Automatic Synthetic Data and Fine-grained Adaptive Feature Alignment for Composed Person Retrieval

Delong Liu[1], Haiwen Li[1], Zhaohui Hou[2], Zhicheng Zhao[1,3,4]*, Fei Su[1,3,4], Yuan Dong[1]
[1]Beijing University of Posts and Telecommunications
[2]SenseTime
[3]Beijing Key Laboratory of Network System and Network Culture
[4]Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism
{liudelong, lihaiwen, zhaozc, sufei, yuandong}@bupt.edu.cn
houzhaohui@sensetime.com

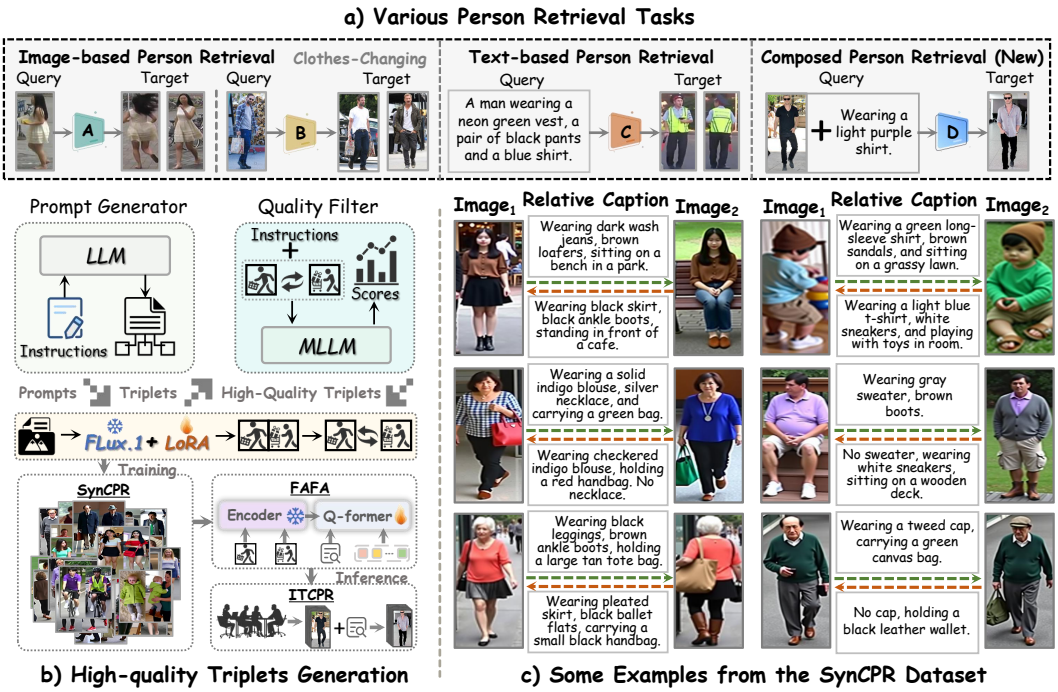

Figure 1: **Overview of our contributions.** (a) Comparison of the proposed composed person retrieval task with several classic person retrieval tasks. (b) Illustration of the proposed automatic high-quality CPR data synthesis pipeline, the proposed training framework FAFA, and the first carefully annotated test set in this domain, ITCPR. (c) Some examples from our fully synthetic SynCPR dataset.

Abstract

Person retrieval has attracted rising attention. Existing methods are mainly divided into two retrieval modes, namely image-only and text-only. However, they are unable to make full use of the available information and are difficult to meet diverse application requirements. To address the above limitations, we propose a new Composed Person Retrieval (CPR) task, which combines visual and textual queries to identify individuals of interest from large-scale person image databases. Nevertheless, the foremost difficulty of the CPR task is the lack of available annotated datasets. Therefore, we first introduce a scalable automatic data synthesis pipeline,

*Corresponding author: Zhicheng Zhao.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

which decomposes complex multimodal data generation into the creation of textual quadruples followed by identity-consistent image synthesis using fine-tuned generative models. Meanwhile, a multimodal filtering method is designed to ensure the resulting SynCPR dataset retains 1.15 million high-quality and fully synthetic triplets. Additionally, to improve the representation of composed person queries, we propose a novel Fine-grained Adaptive Feature Alignment (FAFA) framework through fine-grained dynamic alignment and masked feature reasoning. Moreover, for objective evaluation, we manually annotate the Image-Text Composed Person Retrieval (ITCPR) test set. The extensive experiments demonstrate the effectiveness of the SynCPR dataset and the superiority of the proposed FAFA framework when compared with the state-of-the-art methods. All code and data will be provided at `https://github.com/Delong-liu-bupt/Composed_Person_Retrieval`.

1 Introduction

Person retrieval [1, 2] aims to identify target individuals from large-scale databases and encompasses two primary research directions: image-based person retrieval (IPR) [3] and text-based person retrieval (TPR) [4]. Typically, they rely independently on images or textual queries to identify the intended targets. In fact, in real-world scenarios, visual and textual information are often simultaneously available when searching for specific individuals. For example, when looking for a missing person, people may refer to the past photographs along with a recent verbal description. However, existing methods fail to fully exploit this combined information, resulting in suboptimal retrieval accuracy.

To address this drawback, as shown in Figure 1(a), a novel task named Composed Person Retrieval (CPR) is introduced, which fuses visual and textual information for person retrieval. Similar to Composed Image Retrieval (CIR) [5, 6], the CPR data will also comprise numerous triplets (I_q, T_q, I_t), where each triplet consists of a reference person image (I_q), a relative caption (T_q), and one or more target images4 (I_t). The objective is to effectively locate I_t by exploiting the complementary information between I_q and T_q. Constructing such data requires paired images of individuals with the same identity (ID) and textual descriptions highlighting their differences. However, manual collection and annotation is time-consuming, costly, and often hindered by privacy issues, limiting both the variety and scale of the depicted scenarios. Consequently, this poses a significant challenge to the construction of a comprehensive, high-quality, large-scale training dataset for CPR task.

To cope with these challenges, we propose a scalable automatic CPR data synthesis pipeline, depicted in Figure 1(b). The generation of complex multimodal triplets is achieved by overcoming two key problems: First, how to create pure and diverse textual data. Second, how to leverage the generative models to transform a subset of this text into identity-consistent person images, thus attaining CPR data synthesis. Specifically, this pipeline is decomposed into three stages. First, a Large Language Model (LLM) [7] generates abundant textual quadruples, and each one comprises two image descriptions and two relative captions that connect them. Through carefully designed prompts, the LLM is guided to produce diverse descriptions reflecting a wide range of individuals and states, while effectively capturing relative differences.

In order to solve the second problem, we first fulfill the synthesis of person image-text pairs in the second stage. Considering that directly employing pretrained diffusion models to individually generate I_q and I_t will lead to identity mismatches and discrepancies from real-world distributions. Thus, we fine-tune generative models [8, 9] using real-world data [4] to derive a suitable person image generator firstly. Subsequently, by merging textual prompts, we simultaneously generate a single image containing two related sub-images, which are then cropped into the reference image and the target one, thereby ensuring identity consistency.

In the third stage, rigorous data filtering method is designed to ensure the high quality of triplets. Specifically, a multimodal large language model (MLLM) [10] is applied to evaluates the generated triplets according to four scoring criteria: image quality, identity consistency, text-image alignment, and relative caption quality. After filtering based on these scores, a high-quality, fully synthetic CPR dataset named Synthetic Composed Person Retrieval (SynCPR) can be obtained, and its representative examples are shown in Figure 1(c).

Moreover, we propose a novel framework tailored for CPR task: Fine-grained Adaptive Feature Alignment (FAFA). FAFA strengthens model training by integrating fine-grained dynamic alignment

with bidirectional masked feature reasoning, thereby generating more comprehensive, robust, and fine-grained representations. Finally, in order to conduct an objective evaluation of FAFA's performance, an Image-Text Composed Person Retrieval (ITCPR) test set is carefully constructed and manually annotated, based on widely-used clothes-changing person retrieval datasets such as Celeb-reID [11], LAST [12], and PRCC [13]. Among them, we annotate the relative captions by selecting images of the same identity in different outfits or states, and ultimately form complete triplets. Extensive experiments on ITCPR dataset demonstrate the effectiveness of both the proposed automated triplet synthesis pipeline and the FAFA framework. The main contributions can be summarized as follows:

- A novel cross-modal task, composed person retrieval is proposed for the first time, aiming to address person retrieval by making full use of combined visual-textual information.
- A scalable automatic triplet synthesis pipeline is presented, which greatly alleviates the difficulties in CPR data annotation. Based on this pipeline, the first million-scale, high-quality and fully synthetic CPR dataset named SynCPR is constructed.
- A new CPR framework, called FAFA is proposed, which significantly improves retrieval performance through fine-grained dynamic alignment and bidirectional masked feature reasoning.
- The first carefully annotated test set named ITCPR is constructed, and extensive experiments validate the effectiveness of our proposed methods.

2 Related Work

Person Retrieval. Person retrieval primarily comprises two research directions: IPR and TPR. IPR has been extensively explored from various perspectives, including feature extraction [3, 14, 15], metric learning [16], lightweight architecture design [17–19], multi-branch frameworks [1, 20], and attention mechanisms [21, 22]. A related subtask, clothes-changing image person retrieval (CC-IPR) [11], targets identification across outfit variations and has driven the development of specialized datasets [11, 13, 12] and methods [23–25]. In comparison, TPR emerges later but has progressed rapidly. It focuses on aligning visual and textual features within a unified embedding space. Early TPR approaches emphasize global [26–28] and local [29–33] feature extraction and employ cross-modal matching losses [34] but often have difficulty in balancing efficiency and accuracy. More recently, visual-language pretrained (VLP) models [35–38] have significantly improved retrieval performance through carefully designed auxiliary tasks [39, 40, 2] tailored specifically for TPR fine-tuning. However, despite substantial progress, existing approaches still struggle to effectively integrate visual and textual information for precise identification of specific individuals, which remains an essential and practical requirement. To bridge this gap, we propose the CPR task.

