

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

Anonymous ICME submission

Abstract—Text-to-Image Person Re-Identification faces significant challenges from clothing-induced interference and persistent modality gaps. To address this, we propose a novel and robust person re-identification framework. By leveraging a Multimodal Large Language Model (MLLM), it effectively decouples person identity from other interference features, thereby achieving accurate person identification. Specifically, we introduce a Bidirectional Decoupled Alignment Module (BDAM). By employing MLLM-generated descriptions as fine-grained semantic supervision, it explicitly disentangles visual features into distinct identity and clothing subspaces. This decoupling strategy is enforced by a multi-objective regularization term combining clothing alignment loss with kernel-based orthogonal constraints. Furthermore, our network is based on Mamba, which efficiently models long-range dependencies within image content, avoiding the quadratic costs associated with Transformers. Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid benchmarks demonstrate the effectiveness of our method, showing superior performance and robustness against clothing variations compared to leading contemporary approaches.

Index Terms—Multimodal Learning, Text-to-Image Re-Identification, Feature Decoupling, Semantic Supervision.

I. INTRODUCTION

Text-to-Image Person Re-Identification (T2I-ReID) retrieves a target pedestrian from a large-scale image gallery given a natural-language description [1]. This technology is valuable for video surveillance [2], intelligent security [3], public safety, and social media. Despite recent progress, practical deployment remains challenging due to image factors (pose, viewpoint, illumination) obscuring identity cues and a persistent modality gap hampering fusion. These issues are exacerbated at the fine-grained level, where semantic alignment is particularly difficult. [4]

Due to noisy environments and human morphological variations, a core challenge in T2I-ReID is the substantial semantic gap between images and text. Early work projected global visual and textual features into a shared space [5], but suffered from high intra-class and low inter-class variance. To mitigate this, subsequent studies adopted feature disentanglement to align latent semantics, broadly falling into two categories: explicit methods using auxiliary modules for part-level alignment [6], and implicit approaches leveraging regularizers to associate noun phrases with image regions [7]. This evolution underscores the importance of separating identity-relevant from irrelevant semantics for advancing T2I-ReID.

This pursuit of disentanglement has been propelled by powerful backbones. For instance, models employing ViT [8] cap-

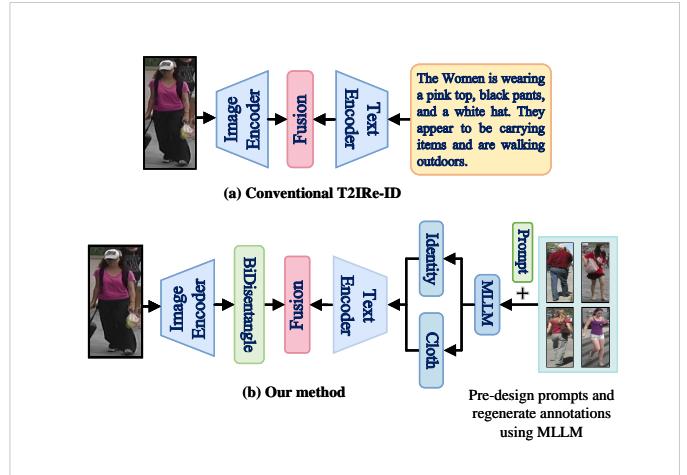


Fig. 1. Comparison of ReID methods. (a) Entangled global feature fusion. (b) *Our method*: MLLM-guided decoupling of identity and clothing.

ture fine-grained details, while methods leveraging CLIP [9] learn a well-aligned joint space. However, CLIP-based methods typically rely on global image-text alignment. This holistic approach often fails to distinguish identity-stable features (e.g., body shape) from transient nuisances (e.g., clothing), leading to entangled representations where salient clothing details overshadow stable identity cues. In contrast, as illustrated in Fig. 1, our framework demonstrates that precise, semantic-aware supervision is more critical than raw pre-training scale. Despite these advances, a critical limitation persists: both lines of work commonly treat the textual description holistically. This overlooks the semantic distinction between content relevant to identity and content irrelevant to identity. This coarse-grained treatment forces the model to entangle these factors, often prioritizing salient clothing details over stable identity cues, which blurs identity features and degrades matching robustness.

