for KMeans [6] and our UPS when clustering style features (DBSCAN automatically determines the number of clusters during clustering). As shown in Figure B, UPS consistently outperforms KMeans under all clustering numbers. Moreover, the performance reaches its best when the cluster number is 1,000.

Number of Tokens in Each Style Prompt. In HAM, we use a style prompt consisting of M tokens to represent a unique style of a specific cluster. When M is too small, each style prompt contains insufficient parameters to represent a specific description style. Conversely, when M is too large, the total number of style prompt parameters becomes excessive, increasing the training overhead. As shown in Figure C, when M exceeds 10, the performance of our HAM is insensitive to changes in M. Therefore, we select M=10 as the final hyper-parameter.

Number of Samples in Each UPS Cluster. In UPS clustering, we assign a fixed number Q of samples that are closest to each cluster center to obtain all clusters. Therefore, the appropriate number of samples in each cluster is crucial for effectively learning the representation of each style. In this experiment, we evaluate three different values: Q = 150, 200, and 250. As shown in the results in Table A, the best average performance is achieved with Q=200, with an average Rank-1 accuracy of 48.34% and average mAP of 36.07% across the three datasets. Therefore, the optimal value of Q is determined to be 200.

Value of the Hyper-Parameter β. In UPS, we use β to control the sampling range for style cluster centers within style feature space. When β is too small, many meaningful cluster centers are ignored, resulting in insufficient exploration of style categories within the style space. When β is too large, the risk of sampling meaningless cluster centers increases. Figure D shows that the optimal value for β is 7.

C. SDM Loss in ReID Model

In Section 3.3, we adopt the Similarity Distribution Matching (SDM) loss proposed by [3] to optimize the ReID model. Specifically, given a mini-batch of B image-text pairs, *i.e.*, $\{(\boldsymbol{v}_i, \boldsymbol{t}_j), y_{i,j}\}(1 \leq i, j \leq B)$, where $\boldsymbol{v}, \boldsymbol{t}$ represent the holistic visual and textual features obtained from the ReID model, respectively. $y_{i,j} = 1$ indicates that $(\boldsymbol{v}_i, \boldsymbol{t}_j)$ is a matched pair according to the pedestrian identity, while $y_{i,j} = 0$ indicates an unmatched pair. We then compute the ground-truth matching probability $q_{i,j}$ and the matching probability $p_{i,j}$ for $(\boldsymbol{v}_i, \boldsymbol{t}_j)$ as follows:

$$p_{i,j} = \frac{\exp(sim(\boldsymbol{v}_i, \boldsymbol{t}_j)/\tau)}{\sum_{b=1}^{B} \exp(sim(\boldsymbol{v}_i, \boldsymbol{t}_b)/\tau)}, \tag{1}$$

$$q_{i,j} = \frac{y_{i,j}}{\sum_{b=1}^{B} y_{i,b}}, \tag{2}$$

where $sim(\boldsymbol{v}_i, \boldsymbol{t}_j) = \boldsymbol{v_i}^\top \boldsymbol{t_j}/\|\boldsymbol{v_i}\|\|\boldsymbol{t_j}\|$ denotes the cosine similarity between $\boldsymbol{v}_i$ and $\boldsymbol{t}_j$, and τ is a temperature coefficient set to 0.02. Then, the SDM loss from image to text

Task Description

<User>: Please replace the phrases describing clothes or appearance of the people in the input sentence. Use 'colored' for color descriptions, 'top', 'bottom', 'shoes', 'accessory', or 'belongings' for clothing-related phrases. Replace held objects with 'object', body parts like 'knee', 'shoulder', 'hand', or 'arms' with 'body part', and substitute gender references with 'person'. Use 'pattern' for any clothing designs or straps, and replace all viewpoint-related terms with 'viewpoint'.
[**Highlight**]: Ensure that other words in the sentence and the structure remain unchanged. Here are some examples.

In-context Examples

<Input>: A dark long haired woman is wearing a striped t-shirt, black pants and a pair of sneakers and is holding a green and white tote bag.
<Output>: A colored-hair person is wearing a patterned top, colored bottom and a pair of shoes and is holding a colored object.

