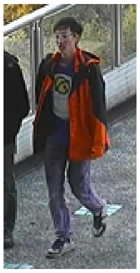
DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID

Supplementary Material

In this supplementary, we include details that we were not able to provide in the main paper due to space constraints. It includes,

- Details on image text descriptions with examples
- Description summarization prompt and experiment details
- Implementation details of the NBDetach module
- Disentangled identity Feature visualization
- Examples of enhanced performance on baseline failure cases.

1. Image Text Description Example



The following is an example of a conversation between the user and the visual-language model CogVLM [3]. The corresponding input image is displayed to the left of the passage . The description of biometric characteristics includes details about the individual's physical attributes, such as age, gender, height, and body type. Non-biometric factors encompass elements like hairstyle, clothing, posture, and background. As demonstrated in the example, the VLM accurately extracts the essential information required for our task but also introduces some errors. For instance, hairstyle information is incorrectly classified as a biometric attribute, and the person's jacket is described as having black and white patches, though the white patches are actually part of the shirt underneath.

- **Input**: *Describe the individual's overall* ***physical appearance****, including estimated age, gender, height (e.g., short, average, tall based on surroundings if applicable), and build(e.g., slender, average, robust).*
- **CogVLM**: *The individual appears to be a young male, possibly in his late teens to early twenties. He has short, dark hair and glasses. Based on the surroundings, he seems to be of average height and build.*
- **Input**: *Provide a detailed description of the* ***hair features****, including hair color, style, and length*
- **CogVLM**: *The individual has short, dark hair.*
- **Human**: *Detail the type of* ***clothing*** *the person is wearing(the style, colors, and any visible logos or patterns), shoes and any accessories (e.g., glasses, watches, jewelry).*
- **CogVLM**: *The individual is wearing a red jacket with black and white patches. Underneath the jacket, he has a blue shirt with a white logo or emblem on it. He is also wearing purple pants and black and white shoes. He is wearing glasses.*
- **Input**: *Describe the person's* ***posture*** *when the image was taken (e.g., standing straight, leaning, walking). Note any characteristics of the gait, such as limping, brisk walking, or any peculiarities that stand out. Mention the alignment and demeanor suggested by the posture (e.g., confident, tired, hurried).*
- **CogVLM**: *The person appears to be walking, with a somewhat brisk gait. His posture is upright, suggesting confidence, and he seems to be moving forward purposefully.*
- **Input**: *Describe the setting or* ***background*** *in which the person is located (e.g., urban street, office, park). Identify any objects or elements in the vicinity that the person is interacting with or that are relevant to the scene. Assess the general atmosphere or mood of the environment, such as busy, tranquil, chaotic, etc.*
- **CogVLM**: *The person appears to be in an urban setting, possibly a street or a pedestrian area. There are glass railings and a metal structure visible in the background, suggesting a public space or a walkway. The general atmosphere seems to be calm and quiet.*

2. Description Summarization

We use GPT-4 [1] to summarize all the biometric descriptions from one person to one prompt. Additionally, we conduct experiments to summarize all clothing descriptions for a specific clothing class into a unified clothing text description, ensuring that all images sharing the same clothing label have identical clothing text features. However, it is important to note that, except for this section, the results presented in this paper do not use summarized clothing descriptions to avoid reliance on extra clothing labels.

2.1. Summary prompts

The following are the summary prompts used to generate biometric and clothing descriptions. These summary prompts were initially created using GPT-4 and refined manually to align with our specific requirements better.