Composed Image Retrieval. CIR [41–43], as a representative compositional learning task [44, 45], jointly leverages image and textual queries for precise image retrieval. CIR has been extensively applied in fashion [5] and real-world domains [6, 46], fostering diverse image-text fusion and training strategies. However, existing supervised CIR methods [47, 48, 42, 43] heavily depend on annotated triplet datasets, inherently limiting their generalizability. To alleviate reliance on annotation, recent zero-shot CIR (ZSCIR) approaches [49] propose techniques such as image-to-pseudo-text conversion [49, 46, 50, 51] or using LLM-generated target descriptions to reformulate CIR as pure text-to-image retrieval [52, 53]. Compared to CIR, CPR imposes stricter constraints on image relevance and places greater emphasis on fine-grained variations during retrieval. Consequently, existing CIR methods generally struggle to maintain effectiveness under the CPR setting.

Diffusion Models. Diffusion models [54, 55] have become the prevailing architecture for image generation, with applications in text-to-image synthesis [56–58], image translation [59–61], and controllable content generation [62–64]. This progress has been accompanied by efficient parameter tuning strategies such as Low-Rank Adaptation (LoRA) [9] and Adapter-based [65] methods, which retain high generation quality while enhancing adaptability. The incorporation of Transformer [66] architectures has led to novel designs like the Diffusion Transformer (DiT) [67], improving scalability and bring about advanced models such as Stable Diffusion 3 [68], PixArt [69], and Flux [8]. Inspired by the above, our work elegantly combines the Flux model with LoRA-based fine-tuning to generate person images that closely resemble visual styles in the real world.

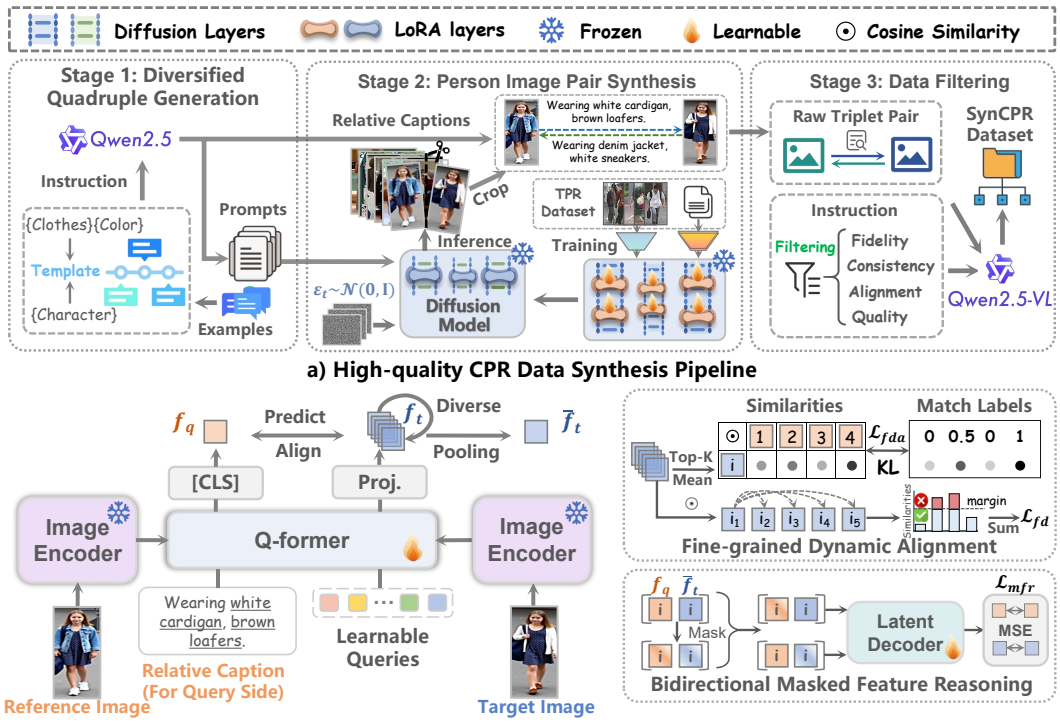

Figure 2: Overall framework of our method. (a) The pipeline for synthesizing high-quality triplets, consisting of three key stages: generation of text quadruples, synthesis of person image pairs, and data filtering. (b) The structure of FAFA. The left part illustrates the training process of the model, while the right part highlights the key objectives employed by FAFA.

3 Method

The overall framework of the proposed CPR method is illustrated in Figure 2, comprising two main components. Section 3.1 introduces the automatic pipeline for synthesizing high-quality CPR data, including textual quadruple generation, identity-consistent image synthesis, and data filtering. Section 3.2 presents the FAFA framework, detailing the model architecture, fine-grained dynamic alignment objectives, bidirectional masked feature reasoning strategies during training, and the inference procedure. To objectively evaluate the proposed method, Section 3.3 outlines the construction of the ITCPR test set.

3.1 High-quality CPR Data Synthesis

Diverse Textual Quadruples Generation. Considering that there is currently a lack of feasible methods for directly generating multimodal triplet data, we propose decomposing this task into the generation of single-modality triplets first and then expanding them into multimodal form, which effectively alleviates this problem. Specifically, an instruction template $\mathcal{P}(Character, Clothes, Color)$ is designed to guide the LLM [7] (denoted as $\mathcal{G}_{llm}(\cdot)$) to produce textual quadruples. Each quadruple comprises two pairs of textual triplets, as expressed in Equation 1:

$$\mathcal{G}_{llm}(p) \rightarrow \langle T_{I_q}, T_{q \rightarrow t}, T_{t \rightarrow q}, T_{I_t} \rangle, \tag{1}$$

where T_{I_q} and T_{I_t} denote the same person with different outfits or states and will later be used to synthesize images I_q and I_t. The relative caption $T_{q \rightarrow t}$ highlights key appearance changes from I_q to I_t, while the reverse caption $T_{t \rightarrow q}$ describes changes in the opposite direction, allowing two usable triplets to be constructed from each quadruple. To enhance diversity and avoid repetitive outputs, each instruction $p \sim \mathcal{P}$ includes multiple descriptive elements and randomly selected high-quality annotated examples. Providing these random elements and examples ensures semantic richness, diversity, output quality, and structural stability (see Appendix A.1 for details). A simplified version of the instruction template is shown below:

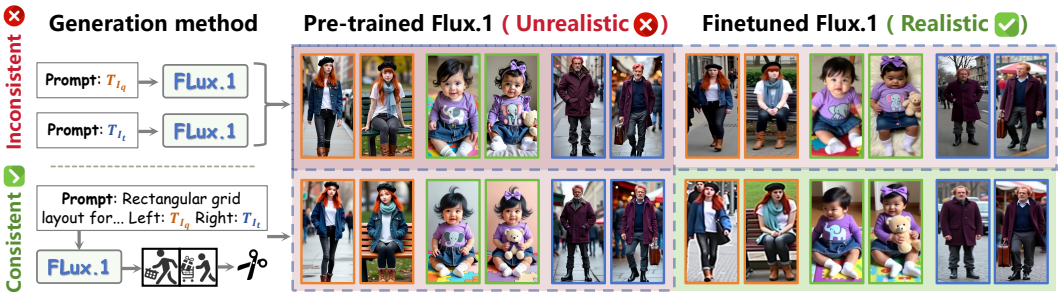

Figure 3: Example pairs of generated person images using different generative models and generation methods under the same text input.

Generate a quadruple satisfying CPR requirements using the following elements: suggested character, clothes, color. Follow the output format and content length from: examples.

Identity-consistent High-quality Image Synthesis. As mentioned before, generative models have been widely applied to text-to-image synthesis. However, most of them are oriented towards natural images and portraits, and methods specifically for generating pedestrian images are still rare. Therefore, person images generated by pretrained models often deviate significantly from the style and distribution of real-world person images encountered in retrieval tasks. To address this, we fine-tune the cross-attention layers of DiT [67] using LoRA [9] on the dataset of person image-text pairs:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad K = \mathbf{W}K\tau\text{txt}(T_{\text{txt}}), \quad V = \mathbf{W}V\tau\text{txt}(T_{\text{txt}}) \tag{2}$$

where Q denotes DiT image features, $\tau_{\text{txt}}(T_{\text{txt}})$ is the text encoder output, and $\mathbf{W}_K$, $\mathbf{W}_V$ are learnable projection matrices. During fine-tuning, only the LoRA components in the cross-attention layers are updated, while all other parameters remain frozen. Given a weight matrix $\mathbf{W} \in \mathbb{R}^{h\times l}$, LoRA introduces trainable matrices $B \in \mathbb{R}^{h\times r}$ and $A \in \mathbb{R}^{r\times l}$, with $r \ll \min(h, l)$, and computes the residual update as $\Delta\mathbf{W} = \beta\gamma BA$, where β controls LoRA strength and γ is a learnable layer-specific scaling factor. The updated weights are then given by $\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}$, enabling parameter-efficient adaptation. Training is guided by a flow-matching objective function [70].

Once fine-tuning is complete, as shown in Figure 3, the generative model's inherent consistency capability, that is, the ability to generate coherent elements within a single image, is ingeniously leveraged to synthesize image pairs with consistent identities, which cannot be achieved through independent generation. Specifically, we first define a layout prefix and merge T_{I_q} and T_{I_t} into a unified prompt. Then, this prompt is input to the model to generate a single image with left and right sub-images. The final images I_q and I_t are obtained by cropping:

Rectangular grid layout for left and right images. Each image is independent, ... Left: T_{I_q}, Right: T_{I_t}.

Furthermore, to maximize textual quadruple utilization, we dynamically adjust β, generating n image pairs for each textual pair (T_{I_q}, T_{I_t}), thus creating $2n$ triplets. Besides, images within the same triplet share a unique ID, while those within groups that share the same relative captions are assigned a common group ID (GID), facilitating label smoothing during training.

