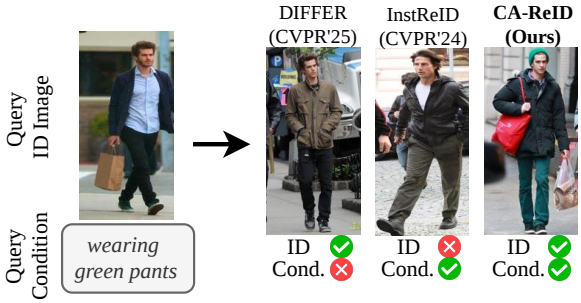
Composite-Attribute Person Re-Identification via Pose-Guided Disentanglement

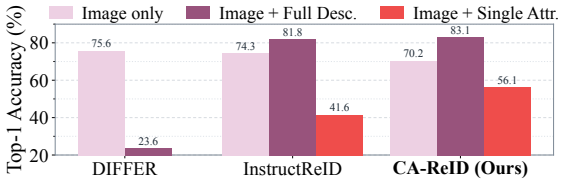
Kartik Patwari[1*] Noranart Vesdapunt[2] Chien-Yi Wang[2] Dawei Li[2]
Cong Phuoc Huynh[2] Ning Zhou[2] Chen-Nee Chuah[1] Kah Kuen Fu[2]
[1]University of California, Davis [2]Amazon

Abstract

Recent advancements in vision-language models have enabled multi-modal person re-identification (Re-ID), where the system takes both an image and a text query to identify matching individuals. While previous state-of-the-art methods perform well with detailed, sentence-level descriptions, we found that their Recall@1 drops by half when using short, keyword-based queries due to ambiguity, training biases, and under-represented attributes. Despite this challenge, short queries provide a more natural and efficient user experience, requiring less effort and allowing for iterative refinement. To address this limitation, we introduce a new problem setting, Composite-Attributes Person Re-ID (CA-ReID), along with a fine-grained composite attribute dataset with queries belonging to varying levels of ambiguity. We further propose two methods: Dense Disentangling Loss to promote attribute-specific embeddings, and Part-Aware Representations that use pose estimation to align textual attributes with relevant body regions. Our method sets a new state of the art on the new CA-ReID benchmark (up to +17% Recall@1) and performs on par with prior methods on existing CC-ReID benchmarks.



(a) Example rank-1 retrievals with composite attributes.



(b) Top-1 accuracy drops off with composite attributes.

Figure 1. **(a)** Existing methods, such as DIFFER [38], perform image-only Re-ID without text control, while multimodal InstructReID [21] underperforms with short attribute queries. **(b)** Top1 accuracy drops on the Celeb-ReID-L [24] dataset with single attribute query vs full description, highlighting the challenge of jointly satisfying identity and short keyword constraints.

1. Introduction

Person re-identification (Re-ID) aims to recognize the same individual across different camera perspectives and varying backgrounds. Traditional Re-ID methods [20, 22, 36, 45, 51, 67, 77, 80, 81] formulate the task as an image-to-image retrieval problem, where the objective is to retrieve instances of the same person without any additional text queries. Recent advancements in vision-language models (VLMs) have enabled multi-modal retrieval [2, 21, 38, 62] using both image and text inputs, allowing more fine-grained control over the text query, such as specifying "wearing a red jacket, purple pants, and black shoes." However, we discovered that state-of-the-art (SOTA) methods [21], while effective with detailed sentence-level descriptions, suffer performance degradation when using short attribute keyword queries. For example, a simple short phrase like "wearing green pants" can reduce Recall@1 of InstructReID [21] by half on COCAS+Real2 dataset [35].

This drop in performance stems from three main issues: (1) short keywords are often ambiguous, for example, over 30% of individuals in the Celeb-ReID-Light dataset [24] match the phrase "long pants" requiring the model to rely on visual identity to disambiguate. (2) Existing VLM-based Re-ID models are primarily trained on full-sentence captions, making them less effective at interpreting brief and ambiguous phrases. (3) Rare keywords, such as "straw hat," are underrepresented in the training data, leading to weak generalization. Despite these challenges, short keyword queries offer a better user experience than full descriptive sentences, as they require less effort and feel more natural; users can start with a simple keyword and iteratively refine

*This work was done at Amazon.

