Disentangling Identity from Clothing: A Semantically-Supervised Decoupling Framework for Robust Person Re-Identification

First Author
Institution1
Institution1 Address
firstauthor@i1.org

Second Author
Institution2
Institution2 Address
secondauthor@i2.org

Abstract

Text-to-Image Person Re-Identification is critically hampered by the difficulty of fine-grained semantic alignment, as retrieval accuracy is degraded by clothing-induced interference and a persistent modality gap. In this paper, we propose a novel framework to resolve this issue, centered on feature decoupling guided by a Multimodal Large Language Model (MLLM). Our framework introduces two core components: a Bidirectional Decoupled Alignment Module and a Mamba State Space Model (SSM) for efficient fusion. To obtain high-quality, fine-grained supervision, we first employ MLLM to automatically generate separate identity and clothing descriptions. These descriptions then guide our decoupling module, which utilizes bidirectional attention and a gated weighting strategy to meticulously disentangle visual features into identity and clothing subspaces. To enforce this separation and ensure identity purity, we design a multi-task loss strategy comprising an alignment loss that actively suppresses the influence of clothing-related features, and a kernel-based orthogonal constraint that ensures statistical independence. Furthermore, we pioneer the integration of the Mamba SSM into cross-modal Re-ID as an efficient fusion module. By leveraging its linear-time complexity and proficiency in modeling long-range dependencies, it facilitates deep contextual interactions across modalities while avoiding the quadratic complexity of Transformers. Comprehensive experiments on multiple benchmark datasets reveal that our proposed method achieves superior performance compared to leading contemporary methods, proving its effectiveness and robustness.

Keywords: Multimodal Learning, Text-to-Image Re-Identification, Feature Decoupling, Semantic Supervision.
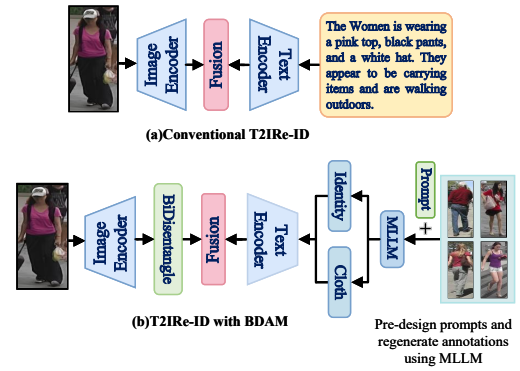

Figure 1. Comparison of ReID methods.(a) Traditional: Direct fusion of image and text features without distinguishing identity from non-identity information limits alignment.(b) Proposed (BDAM): Introduces a decoupling module that aligns separated identity/clothing features with MLLM-generated descriptions for fine-grained matching.

1. Introduction

Text-to-Image Person Re-Identificatio (T2I-ReID) retrieves a target pedestrian from a large-scale image gallery given a natural-language description [7, 15, 31, 40]. It is valuable for video surveillance [3], intelligent security [10], public safety, and social media. Despite recent progress, practical deployment remains challenging: image factors (pose, viewpoint, illumination) obscure identity-relevant cues, and a persistent modality gap hampers effective fusion in a shared space. These issues are exacerbated at the fine-grained level, where semantic alignment is particularly difficult.

A core challenge in T2I-ReID is the semantic gap between images and text. Early work attempted to reduce this discrepancy by projecting global visual and textual features

into a shared space [19, 34, 36], but high intra-class variance and low inter-class variance across both modalities hinder reliable cross-modal matching. To overcome this, subsequent studies introduced feature disentanglement, broadly via explicit or implicit alignment [12, 38]. Explicit methods [7, 37] detect body parts or attributes and align local regions and phrases with auxiliary modules. Implicit methods [15, 31] avoid external tools and use regularizers to associate noun phrases with image regions. This progression demonstrates that distinguishing identity-relevant from identity-irrelevant semantics is essential; therefore, disentanglement has become a key avenue to advance T2I-ReID.

This pursuit of disentanglement has been propelled by powerful backbones. Models employing ViT [8, 14, 21, 33, 42, 47] excel at capturing fine-grained visual details, while methods leveraging CLIP [26, 28, 43] utilize large-scale pretraining to learn a well-aligned joint embedding space. These architectures, often paired with cross-attention, have significantly advanced the state-of-the-art in cross-modal alignment. Despite this progress, a critical limitation persists. Both lines of work commonly treat the textual description holistically. This approach overlooks the semantic distinction between content relevant to identity (e.g., gender, body shape) and content irrelevant to identity (e.g., clothing, hairstyle). In complex scenes, this coarse-grained treatment forces the model to entangle these factors, often prioritizing salient but non-essential clothing details over stable identity cues. This ambiguity blurs identity features, weakens the decoupling process, and ultimately degrades matching robustness.

To address this challenge, we propose a novel framework that moves beyond such holistic feature treatment by introducing explicit, fine-grained semantic supervision. Inspired by the style-clustering paradigm [16], our approach employs MLLM to guide feature decoupling. This is achieved by automatically generating fine-grained, distinct, and mutually exclusive descriptions for both identity and clothing. These decoupled annotations provide the precise supervision for our BDAM to meticulously separate and align identity and clothing information. This disentanglement is enforced by a multi-task loss strategy, including an alignment loss and an orthogonality constraint based on HSIC, as detailed in our Sec. 3. Furthermore, to enhance alignment, we pioneer the integration of the Mamba SSM as an efficient fusion module, adept at capturing long-range cross-modal dependencies with linear complexity.

Our main contributions are threefold: (1) An automatic prompt construction pipeline that combines style clustering with an MLLM to produce fine-grained, decoupled identity and clothing descriptions; (2) The BDAM, which achieves precise decoupling and alignment reinforced by a multi-task

loss strategy combining an alignment loss and an orthogonality constraint based on HSIC; and (3) The novel integration of a Mamba SSM fusion module that models long-range cross-modal dependencies with linear complexity.