Data Filtering. To ensure the quality of generated data, the MLLM [10] is employed to evaluate each generated triplet on a scale from 1 to 10 across four criteria: (1) naturalness of individuals in I_q and I_t (excluding resolution and instead focusing on visual realism, noise, and artifact presence); (2) identity consistency between I_q and I_t; (3) alignment between images and their corresponding descriptions ($I_q \leftrightarrow T_{I_q}$); and (4) CPR task relevance ($I_q + T_q \rightarrow I_t$). Triplets with an average score below a strict threshold of 8.5 are discarded, leading to the removal of approximately 59% of the data.

Based on this pipeline, a large-scale synthetic dataset named SynCPR is constructed, consisting of 1.15 million high-quality triplets. Further implementation details regarding data synthesis (e.g., complete prompt templates and additional visualization examples) can be found in the Appendix A.

3.2 End-to-End Composed Person Retrieval Framework

A new retrieval framework is proposed to achieve end-to-end CPR, where the FAFA is constructed to achieve fine-grained feature alignment.

3.2.1 The FAFA Architecture

Inspired by BLIP-2 [38], the proposed FAFA architecture, as shown in Figure 2(b), integrates a frozen image encoder and a lightweight Query Transformer (Q-Former). The Q-Former enables efficient multimodal representation extraction through a trainable query mechanism. It supports two encoding pathways: one is an image-guided path that combines visual and textual inputs, and the other is a purely visual path.

Given an input triplet $\langle I_q, T_q, I_t \rangle$, the frozen image encoder extracts visual features from the reference image I_q, which are then combined with the relative caption T_q and fed into the Q-Former. The textual [CLS] token, after passing through a text projection layer, yields the query representation $f_q \in \mathbb{R}^d$. Meanwhile, the target image I_t is processed by the same frozen encoder, and its visual features are routed through the purely visual branch of the Q-Former. The learnable query tokens in this branch are projected along the sequence dimension using a visual projection layer, generating the fine-grained feature representation $f_t = \{f_t(1), f_t(2), \ldots, f_t(N)\} \in \mathbb{R}^{N \times d}$, where N denotes the number of learnable queries and d is the feature dimension.

3.2.2 Fine-grained Adaptive Feature Alignment

Fine-grained feature matching is another inherent challenge in the CPR task. To deal with the issue, we propose a fine-grained dynamic alignment mechanism, integrating feature diversity supervision and masked feature reasoning into an end-to-end optimization strategy.

Fine-grained Dynamic Alignment (FDA). Unlike conventional contrastive learning methods [35, 71] that focus on global single-feature matching, the proposed approach dynamically aligns multiple fine-grained features from the target image with the query representation. Specifically, for each input triplet, the similarity between the query representation f_q and the set of target fine-grained features f_t is calculated by using dynamic feature selection and aggregation:

$$\text{Sim}(f_q, f_t) = \frac{1}{k} \sum_{i=1}^{k} \text{TopK}_i \left(\left\{ \frac{f_q^\top f_t(j)}{\|f_q\| \cdot \|f_t(j)\|} \right\}_{j=1}^{N} \right) \tag{3}$$

where $\text{TopK}_i(\cdot)$ denotes the i^{th} highest similarity score. This mechanism allows the model to adaptively select the most relevant fine-grained features for improved precision. During training, distribution matching and label smoothing are incorporated to enhance contextual alignment. For batch size B, the ground-truth matching probability is defined as: $q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{B} y_{i,k}}$, where $y_{i,j} = 1$ for exact matches (with the same ID), $y_{i,j} = \alpha, \alpha \in (0, 1)$ for partial matches (with the same GID), and $y_{i,j} = 0$ for unmatched pairs. The predicted distribution is normalized via softmax: $p_{i,j} = \frac{\exp(\text{Sim}(f_q^i, f_t^j)/\tau)}{\sum_{k=1}^{B} \exp(\text{Sim}(f_q^i, f_t^k)/\tau)}$, where τ is a temperature parameter. Then, the query-to-target alignment loss is defined as:

$$\mathcal{L}_{q2t} = \frac{1}{B} \sum_{i=1}^{B} \text{KL}(\mathbf{p_i} | \mathbf{q_i}) = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} p_{i,j} \log \left(\frac{p_{i,j}}{q_{i,j} + \epsilon} \right) \tag{4}$$

where $\text{KL}(\cdot|\cdot)$ represents Kullback–Leibler divergence and ϵ ensures numerical stability. The reverse loss $\mathcal{L}_{t2q}$ is computed analogously by interchanging f_q and f_t, leading to the overall alignment loss: $\mathcal{L}_{fda} = \mathcal{L}_{q2t} + \mathcal{L}_{t2q}$.

Feature Diversity (FD) Supervision. To reduce redundancy, a feature dispersion loss is introduced:

$$\mathcal{L}_{fd} = \frac{1}{N(N-1)} \sum_{i \neq j} \max \left(\frac{f_t(i)^\top f_t(j)}{|f_t(i)| \cdot |f_t(j)|} - m,\ 0 \right) \tag{5}$$

where m sets the maximum cosine similarity, encouraging diversity among internal representations.

Masked Feature Reasoning (MFR). To exploit complementary information between reference images and relative text, a bidirectional MFR strategy is proposed. Specifically, the random masking operation (30%) is applied to f_q and the average pooled target image feature $\bar{f}_t$, thus producing

Figure 4: Some representative examples from the ITCPR dataset.

masked features $\tilde{f}_q$ and $\tilde{f}_t$. The four features are jointly fed into a lightweight decoder Φ to minimize reconstruction loss:

$$\mathcal{L}_{\text{mfr}} = \mathbb{E}_{(f_q, \bar{f}_t)\sim\mathcal{B}} \left[|f_q - \Phi([\bar{f}_t, \tilde{f}_q])|_2^2 + |\bar{f}_t - \Phi([f_q, \tilde{f}_t])|_2^2 \right] \tag{6}$$

This loss drives the model to recover complete representations and enhance cross-modal alignment. Finally, by combining the above three components, the overall training objective is formulated as: $\mathcal{L} = \mathcal{L}_{fda} + \lambda_1 \mathcal{L}_{fd} + \lambda_2 \mathcal{L}_{mfr}$, where λ_1 and λ_2 are balancing weights for the auxiliary loss terms.

3.2.3 Inference Workflow

During inference, the fine-grained feature sets for all target images in the retrieval dataset are pre-extracted and stored as $\mathcal{V} = {f_t^i}_{i=1}^{N_t}$. Given a combined query feature f_q, the similarity between it and each f_t^i is computed using the same dynamic alignment method employed during training, thus ensuring efficient and reliable retrieval of the most relevant target images.

3.3 ITCPR Dataset

To objectively evaluate CPR methods, we manually construct the ITCPR dataset. Each triplet contains a reference image and a target image sharing the same identity, selected from public clothes-changing datasets including Celeb-reID [11], PRCC [13], and LAST [12], ensuring identity consistency despite variations in clothing or background. Each triplet also includes a relative caption explicitly highlighting differences between the two images, requiring models to jointly leverage visual and textual information for accurate retrieval. To ensure evaluation reliability, gallery images are carefully reviewed to eliminate potential false-negative cases. Ultimately, ITCPR contains 2,225 annotated triplets, comprising 2,202 unique query combinations (I_q, T_q) from 1,199 identities. The gallery consists of 20,510 person images, among which 2,225 correspond directly to queries. Representative examples are illustrated in Figure 4.

4 Experiments

4.1 Experimental Setup

Datasets. For data generation, we fine-tune Flux.1 [8] on the training split of CUHK-PEDES [4], a widely-used real-world person dataset containing 68,126 manually annotated image-text pairs. For the CPR task, FAFA is trained on 1.15 million filtered high-quality triplets from SynCPR, and evaluations are conducted on the manually annotated ITCPR dataset. Detailed descriptions of all datasets are provided in the Appendix B.

Choice of Dataset for Fine-tuning. For fine-tuning the model, we chose CUHK-PEDES over other datasets such as UFine6926 [72], ICFG-PEDES [73], and RSTPReid [74]. CUHK-PEDES was selected primarily because of its greater scene diversity, encompassing images from five surveillance datasets that represent a wide range of real-world scenarios, including urban environments and public spaces. This diversity is crucial for ensuring that the model generalizes well across various contexts, which is essential for practical person retrieval tasks. In contrast, ICFG-PEDES and RSTPReid, which mainly focus on constrained environments like parking lots (MSMT17 [75]), lack the same level of visual variation, potentially limiting the model's adaptability to real-world scenarios. Furthermore, while UFine6926 offers more fine-grained text-image pairings, its higher image quality and controlled video sources do not provide the same environmental diversity as CUHK-PEDES, which could lead to overfitting. By fine-tuning on CUHK-PEDES, we strike a balance between rich textual descriptions

Table 1: **Comparison of methods across different domains and settings.** For all domains other than CPR, models are trained on the most representative dataset within each domain.

