Robust Person Re-Identification via MLLM-Supervised Feature Decoupling

First Author
Institution1
Institution1 Address
firstauthor@i1.org

Second Author
Institution2
Institution2 Address
secondauthor@i2.org

Abstract

The challenge of fine-grained semantic alignment continues to hinder Text-to-Image Person Re-Identification, where clothing-induced interference and a persistent modality gap degrade retrieval accuracy. In this paper, we propose a novel framework to resolve this issue, centered on MLLM-supervised feature decoupling. Our framework introduces two core components: a Bidirectional Decoupling Alignment Module and a Mamba State Space Model for efficient fusion. To obtain high-quality, fine-grained supervision, we first employ a Multimodal Large Language Model 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 adversarial loss that actively suppresses the influence of clothing-related features, and a kernel-based orthogonal constraint that ensures statistical independence. Furthermore, we are the first to integrate the Mamba State Space Model 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: Text-to-Image Re-Identification, Feature Decoupling, MLLM, Mamba State Space Model.

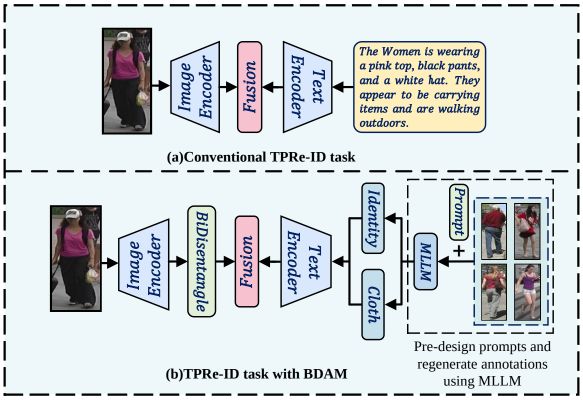


Image Encoder
Fusion
Text Encoder
The Women is wearing a pink top, black pants, and a white hat. They appear to be carrying items and are walking outdoors.
(a)Conventional TPRe-ID task
Image Encoder
BiDisentangle
Fusion
Text Encoder
Identity
Cloth
MLLM
Prompt
+
Pre-design prompts and regenerate annotations using MLLM
(b)TPRe-ID task with BDAM

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-identification (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.

Despite this progress, most recent methods, which often adopt ViT as the image encoder [8, 14, 21, 33, 42, 47] or leverage CLIP to benefit from large-scale contrastive pretraining [26, 28, 43], still face a critical limitation. Both lines commonly treat the description holistically, overlooking the distinction between identity-relevant (e.g., gender, body shape) and identity-irrelevant (e.g., clothing, hairstyle) content. In complex scenes with background clutter, large clothing variations, or redundant details, this coarse treatment blurs identity cues, weakens feature decoupling, and degrades cross-modal matching robustness.

To address this challenge, we propose a novel framework centered on MLLM-supervised feature decoupling. Instead of treating features holistically, we draw on the style-clustering paradigm [16] and leverage an MLLM to automatically generate fine-grained, separate identity and clothing descriptions. These decoupled annotations provide explicit supervision for our Bidirectional Decoupling Alignment Module (BDAM) to explicitly separate and align identity and clothing information. This disentanglement is enforced by a multi-task loss strategy, including an adversarial loss and an HSIC-based orthogonality constraint, as detailed in our Sec. 3. Furthermore, to enhance alignment, we are the first to integrate the Mamba State Space Model (SSM) as an efficient fusion module, capturing long-range cross-modal dependencies with linear complexity.Our main contributions are threefold: (1) A prompt auto-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 adversarial loss and an HSIC-based orthogonality constraint; and (3) The novel integration of a Mamba SSM fusion module that models long-range cross-modal dependencies with linear complexity.