2. Related Work

2.1. Feature Disentanglement

Feature disentanglement [35] aims to separate semantically distinct factors in feature space to improve interpret ability and generalization. Early approaches often leveraged generative models (VAEs, GANs)[23] to partition latent codes into structured factors. More recently, disentanglement has expanded beyond generation to image classification[29], NLP [5], and multimodal learning [25]. Mainstream strategies include minimax multi-task adversarial training that jointly optimizes an encoder and a style discriminator [23]; schemes based on metric learning such as DFR [4] with a Gradient Reversal Layer [11] to decorrelate factors; and orthogonal linear projections that separate visual and textual embeddings under CLIP [25].

In the ReID domain, disentanglement typically separates identity-relevant signals from nuisances [1, 22]. Generative designs model these factors independently. In vehicle Re-ID, DFLNet[2] jointly extracts orientation-specific and generic features. For occlusion, ProFD [6] uses text prompts to isolate body-part features. Re-ID focused on clothing changes often adopts dual-stream architectures to counter appearance shifts and camera bias [22].

While these efforts improve semantic purity and factor independence, two critical limitations persist. First, without an effective interaction mechanism, isolated factors may fail to support robust cross-modal matching, leading to brittle alignment. Second, reliance on manual annotations or external detectors constrains scalability and domain transfer. Furthermore, implicit regularizers can be underconstrained, yielding spurious separations on unseen data. Consequently, recent work emphasizes coupling disentanglement with principled interactions and independence constraints. In this spirit, our framework pairs decoupling at the data level with model-level disentanglement and independence enforcement, providing explicit supervision and controllable separation while preserving cross-modal synergy.

2.2. Feature Fusion

Feature fusion is central to T2I-ReID, with most methods relying on Transformers or CLIP [30]. Cross-modal modules built on multi-head attention process image and text tokens in parallel to capture semantic associations [44], but their quadratic complexity in sequence length causes mem-

ory and latency spikes for high-resolution images or long descriptions. Pipelines built upon CLIP [17] benefit from large-scale contrastive pretraining and well-aligned embeddings, yet global pooling and holistic processing often blur identity versus clothing cues. This leads to semantic confusion under clothing changes or verbose descriptions. Dynamic fusion that reweights modalities via attention can improve adaptivity [9], but introduces higher computational cost, tuning sensitivity, and performance instability across datasets. Graph-based fusion models [18] dependencies through GNNs, offering relational inductive bias. However, assumptions about graph structure and stationarity limit adaptability to free-form text and dynamic visual contexts.

Contemporary evidence suggests that accuracy and robustness improve only when semantic disentanglement and efficient fusion advance in tandem. Practically, an ideal fusion module should respect factorized semantics, such as identity and clothing, to avoid re-coupling nuisances. It must also capture long-range cross-modal dependencies and scale with linear or near-linear complexity to handle long sequences and high-resolution tokens. This motivates our design: BDAM supplies factor-aware representations and decoupled supervision, while a Mamba SSM fusion module models long-horizon interactions with linear complexity, enabling precise alignment without the memory and efficiency bottlenecks of standard Transformers.

3. Method

3.1. Overview

To learn pedestrian representations robust to variations in clothing, pose, and environment, this paper proposes the BDAM. This module disentangles and extracts robust features via contrastive and supervised learning, guided by encoded textual features. As illustrated in Fig. 2, our framework comprises two primary feature extraction modules for vision and text, our core BDAM, and an efficient Mamba SSM Fusion Module.

Specifically, given a pedestrian image $I \in \mathbb{R}^{B \times C \times H \times W}$, a visual encoder first extracts image features f_i. To obtain semantic guidance, we use an MLLM with pre-designed prompts to generate corresponding descriptions for identity and clothing. A text encoder subsequently encodes these into f_{id}^t and f_{clo}^t. During disentanglement, BDAM leverages these textual features to guide the image feature learning process. To ensure the quality of this separation, we introduce a loss based on HSIC to constrain the two resulting feature types towards orthogonality. We also employ an alignment clothing loss, denoted as $\mathcal{L}_{\text{aln}}$, to supervise the learning of visual clothing features using clothing descriptions. Finally, to achieve a deep fusion of visual identity and

textual semantics, we introduce the Mamba SSM as a fusion module. It dynamically models and facilitates interaction among the disentangled multimodal features, enhancing the model's overall representation capability.

3.2. Bidirectional Decoupled Alignment Module

Some studies directly adopt CLIP [28] as a feature extractor for both modalities, aligning global embeddings for retrieval or discrimination. However, this approach presents two key limitations. First, its limited capacity for fine-grained semantics hinders the separation of identity from clothing. Second, its holistic encoding of images and text lacks the modeling of cross-modal structure at the token level, which reduces robustness in complex scenes.

In this paper, we use a pre-trained ViT (ViT-B/16) as the visual encoder E_v [8]. Given an image I_i, E_v outputs token features $f_i \in \mathbb{R}^{B \times L \times D}$ that entangle cues relevant to identity and cues irrelevant to identity. A linear projection with two branches then yields preliminary identity features $f'_{\text{id}} \in \mathbb{R}^{B \times L \times D}$ and clothing features $f'_{\text{clo}} \in \mathbb{R}^{B \times L \times D}$. These are followed by multi-layer self-attention in each branch to enhance local consistency and contextual awareness.

Instead of using the ViT [CLS] token as a global descriptor, we exploit the full patch sequence and introduce cross-attention between the branches to exchange information. In the identity stream, the clothing stream provides auxiliary context, and vice versa, reinforcing semantic distinctions. Each stream then applies global average pooling to produce $\hat{f}_{\text{id}}$ and $\hat{f}_{\text{clo}}$. To enable soft disentanglement that is adaptive to the input, we design a gating mechanism. The two global vectors are concatenated and fed to a lightweight linear network with a Sigmoid output, producing a gate $g \in \mathbb{R}^{B \times D}$. We obtain the final gated features $f_{id}^i = g \odot \hat{f}_{id}$ and $f_{clo}^i = (1 - g) \odot \hat{f}_{clo}$, where $\odot$ denotes element-wise multiplication. This weighting at the dimension level provides a fine degree of control; f_{id}^i is further sent to the fusion module.