- *Summarize the individual's overall* ***physical appearance****, only including estimated age, gender, height (e.g., short, average, tall based on surroundings if applicable), and build (e.g., slender, average, robust) based on the following information. Do not summarize the hairstyle. Only include the information that most sentences agree on.*
- *Summarize the type of* ***clothing*** *the person is wearing(the*

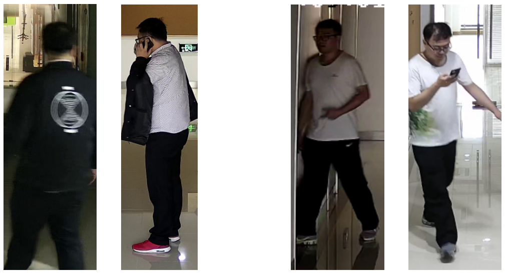


Figure 1. **Input images for the summary descriptions**. The four images, all representing the same individual, are used for summarizing the biometrics and clothing descriptions. The first two images are examples of the person wearing **Cloth1**, and the second two images are for **Cloth2**.

style, colors, and any visible logos or patterns), shoes and any accessories (e.g., glasses, watches, jewelry) based on the following information. Using three to four describing sentences. Only include the information that most sentences agree on.

2.2. Summary Description Example

Here, we present examples of biometric and clothing description summaries, with corresponding example images displayed in Fig. 1. The below text examples illustrate that biometric summarization effectively captures the primary physical traits of the individual, though the descriptions are generally broad and lack fine-grained specificity. In contrast, clothing summaries provide more specific details, such as the red shoes and white patterns on the jacket for Cloth1, along with the overall style. However, these summaries may unintentionally incorporate details from other images that are not visually present in the current image, introducing extraneous information. In conclusion, while biometric summaries enhance consistency and robustness in capturing identity-related features, clothing summaries may lead to inaccuracies by including details not relevant to the image being analyzed.

- **Biometrics summary description**: *The individual appears to be a male, primarily estimated to be in his late 20s to early 30s, with a consensus also leaning towards late teens to early 20s in several descriptions. He consistently has a medium or average build and is of average height based on the surroundings.*
- **Clothing summary description for Cloth1**: *The individual is dressed in a casual style, predominantly featuring black clothing, including a jacket with a white logo on the back and black pants. The jacket's logo is described as either a stylized sand timer or circular. Red shoes or sneakers add a pop of color to the otherwise monochrome outfit. While there is mention of glasses being worn in two descriptions, the presence of other accessories like watches or jewelry is either not mentioned or stated to be absent.*
- **Clothing summary description for Cloth2**: *The individual is dressed in a casual style, wearing a white t-shirt that features a small, greenish logo or design on the left side. They are also wearing black pants, complemented by gray sports shoes. Additionally, the person is accessorized with glasses, adding a practical element to their look.*

Bio	Cloth	LTCC		PRCC	
		top1	mAP	top1	mAP
Image	Image	57.7	31.3	67.3	63.8
Summary	Image	**58.2**	**31.6**	**68.5**	**64.7**
Summary	Summary	57.4	30.5	67.5	**64.7**

Table 1. **Image description and summarized description experiment results**. We compare the results of whether to use the image or summarized description for biometric contrastive loss $\mathcal{L}_{C_b}$ and clothing non-biometric contrastive loss $\mathcal{L}_{C_n}$ in the table. LTCC and PRCC datasets under the cloth-changing setting are used.

2.3. Additional experiment results

We study the effect of using summary text descriptions during training. As shown in Tab. 1, the biometric summary caption could increase the accuracy from 57.7% to 58.2% on LTCC, from 67.3% to 68.5% on PRCC. On the other hand, adding the clothes summary description could decrease the performance by 0.8% on LTCC and 1% on PRCC. The summarized biometric descriptions likely help the model capture missing identity information, especially in cases of blurred or occluded images. Conversely, clothing descriptions depend heavily on the specific image, and summarized descriptions may introduce irrelevant information or overlook critical details, negatively affecting disentanglement performance.

3. NBDetach Module Architecture Details

We present a comprehensive description of the architecture of our NBDetach module. For both the biometric and non-biometric projection heads, H_b and H_n, we employ linear transformations with matrix multiplication and bias addition to execute the projection.

