# 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`

**Project Page**: [https://github.com/Delong-liu-bupt/Composed_Person_Retrieval](https://github.com/Delong-liu-bupt/Composed_Person_Retrieval)

**Presenter: Delong Liu      Date: Nov 5, 2025**

BUPT MCPRL

北京郵電大學
Beijing University of Posts and Telecommunications

商汤
sensetime

NEURAL INFORMATION PROCESSING SYSTEMS

- # Introduction

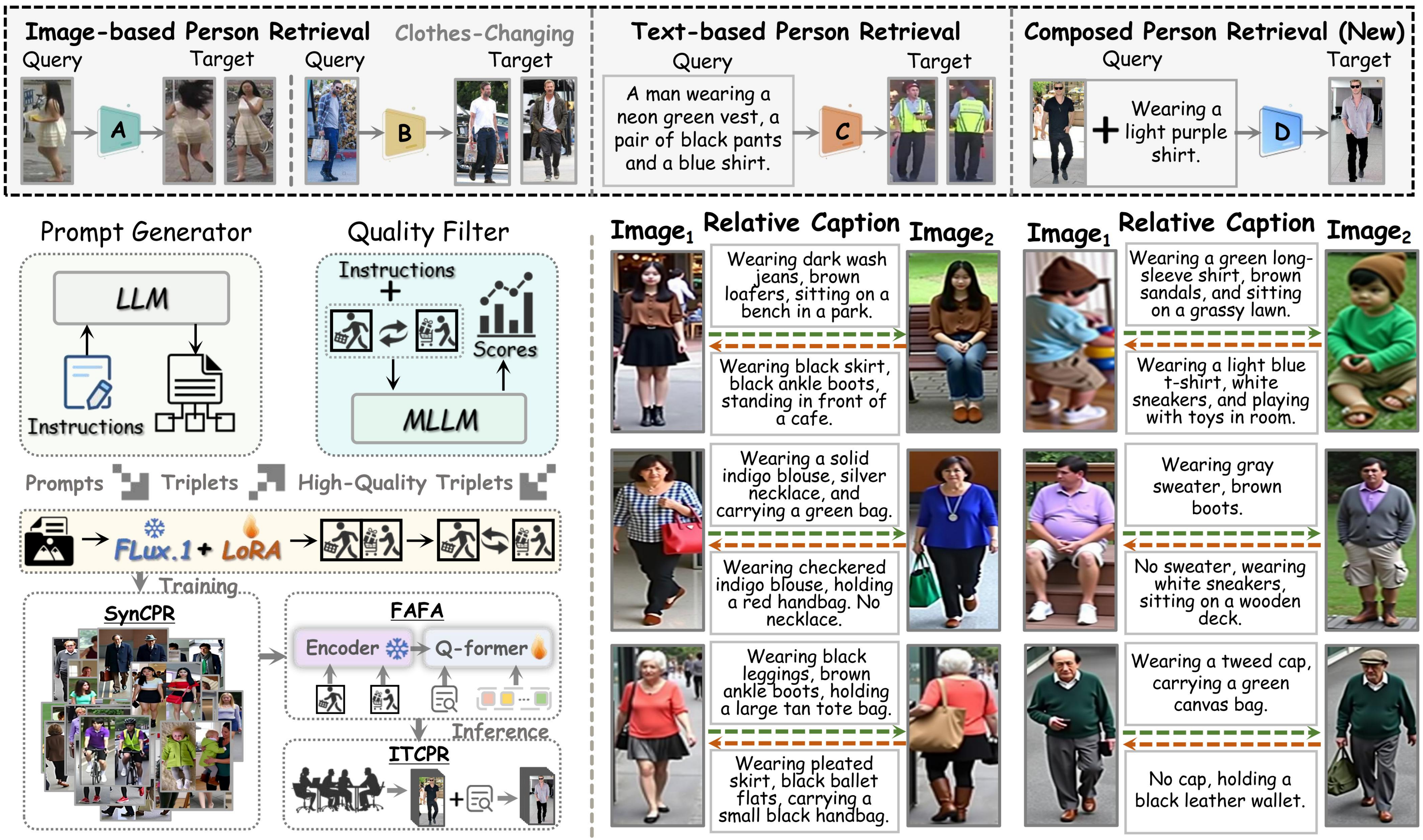


Image-based Person Retrieval
Clothes-Changing
Query
Target
Query
Target
A
B
Text-based Person Retrieval
Query
Target
A man wearing a neon green vest, a pair of black pants and a blue shirt.
C
Composed Person Retrieval (New)
Query
Target
Wearing a light purple shirt.
D
Prompt Generator
LLM
Instructions
Quality Filter
Instructions
Scores
MLLM
Prompts
Triplets
High-Quality Triplets
FLux.1
LoRA
Training
SynCPR
FAFA
Encoder
Q-former
Inference
ITCPR
Image1
Relative Caption
Image2
Wearing dark wash jeans, brown loafers, sitting on a bench in a park.
Wearing black skirt, black ankle boots, standing in front of a cafe.
Wearing a solid indigo blouse, silver necklace, and carrying a green bag.
Wearing checkered indigo blouse, holding a red handbag. No necklace.
Wearing black leggings, brown ankle boots, holding a large tan tote bag.
Wearing pleated skirt, black ballet flats, carrying a small black handbag.
Image1
Relative Caption
Image2
Wearing a green long-sleeve shirt, brown sandals, and sitting on a grassy lawn.
Wearing a light blue t-shirt, white sneakers, and playing with toys in room.
Wearing gray sweater, brown boots.
No sweater, wearing white sneakers, sitting on a wooden deck.
Wearing a tweed cap, carrying a green canvas bag.
No cap, holding a black leather wallet.