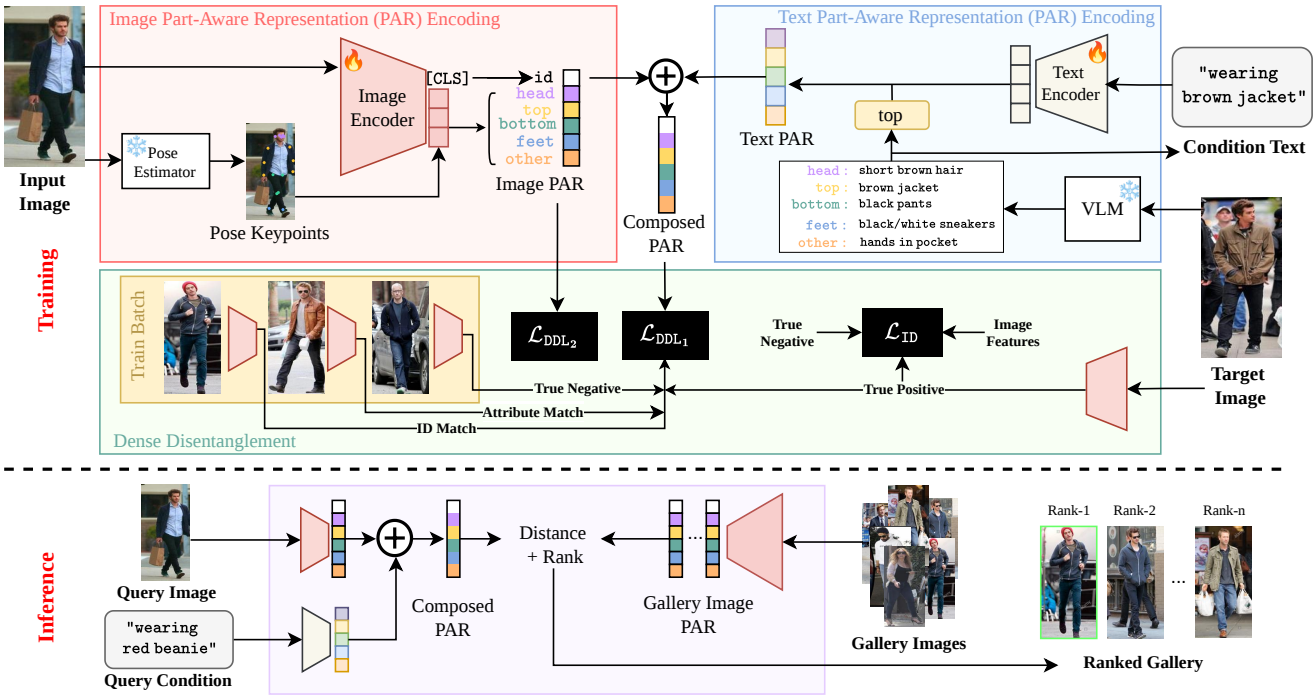


Figure 2. **Overall architecture of the proposed CA-ReID framework.** Given an input image and a set of attributes, we construct *Part-Aware Representations* (PAR) for both image and text modalities. Pose-guided patch pooling and the image encoder generate part-aware image slots (*id*, *head*, *top*, *bottom*, *feet*, and *other*), while a text encoder produces the corresponding part-aware text slots. These PAR features are then fused to form a composed representation. Training jointly optimizes a standard identity loss $\mathcal{L}_{\text{ID}}$ and two Dense Disentangling Losses: (1) $\mathcal{L}_{\text{DDL}_1}$ aligns the composed PAR with target-image features under both identity and attribute conditions, and (2) $\mathcal{L}_{\text{DDL}_2}$ enforces disentanglement across part-aware image features. At inference, a query image (with extracted pose keypoints) and a condition description are converted into PAR, fused into a composed feature, and used for retrieval and re-ranking of gallery images.

their search by adding more terms.

To address the above issues, we first introduce a fine-grained dataset categorized into five attribute groups: headwear/hairstyle, top clothing, bottom clothing, footwear, and other, which includes accessories, belongings, and context information (e.g., shopping bag, holding phone). Based on this, we propose a new task, Composite-Attributes Person Re-identification (CA-ReID, see Fig. 1(a)), which groups attributes into three ambiguity levels: (1) hard, defined as a single attribute that represents a user search with a simple keyword and high ambiguity; (2) medium, consisting of 2–3 attributes that form a composition of keywords with medium ambiguity; and (3) easy, comprising 4+ attributes representing low ambiguity. This dataset motivates our novel Dense Disentangling Loss (DDL) that encourages the model to assign specific embedding dimensions to distinct attributes while suppressing activations in irrelevant dimensions. As a result, each attribute-specific embedding contributes exclusively to its corresponding feature. For example, a query like "jacket" activates only the top clothing embedding, reducing confusion from unrelated attributes such as bottom clothing, footwear, or accessories.

We further propose Part-Aware Representations (PAR, see Fig. 2) by using a human pose estimation model to further reduce the ambiguity from a short query. Our method constructs spatial patches corresponding to individual human body parts and associates each part with a specific embedding. This alignment provides strong spatial priors, enabling more accurate disambiguation. For instance, the head embedding is directly associated with the head region, allowing attributes such as hat or curly hairstyle to be inferred exclusively from the head area. Combined with DDL, our approach improves *Recall@1 by up to +15%* over the previous best method [21] on our benchmark, while maintaining comparable performance on existing CC-ReID datasets. To summarize, our contributions are:

- We introduce *Composite-Attribute Person Re-ID (CA-ReID)*, where a query pairs a reference image with short/composite keywords (e.g., "red jacket, sunglasses") for more intuitive search. In this setting, existing CC-/LI-ReID methods [21, 38] experience sharp performance drops (Recall@1 nearly halves) due to keyword ambiguity, short-phrase vs. full description mismatch, and rare attributes.
- We propose *Part-Aware Representations (PAR)* and a *Dense Disentangling Loss (DDL)* that tie attributes to body regions, encouraging identity–attribute separation; e.g., a

"sunglasses" query links to the head embedding, driving retrieval by head features rather than other parts.

• We achieve up to *+15% Recall@1* over the previous best method [21] on our composite-attribute dataset, while maintaining on-par results on existing re-id benchmarks.

2. Related Work

Person ReID. Traditional person re-identification (ReID) retrieves the same individual across non-overlapping cameras despite changes in pose and background [15, 20, 22, 36, 45, 74, 77, 80, 81]. Clothes-changing ReID (CC-ReID) extends this to large clothing variations, learning clothing-agnostic representations [16, 25, 41, 42, 73], part-based features [6, 43, 61], temporal modeling [51], or lifelong knowledge consolidation [10]. Attribute-guided approaches further improve fine-grained retrieval by incorporating semantic supervision [29, 37, 38]. Text-to-image ReID (T2I-ReID) expands retrieval using natural language queries, from early text-based person search [19, 33] and cross-modal alignment [5, 12, 27, 34, 39, 52, 70, 79, 83] to CLIP-driven and MLLM-based models using large-scale vision–language pretraining, human-in-the-loop refinement, and multi-turn dialogue [2, 3, 28, 53, 62]. Unified multi-modal ReID methods handle RGB/IR/thermal, sketch, or text within one model, often with LLMs for instruction-guided retrieval [17, 32, 69]. Instruct-ReID [21] defines a multi-purpose, instruction-guided ReID benchmark and introduces LI-ReID, where a reference image is paired with a full text description as auxiliary information to retrieve any image of the same identity. In contrast, CA-ReID formulates CC-ReID as composed image retrieval, jointly conditioning on a reference image and a short composite attribute edits so that the retrieved image satisfies both identity and requested attribute(s). To the best of our knowledge, this setting has not been studied in prior person ReID.

Human Part-Aware Feature Disentanglement. Feature disentanglement is widely used in vision to separate identity-related content from appearance or background, including for human body editing and pose/appearance decomposition [44, 60, 68, 75]. In person Re-ID, it helps separate identity-relevant and appearance-specific features (e.g., clothing) to improve robustness under appearance changes [9, 13]. Recent works further exploit 3D body shape, color disentanglement, and generative augmentation to decouple identity from clothing appearance [6, 18, 41, 43, 49, 55, 60]. DIFFER [78] uses VLM-generated descriptions supervision to disentangle identity and clothing, while pose-guided methods [46, 56, 66] improve part-based alignment. PFD [66] uses pose heatmaps to guide feature disentangling for part-aware representations. Such methods typically couple a ReID backbone with an external pose estimator or 3D body model, using these cues at both training and inference to stabilize part alignment [41, 43, 46, 56]. However, these methods focus on visual disentanglement and appearance robustness, without supporting *text composition queries* over part-level features. Our approach explicitly connects part-aware representation learning with language by using disentangled part slots as an interface for compositional text edits in CA-ReID.