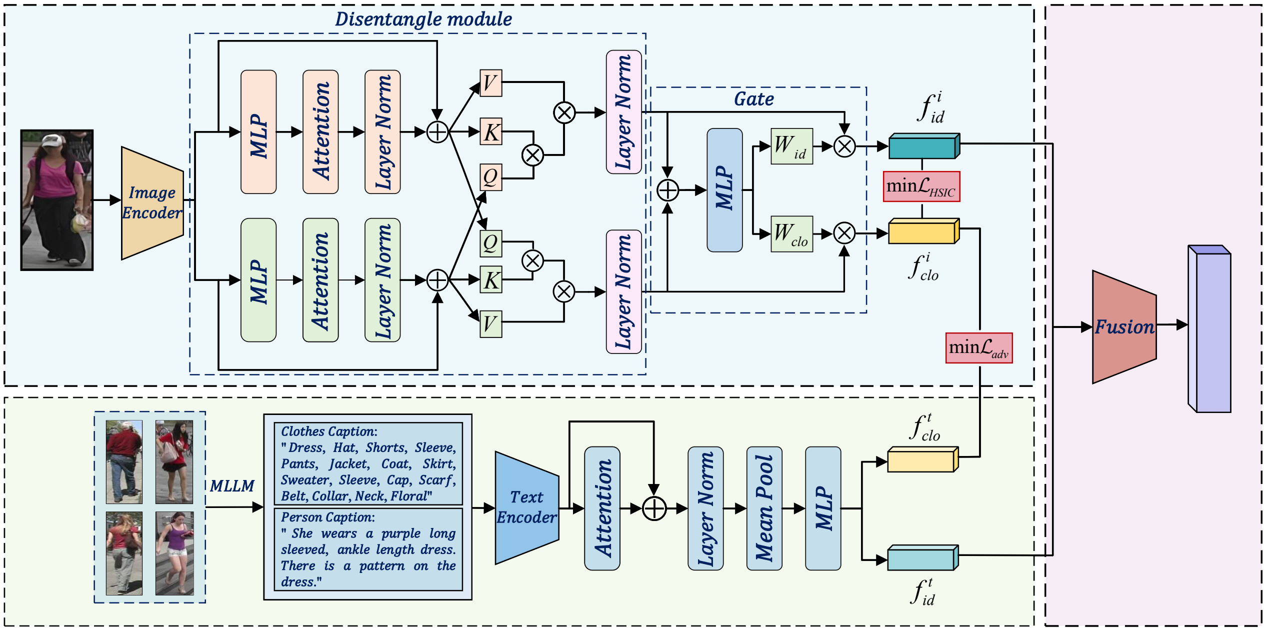
2. Related Work

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 min–max multi-task adversarial training that jointly optimizes an encoder and a style discriminator [23]; metric-learning schemes such as DFR [4] with Gradient Reversal Layer [11] to decorrelate factors; and orthogonal linear projections that separate visual and textual embeddings under CLIP [25]. In Re-ID, 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. Clothing-changing Re-ID often adopts dual-stream architectures to counter appearance shifts and camera bias [22]. These efforts improve semantic purity and factor independence, yet two 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, and 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 data-side decoupling with model-side 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–text tokens in parallel to capture semantic associations [44], but their quadratic complexity in sequence length causes memory and latency spikes for high-resolution images or long descriptions. CLIP-based pipelines [17] benefit from large-scale contrastive pretraining and well-aligned embeddings, yet global pooling and holistic processing often blur identity versus clothing cues, leading 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 [18] models dependencies through GNNs, offering relational inductive bias, whereas assumptions about graph structure and stationarity limit adaptability to free-form, variable-length text and dynamic visual contexts. Contemporary evidence suggests that accuracy and robustness improve only when semantic disentanglement and efficient fusion advance in tandem. Practically, fusion should (i) re-



Clothes Caption:
" Dress, Hat, Shorts, Sleeve, Pants, Jacket, Coat, Skirt, Sweater, Sleeve, Cap, Scarf, Belt, Collar, Neck, Floral"

Person Caption:
" She wears a purple long sleeved, ankle length dress. There is a pattern on the dress."