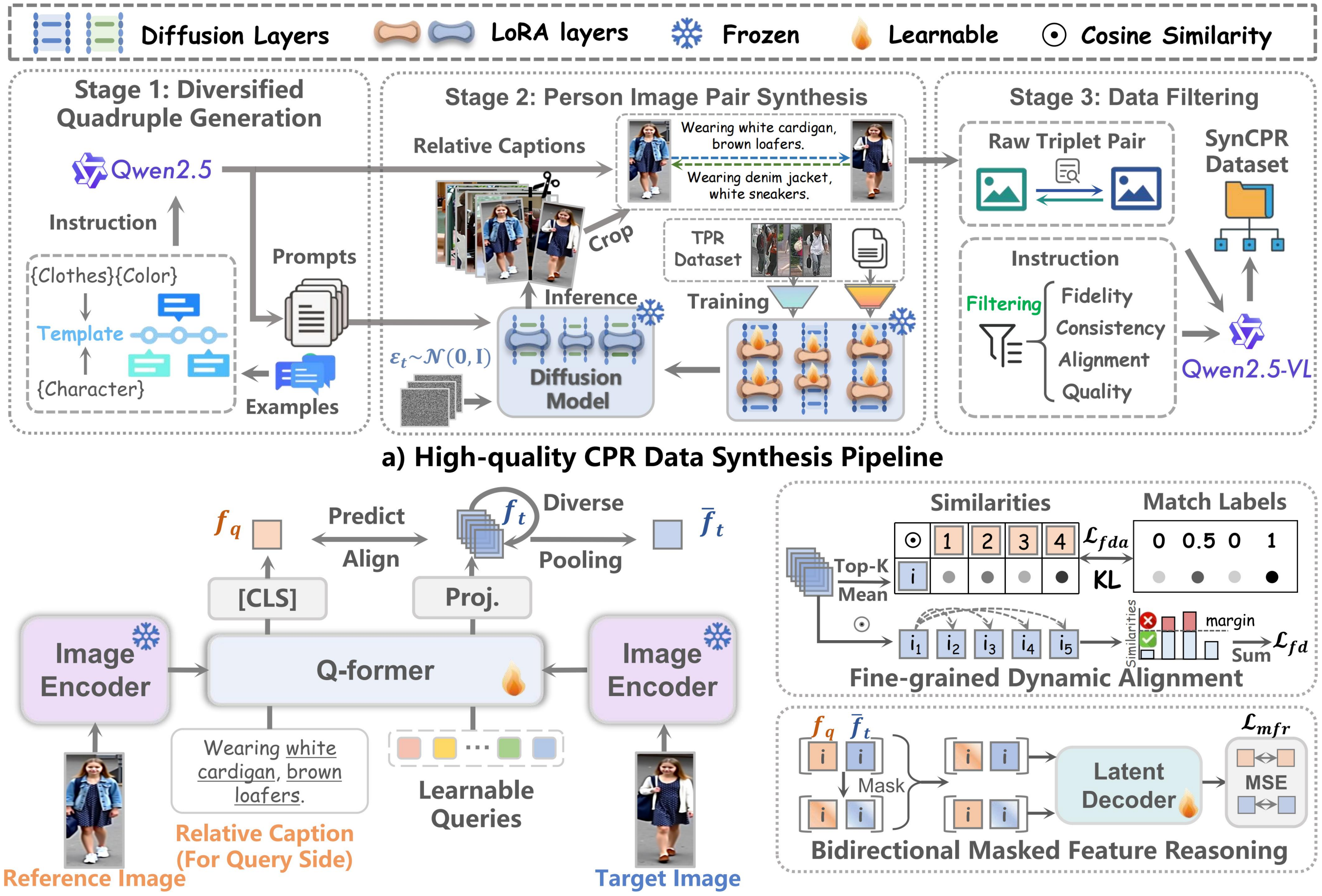
a) Various Person Retrieval Tasks

b) High-quality Triplets Generation

c) Some Examples from the SynCPR Dataset

- **New Task**: **Composed Person Retrieval (CPR)**
- **New Pipeline**: **Automatic Triplet Synthesis**
- **New Dataset**: **Million-Scale Dataset SynCPR**
- **New Benchmark**: **Manually Annotated test set ITCPR**
- **New Framework**: **Retrieval Method FAFA**

- # Method



a) High-quality CPR Data Synthesis Pipeline

b) Fine-grained Adaptive Feature Alignment Framework

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.

## Automatic Synthetic Data Pipeline

- **LLM quadruples**: Generate ⟨ reference text, target text, forward diff, reverse diff⟩ to cover rich identities & states.
- **Identity-consistent images**: LoRA-tuned Flux generates a single left–right image, then crops to (Iq,It)(I_q, I_t)(Iq,It) to guarantee same ID; vary LoRA strength for style diversity.
- **Multimodal filtering**: MLLM scores naturalness, ID consistency, text–image alignment, CPR relevance; keep only samples with mean ≥ 8.5.
- **Result**: SynCPR with 1.15M high-quality triplets.

## FAFA: Fine-grained Adaptive Feature Alignment

FAFA achieves fine-grained and adaptive alignment between visual and textual features by **dynamically matching key tokens ($\mathcal{L}_{fda}$)**, **promoting feature diversity ($\mathcal{L}_{fd}$)**, and using **masked reasoning ($\mathcal{L}_{mfr}$)** to build robust and semantically consistent representations.