Composed Image Retrieval (CIR). CIR retrieves images conditioned on a reference image and a modifying text instruction, and is mainly studied on natural images such as fashion, products, and scenes [4, 8, 31, 48, 63]. Prior work typically operates in CLIP-style vision–language embedding spaces, combining the reference and edit text via feature modulation, residual composition, or attention [4, 30, 58, 64], while more recent work uses large-scale VLMs/LLMs and pretraining for richer compositional reasoning [11, 23, 57, 76]. Existing CIR benchmarks target generic objects and do not enforce maintaining identity, whereas we instantiate CIR for person Re-ID, requiring results to match both identity and the requested attribute edit.

3. Composite Attributes Person Re-ID

3.1. Problem Formulation

Given a query image $\mathbf{x}_q \in \mathbb{R}^{H\times W\times 3}$, attribute text t, and a gallery $\mathcal{G} = \{\mathbf{x}_g^i\}_{i=1}^N$, our goal is to retrieve gallery images that (1) share the same identity as $\mathbf{x}_q$ and (2) satisfy the attributes described in t. The model ranks $\mathcal{G}$ using a similarity score $s(\mathbf{x}_q, \mathbf{x}_g^i, t)$, where higher values indicate better matches in terms of both identity and attributes. The score, $s(\cdot)$ measures similarity in a learned embedding space.

CA-ReID differs from prior settings: CC-ReID performs image-to-image matching across clothing changes without text; T2I-ReID uses text-only queries with full descriptions and no reference image; and LI-ReID [21] combines an image and full-sentence description, optimizing only for identity. In contrast, CA-ReID jointly enforces identity and short or composite attribute conditions. We focus on queries such as [query image] + "with red jacket and sunglasses."

To support such queries, we organize attributes into five regions: *head/hair*, *top*, *bottom*, *feet*, and *other* (accessories, belongings, context). Short phrases (e.g., "brown cap") or composite descriptions (e.g., "red jacket and white sneakers") can mention any subset of these regions, naturally supporting both single- and multi-attribute compositions. Depending on the ambiguity, this may result in either single- or multi-target matches in the gallery.

3.2. Architecture Overview

We use CLIP-based [14, 54] encoders for vision ($\mathcal{E}_I$) and text ($\mathcal{E}_T$) to build Part-Aware Representations (PAR). Given an image $\mathbf{x}$ and attribute text t, PAR decomposes features into slots separating identity from part-level attributes (head, upper body, lower body, feet, other) for training and

retrieval. For visual PAR (Sec. 3.3), $\mathcal{E}_I$ outputs a global token and patch tokens, which we group into semantic regions (head, top, bottom, feet) using a pose estimator [59] which serves as a modular spatial prior for grouping. The identity slot is computed from the global and head-region features to capture face-driven identity cues while remaining robust to clothing changes. For textual PAR (Sec. 3.4), $\mathcal{E}_T$ encodes input attribute t_k into part-aligned text slots, and a gating module activates only the mentioned regions (e.g., "red jacket" updates the top slot). Text and image PAR are fused with FiLM [50] to produce composed PAR, leaving the identity slot untouched. This produces a composed representation that preserves identity and reflects the desired attribute modifications. During training, we optimize both a Dense Disentangling Loss and a standard identity loss (Sec. 3.6). At inference (Sec. 3.7), gallery images are encoded once into image PAR, and retrieval ranks them against the composed query PAR. Fig. 2 shows our pipeline.

3.3. Part-Aware Representations (Image)

We use a vision transformer encoder $\mathcal{E}_I$ to extract a global embedding $\mathbf{c} \in \mathbb{R}^d$ and patch tokens $\{\mathbf{p}_i\}_{i=1}^N$, where each patch token represents a fixed spatial region of the input. To obtain semantic body regions, we run an off-the-shelf pose estimator [59] and map detected keypoints to patch-index sets. For each part $k \in \{\text{head}, \text{top}, \text{bottom}, \text{feet}, \text{other}\}$, we precompute an index set $\mathcal{I}_k \subseteq \{1, \dots, N\}$ indicating which ViT patches belong to that region. These patch–part assignments are cached during training. For each part k, we extract a part feature by mean-pooling the corresponding patch tokens, followed by a lightweight projection:

$$\mathbf{f}_k^I = \text{MLP}_k\left(\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\mathbf{p}_i\right) \in \mathbb{R}^d. \tag{1}$$

To construct an identity feature that benefits from global appearance cues while using discriminative facial regions, we combine the global token with the head-part feature: $\mathbf{f}_{\text{id}}^I = \text{MLP}_{\text{id}}([\mathbf{c}\,;\,\mathbf{f}_{\text{head}}^I])$ where $[\cdot;\cdot]$ denotes concatenation. This preserves identity-relevant information while remaining invariant to clothing changes. The final image part-aware representation consists of five localized part features and the identity vector:

$$\mathbf{F}^I = \left\{\mathbf{f}_{\text{id}}^I, \mathbf{f}_{\text{head}}^I, \mathbf{f}_{\text{top}}^I, \mathbf{f}_{\text{bottom}}^I, \mathbf{f}_{\text{feet}}^I, \mathbf{f}_{\text{other}}^I\right\}. \tag{2}$$

which enables fine-grained reasoning over both identity and localized attributes within a unified feature space.

3.4. Part-Aware Representations (Text)

To enable compositional attribute control, we first transform the global text embedding into part-specific representations that align with the image features. We encode the conditioning text c_t using a frozen CLIP text encoder $\mathcal{E}_T$, which produces a global text embedding $\mathbf{f}^T \in \mathbb{R}^d$. This global representation is then projected into part-specific slots through learned transformations: $\mathbf{f}_k^T = W_k^T\,\mathbf{f}^T$, where $W_k^T \in \mathbb{R}^{d\times d}$ is a learnable projection matrix for each part $k \in \{\text{head}, \text{top}, \text{bottom}, \text{feet}, \text{other}\}$. To ensure that only relevant parts are activated, we use a parsing mechanism that analyzes c_t to identify mentioned attributes. For parts corresponding to unmentioned attributes, we suppress their activations by setting $\mathbf{f}_k^T = \mathbf{0}$; e.g., the query "red jacket" would activate only $\mathbf{f}_{\text{top}}^T$ while keeping other text slots zero. This selective activation prevents spurious cross-part interference and ensures that textual conditioning is applied only where semantically appropriate. During training, frozen vision-language models extract part-specific attributes from target images (e.g., *top*: "red jacket"), supervising these part-aware projections without manual annotations.