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^t_{clo} and f^t_{id}, 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^t_{id} to generate the final fused feature representation.

spect factorized semantics (e.g., identity/clothing) to avoid re-coupling nuisances, (ii) capture long-range cross-modal dependencies, and (iii) scale with linear or near-linear complexity to handle long sequences and high-res 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 between the input image and encoded textual features. As illustrated in Fig. 2, BDAM improves T2I-ReID retrieval performance by effectively separating identity from clothing features, thereby enabling superior cross-modal fusion. The core objective is to learn an identity representation invariant to clothing changes. To achieve this, our model comprises four key components: (1) a Visual Feature Extraction Module, (2) a Textual Feature Extraction Module, (3) the BDAM, and (4) a 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 identity-relevant and identity-irrelevant textual descriptions. A text encoder subsequently encodes these into f^t_{id} and f^t_{clo}. During disentanglement, BDAM leverages these textual features to guide image feature learning. To ensure the quality of this process, we introduce an HSIC loss to constrain the two resulting feature types towards orthogonality. We also employ an adversarial clothing loss, denoted as $\mathcal{L}_{\text{adv}}$, to actively suppress clothing-related features and ensure they do not interfere with identity recognition. Finally, to achieve 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 multi-modal features, enhancing the model's overall representation capability.

3.2. Bidirectional Decoupled Alignment Module

Some studies directly adopt CLIP [28] as a dual-modality feature extractor and align global embeddings for retrieval or discrimination. However, this approach presents two key limitations. First, its limited fine-grained semantics hinder the separation of identity from clothing. Second, its holistic image-text encoding lacks token-level cross-modal structure modeling, 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 identity-relevant and identity-irrelevant cues. A dual-branch linear projection yields preliminary identity features $f'_{\text{id}} \in \mathbb{R}^{B\times L\times D}$, and clothing features $f'_{\text{clo}} \in \mathbb{R}^{B\times L\times D}$, 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-branch cross-attention to exchange information. In the identity branch, the clothing branch provides auxiliary context, and vice versa, reinforcing semantic distinctions. Each branch then applies global average pooling to produce $\hat{f}id$ and $\hat{f}clo$. To enable soft, input-adaptive disentanglement, 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 gated features $f_{id}^{gate} = g \odot \hat{f}_{id}$ and $f_{clo}^{gate} = (1-g) \odot \hat{f}_{clo}$, where $\odot$ denotes element-wise multiplication. This dimension-wise weighting provides fine-grained control; f_{id}^{gate} is further sent to the fusion module. To train BDAM and enforce separation, we introduce a clothing adversarial loss and an HSIC-based identity–clothing decoupling loss. The clothing adversarial loss suppresses correlations between visual clothing features and clothing text, enhancing the purity of identity features:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_i \left[-\log(1 - P_{\text{pos}}(i))\right] \tag{1}$$

where $P_{\text{pos}}(i)$ is the temperature-scaled dot-product similarity between $\hat{f}_{\text{clo}}$ and f^t_{clo}. In practice, clothing features are linearly projected to the text dimension and L2-normalized for stable similarity estimation. This adversarial objective explicitly downweights clothing, while cross-attention implicitly sharpens identity/clothing separation via interaction. By minimizing the matching probability between the visual clothing features and the text, $\mathcal{L}_{\text{adv}}$ encourages the model to weaken the distracting effect of clothing features on identity recognition. This adversarial mechanism and the cross-attention mechanism are complementary. The former explicitly reduces the weight of clothing information by optimizing an objective function, while the latter strengthens the semantic differences between identity and clothing features through interactive modeling. To further encourage statistical independence, we minimize an HSIC-based decoupling loss:

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

Here, $f_{\text{id}}^{\text{gate}} \in \mathbb{R}^{B\times D}$ and $f_{\text{clo}}^{\text{gate}} \in \mathbb{R}^{B\times D}$ are the gated identity and clothing features, respectively. $K_{\text{id}} = f_{\text{id}}^{\text{gate}}(f_{\text{id}}^{\text{gate}})^T$ and $K_{\text{clo}} = f_{\text{clo}}^{\text{gate}}(f_{\text{clo}}^{\text{gate}})^T$ 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(specifically, ChatGPT-4o as detailed in 4.2) to automatically generate fine-grained identity and clothing descriptions for pedestrians, reducing manual annotation and enriching supervision. Prior work, notably HAM [16], shows that modeling annotator styles can steer an MLLM to produce diverse texts. Building on our reproduction of HAM, we adapt and extend the pipeline to meet our model's design goals.