Domain	Method	Ref.	Pretraining Data	Setting	Rank-1	Rank-5	Rank-10	mAP
IPR	TransReID [76]	ICCV21	Market-1501 [77]	*Image-only*	7.27	17.30	22.75	12.57
	SOLIDER [78]	CVPR23			8.45	18.48	23.89	13.74
	CLIP-ReID [79]	AAAI23			7.95	18.12	22.75	13.31
CC-IPR	CAL [80]	CVPR22	LTCC [81]	*Image-only*	9.86	22.34	29.20	16.45
	FIRe2 [82]	TIFS24			10.76	22.84	29.29	17.00
TPR	RaSa [83]	IJCAI23	CUHK-PEDES [4]	*Text-only*	28.02	49.23	57.77	38.04
	IRRA [2]	CVPR23			26.39	46.46	56.27	36.13
	RDE [84]	CVPR24	CUHK-PEDES [4]	*Image-only*	6.31	13.78	18.46	10.43
				Text-only	26.43	47.41	56.45	36.35
				Image + Text	29.79	51.82	60.49	40.10
Fuse	SOLIDER + RaSa	-	-	*Image + Text*	30.97	52.86	61.81	41.22
	FIRe2 + RaSa	-			32.89	54.27	62.03	42.16
ZSCIR	Pic2Word [49]	CVPR23	CC3M [85]	*Combination*	21.21	37.15	44.51	29.11
	CoVR-BLIP [86]	AAAI24	WebVid-CoVR [86]		26.75	47.68	56.36	36.49
	LinCIR (ViT-G) [87]	CVPR24	-		23.93	44.46	53.18	33.95
CIR	CaLa [47]	SIGIR24	CIRR [6]	*Combination*	24.02	44.64	53.45	34.08
			SynCPR (Ours)		39.33	60.85	68.66	49.29
	SPRC [48]	ICLR24	CIRR [6]	*Combination*	25.07	45.73	54.50	35.05
			SynCPR (Ours)		<u>42.27</u>	<u>61.81</u>	<u>69.35</u>	<u>51.62</u>
CPR	**FAFA (Ours)**	-	**SynCPR (Ours)**	*Combination*	**46.54**	**66.21**	**73.12**	**55.60**

**__Bold__ indicates the best performance; <u>Underline</u> indicates the second best.*

and diverse visual settings, ensuring the model's robustness when confronted with the variety of challenges encountered in real-world person retrieval tasks.

Implementation Details. All experiments are conducted using two H800 GPUs. During the SynCPR construction process, we adopt Qwen2.5-70B [7] as the LLM to generate textual quadruples, and use Flux.1 [8] as the base image generation model. This model is fine-tuned by LoRA [9] with its rank $r = 64$, and we set $\beta = 1$ to generate five identity-consistent image pairs per quadruple in the most realistic style. Another five image pairs are generated using random values of $\beta \in (0, 1)$ to ensure stylistic diversity. Qwen2.5VL-32B [10] is employed for data filtering. For training the FAFA framework, we set the total number of epochs to 10 and use a batch size of 256. The soft label strength in FDA is set to $\alpha = 0.5$, the number of selected fine-grained features is $k = 6$, and $\tau = 0.02$. The margin parameter m in $\mathcal{L}_{fd}$ is set to 0.5. The loss balancing hyperparameters are set to $\lambda_1 = 1$ and $\lambda_2 = 0.5$. All comparison methods are implemented using the optimal settings reported by them. Additional implementation details can be found in the Appendix C.1.

Evaluation Metrics. Retrieval performance is measured using Rank-k accuracy and mean average precision (mAP). Rank-k indicates the probability of correct matches in top-k retrievals, while mAP averages precision across all queries.

4.2 Results

To objectively evaluate FAFA and the SynCPR dataset, we extensively compare recent approaches from person retrieval and composed image retrieval. The compared methods are categorized into four settings based on input types: 1) *Image-only*, which relies solely on the reference image and retrieves targets via the visual encoder; 2) *Text-only*, which uses only relative captions and retrieves targets through cross-modal alignment; 3) *Image + Text*, which calculates similarity scores separately via the first two methods and then retrieves targets using their average; and 4) *Combination*, which simultaneously inputs both reference image and relative caption into the model for target retrieval. As shown in Table 1, our method consistently outperforms others across all settings. Specifically, directly applying IPR methods yields the lowest performance due to clothing variations between reference and target images. Even CC-IPR methods trained explicitly on clothes-changing datasets struggle due to limited generalization. In contrast, TPR methods achieve relatively better results, as

Table 2: **Ablation experiments on each component of FAFA.** To validate the effectiveness of FDA, we additionally introduce the image–text contrastive loss (ITC) [71] for comparison.

No.	Components					ITCPR Dataset			
	SynCPR	ITC	FDA	FD	MFR	Rank-1	Rank-5	Rank-10	mAP
1	✓	✓				41.33	61.72	68.94	50.94
2	✓		✓			45.04	64.90	72.21	54.41
3	✓		✓	✓		<u>46.05</u>	<u>65.85</u>	<u>73.02</u>	<u>55.49</u>
4	✓		✓		✓	45.78	65.58	72.62	55.13
5	✓		✓	✓	✓	**46.54**	**66.21**	**73.12**	**55.60**

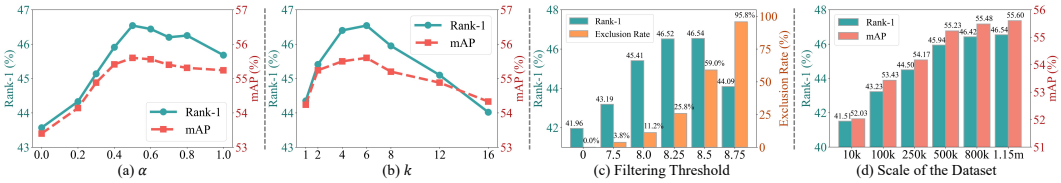

Figure 5: Sensitivity analysis of FAFA on hyperparameters and analysis of the SynCPR dataset.

the relative captions inherently match target images, although some visual information is missing. Among baseline approaches excluding our method, the *Image + Text* strategy achieves the best results, validating the rationality of our ITCPR dataset and emphasizing the necessity of combining visual and textual queries for optimal retrieval.

For CIR methods, although inherently designed for joint image-text queries, their training generally targets natural images involving significant visual modifications, thus lacking fine-grained retrieval capability required by CPR tasks. Notably, supervised CIR methods trained on original CIR datasets perform worse than certain ZSCIR methods on our task, underscoring the need for CPR-specific datasets and methods. Training supervised CIR methods on our SynCPR dataset significantly improves retrieval performance to a practical level. Furthermore, integrating our fine-grained retrieval framework FAFA with SynCPR further substantially enhances retrieval accuracy, confirming the indispensable roles of both the proposed dataset and FAFA.

4.3 Ablation Study

In this section, we conduct comprehensive ablation experiments to investigate the contribution of each component within the FAFA framework. Additionally, we discuss the impact of key hyperparameters in both the FAFA model and the data generation process.

FAFA Model. We train variants of the FAFA model with different components on the SynCPR dataset and evaluate their performance on the ITCPR test set. As shown in Table 2, experimental results demonstrate that due to the specific nature of the CPR task, employing our proposed fine-grained dynamic alignment strategy can substantially improve retrieval performance. Moreover, both supervision strategies, namely the FD strategy for enhancing feature diversity and the MFR strategy for capturing complementary features, contribute effectively to performance gains. The FAFA model equipped with all components achieves the best overall performance.

Hyperparameters of FAFA. Figures 5(a) and 5(b) illustrate the impact of two critical hyperparameters in our proposed FAFA model, namely the soft label strength α and the number of selected fine-grained features k in FDA, on the retrieval performance. Regarding α, lower values mean that the triplets generated from the same textual data will be treated more negatively, thus have an adverse impact on FAFA's semantic understanding. Conversely, higher values will weaken FAFA's ability to maintain identity consistency. This observation is consistent with our experimental results: as α increases, the retrieval performance initially improves and subsequently declines, achieving optimal performance when $\alpha = 0.5$. Similarly, the number of fine-grained features k also exhibits a comparable trend, and the optimal performance can be obtained when $k = 6$. This also aligns with expectations, because smaller k values will restrict the involvement of sufficient fine-grained features in retrieval, whereas excessively large k values may make the training process too homogenized, thus are not suitable for retrieval tasks that require distinctive feature representations.

SynCPR Dataset. Figure 5(c) presents the influence of applying various scoring thresholds on retrieval performance and data filtration ratio after generating all triplet data. Without any filtering, the potential noise in the dataset negatively impacts the FAFA training process, and consequently reduces retrieval performance. The optimal retrieval performance is observed when the threshold is set at 8.25 and 8.5. To enhance training efficiency and ensure the high quality of the SynCPR dataset, we finally adopt the latter. Furthermore, we perform sampling on the retained 1.15 million high-quality triplets via GID to validate the appropriate scale of the SynCPR dataset. As shown in Figure 5(d), as the dataset size increases, the retrieval performance improves rapidly. When the number of samples exceeds 500k, the marginal gains gradually diminish, and it saturates when the number of samples reaches approximately 800k. This confirms that our SynCPR dataset containing 1.15 million triplets is large and challenging enough to train better CPR models and is also convenient for comparison with our baseline method.

5 Conclusion

We introduce a practically significant task of composed person retrieval. Firstly, we put forward a scalable synthetic pipeline to address the data scarcity problem, and construct a high-quality SynCPR dataset at million scale. Secondly, a novel FAFA framework is introduced to enhance fine-grained retrieval accuracy. Extensive experiments on the newly annotated ITCPR benchmark confirm the significant superiority of our approach over the existing IPR, TPR, and CIR methods. Future work will explore composed person retrieval based on multiple images and multiple textual descriptions, as well as retrieval under open-set conditions.

Ethical Considerations. While the CPR task holds significant promise for applications such as locating missing individuals, it also raises critical ethical concerns, particularly regarding privacy and the potential for surveillance misuse. The ability to track individuals across different locations introduces privacy risks, which can be mitigated through various safeguards. For instance, invisible digital watermarks have been embedded in the images of the generated SynCPR dataset to ensure traceability, and access is restricted to academic use under responsible-use agreements. Additionally, biases inherent in synthetic data generated by LLMs have been addressed by ensuring substantial diversity in the generated data. Statistical information and visual representations in Appendix B.2 of the appendix effectively demonstrate the demographic diversity of the SynCPR dataset, encompassing various genders, ages, and ethnicities. These measures ensure the responsible use of the proposed method, adhering to the principle of "Tech for Good" while addressing potential societal risks.