3.5. Composed Part Aware Representations

To integrate attribute conditioning into the visual representation, we fuse image and text part features using an additive residual feature-wise linear modulation network ϕ_{FLM} (a lightweight MLP applied slot-wise) inspired by [50]. For each attribute category k, the fusion module receives the image and text features and predicts a modulation vector:

$$\hat{\mathbf{f}}_k = \mathcal{N}\left(\mathbf{f}_k^I + \boldsymbol{\beta}_k\right), \qquad \boldsymbol{\beta}_k = \phi_{FiLM}\left([\mathbf{f}_k^I;\,\mathbf{f}_k^T]\right), \tag{3}$$

where $\mathcal{N}$ is L-2 feature normalization. This formulation allows the text to selectively modify the appearance of the corresponding visual region while preserving the underlying spatial layout encoded by the vision transformer. The identity feature remains unmodulated to ensure that identity information does not shift under attribute queries: $\hat{\mathbf{f}}_{\text{id}} = \mathbf{f}_{\text{id}}^I$. The final representation is

$$\hat{\mathbf{F}} = \{\hat{\mathbf{f}}_{\text{id}}, \hat{\mathbf{f}}_{\text{head}}, \hat{\mathbf{f}}_{\text{top}}, \hat{\mathbf{f}}_{\text{bottom}}, \hat{\mathbf{f}}_{\text{feet}}, \hat{\mathbf{f}}_{\text{other}}\}, \tag{4}$$

which merges identity and localized attribute cues in a structured, part-aware space.

3.6. Dense Disentangling Loss

To encourage the model to learn part-specific features while maintaining identity, we introduce a Dense Disentangling Loss (DDL) with two complementary objectives, identity–attribute disentanglement and inter-part orthogonality.

3.6.1. Identity-Attribute Disentanglement

The first component, $\mathcal{L}_{\text{DDL}_1}$, learns to separate identity information from attribute variations by considering four types of sample relationships within training batches. For each anchor query with composed features $\hat{\mathbf{F}}$, we identify: (1) **Full Match** ($\mathbf{F}^+$): Target images sharing both the same identity and matching attributes. These should be maximally similar to the anchor. (2) **Attribute-Only Match**

($\mathbf{F}^{A+}$): Images with matching attributes but different identities. These should align in attribute dimensions while differing in identity features. (3) **Identity-Only Match** ($\mathbf{F}^{I+}$): Images of the same person with different attributes. These should maintain identity similarity while allowing attribute variation. (4) **No Match** ($\mathbf{F}^{-}$): Images with neither matching identity nor attributes. These should be maximally dissimilar. We formulate this as a multi-triplet loss:

$$\mathcal{L}_{\text{DDL}_1} = \alpha_1 \mathcal{L}_{\text{tri}}(\hat{\mathbf{F}}, \mathbf{F}^{+}, \mathbf{F}^{A+}) + \alpha_2 \mathcal{L}_{\text{tri}}(\hat{\mathbf{F}}, \mathbf{F}^{+}, \mathbf{F}^{I+}) + \alpha_3 \mathcal{L}_{\text{tri}}(\hat{\mathbf{F}}, \mathbf{F}^{+}, \mathbf{F}^{-}), \tag{5}$$

where $\mathcal{L}_{\text{tri}}(\cdot)$ denotes triplet loss with cosine similarity, and $\{\alpha_1, \alpha_2, \alpha_3\}$ are balancing coefficients. We mine samples globally within each batch using identity labels and part-level attribute annotations. If any of the three required relational samples cannot be found for an anchor, that triplet term is omitted.

3.6.2. Inter-Part Orthogonality

The second component, $\mathcal{L}_{\text{DDL}_2}$, enforces disentanglement between body parts by minimizing their mutual similarity:

$$\mathcal{L}_{\text{DDL}_2} = \sum_{k_i \neq k_j} \left\| \cos(\hat{\mathbf{f}}_{k_i}, \hat{\mathbf{f}}_{k_j}) \right\|^2, \tag{6}$$

where $k_i, k_j \in \{\text{head}, \text{top}, \text{bottom}, \text{feet}, \text{other}\}$ denote different body regions. This regularization prevents feature leakage across parts, ensuring that each slot captures only its designated semantic region.

3.6.3. Overall Training Objective

In addition to DDL, we employ a standard identity loss to maintain person discrimination:

$$\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{tri}}(\hat{\mathbf{f}}_{\text{id}}, \mathbf{f}^{+}_{\text{id}}, \mathbf{f}^{-}_{\text{id}}), \tag{7}$$

where $\mathbf{f}^{+}_{\text{id}}$ and $\mathbf{f}^{-}_{\text{id}}$ denote identity features from same-person and different-person samples, respectively. The complete training objective combines all components:

$$\mathcal{L}_{\text{Total}} = \lambda_1 \mathcal{L}_{\text{ID}} + \lambda_2 \mathcal{L}_{\text{DDL1}} + \lambda_3 \mathcal{L}_{\text{DDL2}}, \tag{8}$$

where $\{\lambda_1, \lambda_2, \lambda_3\}$ are hyperparameters controlling the relative importance of each objective.