To train BDAM and enforce separation, we introduce two specialized loss functions. The first is a **clothing alignment loss**, and the second is a decoupling loss based on HSIC to enforce independence between identity and clothing. The clothing alignment loss supervises the visual clothing features with the MLLM-generated clothing descriptions, ensuring that the model accurately captures clothing semantics:

$$\mathcal{L}_{\text{aln}} = -\mathbb{E}_i \left[\log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ij}/\tau)} \right] \quad (1)$$

where $s_{ij} = \hat{f}_{\text{clo}}^i \cdot (f_{\text{clo}}^t)^j$ is the dot-product similarity between the visual clothing feature of sample i and the textual
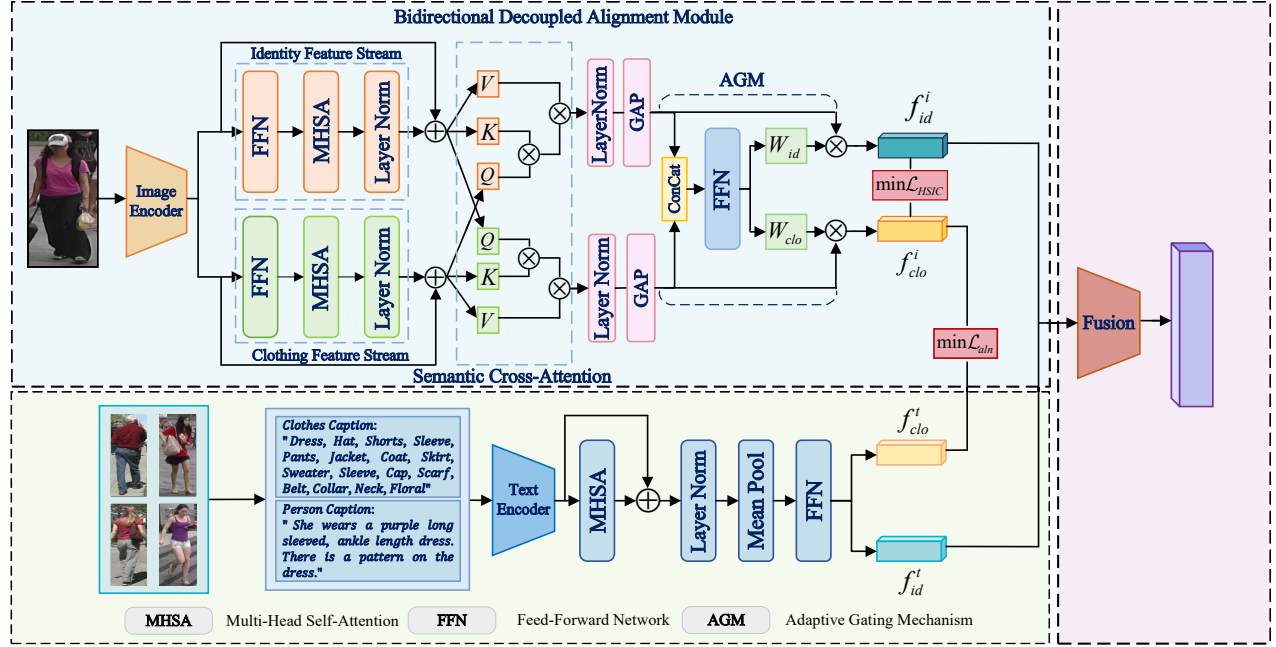
Figure 2. First, we use an MLLM with predefined prompts to generate identity-related and clothing-related descriptions from the pedestrian image, which are encoded as f_{clo}^t and f_{id}^t, respectively. Then, the input pedestrian image is encoded into a visual feature f_i, which still resides in an entangled feature space. Subsequently, the BDAM module, composed of a dual-branch attention mechanism, decouples f_i into identity feature f_{id} and non-identity feature f_{clo}. The decoupling process is supervised and optimized through disentanglement loss and corresponding textual descriptions via contrastive learning. Finally, the fusion module integrates f_{id} and f_{id}^t to generate the final fused feature representation.

clothing feature of sample j, and τ is a temperature parameter. This formulation encourages high similarity (s_{ii}) for positive pairs (same sample) and low similarity ($s_{ij}, i \neq j$) for negative pairs (different samples). In practice, clothing features are linearly projected to the text dimension and normalized using L2 for stable similarity estimation. This alignment objective explicitly ensures the clothing stream learns accurate representations under semantic supervision, which indirectly enhances the purity of the identity features by providing clear guidance on what constitutes clothing information. This works in conjunction with the cross-attention mechanism described earlier, which implicitly sharpens the separation via interaction.

To further encourage statistical independence, we minimize a decoupling loss based on HSIC:

$$
\begin{aligned}
\mathcal{L}_{\text{Decouple}} &= \text{HSIC}(f_{id}^i, f_{clo}^i) \\
&= \frac{1}{(N-1)^2} \text{tr}(K_{id} H K_{clo} H)
\end{aligned}
\tag{2}
$$

Here, $f_{id}^i \in \mathbb{R}^{B \times D}$ and $f_{clo}^i \in \mathbb{R}^{B \times D}$ are the gated identity and clothing features, respectively. $K_{id} = f_{id}^i (f_{id}^i)^T$ and $K_{clo} = f_{clo}^i (f_{clo}^i)^T$ are their respective kernel matrices. $H = $

$I_N - (1/N)\mathbf{1}_N \mathbf{1}_N^T$ is the centering matrix, where I_N is the N-dimensional identity matrix and $\mathbf{1}_N$ is a column vector of all ones. HSIC measures the statistical dependence between features by calculating the mean trace of the product of their kernel matrices and the centering matrix. By minimizing this value, the loss encourages the features to be statistically independent.