<Input>: The man is wearing a pair of dark sneakers. He is mainly bald. He has on dark shorts with a belt and a blue t-shirt with white writing on the back.
<Output>: The person is wearing a pair of colored shoes. He/She has a hairstyle. He/She has on colored bottom with accessories and a colored top with colored pattern.

<Input>: She is seen from behind with long hair wearing a light colored t-shirt with a pair of dark capris, and a tan purse slung across her body from her left shoulder to her right hip.
<Output>: He/She is seen from a viewpoint with a hairstyle wearing a colored top with a pair of colored bottom, and a colored belongings slung across his/her body from his/her body part to his/her body part.

<Input>: This man with dark hair has his back to the camera and is wearing a coat with a hood and puffy sleeves, a pair of shorts with red lines going through them.
<Output>: This person with a hairstyle has his/her body part to the camera and is wearing a top with designed elements and designed sleeves, a pair of bottom with colored pattern going through them.

<Input>: She has her hair pulled back in a tight braid. She is wearing an officer hat and yellow vest. She is in her early teens or twenties and is fair-skinned.
<Output>: He/She has a hairstyle. He/She is wearing an accessory and colored top. He/She is in a certain age and is of a certain skin tone.

Actual Question

<User>: Please process the following sentences according to the given example:"{caption}".
Just output the processed sentence without any explanation.
<Assistant>:

Figure A. The instruction we adopted for Qwen2.5-7B.

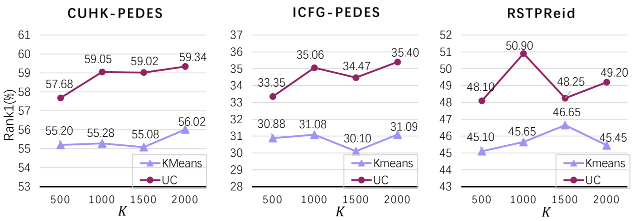


Figure B. Results of different clustering numbers in KMeans and UPS.

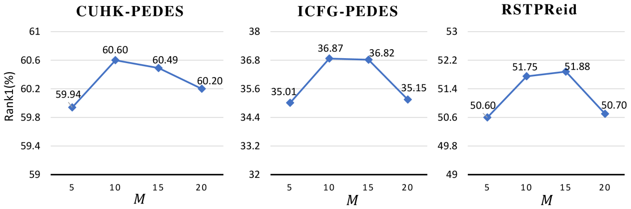


Figure C. Results of different values of M.

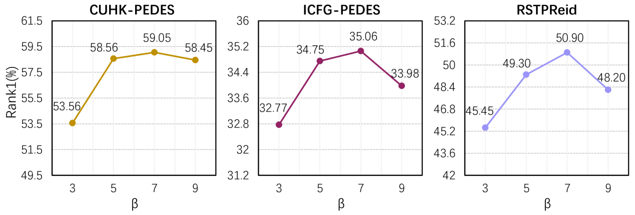


Figure D. Results of different values of β.

can be computed by:

$$\mathcal{L}_{i2t} = \frac{1}{B}\sum_{i=1}^{B} KL(\mathbf{p}_i \| \mathbf{q}_i) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\epsilon}\right), \tag{3}$$

where $\mathbf{p}_i$, $\mathbf{q}_i$ denote the predicted probability distribution and ground truth matching distribution for the i-th image, respectively. ϵ is a small number to avoid numerical problems and we set ϵ as 1e-8. The SDM loss $\mathcal{L}_{t2i}$ from text to image can be computed in a similar way by exchanging their roles in Eq.(1). Finally, the complete SDM loss is rep-

Table A. Results of different values of Q.

Q	CUHK-PEDES		ICFG-PEDES		RSTPReid		*Avg.*	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
150	57.24	51.51	36.45	19.52	47.20	35.94	46.96	35.66
200	59.05	52.92	35.06	18.34	50.90	36.94	48.34	36.07
250	58.28	52.57	36.35	18.25	49.95	36.71	48.19	35.84

resented as follows:

$$\mathcal{L}_{sdm} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}. \tag{4}$$

D. Details in Ablation Study

Extraction of Descriptive Templates. In Section 4.3, we extract 68,126 templates according to all human-annotations in the training set of CUHK-PEDES, and use these templates to enhance the diversity of MLLM-generated captions. In this section, we will provide more details on how the templates are extracted.