3.7. Inference

At test time, no VLM annotations are needed. Given a query image $\mathbf{x}_q$, conditioning text t, and gallery images $\{\mathbf{x}^i_g\}$, we first obtain image PARs $\mathbf{F}^I_q$ and $\mathbf{F}^I_g$ using $\mathcal{E}_I$. We encode t with $\mathcal{E}_T$ to get text PARs ($\mathbf{F}^T$), and form a slot-wise composed query PAR following Sec. 3.5

$$\hat{\mathbf{F}}_q = \mathcal{N}\left(\mathbf{F}^I_q + \phi_{\text{FiLM}}(\mathbf{F}^I_q; \mathbf{F}^T)\right). \tag{9}$$

where $\hat{\mathbf{F}}_q$ is the composed query PAR. Retrieval is done by cosine similarity between $\hat{\mathbf{F}}_q$ and cached gallery PARs:

$$s(\mathbf{x}_q, \mathbf{x}_g, t) = \frac{\hat{\mathbf{F}}_q \cdot \mathbf{F}^I_g}{\|\hat{\mathbf{F}}_q\| \|\mathbf{F}^I_g\|}. \tag{10}$$

For standard CC-ReID (image-only), we omit t and directly compare $\mathbf{F}^I_q$ and $\mathbf{F}^I_g$ with cosine similarity.

4. Composed Attribute Dataset

4.1. Motivation

Recent multimodal Re-ID work, particularly Instruct-ReID [21], introduced Language-Instructed Re-ID (LI-ReID): pairing reference images with full-sentence descriptions from GPT-4 labeled COCAS+Real2 [35] to retrieve same-identity images. While promising, LI-ReID evaluation remains identity-centric and inherits three critical limitations from its dataset: (1) *Identity-only evaluation*: retrieval success requires only correct identity, regardless of whether text conditions (e.g., "with black bag") are satisfied; the text functions as auxiliary guidance rather than a binding constraint; (2) *Limited diversity*: COCAS+Real2 contains merely 101 identities with few outfit variations across multiple views, while its GPT-4 captions, generated from fixed attribute lists, produce repetitive phrases with restricted coverage; (3) *No compositional support*: the benchmark requires full-sentence descriptions paired with reference images, failing to evaluate keyword-based or compositional queries (e.g., "red jacket", "holding cup") that demand joint identity-attribute reasoning.

To address these limitations and enable CA-ReID, we construct a comprehensive dataset with explicit identity-condition retrieval protocol. We leverage a diverse Re-ID base [24] and annotate images with open-vocabulary attributes spanning new categories (appearance, accessories, belongings, and contextual actions; see Fig. 3) via an MLLM-as-judge pipeline. We compose queries at three granularities (Easy/Medium/Hard) ranging from full descriptions to single keywords (shown in Fig. 4), and evaluate by enforcing joint identity and attribute satisfaction, supporting both single and multi-target scenarios.

4.2. Dataset Construction

We build our CA-ReID benchmark on Celeb-ReID-Light [24] (590 identities), which provides rich appearance variation for compositional queries. Celebrities' natural variability enables both expressive and ambiguous attribute-conditioned retrieval; for example, "red jacket" may correspond to multiple valid images of one or several identities, better reflecting realistic search than clothing-template datasets [35]. We annotate images with open-vocabulary attributes across seven categories: appearance

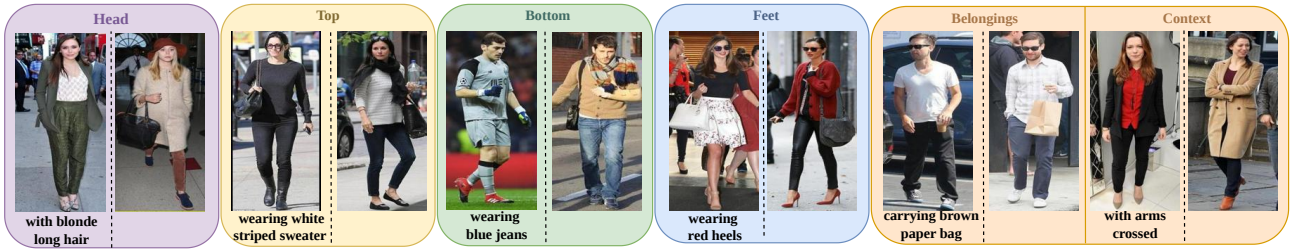


Figure 3. Single-attribute query examples (hard setting) from our dataset. Dashed line separates the image and condition text query (left) from its valid target image (right). Our dataset contains both single- and multi-target queries. More examples in the *Supplementary*

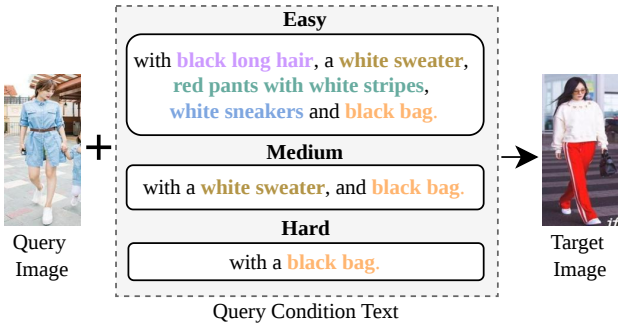


Figure 4. Proposed dataset settings. From structured attribute categories: head, top, bottom, feet, and other (accessories, belongings, context), we generate three text query settings: *Easy* (all attributes), *Medium* (2 attributes), and *Hard* (single attribute).

(head/hairstyle, top, bottom, feet), objects (accessories, belongings), and context (actions such as holding a phone or crossing arms). Context attributes capture conditions not tied to a single body part (e.g., crossed arms or hands in pockets), while body attributes remain grounded in PAR.

Annotation Pipeline. VLMs are increasingly used in recent Re-ID works [7, 21, 38, 62] for scalable image annotation. We use InternVL3-38B [82] and Qwen2.5-VL-34B [1], with structured prompts (see *Supplementary*). Both models independently annotate all images; agreement ($\sim$90% of samples) was automatically accepted, while disagreements ($\sim$10%) were manually audited adn reviewed. We found InternVL3 more robust and detailed in ambiguous cases and use it as default during disagreement resolution. A small subset of images ($\sim$100 per dataset) with empty or invalid outputs were fully labeled by human annotators.

Multi-Granularity Query Generation. We synthesize three levels of query granularity (Fig. 4) from structured attributes: (1) *Easy*, consisting of full sentences that combine all categories to minimize ambiguity; (2) *Medium*, consisting of phrases mentioning two categories (e.g.,"white t-shirt and sunglasses") that balance brevity and specificity; and (3) *Hard*, consisting of keywords mentioning a single attribute (e.g., "with black bag") to probe identity and attribute disentanglement under high ambiguity. The conditioning text c_t is derived from gallery attributes at the chosen difficulty. Easy queries typically yield a single target, whereas Hard queries often produce multiple matches in the diverse appearance space of Celeb-ReID-L (see *Suppl.*).