We use the CLIP text encoder (ViT-L/14) to embed the original descriptions into fixed-dimensional vectors.With prompts, an MLLM generalizes and substitutes entity attributes to emphasize expression style rather than content. We then cluster these style embeddings with DBSCAN [13], which adaptively discovers dense regions without predefining the cluster count.To stabilize clusters, we reassign noise points and merge small clusters. This setup aligns the learned style categories with the identity/clothing disentanglement expected by BDAM.

A dual-prompt generator guides the MLLM to output two texts per image: an identity description (biological traits such as race, gender, age, body type) and a clothing description (apparel type, color, pattern, material, accessories). We control generation with length and temperature and apply syntax checks plus attribute-coverage validation so that outputs remain grammatical, structured, and parsable. Summary. Our adaptations deliver flexible style modeling via DBSCAN and a decoupled dual-description mechanism that strengthens the distinctiveness and diversity of identity and clothing semantics. The resulting supervision improves data expressiveness and provides richer training signals for BDAM.

3.4. Feature Fusion

For efficient and semantically-sensitive cross-modal feature fusion, we introduce the Mamba SSM.The core objective is to preserve the semantic integrity of the purified image and text identity features, enabling a fine-grained fusion that is robust to the clothing variations previously isolated by the BDAM.The module first aligns features from both modalities, then applies a pre-fusion gating mechanism to dynamically adjust their weights, and finally employs a multi-layer Mamba SSM for deep semantic interaction to generate the final fused representation.Initially, to mitigate distributional discrepancies between modalities, MLP performs dimensional alignment.It processes the decoupled visual features, $f_{\text{id}}^{\text{gate}}$, and the textual features, f_{id}^{t}, to generate aligned features, f_{img} and f_{txt}.Notably, the decoupled visual clothing feature, $f_{\text{clo}}^{\text{gate}}$, is *intentionally discarded* during fusion.This design is central to our goal: the BDAM, supervised by $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{Decouple}}$, is tasked with purging identity-irrelevant information into $f_{\text{clo}}^{\text{gate}}$.By excluding this feature from the final fusion, the model is forced to learn a representation based purely on stable identity semantics, thereby achieving robustness against clothing changes.Following this, a gating mechanism achieves dynamic weighted fusion. It outputs a weight vector $w \in \mathbb{R}^{B\times 2}$, which is normalized via a SoftMax layer to produce image w_1 and text w_2, satisfying $w_1 + w_2 = 1$.The resulting fusion is computed as: $f_{\text{fusion}} = w_1 \cdot f_{\text{img}} + w_2 \cdot f_{\text{txt}}$.This 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 $w \in \mathbb{R}^{B\times 2}$, for the modalities, whereas the disentanglement gate provides a dimension-wise weight vector $g \in \mathbb{R}^{B\times D}$ for fine-grained feature control.The fused features, f_{fusion}, are then processed by the Mamba SSM to enhance inter-modal interaction.Leveraging its long-range dependency modeling capabilities, Mamba effectively captures complex sequential relationships between the modalities.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 long-range semantic modeling,providing robust support for downstream tasks like person re-identification.

3.5. Loss Function

To achieve fine-grained cross-modal alignment, we adopt the InfoNCE loss, which maximizes similarity for positive image–text pairs (same identity) while separating negatives.It is defined as:

$$\mathcal{L}_{\text{info}} = -\log \frac{\exp(\mathbf{v}^\top \mathbf{t}i/\tau)}{\sum_j \exp(\mathbf{v}^\top \mathbf{t}j/\tau)} \tag{3}$$

Here, v_i is the L2-normalized image identity feature from the visual encoder, t_i is the text feature for the i-th identity, and τ controls distribution sharpness.In-batch negatives reduce the cross-modal semantic gap and promote alignment in a shared space.To enhance intra-modal identity discrimination, we include a triplet loss that enforces intra-class compactness and inter-class separation:

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

Here, (a,p,n) represent the 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 multi-task training, differing loss scales can cause one task to dominate.We adopt GradNorm to balance training by dynamically adjusting each task's gradient norm:

$$\mathcal{L}_{\text{GradNorm}} = \sum_k |\nabla\theta(w_k\mathcal{L}_k) - \tilde{r}kG\text{ref}|1 \tag{5}$$

Here, $\nabla\theta(w_k\mathcal{L}k)$ represents the gradient of the weighted loss of the k-th task, $w_k\mathcal{L}_k$, with respect to the shared parameters θ;$||\cdot||_1$ 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 beginning of training (used as a baseline), and $\bar{r}$ is the average of the loss ratios over all tasks.$Gref$ is a reference gradient norm, typically set to the gradient norm of the first task, $||\nabla\theta(w_1\mathcal{L}1)||$, as a baseline.This mechanism enforces a gradient balance among tasks during training by minimizing the L1 deviation between the actual gradient norm and a target value, $\tilde{r}_k G_{ref}$.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 the stability of the training process.Finally, the overall loss function is defined as follows:

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

Here, $\mathcal{L}_k$ represents the loss term for the k-th task (which includes $\mathcal{L}$info, $\mathcal{L}_{\text{triplet}}$, $\mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{Decouple}}$), and $w_k = \exp(s_k)$ is the task weight, which is dynamically adjusted by learning the task uncertainty parameter s_k.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 pre-trained bert-base-uncased as the text encoder with hidden dimensionality of 768, and vit-base-patch16-224 as the visual encoder. All images are resized to 224 × 224 pixels. The BDAM module processes visual tokens through multi-layer self-attention and cross-attention, disentangling them into identity and clothing features of 768 dimensions each. A two-layer Mamba fusion module with input/output dimension 256, state dimension 16, and convolutional kernel size 4 integrates the multimodal representations. Dropout rate is set to 0.1 throughout. Training utilizes the Adam optimizer with learning rate 1×10^{-4}, weight decay 1×10^{-3}, and cosine annealing scheduling. The multi-task loss combines InfoNCE, triplet, clothing adversarial, HSIC decoupling, and gate regularization terms. We employ GradNorm for dynamic task weighting with $\alpha = 1.5$ and weight learning rate 0.025; task weights are normalized to sum to the number of tasks and clipped to $[10^{-4}, 10]$. Gradient norms are computed on the final shared layer, with an additional log-variance regularization coefficient of 1×10^{-3}.

For data augmentation, we extract style embeddings using clip-vit-base-patch32, cluster them via DBSCAN, and generate style-consistent identity and clothing descriptions through ChatGPT-4o. 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 three runs.

4.3. Parameter Analysis

Although the identity and clothing branches in BDAM adopt identical architectures, they learn distinctly different representations through independent parameterization and asymmetric supervision. The BDAM gating mechanism generates dimension-wise weights $g_{\text{dis}} \in \mathbb{R}^{B \times D}$ via $g_{\text{dis}} = \sigma(W_g[\hat{f}_{id}; \hat{f}_{clo}] + b_g)$, yielding complementary coefficients $W_{id} = g_{\text{dis}}$ and $W_{clo} = 1 - g_{\text{dis}}$ that satisfy the zero-sum constraint $W_{id} + W_{clo} = \mathbf{1}$. This constraint introduces competitive pressure between branches, encouraging each to specialize in unique semantic patterns. The learning dynamics are governed by asymmetric loss signals: the identity branch receives supervision from InfoNCE and triplet losses, steering it toward person-invariant attributes such as body shape and posture, while the clothing branch is guided by matching and adversarial losses toward appearance-centric features. The HSIC-based decoupling loss 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 $g_{\text{fus}} = \text{Softmax}(W_f[f_{\text{img}}^{\text{gate}}; f_{\text{txt}}])$, 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 sample-level gating mechanism operates at the instance level, dynamically adjusting modality importance based on input characteristics, whereas the disentanglement gate provides dimension-level control within the visual stream.