3.3. Semantic enhancement

We employ an MLLM, as detailed in Section 4.2, to automatically generate fine-grained identity and clothing descriptions for pedestrians. This approach reduces the burden of manual annotation and enriches the available supervision. Figure 3 illustrates this generation pipeline. Prior work, notably HAM [16], shows that modeling annotator styles can steer an MLLM to produce diverse texts. Adapting this core insight, we extend the pipeline to meet our model's design goals.

We first use the CLIP text encoder to embed the original descriptions into vectors of a fixed dimension. Using prompts, an MLLM generalizes and substitutes entity attributes to emphasize expression style rather than content. We then
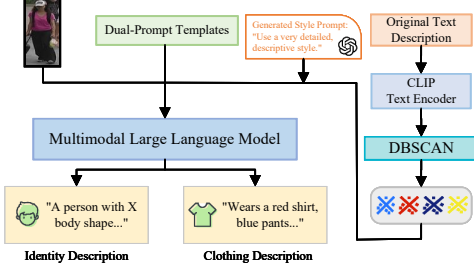
Figure 3. Overview of the offline pipeline using an MLLM to generate decoupled identity and clothing descriptions. Style prompts are derived via CLIP and DBSCAN to enhance diversity.

cluster these style embeddings with DBSCAN [13], which adaptively discovers dense regions without predefining the cluster count. To stabilize the clusters, we reassign noise points and merge small clusters. These discovered style categories are then used to formulate textual prompts, such as "Use a very detailed, descriptive style," which guide the MLLM's generation tone. This setup aligns the learned style categories with the identity and clothing disentanglement expected by BDAM.

A dual prompt generator, using content-specific templates (e.g., "Describe the person's identity" and "Describe the person's clothing"), guides the MLLM to output two distinct texts per image: one description for identity, covering biological traits, and another for clothing, detailing apparel, colors, and patterns. We control the generation process with length and temperature constraints. We also apply syntax checks and validation for attribute coverage to ensure the outputs remain grammatical, structured, and parsable.

3.4. Feature Fusion

For efficient feature fusion that is sensitive to semantics, we introduce the Mamba SSM. The core objective is to preserve the semantic integrity of the purified identity features from both image and text, enabling a fusion that is robust to clothing variations previously isolated by the BDAM. The process begins with an FFN performing dimensional alignment to mitigate distributional discrepancies between modalities. It processes the decoupled visual features f_{id}^i and the textual features f_{id}^t to generate aligned features, f_{img} and f_{txt}. Notably, the decoupled visual clothing feature f_{clo}^i is *intentionally discarded* during fusion. This design is central to our goal: the BDAM, supervised by $\mathcal{L}_{aln}$ and $\mathcal{L}_{Decouple}$, is tasked with purging information irrelevant to identity into f_{clo}^i. By excluding this feature from the final fusion, the

model is forced to learn a representation based purely on stable identity semantics.

Following this alignment, a gating mechanism achieves dynamic weighted fusion. It outputs a weight vector $g_{fus} \in \mathbb{R}^{B \times 2}$, which is normalized via a SoftMax layer to produce image W_{img} and text W_{txt} weights, satisfying $W_{img} + W_{txt} = 1$. The resulting fusion is computed as: $f_{fusion} = W_{img} \cdot f_{img} + W_{txt} \cdot f_{txt}$. This mechanism allows the model to adaptively balance modal contributions based on context. This fusion gate is distinct from the one in the disentanglement module; it outputs a global, two-dimensional weight vector $g_{fus} \in \mathbb{R}^{B \times 2}$ for the modalities, whereas the disentanglement gate provides a vector $g \in \mathbb{R}^{B \times D}$ for feature control at the dimension level.

The resulting f_{fusion} features are then processed by the Mamba SSM to enhance interaction between modalities. Leveraging its capability to model dependencies over long ranges, Mamba effectively captures complex sequential relationships. We employ a stack of Mamba layers, where each layer updates its input $f_{fusion}^{(l)}$ using a residual connection: $f_{fusion}^{(l+1)} = \text{Mamba}(f_{fusion}^{(l)}) + f_{fusion}^{(l)}$. This structure mitigates the vanishing gradient problem and improves information flow. Finally, the output from the last Mamba layer is projected to produce the final representation, $f_{final} \in \mathbb{R}^{B \times D_{out}}$. The resulting feature is highly adaptive in its modal weighting and benefits from Mamba's semantic modeling, providing robust support for downstream tasks like person re-identification.

3.5. Loss Function

To achieve alignment between modalities at a fine-grained level, we adopt the InfoNCE loss. This loss maximizes similarity for positive image and text pairs (representing the same identity) while separating negatives. It is defined as:

$$\mathcal{L}_{info} = -\log \frac{\exp(v_i^\top t_i / \tau)}{\sum_j \exp(v_i^\top t_j / \tau)} \qquad (3)$$

Here, v_i is the final fused representation, normalized using the L2 norm; t_i is the text feature for the i-th identity; and τ controls distribution sharpness. Negatives within the batch help reduce the semantic gap between modalities and promote alignment in a shared space.

To enhance identity discrimination within a single modality, we include a triplet loss. This loss enforces compactness within classes and separation between classes:

$$\mathcal{L}_{triplet} = \mathbb{E}(a, p, n) \left[\max \left(|f_a - f_p|_2^2 - |f_a - f_n|_2^2 + m, 0 \right) \right] \qquad (4)$$

Here, f_a, f_p, f_n decoupled visual identity features of the anchor, positive, and negative samples, respectively; $|| \cdot ||_2$ denotes the L2 norm; and m is the margin parameter used

to enforce a minimum distance gap between positive and negative pairs.