Extending Existing Benchmarks. While Celeb-ReID-Light contains rich clothing variation, it offers limited cross-camera diversity. We therefore use COCAS+Real2 [35] as a complementary dataset. Instruct-ReID [21] augmented this dataset with full GPT-4-generated image captions; we then use Qwen3-8B-Instruct [71] as a text-only parser to extract our predefined attribute categories directly from these captions. Unlike Celeb-ReID-L, which we annotate directly from images using VLMs, COCAS+Real2 is constructed entirely from caption parsing (details in *Suppl.*). Applying our Easy/Medium/Hard generation reveals that LI-ReID [21] degrades under CA-ReID's compositional queries. Celeb-ReID-Light and COCAS+Real2 thus serve as complementary testbeds with varying attributes and camera angles.

Vocabulary and Evaluation. Unlike COCAS+Real2's fixed-list captions, our free-form VLM annotations yield open-vocabulary descriptions ranging from generic ("black top") to fine-grained ("checked white blazer") phrases. We apply light post-processing through punctuation/filler cleanup, lexical normalization, synonym merging, and category reassignment (e.g., headwear and facewear are mapped to the head slot). Celeb-ReID-L's diversity naturally produces single- or multi-target queries, while COCAS+Real2's constrained space yields repeated phrases (Table 1). For evaluation, we precompute gallery embeddings and compute a distance matrix with composed query embeddings. Retrieval succeeds if any valid target appears at rank k, measuring joint identity-condition matching.

5. Experiments

5.1. Experimental Setup

We evaluate in two settings: (1) CA-ReID where queries are image and attribute text; matches must satisfy identity and the attribute, and (2) standard CC-ReID where only query image is used (identity-only retrieval).

Datasets. For CA-ReID, we build composed-attribute splits (see Sec. 4) on Celeb-ReID-Light [24] (10,842 im-

Attribute	Celeb-ReID-L		COCAS+Real2	
	Train	Test	Train	Test
head	447 / 12,014	254 / 6,829	68 / 8,063	28 / 876
top	2,948 / 11,345	572 / 7,987	143 / 9,086	75 / 1,884
bottom	909 / 7,087	201 / 5,191	38 / 10,262	14 / 324
feet	719 / 7,717	181 / 5,955	37 / 10,552	28 / 1,046
accessories	403 / 2,118	97 / 1,941	– / –	– / –
belongings	831 / 4,454	202 / 3,799	52 / 3,250	19 / 709
context	802 / 3,784	144 / 2,633	140 / 20,480	67 / 983

Table 1. CA-ReID attribute statistics. Each cell shows *unique / total* attribute instances in train and test splits. Celeb-ReID-L uses free-form VLM text (many unique attributes), whereas COCAS+Real2 relies on a fixed list (fewer unique but many repeated).

ages, 590 identities) and COCAS+ Real2 [35] (21,319 images, 101 identities). Each query consists of a query image, a conditioning text, and one or more valid target images, grouped into three difficulty levels (Easy / Medium / Hard, see Sec. 4) based on the number and complexity of attributes; Table 1 summarizes the resulting query and attribute statistics. For completeness, we also evaluated the standard CC-ReID setting on LTCC [51] (17,119 images, 152 identities), PRCC [61] (33,698 images, 221 identities).

Implementation Details. Following [38], we use EVA02-CLIP-L/14 as the visual encoder backbone and its paired text encoder. All input images are resized to 224×224. For pose estimation in PAR, we use HRNet-W32 [65] (75.8/74.9 AP on COCO [40] val/test) from MMPose [47]; different off-the-shelf estimators yield similar ReID performance [66]. All attribute descriptions are generated with InternVL3 and Qwen2.5-VL (details in the *Suppl.*). The batch size is set to 32 for all experiments. The EVA02-CLIP backbone base learning rate is set to 1×10^{-6} and weight decay of 5×10^{-4}, and all MLP heads use a learning rate of 3×10^{-4} and weight decay of 5×10^{-4}. We use SGD with a cosine-annealing scheduler. We apply a warm-up phase with an initial learning rate of 1×10^{-7} and train for 70 epochs on 2 NVIDIA A100 GPUs. We set the loss weights to λ_{ID}, λ_{DDL1}, and λ_{DDL2} to 1.0, 1.0, and 0.02, and the margin hyperparameters α_1, α_2, α_3 to 1.0, 1.0, and 1.5.

Evaluation Metrics. For CA-ReID, each query q is paired with a set of valid targets $\mathcal{T}(q)$, defined by *both* identity and attribute satisfaction. Following CIR-style evaluation [26], we report the Recall@K (R@K): a query is counted as correct if any $t \in \mathcal{T}(q)$ appears within the top-K matches. In addition, we report mAP to compare with conventional Re-ID setups. For CC-ReID benchmarks, we follow standard protocol [21, 38] and report identity-only CMC top-k ($k \in \{1, 5\}$) and mAP metrics.

5.2. Evaluation in CA-ReID Setting

Table 2 summarizes CA-ReID performance across Easy / Medium / Hard query settings. We compare against DIFFER [38], a strong image-only CC-ReID baseline, and the multi-modal model from Instruct-ReID [21]. Since DIFFER ignores the conditioning text, it cannot effectively control composite attributes in the CA-ReID setting. Therefore, its performance is poor for both datasets. Our method consistently outperforms Instruct-ReID across all difficulty levels and datasets. Instruct-ReID performs reasonably well for Easy / Medium queries with clearer descriptions, but degrades on Hard queries with more ambiguous composite attributes. Notably, on Hard queries, we observe gains of roughly +17% R@1 and +5.7% mAP on the Celeb-ReID-Light dataset, and +6.4% R@1 and +1.4% mAP on COCAS+Real2. These results indicate that Part-Aware Representations and Dense Disentangling Loss lead to more reliable grounding of short, compositional attribute edits.