In contrast to the textual descriptions we extract by the Qwen2.5 in the HAM approach, a pedestrian descriptive template requires explicit placeholders for specific attributes descriptions, such as the template used in [7]: *"With [hair description], the [person/woman/man] is wearing [clothing description] and is also carrying [belongings description]"*. To achieve this, we employ an advanced LLM (*i.e.*, Llama3-8B [1] in our experiment) and design appropriate instructions to prompt the LLM to extract a template from a pedestrian description. Moreover, to better align the LLM's output with our expectations, we also provide some in-context examples in the instruction. The detailed instruction and in-context examples are shown in Figure E.

DBSCAN Clustering. In the ablation study in Section 4.3, we also compare the performance of the DBSCAN [2] clutering method used in our HAM. DBSCAN is a density-based clustering algorithm that does not require specifying the number of clusters in advance. Instead, it automatically determines the number of clusters and clustering results based on the density distribution of samples. However, the algorithm is highly sensitive to the choice of two hyper-parameters: the neighborhood radius (*i.e.*, ε) and the minimum number of points (*i.e.*, $MinPts$). These parameters define the distance threshold for two samples to be considered density-connected and the minimum number of neighbors required within a cluster, respectively.

For the high-dimensional style features (*i.e.*, 512-dimensional vectors), selecting appropriate values for these hyper-parameters becomes challenging. Furthermore, the significant variations in the density distribution of style features result in DBSCAN classifying a large portion of the samples as noise points. Consequently, DBSCAN is not well-suited for our HAM framework.

In the experiment (7) in Table 1, we conduct extensive tuning of the two hyper-parameters and identify the optimal values as ε=1.5 and $MinPts$=3. Under these settings, the number of clusters formed is 823. Out of a total of 68,126 style feature samples, 37,273 are classified as noise points.

E. Visualization of Obtained Captions

To visualize the diverse captions by specifying different style prompts in MLLM, we utilize the LLaVA1.6 [5] which is trained on the CUHK-PEDES and ICFG-PEDES datasets by our HAM and randomly select five style prompts to generate five different textual descriptions for the same pedestrian images, as shown of (1)-(5) in Figure F. The results show that our method can mimic different description styles of human annotators and enhance the diversity of MLLM-generated captions.

References

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 1, 3

[3] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2023. 1

[4] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 1

[5] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: improved reasoning, ocr, and world knowledge (2024). *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*. 1, 3

[6] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967. 1

[7] 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 *CVPR*, 2024. 3

[8] Qwen Team. Qwen2. 5: A party of foundation models, 2024. 1

Task Description

<**User**>: Please extract a pedestrian descriptive template from the input pedestrian description. Replace the gender description as "[person]", hair description as "[hair description]", upper clothes description as "[upper clothes description]" and lower clothes description as "[lower clothes description]“, shoes description as "[shoes description]" and accessory description as "[accessory description]“. If these descriptions exist, please replace them, otherwise please ignore them.
[**Highlight**]: Ensure that other words in the sentence and the structure remain unchanged. Here are some examples.

In-context Examples

<**Input**>: The woman is looking down and carrying a plant. She is wearing a light-colored dress and dark shoes.
<**Output**>: The [person] is looking down and carrying a [accessory description]. [person] is wearing a [lower clothes description] and [shoes description].

<**Input**>: The man has a white shirt on and black pants. He is wearing blue and white Nike's and also has glasses.
<**Output**>: The [person] has a [upper clothes description] on and [lower clothes description]. He is wearing [shoes description] and also has [accessory description].

<**Input**>: A man wearing a blue shirt, a pair of black pants, a pair of brown shoes and a hat on his head and a scarf around his neck.
<**Output**>: A [person] wearing a [upper clothes description], a pair of [lower clothes description], a pair of [shoes description] and a [accessory description] and a [accessory description].

<**Input**>: The woman wears a black and white patterned shirt. She wears a skirt with black and white horizontal stripes. Her hair is in a bun and she pushes a red stroller.
<**Output**>: The [person] wears a [upper clothes description]. [person] wears a [lower clothes description]. [person's] [hair description] and [person] pushes a [accessory description].