The input image feature $\mathbf{f}^i$ has a dimensionality of 1024 for the EVA2-CLIP-L model. The projected biometric and non-biometric image features, $\mathbf{f}_b^i$ and $\mathbf{f}_n^i$, are designed to match the dimensions of their corresponding textual features $\mathbf{f}_b^t$ and $\mathbf{f}_n^t$, which can be 512, 768 or 1024 dimensions. In our experiments, we use the 512-dimensional textual feature for LTCC and 768-dimensional for other datasets. In the gradient reversal layer (GRL), we set the negative scalar α to -1, which changes the gradient sign to negative, i.e. multiplies it by -1.

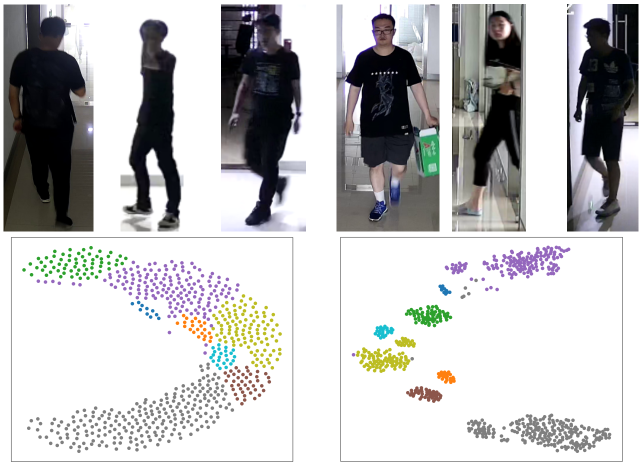


Figure 2. **Identity disentanglement visualization. Top**: Examples of different people with similar outfits. These ID groups are used for identity feature visualization in bottom row. **Bottom**: Identity feature cluster visualization results, with baseline results on the left and DIFFER results on the right. Different colors represent different person IDs. The baseline features exhibit greater dispersion, whereas the features produced by DIFFER demonstrate tighter clustering within individual identity groups.

4. Feature visualization for disentangled identity

We visualize the disentangled identity image feature with t-SNE [2] of the same clothing classes from the clothing textual feature cluster in Fig. 2. The LTCC test dataset is used. As shown in the first row in Fig. 2, all the identities are dressed in a similar all-black style. Features from the baseline method are more dispersed without clear boundaries between different identity classes. On the contrary, the features from different ID groups in the proposed method are separated with clear boundaries. This demonstrates the proposed disentangle method successfully differentiates the identity feature from the encoded image feature space without the influence of similar clothing features.

5. Examples of enhanced performance on baseline failure cases

This section illustrates the effectiveness of DIFFER compared to the baseline model through examples from three benchmark datasets: LTCC, PRCC, and Celeb-reID-light, as shown in Fig. 3. Each row corresponds to examples from one dataset, where (a) represents the query image, (b) shows the top-1 incorrect match by the baseline model, and (c) displays the top-1 correct match by DIFFER.

As seen in the examples, DIFFER demonstrates its ability to address significant challenges that often hinder the baseline model. These challenges include changes in clothing, where individuals are dressed in entirely different outfits; similar poses, where non-biometric factors such as body orientation or posture confuse the baseline model; and overlapping characteristics like hairstyles or environmental settings that lead to false matches. For instance, in the LTCC dataset examples, DIFFER correctly matches individuals despite substantial changes in their attire, where the baseline confuses individuals with similar clothing patterns. Similarly, in the PRCC dataset, DIFFER handles challenging cases where pose similarity between different individuals misleads the baseline. In the Celeb-reID-light examples, DIFFER effectively overcomes confounding factors such as varying environmental contexts and subtle similarities in appearance that the baseline fails to disentangle.