Method	Query	Celeb-ReID-L			COCAS+ Real2		
		R@1	R@5	mAP	R@1	R@5	mAP
DIFFER [38]*	E	23.6	25.2	12.5	31.6	39.2	14.9
Inst-ReID [21]	E	81.8	95.6	20.8	82.7	93.5	36.4
	M	74.0	81.8	19.0	51.8	78.2	17.9
	H	41.6	71.0	14.7	44.0	67.9	19.9
CA-ReID (Ours)	E	**83.1**	**97.8**	**24.5**	**83.9**	**94.7**	**38.3**
	M	**78.9**	**86.2**	**23.3**	**55.1**	**82.4**	**21.2**
	H	**58.6**	**79.4**	**20.4**	**50.4**	**74.2**	**21.3**

Table 2. Results on the proposed CA-ReID setting. E/M/H denote Easy, Medium, and Hard query settings. *Image only ReID.

Category	R@1	R@5	mAP
head	59.4	79.3	18.4
top	62.4	85.2	23.5
bottom	60.3	83.2	21.9
feet	58.3	80.2	21.0
accessories	55.4	76.1	19.5
belongings	58.2	77.8	19.9
context	56.2	74.1	18.4

Table 3. Results by attribute category on Celeb-ReID-L (Hard).

Table 3 further breaks down performance for Hard queries by attribute sub-category (head, top, bottom, feet, accessories, belongings, context) on the Celeb-ReID-Light dataset. The results highlight varying levels of difficulty in localizing and disentangling different attribute types. Attributes with strong, localized visual grounding, such as top and bottom clothing, achieve the highest recall, likely due to their distinct appearance cues and lower ambiguity. In contrast, attributes that are diffuse, context-dependent, or weakly localized (e.g., accessories, belongings, and scene context) remain more challenging. The top category achieves the highest recall, potentially because many top-clothing attributes are more fine-grained and distinctive.

Losses			Celeb-ReID-L		LTCC	
$\mathcal{L}_{\text{ID}}$	$\mathcal{L}_{\text{DDL}_2}$	$\mathcal{L}_{\text{DDL}_1}$	R@1	mAP	Top1	mAP
✓			54.7	17.2	60.2	52.3
✓	✓		55.0	18.5	62.7	52.6
✓	✓	✓	**56.1**	**20.7**	**63.8**	**53.7**

Table 4. Ablation on loss terms. $\mathcal{L}_{\text{DDL}_1}$ is the identity-attribute disentanglement loss, $\mathcal{L}_{\text{DDL}_2}$ is the inter-part orthogonality loss.

5.3. Ablation Study

Table 4 shows the contribution of each loss term on both composed-attribute retrieval (Celeb-ReID-L, R@1/mAP) and standard clothes-changing Re-ID (LTCC, Top-1/mAP). The identity-only baseline ($\mathcal{L}_{\text{ID}}$) serves as a no-PAR setting, relying on global features. Adding $\mathcal{L}_{\text{DDL}_2}$, applied on PAR representations, provides consistent improvements, indicating that separating part features stabilizes the representation. The identity–attribute disentanglement loss $\mathcal{L}_{\text{DDL}_1}$ further improves performance across all metrics, yielding the best results. Overall, $\mathcal{L}_{\text{DDL}_2}$ and $\mathcal{L}_{\text{DDL}_1}$ are complementary: the former enforces structural separation across body parts, while the latter aligns identity and attribute factors to produce attribute-aware yet identity-preserving retrieval.

Figure 5. Sample Top-5 retrieval results for the hard setting from the Celeb-ReID-L dataset. Target image highlighted in green.

Method	Venue	LTCC [51]		PRCC [72]	
		Top1	mAP	Top1	mAP
TransReID [20]	CVPR'21	46.6	44.8	34.4	17.1
CAL [16]	CVPR'22	55.2	55.8	40.1	18.0
AIM [73]	CVPR'23	57.9	58.3	40.6	19.1
LDF [6]	ACM'23	58.4	58.6	32.9	15.4
3DInv [41]	ICCV'23	40.9	18.9	56.5	57.2
CCFA [18]	CVPR'23	45.3	22.1	61.2	**58.4**
CLIP3D [43]	CVPR'24	42.1	22.9	**61.8**	58.3
Inst-ReID [21]	CVPR'24	66.7	46.7	54.2	52.3
DIFFER [38]	CVPR'25	**68.5**	**64.7**	58.2	31.6
CA-ReID (Ours)	–	63.8	53.7	55.2	43.4

Table 5. Standard CC-ReID benchmark comparisons.

5.4. Qualitative Results

Fig. 5 shows example retrievals on Celeb-ReID-L (Hard). The first two rows are successful, the correct target image appears within the Top-5. Retrieved results typically satisfy the attribute or close variants (e.g., straw hat query retrieves hats), while identity-similar impostors cluster nearby. This indicates the composed representation enforces the attribute constraint strongly and largely preserves identity for common, well-localized cues. The last row is a failure: the model returns attribute-consistent distractors but misranks the true identity, likely due to the small/occluded region and class imbalance. Overall we observe an attribute-first bias, clear and frequent attributes help, whereas rare or fine-grained attributes lead to confusion.

5.5. Evaluation in CC-ReID Setting

Table 5 compares our method against leading CC-ReID methods on LTCC and PRCC. Although our model is designed for compositional retrieval rather than purely clothes-agnostic ReID, it remains competitive with specialized CC-ReID methods: on LTCC it surpasses many prior baselines [18, 41, 43] and is close to the best reported methods [21, 38], and on PRCC it approaches the performance of recent clothes-invariant methods [41] while outperforming [21]. This performance gap is expected, as our model is designed for attribute-conditioned retrieval, which we believe introduces an inductive bias toward attribute compliance.

6. Conclusion

This work presents *Composite-Attributes Person Re-ID* (CA-ReID), a setting that matches a reference identity to compositional text queries. We construct a dataset with varying ambiguity (Easy/Medium/Hard) and fine-grained attributes for real-world applications. We propose pose-guided Part-Aware Representations (PAR) and Dense Disentangling Loss (DDL) to align text with body regions and disentangle identity from part-level attributes. Experiments show consistent gains on CA-ReID and competitive performance on standard CC-ReID benchmarks.