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

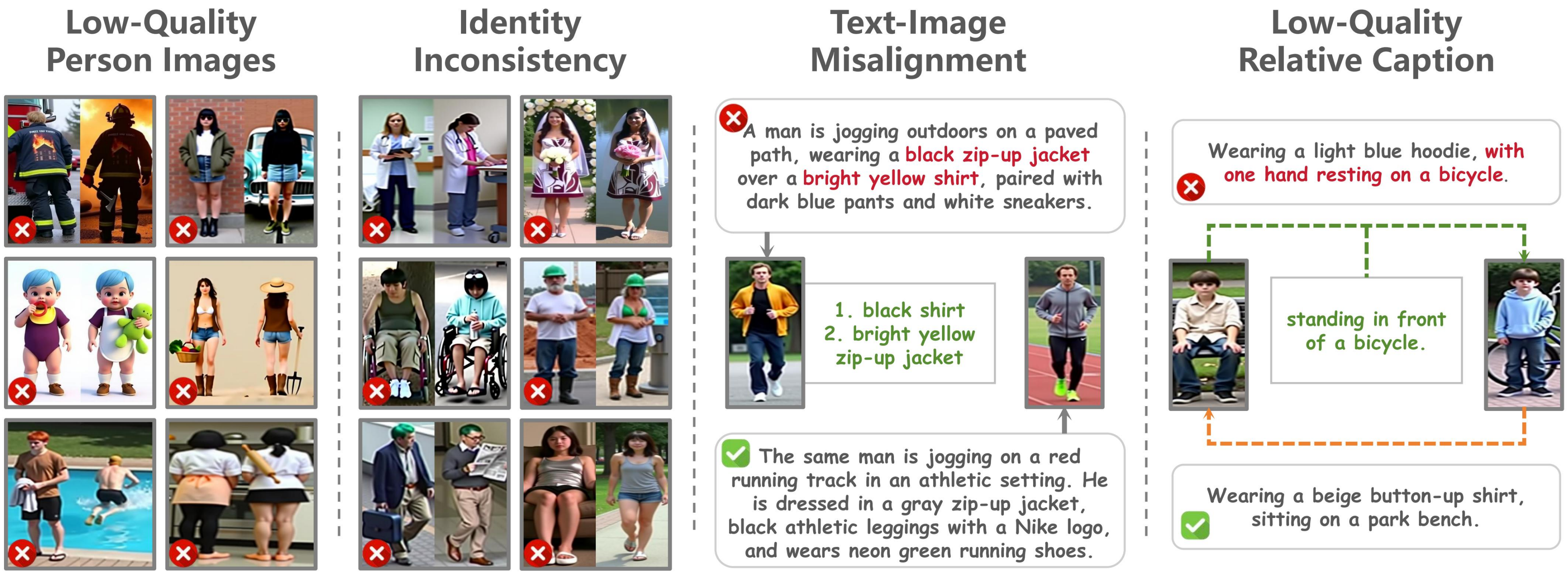
$$\mathcal{L}_{fda} = \mathcal{L}_{q2t} + \mathcal{L}_{t2q}$$

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

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

$$\mathcal{L} = \mathcal{L}_{fda} + \lambda_1\mathcal{L}_{fd} + \lambda_2\mathcal{L}_{mfr}$$

- **Data Generation**



Low-Quality Person Images

Identity Inconsistency

Text-Image Misalignment

A man is jogging outdoors on a paved path, wearing a black zip-up jacket over a bright yellow shirt, paired with dark blue pants and white sneakers.

1. black shirt
2. bright yellow zip-up jacket

The same man is jogging on a red running track in an athletic setting. He is dressed in a gray zip-up jacket, black athletic leggings with a Nike logo, and wears neon green running shoes.

Low-Quality Relative Caption

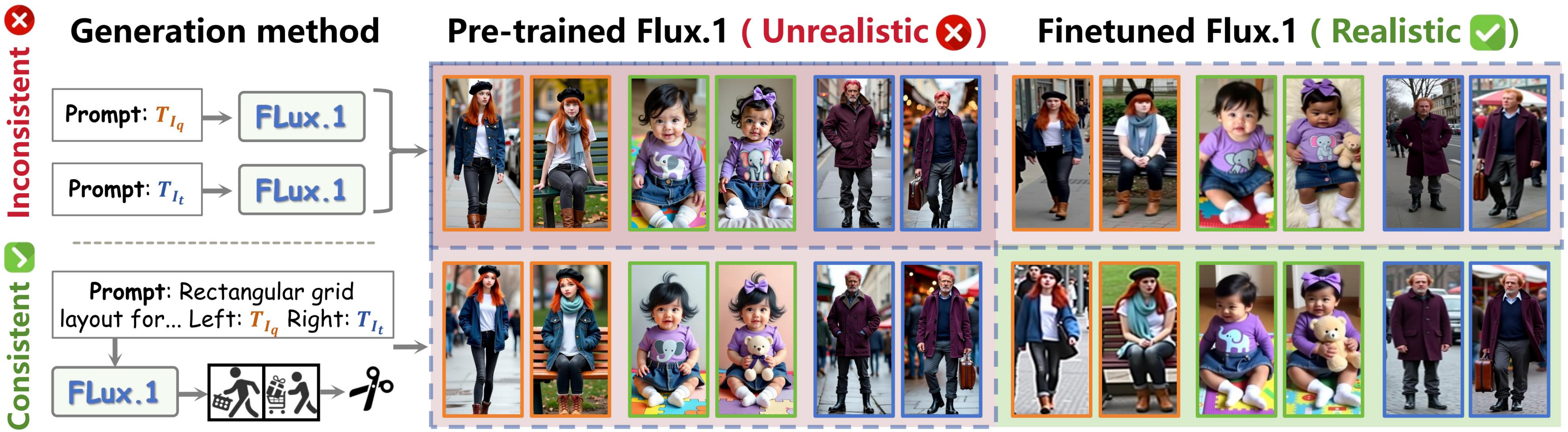
Wearing a light blue hoodie, with one hand resting on a bicycle.

standing in front of a bicycle.

Wearing a beige button-up shirt, sitting on a park bench.

## MLLM Filtering

Evaluates each triplet on four dimensions—**image naturalness, identity consistency, text–image alignment, and caption relevance**—to remove low-quality samples and retain only accurate, realistic, and semantically coherent data.