In training with multiple tasks, differing loss scales can cause one task to dominate. We adopt GradNorm to balance the training process by dynamically adjusting the gradient norm of each task:

$$\mathcal{L}_{\text{GradNorm}} = \sum_k |\nabla\theta(w_k\mathcal{L}_k) - \tilde{r}kG_{\text{ref}}| \quad (5)$$

Here, $\nabla_\theta(w_k\mathcal{L}k)$ represents the gradient of the weighted loss of task k, $w_k\mathcal{L}_k$, with respect to the shared parameters θ. The term $||\cdot||_1$ denotes the L1 norm, which emphasizes a linear penalty on the deviation. $\tilde{r}_k = (\mathcal{L}_k/\mathcal{L}_k^0)/\bar{r}$ is the normalized loss ratio, where $\mathcal{L}_k^0$ is the initial loss of task k at the start of training, serving as a baseline, and $\bar{r}$ is the average of the loss ratios over all tasks. G_{ref} is a reference gradient norm, typically set to the gradient norm of the first task, $||\nabla\theta(w_1\mathcal{L}1)||$.

This mechanism enforces a gradient balance among tasks during training by minimizing the L1 deviation between the actual gradient norm and a target value, $\tilde{r}_kG_{ref}$. Furthermore, to prevent instability caused by the abnormal scaling of task weights w_k, we add a regularization term, $\lambda\sum_k(\log w_k)^2$, to the total loss. This term effectively suppresses large discrepancies among the task weights by penalizing the square of their logarithms, thereby improving training stability. Finally, the overall loss function is defined as follows:

$$\mathcal{L}_{\text{Total}} = \sum_k w_k\mathcal{L}_k + \alpha\mathcal{L}_{\text{GradNorm}} + \lambda\sum_k(\log w_k)^2 \quad (6)$$

Here, $\mathcal{L}_k$ represents the loss term for task k (which includes $\mathcal{L}_{\text{info}}$, $\mathcal{L}_{\text{triplet}}$, $\mathcal{L}_{\text{aln}}$, and $\mathcal{L}_{\text{Decouple}}$), and $w_k = \exp(s_k)$ is the task weight, where s_k is a learnable parameter initialized to zero and optimized during training to capture task-specific uncertainty. The hyperparameter α is used to control the strength of the GradNorm loss, while λ serves as the regularization coefficient.

4. Experiments

4.1. Datasets and Metrics

CUHK-PEDES [20] is the first benchmark for text-to-image person re-identification, comprising 40,206 images of 13,003 identities with two textual descriptions per image. Following the official protocol, we partition the dataset into training (11,003 identities), validation (1,000 identities), and testing (1,000 identities) sets.

ICFG-PEDES [46] consists of 54,552 images across 4,102 identities, each annotated with a single human-written description. The official split allocates 3,102 identities for training and 1,000 for testing.

RSTPReid [7] captures 20,505 images of 4,104 identities from 15 cameras, with five viewpoint-diverse images and two descriptions per identity. We adopt the standard split: 3,701 identities for training and 200 each for validation and testing.

Evaluation Metrics. We report Rank-k accuracy (R-1, R-5, R-10) and mean Average Precision (mAP), consistent with prior works in this domain.

4.2. Implementation Details

Our model employs the `bert-base-uncased` model, which is pre-trained, as the text encoder with a hidden dimensionality of 768. We use `vit-base-patch16-224` as the visual encoder. All images are resized to 224×224 pixels. The BDAM module processes visual tokens through multiple layers of attention, disentangling them into identity and clothing features of 768 dimensions each. A Mamba fusion module with two layers, configured with an input and output dimension of 256, a state dimension of 16, and a convolutional kernel size of 4, integrates the multimodal representations. The dropout rate is set to 0.1 throughout.

Training utilizes the Adam optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-3}, and cosine annealing scheduling. The loss function for multiple tasks combines InfoNCE, triplet, clothing alignment, HSIC decoupling, and gate regularization terms. We employ GradNorm for dynamic task weighting with $\alpha = 1.5$ and a weight learning rate of 0.025. Task weights are normalized to sum to the number of tasks and clipped to the range $[10^{-4}, 10]$. Gradient norms are computed on the final shared layer, with an additional regularization coefficient for log variance of 1×10^{-3}.

For data augmentation, we extract style embeddings using `clip-vit-base-patch32`, cluster them via DBSCAN, and generate descriptions consistent with the style for identity and clothing through ChatGPT-4. All splits are enforced at the identity level to prevent data leakage. Each experiment is repeated with random seeds 0, 1, and 2; we select the best checkpoint based on validation performance and report the mean across the three runs.

4.3. Parameter Analysis

Although the identity and clothing stream in BDAM adopt identical architectures, they learn distinctly different representations through independent parameterization and asymmetric supervision. The BDAM gating mechanism generates weights at the dimension level $g_{\text{dis}} \in \mathbb{R}^{B \times D}$ via $g_{\text{dis}} = \sigma(W_g[\hat{f}_{id}; \hat{f}_{clo}] + b_g)$. This yields complementary coefficients $W_{id} = g_{\text{dis}}$ and $W_{clo} = 1 - g_{\text{dis}}$ that satisfy the constraint $W_{id} + W_{clo} = \mathbf{1}$. This constraint introduces competitive pressure between stream, encouraging each to

specialize in unique semantic patterns.

The learning dynamics are governed by asymmetric loss signals. The identity stream receives supervision from InfoNCE and triplet losses, steering it toward attributes that are invariant to the person, such as body shape and posture. In contrast, the clothing stream is guided by matching and alignment losses toward features centered on appearance. The decoupling loss based on HSIC enforces statistical independence by orthogonalizing the two feature subspaces in a Reproducing Kernel Hilbert Space, thereby preventing representational collapse.