References

[1] 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. 6

[2] 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. *arXiv preprint arXiv:2305.13653*, 2023. 1, 3

[3] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3952–3962, 2025. 3

[4] 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. 3

[5] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 465–473, 2024. 3

[6] Patrick PK Chan, Xiaoman Hu, Haorui Song, Peng Peng, and Keke Chen. Learning disentangled features for person re-identification under clothes changing. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–21, 2023. 3, 8

[7] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 6

[8] Jingyuan Chen et al. Image retrieval with text feedback using attribute-based compositionality. In *SIGIR*, 2020. 3

[9] Can Cui, Siteng Huang, Wenxuan Song, Pengxiang Ding, Min Zhang, and Donglin Wang. Profd: Prompt-guided feature disentangling for occluded person re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1583–1592, 2024. 3

[10] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Dkc: Differentiated knowledge consolidation for cloth-hybrid lifelong person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3573–3582, 2025. 3

[11] Wei Ding et al. Storyvisual: Using llms for compositional visual retrieval. In *CVPR*, 2023. 3

[12] 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. 3

[13] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Advances in neural information processing systems*, 32, 2019. 3

[14] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 3

[15] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 3

[16] 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. 3, 8

[17] Ruiyang Ha, Songyi Jiang, Bin Li, Bikang Pan, Yihang Zhu, Junjie Zhang, Xiatian Zhu, Shaogang Gong, and Jingya Wang. Multi-modal multi-platform person re-identification: Benchmark and method. *arXiv preprint arXiv:2503.17096*, 2025. 3

[18] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22066–22075, 2023. 3, 8

[19] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021. 3

[20] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 1, 3, 8

[21] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024. 1, 2, 3, 5, 6, 7, 8

[22] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9317–9326, 2019. 1, 3

[23] Xinyu Huang et al. Parameter-efficient compositional retrieval using llms. In *NeurIPS*, 2023. 3

[24] 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. 1, 5, 6

[25] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11895–11904, 2021. 3

[26] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav

Shrivastava. Collm: A large language model for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3994–4004, 2025. 7

[27] 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

[28] Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. Modeling thousands of human annotators for generalizable text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9220–9230, 2025. 3

[29] Xiaoshuai Jing et al. Personalized attribute-guided re-identification. In *CVPR*, 2023. 3

[30] Kimin Lee et al. Cosmo: Compositional modality learning for retrieval. In *NeurIPS*, 2021. 3

[31] Chao Li et al. Composing text and image for image retrieval with neural modules. In *CVPR*, 2020. 3

[32] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17459–17469, 2024. 3

[33] 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. 3

[34] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022. 3

[35] Shihua Li, Haobin Chen, Shijie Yu, Zhiqun He, Feng Zhu, Rui Zhao, Jie Chen, and Yu Qiao. Cocas+: Large-scale clothes-changing person re-identification with clothes templates. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1839–1853, 2022. 1, 5, 6, 7

[36] 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*, pages 1405–1413, 2023. 1, 3

[37] Wei Li et al. Diverse attribute-based person re-identification. In *ICCV*, 2021. 3

[38] Xin Liang and Yogesh S Rawat. Differ: Disentangling identity features via semantic cues for clothes-changing person re-id. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13980–13989, 2025. 1, 2, 3, 6, 7, 8

[39] Dixuan Lin, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Cross-modal adaptive dual association for text-to-image person retrieval. *IEEE Transactions on Multimedia*, 26:6609–6620, 2024. 3

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[41] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19617–19626, 2023. 3, 8

[42] Fangyi Liu, Mang Ye, and Bo Du. Dual level adaptive weighting for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 32:5075–5086, 2023. 3

[43] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 256–266, 2024. 3, 8

[44] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2080–2089, 2018. 3

[45] 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. 1, 3

[46] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019. 3

[47] OpenMMLab. Mmpose: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 7

[48] Jinwoo Park et al. Position-aware composed image retrieval with clip. In *CVPR*, 2024. 3

[49] Priyank Pathak and Yogesh S Rawat. Colors see colors ignore: Clothes changing reid with color disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16797–16807, 2025. 3

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[51] 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. 1, 3, 7, 8

[52] Yifan Qin, Yao Yang, Mengde Xu, Xiaoyi Tan, et al. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, 2024. RDE. 3

[53] Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14390–14399, 2025. 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

ing transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[55] Nyle Siddiqui, Florinel Alin Croitoru, Gaurav Kumar Nayak, Radu Tudor Ionescu, and Mubarak Shah. Dlcr: A generative data expansion framework via diffusion for clothes-changing person re-id. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1608–1617. IEEE, 2025. 3

[56] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *European Conference on Computer Vision*, pages 216–233. Springer, 2024. 3

[57] Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. A comprehensive survey on composed image retrieval. *arXiv preprint arXiv:2502.18495*, 2025. 3

[58] Yan Song et al. Clip for compositional image retrieval. In *ECCV*, 2022. 3

[59] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 4

[60] Xiaokun Sun, Qiao Feng, Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, and Kun Li. Learning semantic-aware disentangled representation for flexible 3d human body editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16994, 2023. 3

[61] 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). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 3, 7

[62] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17127–17137, 2024. 1, 3, 6

[63] Nam Vo et al. Composing text and image for image retrieval - an empirical odyssey. In *CVPR*, 2019. 3

[64] Nam Vo et al. Parameterizing compositionality in image retrieval. In *CVPR*, 2021. 3

[65] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 7

[66] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2540–2549, 2022. 3, 7

[67] 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. 1

[68] Yan Xie, Zequn Zeng, Hao Zhang, Yucheng Ding, Yi Wang, Zhengjue Wang, Bo Chen, and Hongwei Liu. Discovering fine-grained visual-concept relations by disentangled optimal transport concept bottleneck models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 30199–30209, 2025. 3

[69] Jincheng Yan, Yun Wang, Xiaoyan Luo, and Yu-Wing Tai. Fusionsegreid: Advancing person re-identification with multimodal retrieval and precise segmentation. *arXiv preprint arXiv:2503.21595*, 2025. 3

[70] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023. 3

[71] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6

[72] 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. 8

[73] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1472–1481, 2023. 3, 8

[74] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 3

[75] Wenlong Yu, Ruonan Liu, Dongyue Chen, and Qinghua Hu. Explainability enhanced object detection transformer with feature disentanglement. *IEEE Transactions on Image Processing*, 2024. 3

[76] Xia Yu et al. Composed image retrieval with large-scale pre-training. In *ICCV*, 2023. 3

[77] Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. From poses to identity: Training-free person re-identification via feature centralization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24409–24418, 2025. 1, 3

[78] Zhe Zhang et al. Differ: Disentangling identity features via semantic cues for clothes-changing person re-id. In *CVPR*, 2025. 3

[79] Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7534–7542, 2024. 3

[80] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 1, 3

[81] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 1, 3

[82] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

[83] 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. 3