<**Input**>: A man in blue jeans, leather shoes and a red jacket with yellow sleeves holds a briefcase and a back pack in an airport.
<**Output**>: A [person] in [lower clothes description], [shoes description] and a [upper clothes description] holds a [accessory description] and a [accessory description] in an airport.

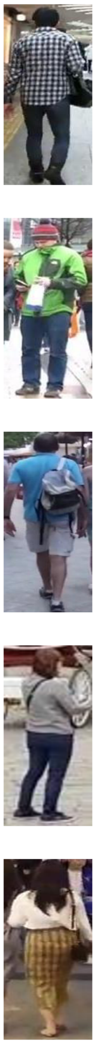
<**Input**>: The man is facing away and his back is visible. He has short black hair and he is wearing black pants. He has on a brown pullover and is wearing black shoes.
<**Output**>: The [person] is facing away and [person's] back is visible. [person] has [hair description] and [person] is wearing [lower clothes description]. [person] has on a [upper clothes description] and is wearing [shoes description].

<**Input**>: The man wears a dark blue shirt with white lettering on the back. He wears black shorts and black and white tennis shoes.
<**Output**>: The [person] wears a [upper clothes description]. [person] wears [lower clothes description] and [shoes description].

Actual Question

<**User**>: Please process the input description according to the given example:“{caption}”.
Please just return the template without any explanation.
<**Assistant**>:

Figure E. The instruction we adopted for Llama3-8B.



(1) The man is wearing a plaid shirt, jeans, and black shoes.
(2) He has short black hair and is wearing a plaid shirt. He is also holding a bag in his hand.
(3) A man in his thirties is wearing a black and white checkered shirt, blue jeans, black shoes, and is carrying a black bag.
(4) He is wearing a plaid shirt, dark blue jeans, and black shoes. He is carrying a black bag in his hand.
(5) A young man with straight short black hair is wearing a plaid shirt with black and white checkered pattern. He is also wearing blue jeans and black shoes. He is carrying a black bag in his hand.

(1) The man is wearing a green jacket, blue jeans, and brown shoes.
(2) He has a red beanie and is wearing a green jacket. He is also holding a white bag in his hand .
(3) A man in his mid-30s is wearing a green jacket, blue jeans, and brown shoes. He is carrying a white bag in his hand.
(4) He is wearing a green jacket, blue jeans, and brown shoes. He is carrying a white bag in his left hand.
(5) A young man with a red beanie is wearing a green jacket with black pants. He is also wearing brown shoes.

(1) The man is wearing a blue polo shirt, khaki shorts, and flip flops.
(2) He has short black hair and is wearing a blue shirt. He is also carrying a backpack on his back.
(3) A man in his mid-30s is wearing a blue shirt and grey shorts. He is carrying a black and white bag.
(4) He is wearing a blue shirt and shorts, and he is carrying a backpack.
(5) A man with straight short black hair is wearing a blue shirt and white shorts. He is also wearing black shoes and carrying a black bag on his shoulder.

(1) The woman is wearing a grey sweatshirt, dark blue jeans, and black sneakers.
(2) She has long dark hair and is wearing a gray jacket. She is also holding a bag in his hand.
(3) A woman in her twenties is wearing a grey sweatshirt and blue jeans. She is carrying a black handbag.
(4) She is wearing a grey sweatshirt, blue jeans, and black shoes. She is carrying a black bag.
(5) A woman with straight shoulder-length brown hair is wearing a grey sweatshirt and blue jeans. She is also wearing black sneakers with black soles. She is carrying a black bag on her shoulder.

(1) The woman is wearing a white blouse with a black and gold patterned skirt.
(2) She has long black hair and is wearing a white shirt. She is also carrying a black bag in her hand .
(3) A young woman in her early-20s is wearing a white blouse and a brown and white checkered skirt. She is carrying a black handbag.
(4) She is wearing brown sandals, a white shirt, and a long brown skirt with a white top. She is carrying a black purse.
(5) A young woman with straight shoulder-length black hair is wearing a white blouse with a black and brown checkered skirt. She is also wearing brown sandals and carrying a black purse.

Figure F. Diverse captions with different style prompts.