Generation method

Pre-trained Flux.1 ( Unrealistic ☒)

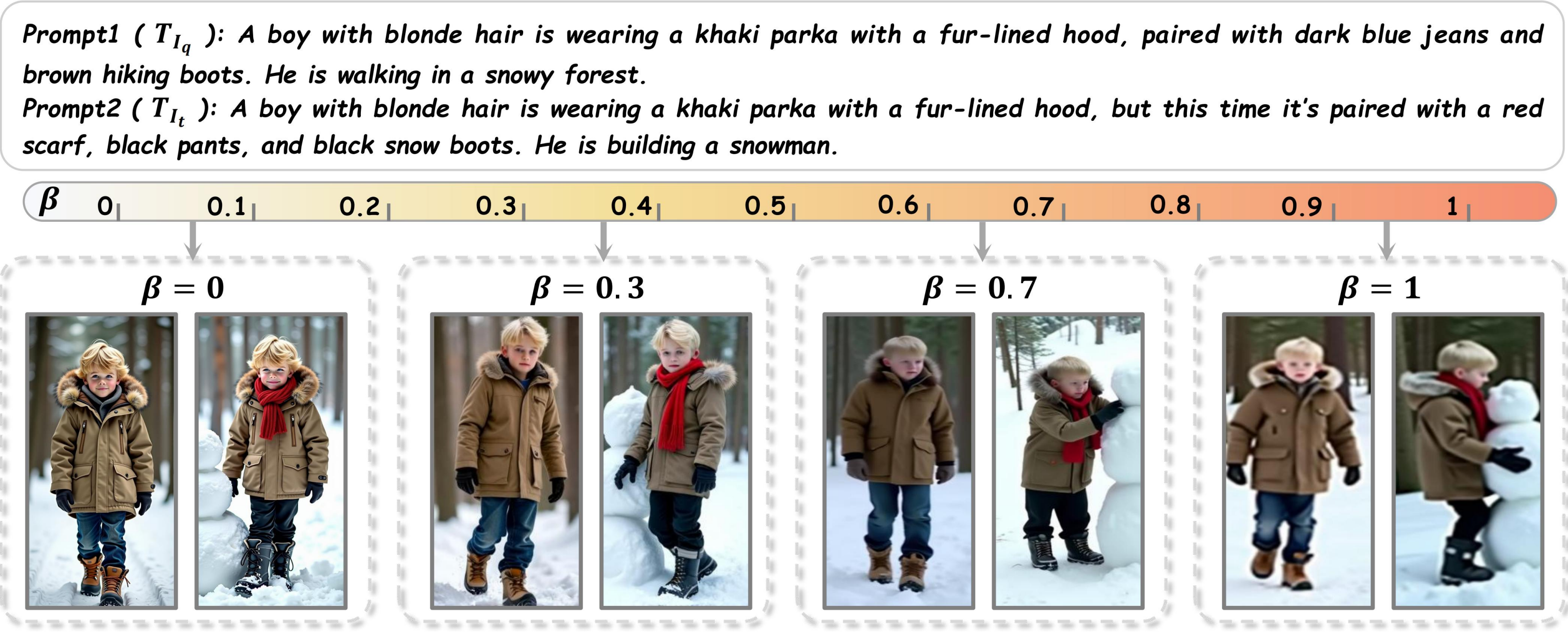
Finetuned Flux.1 ( Realistic ☑)

Prompt: Rectangular grid layout for... Left: $T_{I_q}$ Right: $T_{I_t}$

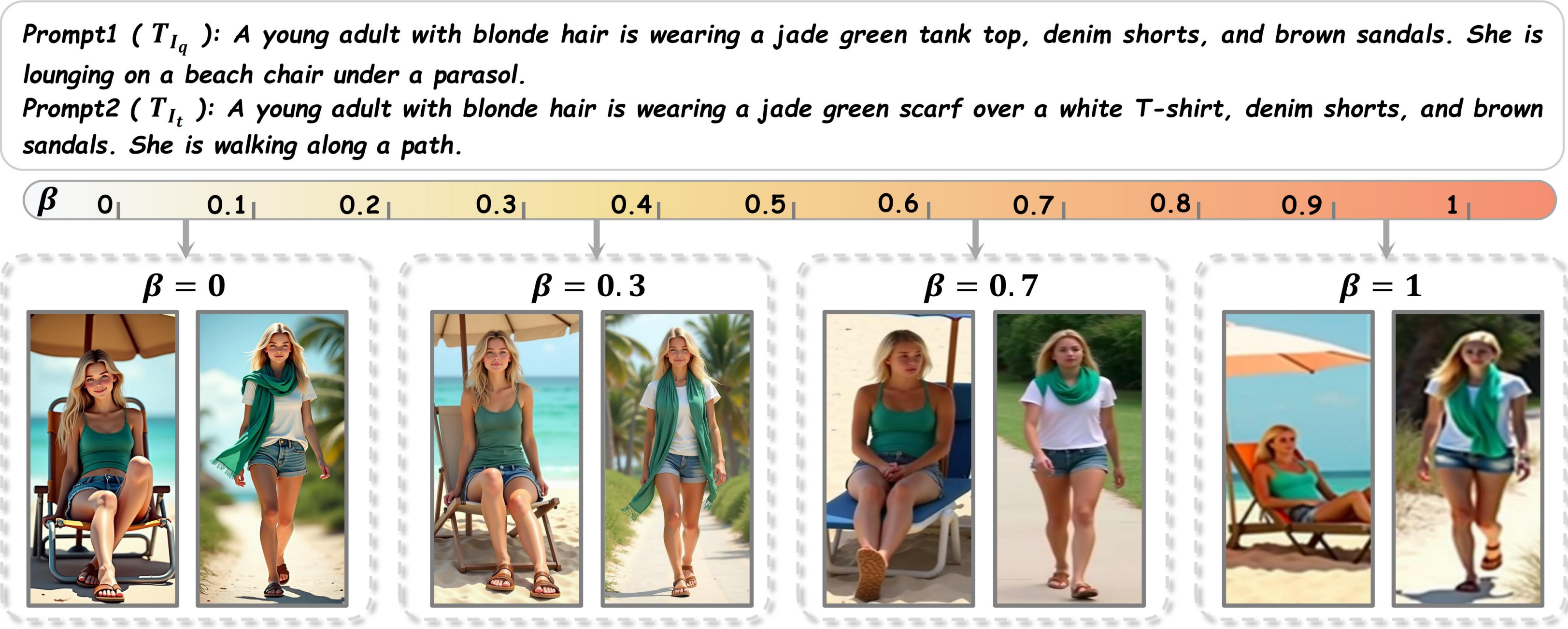
## Dual-Panel Generation

Generating two sub-images of the same person within one image **leverages the model's internal coherence to naturally preserve identity consistency while allowing controlled variations in appearance or state.**