To empirically validate these mechanisms, we analyze the distributions of learned gate weights across the test set, as shown in Fig. 3. The BDAM identity weights exhibit a right-skewed distribution with mean 0.62 and standard deviation 0.18, substantially exceeding the clothing weights which concentrate around 0.38. This asymmetry confirms that identity-relevant features dominate the final representation, aligning with our design objective. 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 near-equal modality contributions with means of 0.53 for image and 0.47 for text, both with moderate variance around 0.12. This balanced distribution,

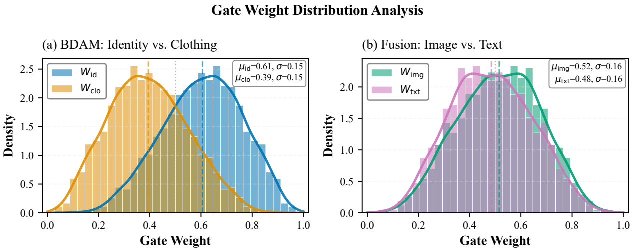


Figure 3. Learned gate weight distributions on the CUHK-PEDES test set. (a) BDAM assigns significantly higher weights to identity features than clothing features, validating the identity-centric design. (b) Fusion module maintains balanced modality contributions, confirming effective cross-modal integration without modality dominance.

centered near the uniform baseline of 0.5, demonstrates stable cross-modal alignment without collapse toward either modality. These empirical observations corroborate our theoretical design: BDAM enforces semantic disentanglement through asymmetric supervision and HSIC regularization, 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 across all metrics, demonstrating the effectiveness of explicit identity-clothing disentanglement. Within the BDAM architecture, both bidirectional cross-attention and the gating mechanism prove essential: removing cross-attention results in significant performance degradation, confirming its role in enabling semantic interaction between branches; ablating the gate mechanism similarly impairs retrieval accuracy. Furthermore, reducing the network depth to three layers degrades performance, indicating that deeper semantic modeling is necessary to achieve robust disentanglement. The learned gate weight distributions in Fig. 3 provide additional evidence, showing that the gating mechanism automatically achieves semantic specialization in BDAM and modality balance in fusion through end-to-end training.

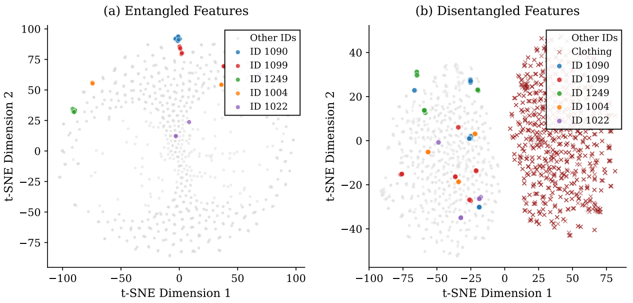


Figure 4. t-SNE visualization comparing feature distributions. (a) The baseline model's 'Entangled Features' show poor separation, with clusters for different identities (e.g., ID 1090 and ID 1099) positioned very closely. (b) With BDAM and HSIC regularization, the 'Disentangled Features' show a clear separation: identity features (dots) form a distinct cluster [cite: 165, 167-172], while 'Clothing' features (marked by 'x') occupy an independent region on the right.

To qualitatively assess disentanglement quality, we visualize learned features using t-SNE dimensionality reduction in Fig. 4. The baseline model produces a highly mixed embedding space with extensive identity overlap, indicating severe entanglement between identity and appearance attributes. In contrast, incorporating BDAM with HSIC regularization yields a well-structured feature space wherein identity embeddings form tight, well-separated clusters, while clothing features occupy independent regions. Within the identity cluster, same-person samples exhibit strong intra-class cohesion with clear inter-class boundaries. 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. Tab. 2 demonstrates that the complete Mamba-based fusion module substantially outperforms the baseline. Among the three components, removing the Mamba SSM backbone incurs the largest performance drop, highlighting its critical role in modeling long-range dependencies across modalities. Ablating either the gating mechanism or modality alignment layers results in more modest degradation, suggesting these components provide complementary benefits by enabling adaptive weighting and distributional alignment, respectively.