Limitations

While the proposed CPR task demonstrates significant potential, several limitations remain. The SynCPR dataset, although highly diverse, relies on synthetic data, which may still introduce a domain gap when applied to real-world scenarios. Despite efforts to minimize this gap through adjustments in the generation strategy and fine-tuning of the generative model, the synthetic nature of the training data may not fully capture all the variations found in real-world images. Additionally, the ITCPR test set primarily focuses on clothing changes, which limits the model's ability to generalize to other types of variations, such as changes in scenes or hairstyles. Expanding the dataset to encompass a broader range of person-related variations will be an important area for future improvement. Furthermore, the current FAFA framework and the ITCPR test set focus mainly on scenarios where only a single image and textual description are provided, which may not align with more complex real-world situations. Addressing this limitation will be a key focus for future work.

Acknowledgements

This work was supported by the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing (GJJ-24-021) and the BUPT Innovation and Entrepreneurship Support Program (2025-YC-T043).

References

[1] Yan Zhang, Binyu He, Li Sun, and Qingli Li. Progressive multi-stage feature mix for person re-identification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

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

[3] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[4] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017.

[5] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021.

[6] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021.

[7] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[8] Black Forest Labs. Flux: Official inference repository for flux.1 models. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2024-11-12.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[10] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[11] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[12] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.

[13] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019.

[14] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. *Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline)*, page 501–518. 2018.

[15] Jingjing Qian, Wei Jiang, Hao Luo, and Hongyan Yu. Stripe-based and attribute-aware network: a two-branch deep model for vehicle re-identification. *Measurement Science and Technology*, page 095401, 2020.

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2017.

[17] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[18] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[19] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] Pingyu Wang, Zhicheng Zhao, Fei Su, and Honying Meng. Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification. *IEEE Transactions on Multimedia*, 2022.

[21] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, 2019.

[22] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[23] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[24] Peng Xu and Xiatian Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11196–11205, 2023.

[25] Xiangzeng Liu, Kunpeng Liu, Jianfeng Guo, Peipei Zhao, Yining Quan, and Qiguang Miao. Pose-guided attention learning for cloth-changing person re-identification. *IEEE Transactions on Multimedia*, 26:5490–5498, 2024.

[26] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824, 2019.

[27] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer, 2022.

[28] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023.

[29] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021.

[30] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931, 2023.

[31] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 2023.

[32] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022.

[33] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, pages 5566–5574, 2022.

[34] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018.

[35] 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*, pages 8748–8763. PMLR, 2021.

[36] 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.

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[39] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021.

[40] Delong Liu, Haiwen Li, Zhicheng Zhao, and Yuan Dong. Text-guided image restoration and semantic enhancement for text-to-image person retrieval. *Neural Networks*, 184:107028, 2025.

[41] Ginger Delmas, Rafael S Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2024.

[42] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474, 2022.

[43] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5753–5762, January 2024.

[44] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C.Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv: Computation and Language,arXiv: Computation and Language*, 2015.

[45] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[46] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023.

[47] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting comoposed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2177–2187, 2024.

[48] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

[50] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*.

[51] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer, 2022.

[52] S Karthik, K Roth, M Mancini, Z Akata, et al. Vision-by-language for training-free compositional image retrieval. In *The Twelfth International Conference on Learning Representations*. OpenReview. net, 2024.

[53] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90, 2024.

[54] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[55] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[59] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[60] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[61] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024.

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[63] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[64] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

[65] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*, 2017.

[67] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[68] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[69] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

[70] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[71] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[72] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024.

[73] 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.

[74] 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*, pages 209–217, 2021.

[75] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[76] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[77] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[78] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[79] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1405–1413, 2023.

[80] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069, 2022.

[81] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[82] Qizao Wang, Xuelin Qian, Bin Li, Xiangyang Xue, and Yanwei Fu. Exploring fine-grained representation and recomposition for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:6280–6292, 2024.

[83] 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. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 555–563, 2023.

[84] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024.

[85] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[86] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279, 2024.

[87] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234, 2024.

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

[89] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[90] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019.

[91] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923, 2023.

[92] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2991–2999, 2024.

[93] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*, 2024.

Appendix

A More Details for High-Quality Triplet Synthesis

The triplet data required for Composed Person Retrieval (CPR) consists of three key elements: a reference image I_q, a relative caption T_q, and a target image I_t. Two significant challenges hinder the complete synthetic generation of such triplets. Firstly, generating a pair of person images (I_q, I_t) consistent with real-world distributions while preserving identity. Secondly, providing accurate textual descriptions of relative changes between the two images. To address these challenges, as shown in Figure S6, we effectively divide the data synthesis process into three steps. First, a Large Language Model (LLM) [7] generates textual data comprising descriptions for synthesizing image pairs and relative captions. Second, generative models [8] utilize the textual descriptions to produce realistic and identity-consistent image pairs, thereby forming the required triplets. To ensure realism in the generated images, we further fine-tune the generative models. Finally, a multimodal large language model (MLLM) [10] evaluates the synthesized triplets across multiple dimensions, filtering out lower-quality data.

A.1 Diverse Textual Quadruples Generation

In this step, we simplify the multimodal triplet generation objective (I_q, T_q, I_t) to purely textual quadruple generation $(T_{I_q}, T_{q\to t}, T_{t\to q}, T_{I_t})$ using an LLM. Each quadruple comprises a reference description T_{I_q}, a target description T_{I_t}, a relative caption describing changes from T_{I_q} to T_{I_t} ($T_{q\to t}$), and another describing the reverse changes ($T_{t\to q}$). To achieve this, we select QWen2.5-72B [7] as the LLM and carefully design structured instructions to generate quadruples meeting the desired criteria.

The instruction format, illustrated in Figure S6, initially provides an overview of the task and fundamental requirements for the LLM. It then specifies detailed guidelines for generating each element within the quadruple. Additionally, the instructions include three high-quality example outputs randomly selected from 100 manually annotated cases to enhance the quality and stability of the LLM outputs. Random sampling of examples promotes diversity in instructions, preventing repetitive outputs caused by similar inputs.

Notably, the instructions suggest the primary character, clothing items, and colors, randomly drawn from candidate lists. These lists are derived from relevant datasets [4, 73, 74] containing person descriptions, supplemented by additional related elements. Such a design ensures generated data closely aligns with real-world distributions while maximizing its diversity and comprehensiveness. The LLM subsequently generates structured prompts for image synthesis and composed retrieval based on these provided elements.

A.2 Identity-consistent High-quality Image Synthesis

Once the textual quadruples $(T_{I_q}, T_{q\to t}, T_{t\to q}, T_{I_t})$ are obtained, we convert T_{I_q} and T_{I_t} into their corresponding images, I_q and I_t, thus forming the desired triplet data. For this purpose, we adopt FLUX.1 [8], an advanced generative model, as the base model and fine-tune it using Low-Rank Adaptation (LoRA) [9] on a text-based person retrieval (TPR) dataset [4] to generate person images consistent with real-world distributions. Due to the inherent randomness in diffusion model image generation, a critical challenge remains ensuring identity consistency between paired images. However, diffusion models intrinsically possess the capability to generate two identical or similar objects within one image. We leverage this internal consistency capability by generating two sub-images within a single image, thereby ensuring detailed consistency of shared elements in I_q and I_t. To achieve this, we specifically design the image generation prompt template to include two equally sized sub-images with the same identity.

During this stage, prompts generated by the LLM are input into FLUX.1 to produce images with a resolution of 400×400. This configuration allows each generated image to be split into two sub-images (192×384), forming the pair (I_q, I_t). This padding strategy mitigates inaccuracies that may arise during image generation and avoids artifacts caused by image cropping. As depicted in Figure S6, this method effectively maintains identity consistency between the sub-images while varying their appearances and states, demonstrating clear advantages over separate generation. Consequently,