Similarly, the fusion module employs an adaptive gating mechanism to balance modality contributions. The fusion gate $g_{\text{fus}} \in \mathbb{R}^{B \times 2}$ is computed as $g_{\text{fus}} = \text{Softmax}(W_f[f_{\text{img}}; f_{\text{txt}}])$, producing normalized modality weights $W_{\text{img}} = g_{\text{fus}}[:, 0]$ and $W_{\text{txt}} = g_{\text{fus}}[:, 1]$ satisfying $W_{\text{img}} + W_{\text{txt}} = 1$. This mechanism operates at the instance level, dynamically adjusting modality importance based on input characteristics, whereas the disentanglement gate provides control at the dimension level within the visual stream. To empirically validate these mechanisms, we analyze the
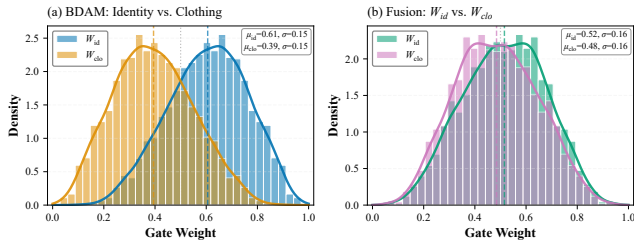

Figure 4. Learned gate weight distributions on the CUHK-PEDES test set. (a) BDAM assigns significantly higher weights to identity features than clothing features, validating the identity-centric design. (b) Fusion module maintains balanced modality contributions, confirming effective cross-modal integration without modality dominance.

distributions of learned gate weights across the test set, as shown in Fig. 4. The BDAM identity weights exhibit a distribution skewed to the right, with a mean of 0.61 and standard deviation of 0.17, substantially exceeding the clothing weights which concentrate around 0.38. This asymmetry confirms that features relevant to identity dominate the final representation, aligning with our design centered on identity. The bimodal separation pattern indicates the gate has learned to discriminate between samples with strong identity cues versus those requiring additional contextual information. In contrast, the fusion module maintains nearly equal modality contributions, with means of 0.52 and 0.48, both with moderate variance around 0.10. This balanced distribution, centered near the uniform baseline of 0.5, demonstrates stable alignment between modalities without collapse toward

either one.

These empirical observations corroborate our theoretical design: BDAM enforces semantic disentanglement through asymmetric supervision and regularization via HSIC, while the fusion gate achieves dynamic equilibrium via Softmax normalization and balanced loss weighting.

4.4. Ablation Study

We conduct systematic ablation studies to validate the contribution of each architectural component and loss term. All experiments are performed on CUHK-PEDES unless otherwise specified.

Table 1. Ablation study on the BDAM module.

Method	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Baseline(w/o BDAM)	59.81	70.54	85.49	91.26
+ BDAM	66.74	76.27	89.30	94.02
w/o Cross-Attn	62.56	71.39	87.05	92.98
w/o Gate	65.11	74.63	88.77	93.56
Shallow(3-layer)	64.27	73.74	88.09	93.32

Disentanglement Module. As shown in Tab. 1, incorporating BDAM yields substantial improvements over the baseline, demonstrating the effectiveness of explicit disentanglement between identity and clothing. The subsequent analysis of its components confirms the necessity of each design choice. Removing the bidirectional cross attention, ablating the gating mechanism, or reducing the network depth to three layers all result in significant performance degradation. This indicates that the semantic interaction between stream, the adaptive feature control, and sufficient model depth are all essential for achieving robust disentanglement.
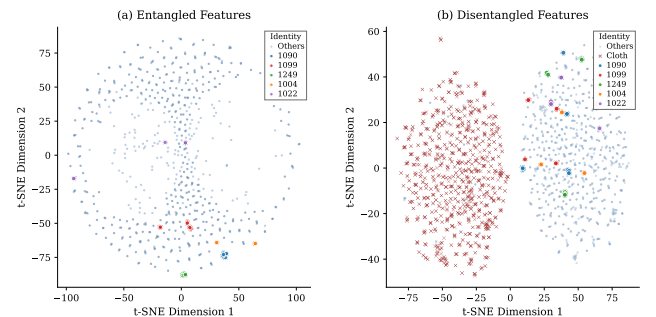

Figure 5. Visualization of feature distributions using t-SNE. (a) The baseline model exhibits highly entangled features with poor identity separation. (b) Our method (BDAM with HSIC) yields a highly organized space, clearly decoupling identity (dots) from clothing (crosses) features.

To provide a qualitative assessment of disentanglement quality, we visualize the learned feature space using dimensionality reduction via t-SNE in Fig. 5. The visualization reveals

a stark contrast. The baseline model (a) produces a highly mixed embedding space where identities overlap, indicating severe entanglement between identity and appearance attributes. (b) Conversely, our method, incorporating BDAM with HSIC regularization, yields a highly organized feature space. The identity embeddings form tight clusters, exhibiting strong cohesion for samples of the same person and clear separation between different identities. Concurrently, clothing features are projected to an independent region. This geometric structure confirms the orthogonality enforced by our disentanglement mechanism and provides visual evidence for the quantitative performance gains observed in Tab. 1.

Table 2. Ablation study on the fusion module.

Method	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Baseline(w/o Fusion)	59.81	70.54	85.49	91.26
Full Fusion	69.58	78.42	90.74	95.11
w/o Mamba	66.89	75.73	89.06	93.92
w/o Gate	68.64	77.58	90.11	94.87
w/o Alignment	68.15	77.09	89.84	94.53

Fusion Module. The analysis in Tab. 2 confirms that the effectiveness of our complete fusion module, which substantially outperforms the baseline, stems from the synergy of its three components. The Mamba SSM backbone is identified as the most critical element; its removal incurs the largest performance drop, highlighting the importance of modeling dependencies over long ranges across modalities. Furthermore, the gating mechanism and modality alignment layers provide essential complementary benefits. Ablating either component results in a clear degradation, confirming their respective roles in enabling adaptive weighting and achieving distributional alignment.

Table 3. Ablation study on individual loss components.

Method	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Full Model	72.61	79.93	92.95	96.47
w/o InfoNCE	28.14	36.55	55.21	65.83
w/o Triplet	67.22	74.89	88.15	93.12
w/o Alignment	69.15	76.92	89.53	94.22
w/o Decoupling	70.03	77.81	90.11	94.98
w/o Gate Reg.	71.98	79.23	91.35	95.71

Loss Functions. To validate the necessity of each loss component, we systematically remove individual terms while keeping all other settings fixed. As shown in Tab. 3, every loss term contributes positively to the final performance. The InfoNCE and triplet losses emerge as foundational. Removing InfoNCE causes a catastrophic performance collapse, underscoring the critical importance of contrastive alignment between modalities. Ablating the triplet loss also significantly degrades retrieval accuracy, confirming the necessity of strong identity discrimination within a modality.