- # Data Generation



*Prompt1 ( $T_{I_q}$ ): A boy with blonde hair is wearing a khaki parka with a fur-lined hood, paired with dark blue jeans and brown hiking boots. He is walking in a snowy forest.*
*Prompt2 ( $T_{I_t}$ ): A boy with blonde hair is wearing a khaki parka with a fur-lined hood, but this time it's paired with a red scarf, black pants, and black snow boots. He is building a snowman.*

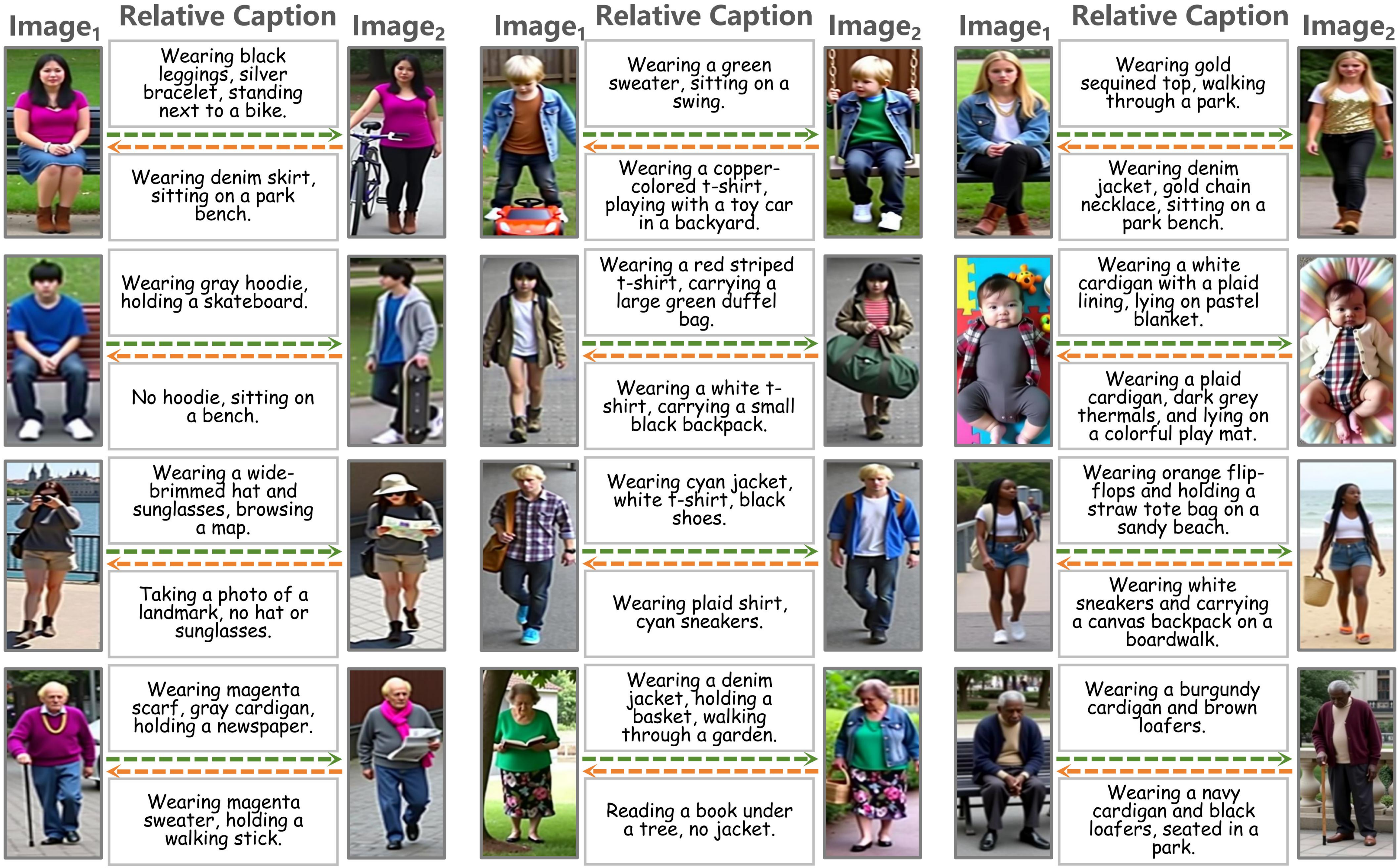


*Prompt1 ( $T_{I_q}$ ): A young adult with blonde hair is wearing a jade green tank top, denim shorts, and brown sandals. She is lounging on a beach chair under a parasol.*
*Prompt2 ( $T_{I_t}$ ): A young adult with blonde hair is wearing a jade green scarf over a white T-shirt, denim shorts, and brown sandals. She is walking along a path.*

## Dynamic β Generation

Adjusting the LoRA strength β during image synthesis produces diverse visual styles for the same identity, enriching data variability and improving the model's generalization ability.

# • SynCPR Dataset



Wearing black leggings, silver bracelet, standing next to a bike.

Wearing denim skirt, sitting on a park bench.

Wearing gray hoodie, holding a skateboard.

No hoodie, sitting on a bench.

Wearing a wide-brimmed hat and sunglasses, browsing a map.

Taking a photo of a landmark, no hat or sunglasses.

Wearing magenta scarf, gray cardigan, holding a newspaper.

Wearing magenta sweater, holding a walking stick.

Wearing a green sweater, sitting on a swing.

Wearing a copper-colored t-shirt, playing with a toy car in a backyard.

Wearing a red striped t-shirt, carrying a large green duffel bag.

Wearing a white t-shirt, carrying a small black backpack.

Wearing cyan jacket, white t-shirt, black shoes.

Wearing plaid shirt, cyan sneakers.

Wearing a denim jacket, holding a basket, walking through a garden.

Reading a book under a tree, no jacket.

Wearing gold sequined top, walking through a park.

Wearing denim jacket, gold chain necklace, sitting on a park bench.

Wearing a white cardigan with a plaid lining, lying on pastel blanket.

Wearing a plaid cardigan, dark grey thermals, and lying on a colorful play mat.

Wearing orange flip-flops and holding a straw tote bag on a sandy beach.