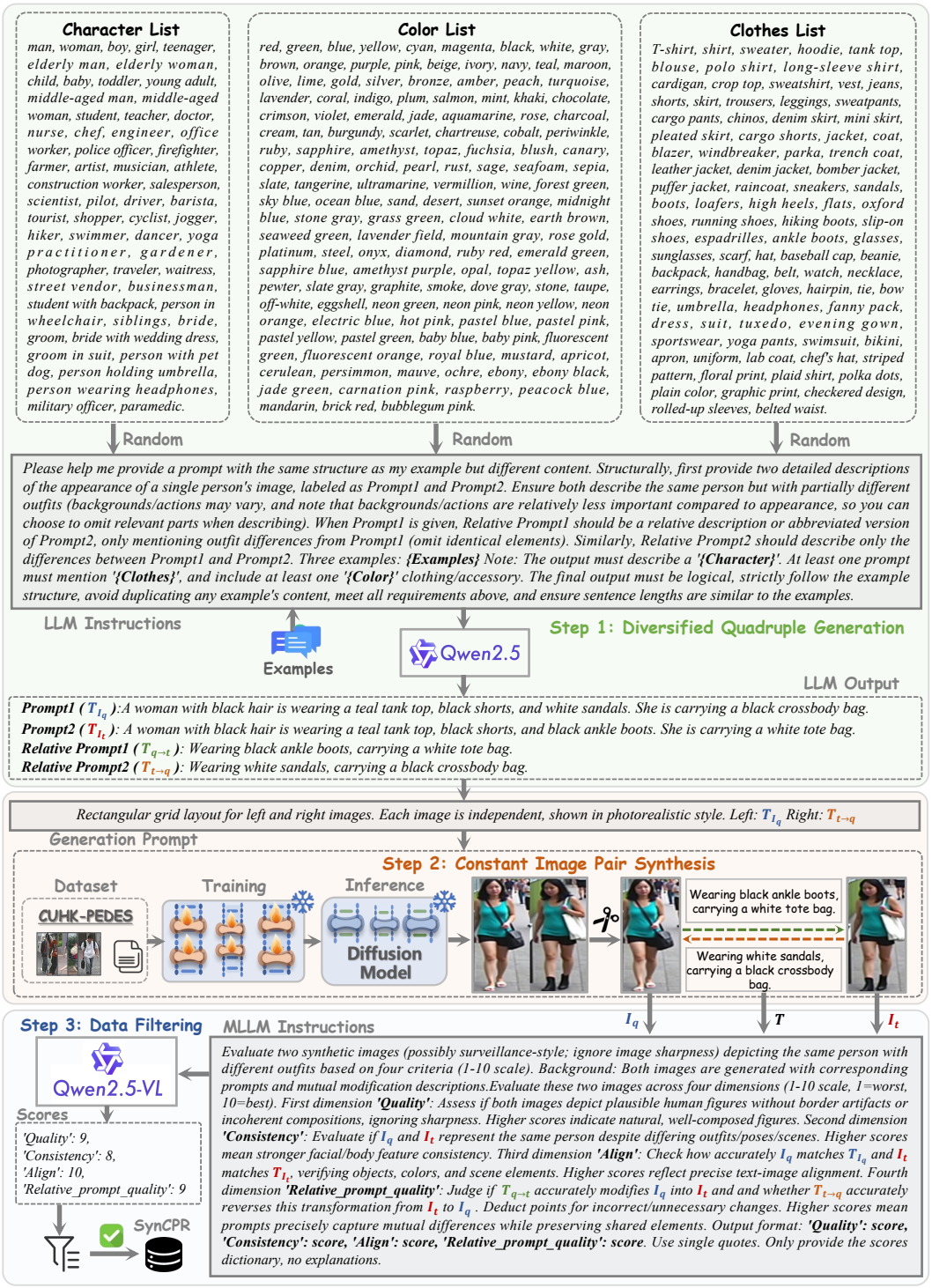

Figure S6: Pipeline of high-quality CPR triplet construction with detailed instruction design.

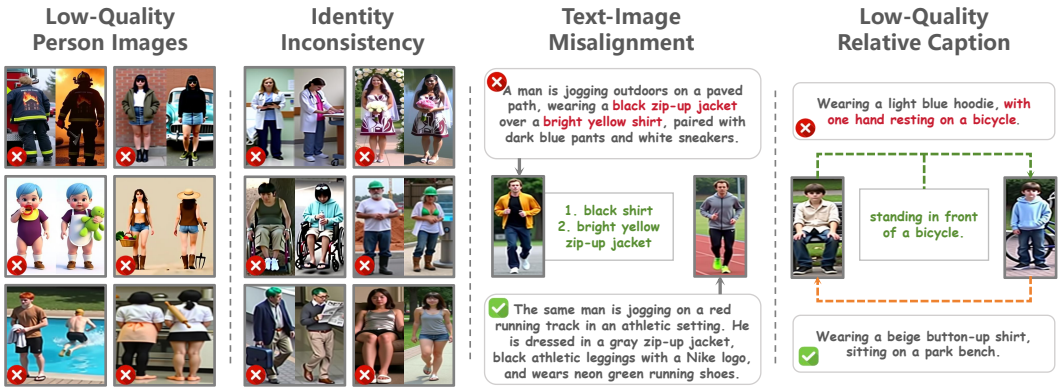

Figure S7: Representative examples of samples filtered out during the data filtering process. From left to right, each panel corresponds to one of the four evaluation dimensions, and the samples are excluded due to low scores in their respective dimensions.

we acquire the desired dataset, and swapping the reference and target images yields two sets of triplet annotations. By dynamically adjusting the LoRA strength β, each textual quadruple generates ten image pairs. Images with identical relative captions are defined under the same group identity (GID), thus forming strong positive samples within triplets and weak positive samples within groups, with other instances treated as negative samples.

A.3 Data Filtering

To further ensure the quality of the synthesized data, we employ Qwen2.5-VL 32B [10] combined with carefully designed instructions to score and filter all generated data. Four dimensions are considered: first, assessing the fidelity of person images I_q and I_t, focusing on naturalness, noise, and artifacts while ignoring image clarity; second, evaluating identity consistency between I_q and I_t; third, assessing the alignment between images and their descriptions (e.g., $I_q \leftrightarrow T_{I_q}$); fourth, evaluating the overall quality of the triplet ($I_q + T_q \rightarrow I_t$), where higher scores indicate the ability to accurately infer I_t from the combination of I_q and T_q, with minimal overlap and high complementarity. Qwen2.5-VL rates each dimension from 1 to 10, and the final score is the average of these dimensions. Triplets scoring higher than 8.5 are retained and included in the SynCPR dataset. Representative examples of discarded low-quality data are shown in Figure S7.

B Additional Datasets Details

B.1 ITCPR Dataset

In contrast to existing CIR datasets [5, 6], where reference and target images only need to be loosely related, the CPR datasets subject to the constraint that both of them depict the same person. Therefore, when constructing the ITCPR dataset, we ask the selected images to have the same identity, but wear different clothes or be in different scenes. In our implementation, publicly available clothes-changing datasets such as Celeb-reid [11], PRCC [13], and LAST [12] are utilized as our image sources.

B.1.1 Dataset Annotation Process

The annotation process, as shown in Figure S8, primarily consists of three steps. The first step involves selecting identities from the image data sources that have multiple images with different outfits, ensuring a diverse selection for subsequent steps. In the second step, a pair of images associated with each chosen identity is selected and denoted as I_q and I_t. It is worth noting that, ideally, these two images should depict partially matching outfits, allowing I_q to provide additional clothing-related information beyond facial features and body posture. This additional clothing information is not mentioned in the corresponding annotation T_q, ensuring that CPR methods can only correctly identify I_t by utilizing both I_q and T_q. Once the image pair is selected, the process moves to the third step, where manual annotations are created to specifically capture the differences between I_q and I_t. For

Figure S8: The annotation process of the ITCPR dataset. The annotation process can be summarized in three steps: the first step is selecting identities from the clothes-changing datasets, the second step is choosing pairs of reference and target images for each identity, and the third step is manually annotating the relative captions.

instance, as shown in case 1 of Figure S8, if the skirt is the same in both the target image and the reference image, it does not need to be described; the annotation focuses only on differences in the top and shoes. After manually annotating T_q, a complete triplet annotation process is finalized. Repeating this process, a batch of triplets (I_q, T_q, I_t) can be generated for testing CPR methods.

B.1.2 Re-Annotation of the ITCPR Dataset

The gallery contains a large number of noisy images, which may introduce false negatives. For example, for certain queries, some images may be potential ground truth but remain unlabeled. Including such cases would reduce the reliability of the evaluation metrics. To address this issue, all images in the gallery are screened, and any false negative images identified in the dataset are re-annotated. The re-annotation process is illustrated in Figure S9. After completing the dataset annotation and adding noise images to the gallery, we use a well-trained visual encoder [78] to search for the most similar images for each target image, followed by manual inspection and verification. Through this approach, we effectively eliminate false negatives in the ITCPR dataset.

B.1.3 Statistics of the ITCPR Dataset

In summary, ITCPR comprises a total of 2,225 annotated triplets. These triplets encompass 2,202 unique combinations (I_q, T_q) as queries. ITCPR contains 1,151 images and 512 identities from Celeb-reID [11], 146 images and 146 identities from PRCC [13], and 905 images and 541 identities from LAST [12]. In the target gallery, there are a total of 20,510 images of persons from the three datasets, with 2,225 corresponding ground truths for the queries. The textual annotations have an average sentence length of 9.54 words. The longest sentence contains 32 words, while the shortest sentence only contains 3 words. These annotations are exclusively designated for testing in the ZS-CPR task, which expects to achieve substantial performance without utilizing any data from the three datasets mentioned above.

B.2 SynCPR Dataset

Using our proposed automated construction pipeline, we successfully build the SynCPR dataset, which is a fully synthetic dataset specifically designed for the composed person retrieval task. In the first stage, we utilize Qwen2.5-70B [7] to generate a total of 140,500 textual quadruples. In the second stage, by employing fine-tuned LoRA [9] combined with Flux.1 [8] and setting $\beta = 1$ for the most realistic person image style, we generate five image pairs per quadruple. Additionally, we create another five image pairs using randomly selected $\beta \in (0, 1)$, ensuring diverse styles across generated images. Combining these images with two relative captions from each quadruple yields a total of 2,810,000 valid triplets. In the third stage, under stringent data filtering criteria, 1,153,220 high-quality triplets are retained. Among the retained samples, a total of 177,530 unique GIDs are involved. The average length of the relative caption sentences is 13.3 words, excluding punctuation.

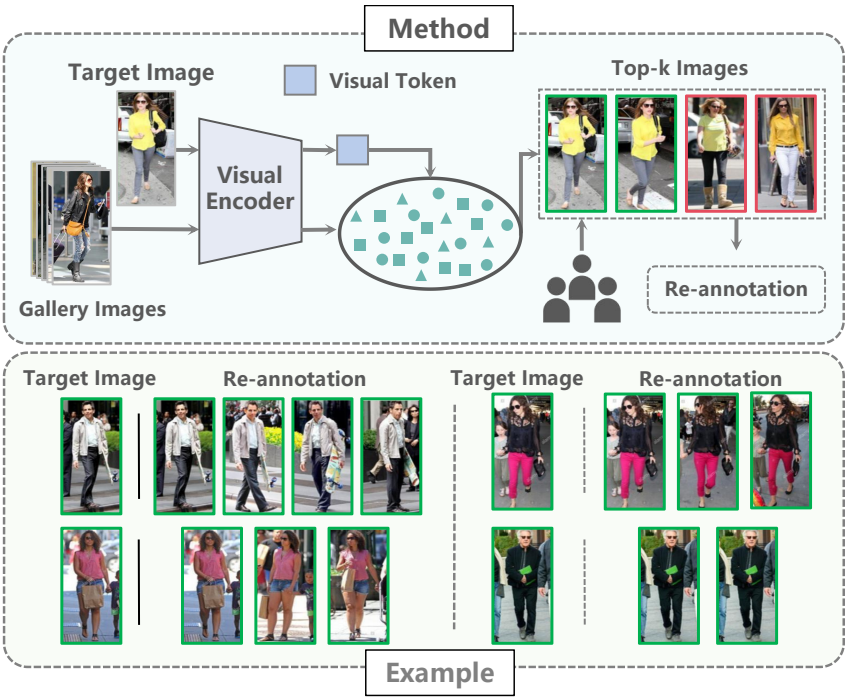

Figure S9: False negative elimination scheme in ITCPR. **Top**: The method of eliminating false negative images and adding annotations in the dataset. **Bottom**: Examples of false negative images re-annotated in ITCPR.