The losses specific to disentanglement, namely the alignment and HSIC decoupling losses, are likewise essential. The performance degradation upon removing either term demonstrates that architectural design alone is insufficient; explicit constraints imposed by the loss function are required to enforce the separation of identity and clothing. Finally, the minor but consistent degradation from removing gate regularization validates its auxiliary role in stabilizing the learned gating mechanisms.

Table 4. Ablation study on clustering strategies for style prompt generation.

Method	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Baseline (w/o Style)	71.22	79.18	91.51	95.79
+ Random	71.95	79.54	91.66	95.82
+ K-Means	71.85	79.20	91.60	95.90
+ GMM	71.70	79.05	91.50	95.75
+ Agglomerative	71.78	79.10	91.55	95.80
+ HDBSCAN	72.20	79.40	91.95	96.10
+ DBSCAN (Ours)	72.61	79.93	92.95	96.47

Style Prompt Generation. We compare various clustering algorithms for generating prompts consistent with style, as detailed in Tab. 4. While random sampling provides minimal gains over the baseline, demonstrating the importance of a structured prompt design, traditional methods such as K-Means, Gaussian Mixture Models, and Hierarchical Agglomerative clustering offer only modest improvements. These methods are constrained by limitations such as the manual specification of cluster counts and difficulty handling irregular density distributions. In contrast, approaches based on density, particularly DBSCAN, achieve superior performance. This is because DBSCAN automatically discovers clusters of arbitrary shape while identifying and filtering noise. DBSCAN's ability to adaptively determine the number of style categories without requiring hyperparameter tuning makes it the optimal choice for this task.

4.5. Efficiency Analysis

As shown in Fig. 6, our fusion module employing Mamba achieves an optimal balance in the trade-off between accuracy and efficiency. The visualization reveals that Mamba maintains performance near the peak across all dimensions. Compared to the four-layer Transformer baseline, our method delivers competitive retrieval accuracy (reflected in comparable mAP scores) while achieving substantially higher efficiency scores in Params, FLOPs, and Memory, which confirms lower computational overhead. Against the Simple baseline, Mamba demonstrates marked mAP improvements while maintaining comparable efficiency scores, validating superior resource utilization.

Table 5. Performance comparison with state-of-the-art methods on three benchmark datasets.

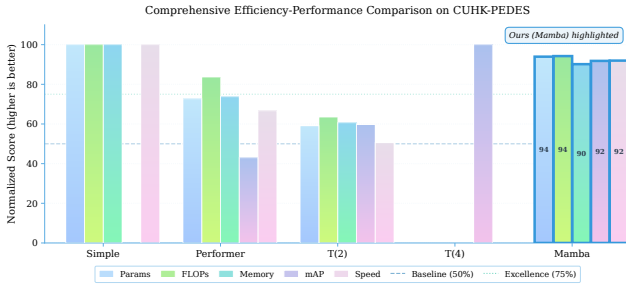| Method	Backbone	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Methods with CLIP backbone:													
IRRA [15]	CLIP-ViT	73.38	89.93	93.71	66.10	63.36	80.82	85.82	38.06	60.20	81.30	88.20	47.17
IRLT [24]	CLIP-ViT	73.67	89.71	93.57	65.94	63.57	80.57	86.32	38.34	60.51	82.85	89.71	47.64
CFAM [48]	CLIP-ViT	74.46	90.19	94.01	-	64.72	81.35	86.31	-	61.49	82.26	89.23	-
Propot [41]	CLIP-ViT	74.89	89.90	94.17	67.12	65.12	81.57	86.97	42.93	61.87	83.63	89.70	47.82
RDE [27]	CLIP-ViT	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
HAM [16]	CLIP-ViT	77.99	91.34	95.03	69.72	69.95	83.88	88.39	42.72	72.50	87.70	91.95	55.47
Methods with ViT backbone:													
CPCL [45]	ViT	70.03	87.28	91.78	63.19	62.60	79.07	84.46	36.16	58.35	81.05	87.65	45.81
PDReid [21]	ViT	71.59	87.95	92.45	65.03	60.93	77.96	84.11	36.44	56.65	77.40	84.70	45.27
SSAN [7]	ViT	61.37	80.15	86.73	-	54.23	72.63	79.53	-	43.50	67.80	77.15	-
CFine [39]	ViT	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
IVT [32]	ViT	65.59	83.11	89.21	-	56.04	73.60	80.22	-	46.70	70.00	78.80	-
Ours	ViT	**79.93**	**92.95**	**96.47**	**72.61**	**68.68**	**84.29**	**89.74**	**41.78**	**74.33**	**88.85**	**92.95**	**57.68**

Figure 6. Comparison of efficiency and performance on CUHK-PEDES. All metrics are normalized to [0, 100], where higher is better. Cost metrics (Params, FLOPs, Memory) are inverted, meaning lower resource usage yields a higher score. Our fusion module employing Mamba achieves competitive accuracy with reduced computational overhead.

4.6. Comparisons with State-of-the-Art Methods

We compare our method against recent state-of-the-art approaches across three benchmarks, as summarized in Tab. 5. On CUHK-PEDES, our method establishes new state-of-the-art results, substantially surpassing both ViT-based methods and the recent CLIP-based HAM baseline. On ICFG-PEDES, our approach remains competitive with HAM while significantly outperforming other methods. On RSTPReid, we achieve the best reported results to date, surpassing all ViT-based methods and establishing clear improvements over HAM.

These consistent gains across diverse datasets underscore the generalization capability of our approach. Notably, this high performance is achieved without leveraging large-scale pre-trained vision-language models, demonstrating that our identity-clothing disentanglement architecture and efficient fusion strategy provide effective cross-modal alignment through task-specific design. In summary, our method delivers superior or highly competitive performance across all three benchmarks, validating the effectiveness of our proposed components for text-to-image person re-identification.