Wearing white sneakers and carrying a canvas backpack on a boardwalk.

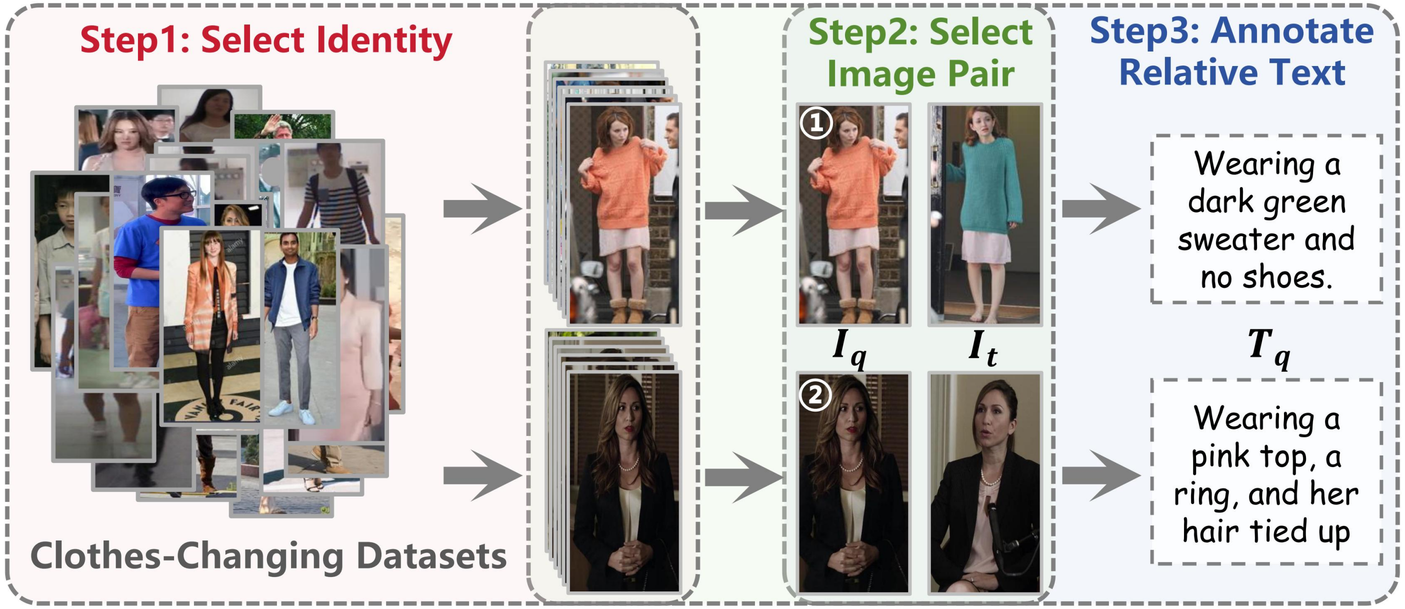
Wearing a burgundy cardigan and brown loafers.

Wearing a navy cardigan and black loafers, seated in a park.

## SynCPR Dataset Summary

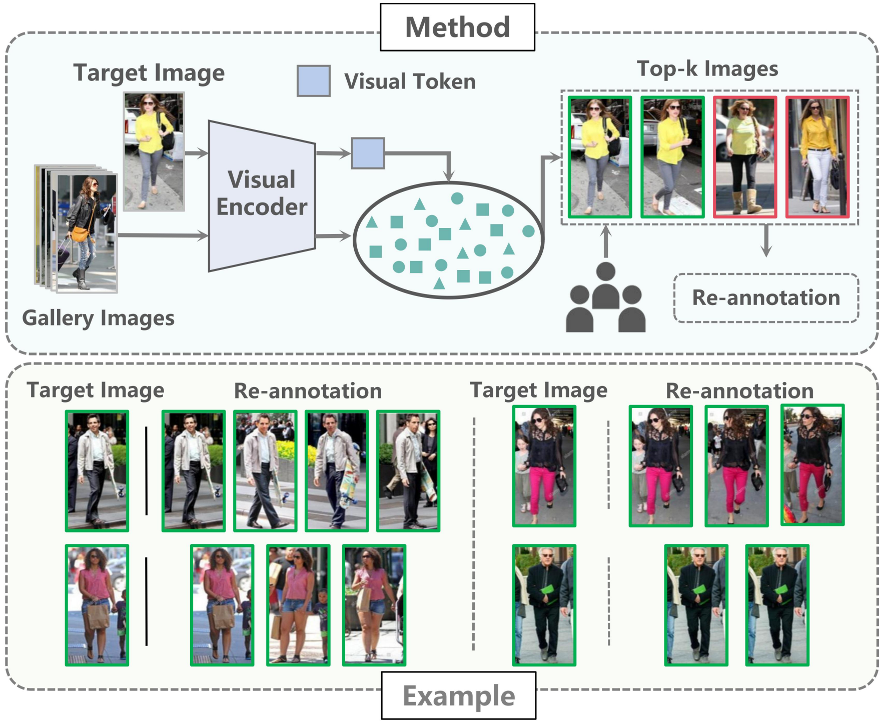
- SynCPR is a **fully synthetic, large-scale dataset** for composed person retrieval.
- It includes **1.15M high-quality triplets** generated from **140.5K textual quadruples** using LoRA-tuned Flux with dynamic β for style diversity.
- Each sample is filtered by an MLLM, **covering 177.5K group IDs**, with captions **averaging 13.3 words and a vocabulary of 4,370**.
- The dataset is balanced by **gender (51.2% male) and features rich variation in age, clothing, and scenes,** ensuring high realism and diversity.