Figure S10: Representative examples of samples from the SynCPR dataset.

In total, 4,370 distinct words appear across all sentences, further highlighting the diversity of the SynCPR dataset.

The samples from SynCPR dataset are visualized in Figure S10. Thanks to our diversified textual generation strategy, realism-oriented fine-tuning of generative models, and rigorous data filtering

mechanisms, the SynCPR dataset ensures high quality, realism, and diversity of person images. By leveraging varied image generation prompts and the zero-shot generation capability of generative models, the SynCPR dataset encompasses rich scenarios, broad age coverage, diverse image clarity, varied attire and states of individuals, and comprehensive ethnic representation. Furthermore, the gender ratio in the generated dataset is highly balanced, with males accounting for 51.2%. Although SynCPR is entirely synthetic, its comprehensiveness significantly surpasses other manually annotated datasets within the person retrieval domain.

B.3 CUHK-PEDES Dataset

CUHK-PEDES [4] is a widely used benchmark for text-to-person retrieval, comprising 40,206 pedestrian images and 80,412 corresponding textual descriptions annotated across 13,003 unique identities. The dataset is divided into three subsets: a training set containing 34,054 images and 68,108 descriptions for 11,003 identities, a validation set with 3,078 images and 6,158 descriptions for 1,000 identities, and a test set comprising 3,074 images and 6,156 descriptions for a separate set of 1,000 identities. Each image is paired with two independent human-written descriptions, and the average length of the descriptions exceeds 23 words.

C Additional Experiments and Results

C.1 Additional Implementation Details.

All experiments are conducted on two NVIDIA H800 GPUs. For training the generative models, we select FLUX.1-dev [8] as the base model and apply LoRA with a rank of $r = 64$ specifically to the cross-attention layers. The training is performed with a per-GPU batch size of 1, utilizing gradient accumulation over eight steps, and optimized using AdamW [88] with an initial learning rate of 1×10^{-5}, 10-step warm-up, and weight decay of 0.01, for a total of 20,000 steps. The training employs bfloat16 mixed precision. Input person images from CUHK-PEDES are resized to 192×384 during training. During inference, each prompt dynamically adjusts LoRA strength, generating ten paired sub-images of size 400×400 via a 25-step beta noise reduction. Each generated image is first centrally cropped horizontally into two separate images, then each resulting sub-image is further centrally cropped to form a pair of person images sized 192×384.

For the Fine-grained Adaptive Feature Alignment (FAFA) framework, we use BLIP-2 [38] and a frozen ViT-G/14 [89] with an input resolution of 224 pixels. Input images undergo random horizontal flipping, random cropping with padding, and random erasing, followed by scaling the longer side to 224 pixels while preserving the aspect ratio. The images are then symmetrically padded horizontally to a final size of 224×224 before being input into FAFA. The model is trained on the SynCPR dataset using a single NVIDIA H800 GPU with a batch size of 256 for 10 epochs. Optimization is performed using AdamW [88] with an initial learning rate of 2×10^{-6}. In FAFA, the soft label strength parameter α is set to 0.5, the number of selected fine-grained features k is set to 6, and the temperature parameter τ is 0.02. The margin parameter m in the feature difference loss $\mathcal{L}_{fd}$ is 0.5. The loss balancing hyperparameters are set as $\lambda_1 = 1$ and $\lambda_2 = 0.5$. Additionally, for subsequent training from scratch on the Composed Image Retrieval (CIR) dataset CIRR [6], the model is trained for 50 epochs with an initial learning rate of 1×10^{-5}, while all other settings remain consistent.

In our experiments on CIRR, Rank-K serves as the primary metric, measuring the likelihood of finding the target image within the top-K retrieved candidates. For CIRR, we additionally report Rank$_s$-K on visually similar subsets, with overall performance summarized as $Avg. = \frac{\text{Rank-5+Rank}_s\text{-1}}{2}$.

C.2 Additional Quantitative Results

C.2.1 FAFA for Composed Image Retrieval

The CPR task can be viewed as a more constrained and finer-grained variant of the CIR task, involving stricter alignment requirements between the reference and target images. Consequently, the FAFA framework, originally designed for CPR, can naturally be applied to CIR scenarios. We thus conduct experiments on CIRR, the most representative dataset within the CIR domain. The results, summarized in Table R1, demonstrate that our FAFA framework achieves comprehensive state-of-the-art performance with significant advantages, even in the context of CIR. Specifically,

Table S3: Performance comparison with existing supervised CIR methods on CIRR dataset only. The best results are marked in **bold**, and the second-best results are underlined. † indicates that the method is pretrained on its own constructed triplet dataset.

Method	Ref.	Rank-K			Rank$_s$-K		Avg.
		K=1	K=5	K=10	K=1	K=3	
TIRG [90]	CVPR19	14.61	48.37	64.08	22.67	65.14	35.52
CIRPLANT [6]	ICCV21	19.55	52.55	68.39	39.20	79.49	45.88
ARTEMIS [41]	ICLR22	16.96	46.10	61.31	39.99	75.67	43.05
CLIP4CIR [42]	CVPR22	38.53	69.98	81.86	68.19	94.17	69.09
TG-CIR [91]	MM23	45.25	78.29	87.16	72.84	95.13	75.57
BLIP4CIR+Bi [43]	WACV24	40.15	73.08	83.88	72.10	95.93	72.59
CASE† [92]	AAAI24	48.68	79.98	88.51	76.39	95.86	78.19
CoVR-BLIP† [86]	AAAI24	49.69	78.60	86.77	75.01	93.16	76.81
CompoDiff† [93]	TMLR24	32.39	57.61	77.25	67.88	94.07	62.75
CaLa [47]	SIGIR24	49.11	81.21	89.59	76.27	96.46	78.74
SPRC [48]	ICLR24	51.96	82.12	89.74	80.65	96.60	81.39
FAFA (Ours)	-	**54.48**	**84.07**	**91.48**	**81.05**	**97.11**	**82.56**

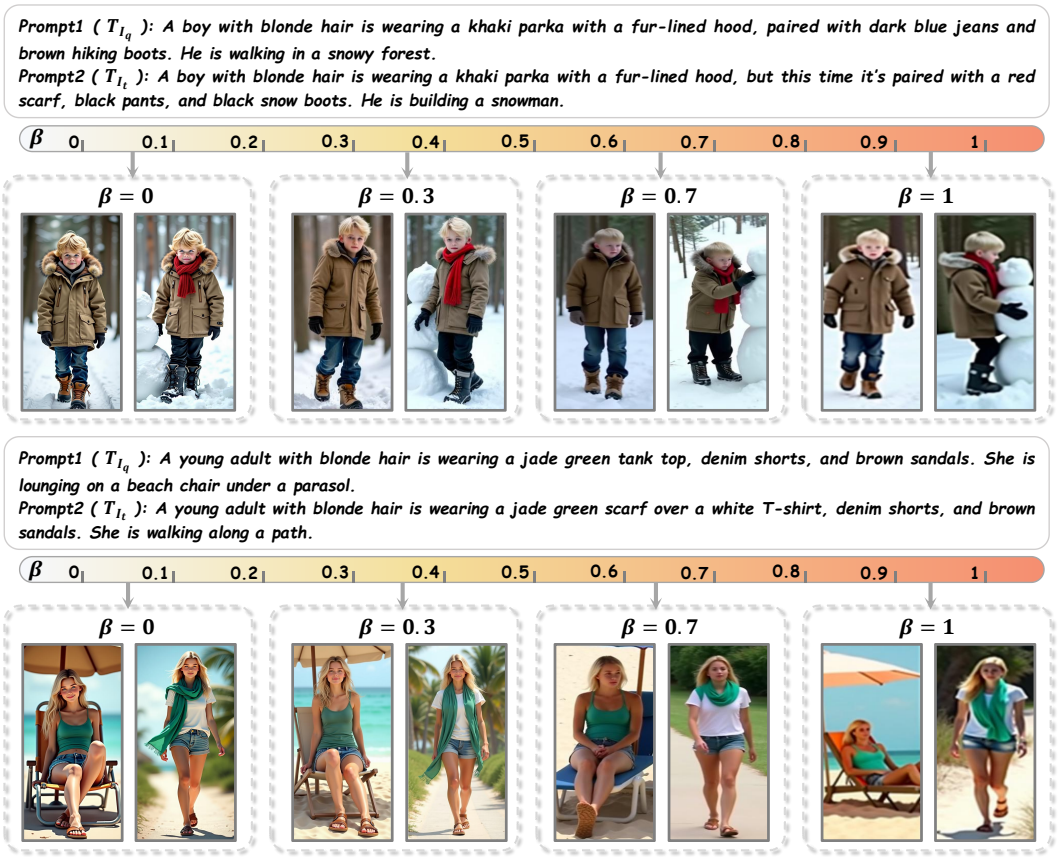

Figure S11: Person image generation results under different LoRA strengths.

FAFA outperforms the second-ranked method SPRC, which also utilizes BLIP-2 as its backbone, by 2.51% in Rank-1 accuracy on the CIRR dataset. When compared with CaLa, another BLIP-2-based method, FAFA achieves an even more notable improvement, surpassing it by 5.37% in Rank-1.