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 Adversarial	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. Tab. 3 reveals that every loss term contributes positively to final performance. The InfoNCE and triplet losses emerge as foundational: removing InfoNCE causes catastrophic performance collapse, underscoring the critical importance of cross-modal contrastive alignment; ablating the triplet loss also significantly degrades retrieval accuracy, confirming the necessity of strong intra-modal identity discrimination. The disentanglement-specific losses—adversarial and HSIC decoupling—are likewise essential; removing either term impairs performance, demonstrating that architectural design alone is insufficient and explicit loss-based constraints are required to enforce identity-clothing separation. Finally, removing gate regularization causes minor but consistent degradation, validating 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 style-consistent prompts, as detailed in Tab. 4. While random sampling provides minimal gains over the baseline, demonstrating the importance of structured prompt design, centroid-based methods such as K-Means and Gaussian Mixture Models offer modest improvements but require manual specification of cluster counts and struggle with irregular density distributions. Hierarchical agglomerative clustering exhibits similar limitations. In contrast, density-based approaches—particularly DBSCAN—achieve superior performance by automatically discovering arbitrarily-shaped clusters while identifying and filtering noise. DBSCAN's ability to adaptively determine the number of style categories without hyperparameter tuning makes it the optimal choice for this task.

4.5. Efficiency Analysis

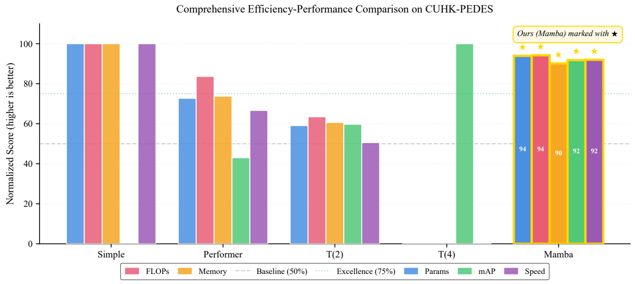


Figure 5. Efficiency-performance comparison on CUHK-PEDES. Our Mamba-based fusion achieves competitive accuracy with minimal computational overhead. All metrics normalized where higher is better.

As shown in Fig. 5, our Mamba-based fusion module achieves an optimal balance in the accuracy-efficiency trade-off space. The normalized comparison demonstrates near-peak performance across all dimensions while maintaining computational costs comparable to lightweight baselines. This efficiency advantage stems from the Mamba state-space architecture's ability to model long-range dependencies without the quadratic complexity of self-attention. Compared to Transformer-4Layer, our method delivers competitive retrieval accuracy with substantially lower resource consumption. Against the Simple baseline, Mamba achieves marked accuracy improvements with only modest computational overhead, confirming its suitability for practical deployment.

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

We compare our method against recent state-of-the-art approaches across three widely-adopted benchmarks, as summarized in Tab. 5. On CUHK-PEDES, our method establishes new state-of-the-art results, substantially surpassing both traditional ViT-based methods and the recent CLIP-based HAM baseline. Notably, this 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. On ICFG-PEDES, our approach remains competitive with HAM while significantly outperforming other recent methods, further validating the robustness of decoupled modeling under clothing variation. On RSTPReid, we achieve the best reported results to date, surpassing all ViT-based meth-

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**

ods and establishing clear improvements over HAM. These consistent gains across diverse datasets—varying in scale, camera setup, and annotation density—underscore the generalization capability of our disentanglement and fusion mechanisms in complex, multi-camera, cross-scene scenarios. In summary, our method delivers superior or highly competitive performance across all three benchmarks, validating the effectiveness of the Bidirectional Decoupling Alignment Module, Mamba-based fusion, and dynamic multi-task optimization for text-to-image person re-identification.

5. Conclusion and Limitations

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

Nonetheless, limitations remain, primarily concerning the potential for noise in MLLM-generated descriptions and the need for further optimization for large-scale deployment. Future work will focus on improving description reliability to mitigate noise and enhancing inference efficiency. We also plan to explore the framework's applicability to more challenging scenarios, such as clothing-changing person Re-ID, 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, 4

[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. 2

[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. 1

[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. 4

[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. 2

[18] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion, 2023. 2

[19] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person Search with Natural Language Description, 2017. 1

[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 Reidentification. 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. Multitask 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, 4

[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. 1

[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. 1

[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. 1

[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