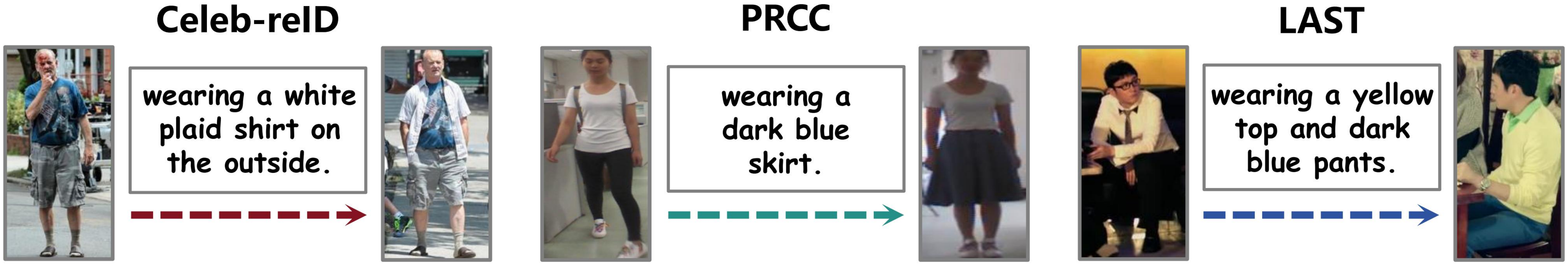
- # ITCPR Dataset



Wearing a dark green sweater and no shoes.

Wearing a pink top, a ring, and her hair tied up





wearing a white plaid shirt on the outside.

wearing a dark blue skirt.

wearing a yellow top and dark blue pants.

## ITCPR Dataset Summary

- ITCPR is a **manually annotated test set** for the CPR task, containing **2,225 triplets with 2,202 unique (image–text) queries**.
- It includes 1,151 images / 512 IDs from Celeb-reID, 146 / 146 from PRCC, and 905 / 541 from LAST, forming a gallery of 20,510 images with 2,225 ground truths. Captions average 9.54 words (range: 3–32).
- The dataset is used exclusively for zero-shot testing, ensuring no overlap with training data.

- **Results**

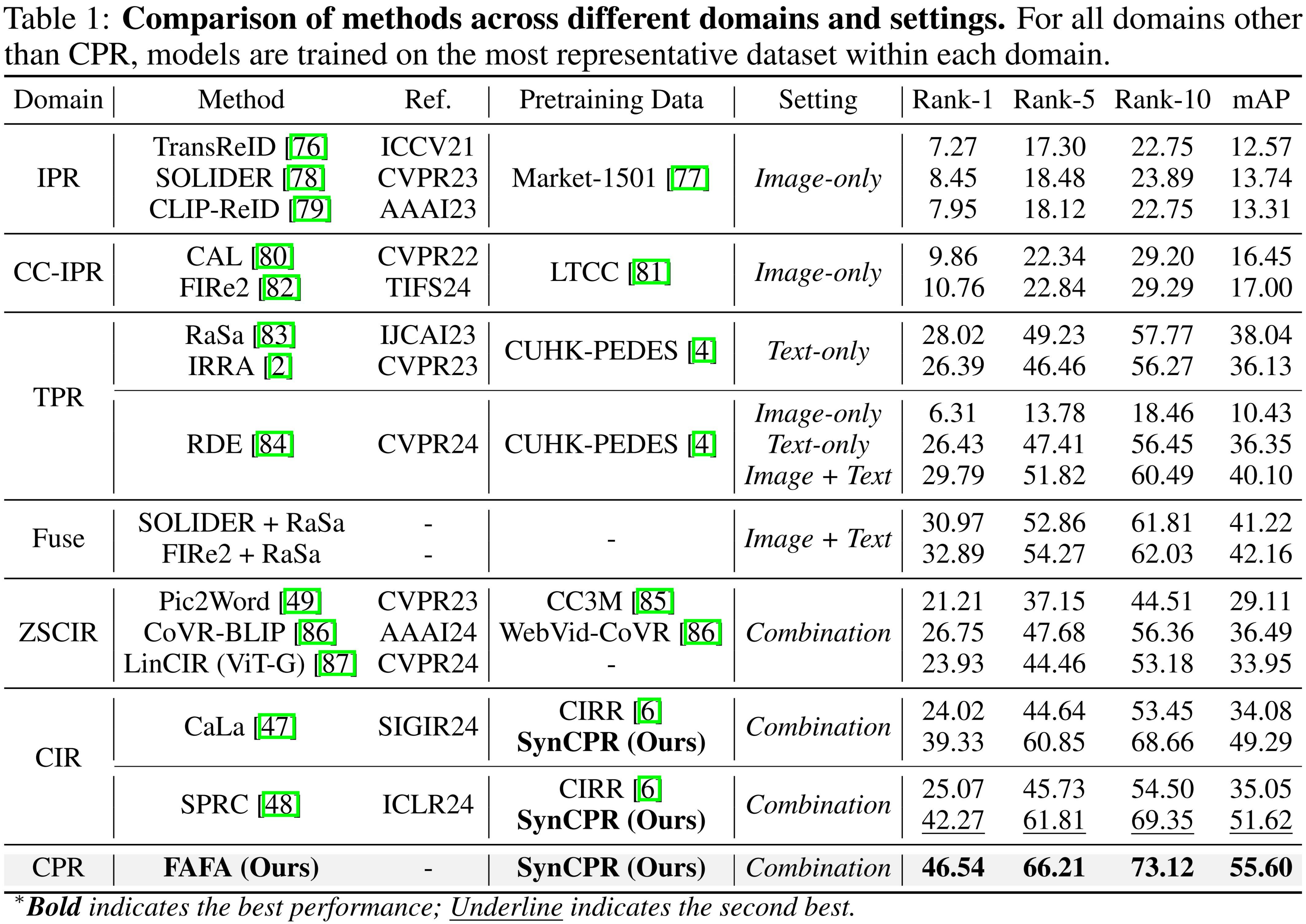


Table 1: **Comparison of methods across different domains and settings.** For all domains other than CPR, models are trained on the most representative dataset within each domain.