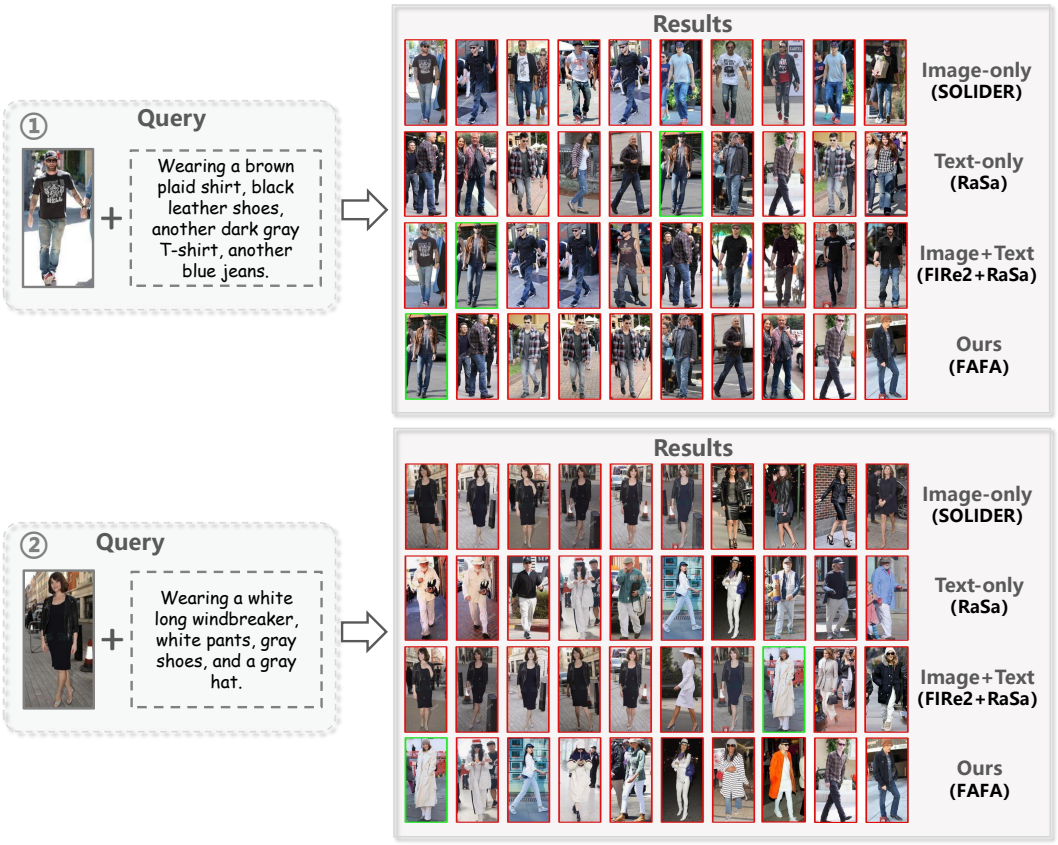

Figure S12: Comparative visualization of Top-10 retrieval results across different methods on the ITCPR dataset.

C.3 Additional Qualitative Results

C.3.1 Effects of Different LoRA Strengths on Person Image Generation

During the person image generation process, in order to achieve more comprehensive and realistic styles, multiple groups of images are generated for each textual prompt by dynamically adjusting the LoRA strength $\beta \in (0, 1]$. Figure S11 illustrates the effects of different LoRA strengths on generated person images, clearly demonstrating that the same textual input combined with varying values of β can yield distinct image styles. Specifically, when $\beta = 0$, the pre-trained generative model is employed directly, resulting in high-quality images but with noticeable discrepancies from real-world styles. As the β value increases gradually, the realism of the generated images correspondingly improves, ultimately reaching a style closely aligned with that of real-world person retrieval datasets at $\beta = 1$.

C.3.2 Visualization of Results from Different Methods

Figure S12 presents two illustrative examples of the Top-10 retrieval results obtained by various representative methods from Table 1 on the ITCPR dataset. It is evident that the *Image-only* retrieval method yields the poorest performance, primarily because it tends to retrieve images with visually similar pixel distributions. Given that the dataset contains numerous instances involving clothing changes, this leads to suboptimal performance. *Text-only* retrieval also falls short of expectations, as most annotations in the dataset provide brief descriptions of clothing differences between the reference and target images, while the retrieval database includes many images with similar clothing. Combining both modalities typically retrieves the target image within the Top-10 results; however, its inability to dynamically complement multimodal query information leads to scenarios where an excessively high match in one modality adversely affects the final retrieval results. For instance, in

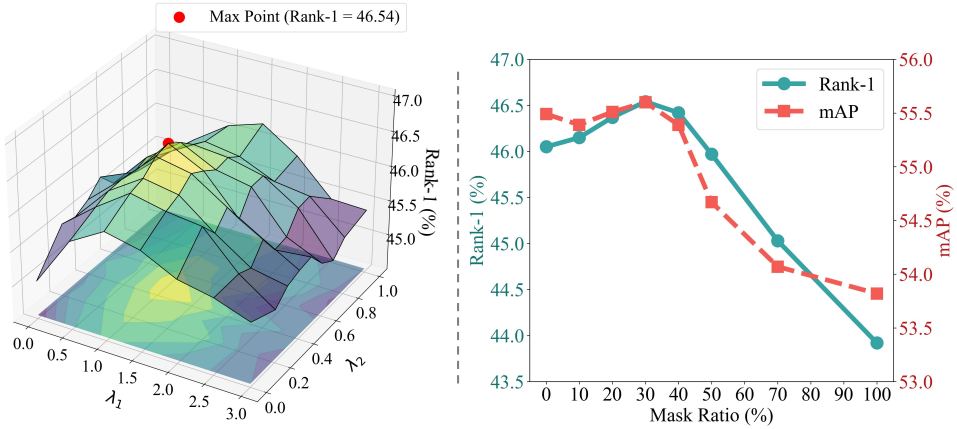

Figure S13: **Left**: Variations in FAFA's Rank-1 performance under different balancing weights of auxiliary loss terms. **Right**: Relationship between FAFA's performance and the feature mask ratio in $\mathcal{L}_{mfr}$.

Example ② of Figure S12, the high visual similarity causes the *Image + Text* method's most confident retrieval results to closely resemble those of the *Image-only* method. In contrast, our proposed FAFA method dynamically extracts complementary information from multimodal queries, consistently identifying the target person's image among the top-ranked retrieval results.

C.4 Additional Ablation Study

C.4.1 Balancing Weights of Auxiliary Loss Terms

To fully leverage the synergistic effects of the proposed loss functions, extensive experiments on balancing the weights of FAFA losses are conducted. As illustrated in Figure S13, when fixing the value of λ_2, the retrieval performance of FAFA initially rises and subsequently declines with increasing λ_1, achieving its highest performance at $\lambda_1 = 1$. Similarly, fixing λ_1 and varying λ_2 reveals the same trend, ultimately attaining the optimal performance at $\lambda_1 = 1$ and $\lambda_2 = 0.5$, which corresponds to our final selected configuration.

C.4.2 Feature Masking Ratio in $\mathcal{L}_{mfr}$

As demonstrated in Figure S13, the optimal performance of FAFA in the $\mathcal{L}_{mfr}$ setting is achieved when the feature masking ratio is set to 30%. When the masking ratio is set to 0, it is equivalent to disabling the $\mathcal{L}_{mfr}$ loss. As the masking ratio increases, performance first improves and then declines. When the masking ratio exceeds 50%, the complexity of masked feature reasoning becomes excessively high, resulting in elevated loss values that negatively impact overall training stability, thereby diminishing the effectiveness of the $\mathcal{L}_{mfr}$ component.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's core contributions. Specifically, the abstract clearly introduces the novel Composed Person Retrieval (CPR) task, the automated synthetic data generation pipeline, the construction of the SynCPR dataset, the proposal of the FAFA framework, and the ITCPR benchmark. These claims are further expanded in the introduction with detailed descriptions of the task formulation, methodology, and motivation. All major contributions are consistently described and substantiated throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the current work are discussed as part of the future work section in the conclusion. Specifically, the current setting of the CPR task restricts each query to contain only a single image and a single textual description.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper primarily addresses a practical application problem and does not involve any theoretical assumptions or formal theoretical results that would require proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main information required to reproduce the experimental results is provided in the main paper, while detailed configuration settings and additional implementation details are included in the appendix. Furthermore, all related code will be open-sourced to ensure full reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code and data will be made publicly available on GitHub after the review process. To preserve anonymity and ensure a rigorous double-blind review, no links are included in the current submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main training and test details are provided in Section 4.1 of the main paper, and a more comprehensive version with additional implementation details is included in Appendix C.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The proposed composed person retrieval task is a subtask of the broader person retrieval domain, where standard evaluation metrics are well established and widely adopted. These metrics, such as Rank-1 and mAP accuracy, do not typically include statistical significance measures such as error bars or confidence intervals.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The relevant information about compute resources is included in Section 4.1 (Implementation Details), with additional details provided in Appendix C.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have strictly followed the anonymization requirements and fully complied with the NeurIPS Code of Ethics throughout the research and submission process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed composed person retrieval task is of practical significance and has the potential to improve real-world person retrieval systems, particularly in scenarios where both visual and textual information are available. The positive societal impact of this work is discussed in the introduction. To the best of our knowledge, the task does not pose any obvious negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not involve any content that poses high risk of misuse requiring safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and datasets used in the paper have been properly credited to their original creators and used in accordance with their respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The main paper and appendix provide detailed information about the newly introduced SynCPR and ITCPR datasets, including their composition and usage details. All underlying models and base data used in constructing these datasets are used within the scope of their original creators' licenses. The new datasets will be publicly released after the review process along with comprehensive documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This paper uses an open-source large language model in the construction of the SynCPR dataset, and the usage is described in detail in the paper. All usage complies with the license and terms specified by the model's original creators, and adheres to the NeurIPS LLM usage policy.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.