5. Conclusion and Limitations

In this work, we proposed a novel framework to address interference caused by clothing in text-to-image person re-identification. Our approach is centered on feature decoupling guided by a Multimodal Large Language Model. We successfully demonstrated that an MLLM can provide supervision at a fine-grained level to guide our Bidirectional Decoupling Alignment Module. By design, this module explicitly isolates features relevant to identity while actively suppressing interference related to clothing through a combined alignment and orthogonal loss strategy based on a kernel function. Furthermore, our integration of a Mamba State Space Model proved to be an effective and efficient fusion strategy, adept at capturing dependencies between modalities. Our method's effectiveness was validated by achieving new state-of-the-art results on the CUHK-PEDES and RSTPReid benchmarks.

Nonetheless, limitations remain. These primarily concern the potential for noise in descriptions generated by the MLLM and the need for further optimization for deployment at a large scale. Future work will focus on improving description reliability to mitigate noise and on enhancing inference efficiency. We also plan to explore the framework's applicability to more challenging scenarios, such as person Re-ID involving clothing changes, to further test the robustness of our decoupling mechanism.

References

[1] Shehreen Azad and Yogesh Singh Rawat. Activity-Biometrics: Person Identification from Daily Activities, 2024. 2

[2] Yan Bai, Yihang Lou, Yongxing Dai, Jun Liu, Ziqian Chen, and Ling-Yu Duan. Disentangled Feature Learning Network for Vehicle Re-Identification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 474–480, Yokohama, Japan, 2020. 2

[3] Maryam Bukhari, Sadaf Yasmin, Sheneela Naz, Muazzam Maqsood, Jehyeok Rew, and Seungmin Rho. Language and vision based person re-identification for surveillance systems using deep learning with LIP layers. *Image and Vision Computing*, 132:104658, 2023. 1

[4] Hao Cheng, Yufei Wang, Haoliang Li, Alex C. Kot, and Bihan Wen. Disentangled Feature Representation for Few-shot Image Classification, 2021. 2

[5] Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. Improving Disentangled Text Representation Learning with Information-Theoretic Guidance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7530–7541, Online, 2020. 2

[6] Can Cui, Siteng Huang, Wenxuan Song, Pengxiang Ding, Min Zhang, and Donglin Wang. ProFD: Prompt-Guided Feature Disentangling for Occluded Person Re-Identification, 2024. 2

[7] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification, 2021. 1, 2, 6, 9

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. 2, 3

[9] Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis, 2024. 3

[10] Hiren Galiyawala and Mehul S. Raval. Person Retrieval in Surveillance Using Textual Query: A Review, 2021. 1

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation, 2015. 2

[12] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search, 2021. 2

[13] Nooshin Hanafi and Hamid Saadatfar. A fast DBSCAN algorithm for big data based on efficient density calculation. *Expert Systems with Applications*, 203:117501, 2022. 5

[14] Weizhen He, Yiheng Deng, Yunfeng Yan, Feng Zhu, Yizhou Wang, Lei Bai, Qingsong Xie, Donglian Qi, Wanli Ouyang, and Shixiang Tang. Instruct-ReID++: Towards Universal Purpose Instruction-Guided Person Re-identification, 2025. 2

[15] Ding Jiang and Mang Ye. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval, 2023. 1, 2, 9

[16] 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, 2025. 2, 4, 9

[17] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending CLIP's Image-Text Alignment to Referring Image Segmentation, 2024. 3

[18] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion, 2023. 3

[19] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person Search with Natural Language Description, 2017. 2

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person Search with Natural Language Description, 2017. 6

[21] Weihao Li, Lei Tan, Pingyang Dai, and Yan Zhang. Prompt Decoupling for Text-to-Image Person Re-identification, 2024. 2, 9

[22] Yubo Li, De Cheng, Chaowei Fang, Changzhe Jiao, Nannan Wang, and Xinbo Gao. Disentangling Identity Features from Interference Factors for Cloth-Changing Person Re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2252–2261, Melbourne VIC Australia, 2024. 2

[23] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Multi-task Adversarial Network for Disentangled Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3743–3751, Salt Lake City, UT, USA, 2018. 2

[24] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:14052–14060, 2024. 9

[25] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in CLIP, 2022. 2

[26] Qun Niu, Tao Chen, Xing Zhang, Yifan Wang, and Ning Liu. LLM-Loc: Bootstrap single-image indoor localization with large language model. *Expert Systems with Applications*, 291:128368, 2025. 2

[27] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-Correspondence Learning for Text-to-Image Person Re-identification, 2024. 9

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. 2, 3

[29] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning Disentangled Representations via Mutual Information Estimation, 2019. 2

[30] Johann Schmidt and Sebastian Stober. Robust Canonicalization through Bootstrapped Data Re-Alignment, 2025. 2

[31] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification, 2022. 1, 2

[32] 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, 2022. 9

[33] Wentao Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID, 2024. 2

[34] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings, 2016. 2

[35] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled Representation Learning, 2024. 2

[36] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language Person Search with Mutually Connected Classification Loss. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2057–2061, Brighton, United Kingdom, 2019. 2

[37] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language, 2020. 2

[38] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022. 2

[39] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. CLIP-Driven Fine-grained Text-Image Person Re-identification, 2022. 9

[40] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning Comprehensive Representations with Richer Self for Text-to-Image Person Re-Identification, 2023. 1

[41] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. Prototypical Prompting for Text-to-image Person Re-identification, 2024. 9

[42] Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark, 2023. 2

[43] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training, 2021. 2

[44] Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. GraFT: Gradual Fusion Transformer for Multimodal Re-Identification, 2023. 2

[45] Yanwei Zheng, Xinpeng Zhao, Chuanlin Lan, Xiaowei Zhang, Bowen Huang, Jibin Yang, and Dongxiao Yu. CPCL: Cross-Modal Prototypical Contrastive Learning for Weakly Supervised Text-based Person Re-Identification, 2024. 9

[46] 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, 2021. 6

[47] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. PLIP: Language-Image Pre-training for Person Representation Learning, 2024. 2

[48] 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, 2024. 9