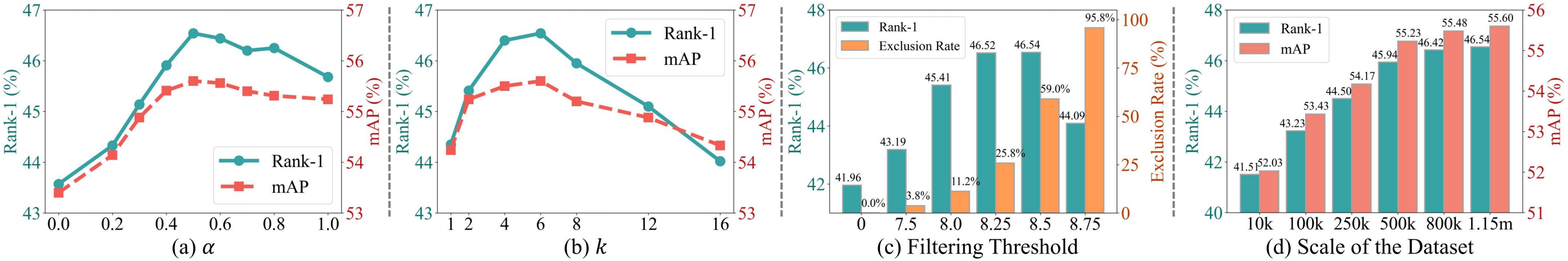
| Domain | Method | Ref. | Pretraining Data | Setting | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|---|---|---|---|
| IPR | TransReID [76] | ICCV21 | Market-1501 [77] | *Image-only* | 7.27 | 17.30 | 22.75 | 12.57 |
| | SOLIDER [78] | CVPR23 | | | 8.45 | 18.48 | 23.89 | 13.74 |
| | CLIP-ReID [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.*

- **Ablation Study**

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 | ✓ | | ✓ | ✓ | | 46.05 | 65.85 | 73.02 | 55.49 |
| 4 | ✓ | | ✓ | | ✓ | 45.78 | 65.58 | 72.62 | 55.13 |
| 5 | ✓ | | ✓ | ✓ | ✓ | **46.54** | **66.21** | **73.12** | **55.60** |



(a) $\alpha$

(b) $k$

(c) Filtering Threshold

(d) Scale of the Dataset

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

- **Ablation Study**

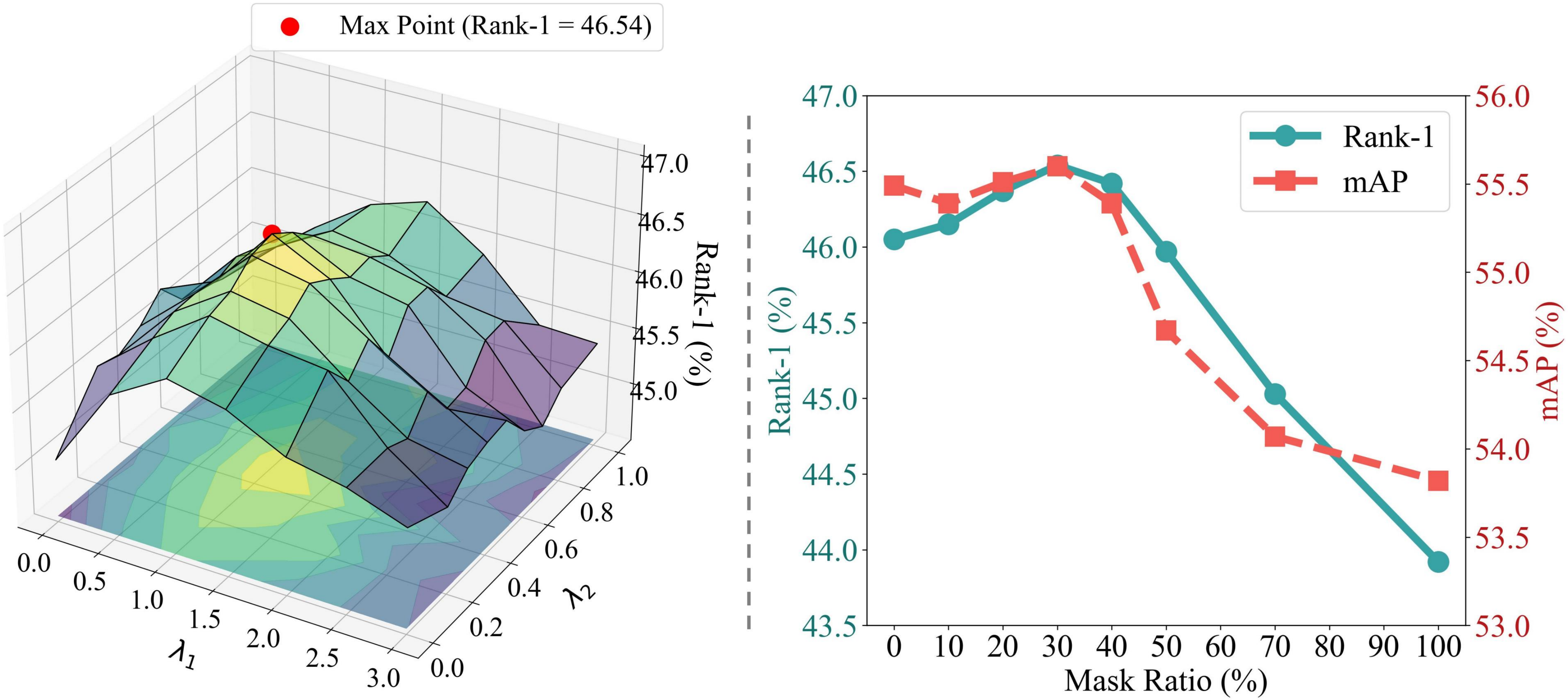


Figure S13: **Left**: Variations in FAFA's Rank-1 performance under different balancing weights of auxiliary loss terms. **Right**: Relationship between FAFA's performance and the feature mask ratio in $\mathcal{L}_{mfr}$.

# Thank you！

If you are interested, you can **visit and star our project page**, where **we provide access to all datasets and the implementation code** of our method.

**Project Page**: https://github.com/Delong-liu-bupt/Composed_Person_Retrieval

Presenter: Delong Liu　　Date: Nov 5, 2025

BUPT MCPRL

北京郵電大學
Beijing University of Posts and Telecommunications

商汤 sensetime

NEURAL INFORMATION PROCESSING SYSTEMS