These results highlight DIFFER's robustness in disentangling biometric information from non-biometric factors and its ability to leverage semantic descriptions effectively. By successfully overcoming the limitations of the baseline model, DIFFER significantly improves identification accuracy across diverse and challenging scenarios. This demonstrates the practical applicability of the proposed method in real-world settings where non-biometric interference is prevalent.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[3] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 1

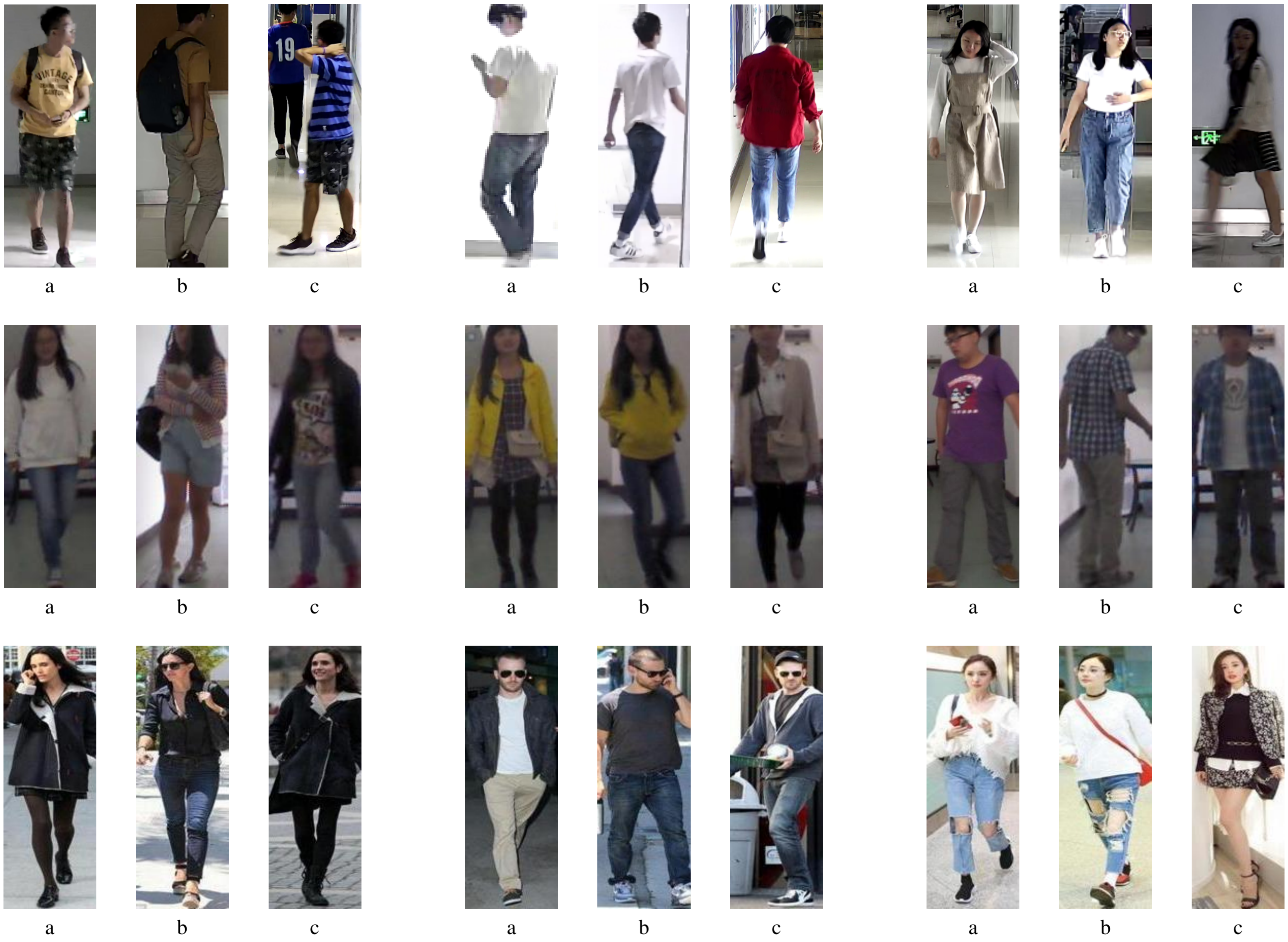


Figure 3. **Example of improvement of DIFFER on different datasets**. From the top to bottom rows are examples from LTCC, PRCC, and Celeb-reID-light respectively. (a: query image; b: baseline method top1 matched error image; c: DIFFER top1 matched correct image)