To address this challenge, we propose a novel framework with explicit, fine-grained semantic supervision. However, a primary obstacle is that existing datasets lack specific annotations separating identity attributes from clothing details. To bridge this gap, we introduce a Style-Guided Semantic Enhancement strategy. Drawing inspiration from the style-clustering paradigm [10], this strategy leverages an MLLM to automatically generate distinct descriptions for identity

and clothing. These decoupled annotations provide precise supervision for our Bidirectional Decoupled Alignment Module (BDAM) to meticulously separate and align identity and clothing information. This separation is enforced by a multi-objective loss strategy, including an alignment loss and an HSIC-based orthogonality constraint. Furthermore, we pioneer the integration of the Mamba as an efficient fusion module, adept at capturing long-range cross-modal dependencies with linear complexity.

Our main contributions are summarized as follows:

- We introduce a Style-Guided Semantic Enhancement strategy that leverages clustering and MLLMs to generate fine-grained, decoupled identity and clothing descriptions, providing precise supervision for explicit feature disentanglement.
- We design the Bidirectional Decoupled Alignment Module (BDAM), which achieves precise decoupling and alignment reinforced by a multi-objective loss strategy combining an alignment loss and an orthogonality constraint based on HSIC.
- We empirically validate the proposed Mamba-based fusion strategy through comprehensive experiments, which demonstrate its capability to efficiently model cross-modal dependencies and outperform contemporary methods in handling clothing interference.

II. RELATED WORK

A. Feature Decoupling for Person Re-Identification

Feature disentanglement [11] aims to separate semantically distinct factors to improve generalization. Early work used generative models [12], while modern strategies employ adversarial training [12], metric learning [13], or orthogonal projections [14]. In the ReID domain, this is used to separate identity-relevant signals from nuisances [15], [16], such as occlusion [17] or clothing changes [15].

While these efforts improve semantic purity, critical limitations persist. Without an effective interaction mechanism, isolated factors may fail to support robust cross-modal matching [16]. Reliance on manual annotations or external detectors constrains scalability [17], and implicit regularizers can be underconstrained, yielding spurious separations on unseen data [7]. Consequently, recent work emphasizes coupling disentanglement with principled interaction and independence constraints [16]. Our framework addresses this by pairing explicit supervision with model-level disentanglement to preserve cross-modal synergy.

B. Feature Fusion

Feature fusion is central to T2I-ReID [18], but standard methods have drawbacks. Cross-modal modules built on multi-head attention [19] suffer from quadratic complexity, causing latency spikes. Pipelines built on CLIP [20] benefit from pretraining but often blur identity versus clothing cues [21]. Other approaches like dynamic fusion [22] or GNNs [23] introduce higher computational costs or limiting structural assumptions.

Contemporary evidence suggests that accuracy improves only when semantic disentanglement and efficient fusion advance in tandem [24]. Practically, an ideal fusion module should respect factorized semantics (i.e., identity and clothing) and capture long-range cross-modal dependencies with linear complexity [25]. This motivates our design: BDAM supplies factor-aware representations, while a Mamba fusion module models long-horizon interactions with linear complexity, enabling precise alignment without the memory and efficiency bottlenecks of standard Transformers.

III. METHODOLOGY

A. Overview

To learn robust pedestrian representations, we propose the BDAM, which disentangles features using textual guidance. As illustrated in Fig. 2, our framework comprises vision and text encoders, the core BDAM, and an efficient Mamba SSM Fusion Module. Given an image $I \in \mathbb{R}^{B \times C \times H \times W}$, a visual encoder extracts features f_i . Concurrently, we use an MLLM to generate separate identity and clothing descriptions, which a text encoder encodes into f_{id}^t and f_{clo}^t . The BDAM leverages these textual features to guide the disentanglement of f_i . To enforce this separation, we employ two losses: an alignment loss (\mathcal{L}_{aln}) to supervise the visual clothing features using clothing descriptions, and an HSIC-based loss to ensure orthogonality. Finally, the Mamba efficiently fuses the disentangled visual identity and textual semantics, enhancing the model's overall representation.

B. Semantic Enhancement

We employ an MLLM (specifically, GPT-4) to automatically generate fine-grained identity and clothing descriptions for pedestrians, reducing the burden of manual annotation and enriching the available supervision. Fig. 3 illustrates this generation pipeline. Inspired by prior work on modeling annotator styles [10], we first derive style categories to guide the MLLM's generation tone.

To achieve this, we use the CLIP text encoder to embed original descriptions and then cluster these style embeddings with DBSCAN. These discovered style categories are used to formulate textual prompts (e.g., "Use a very detailed, descriptive style"). A dual prompt generator, using content-specific templates (e.g., "Describe the person's identity" and "Describe the person's clothing"), then guides the MLLM to output two distinct texts per image: one description for identity (biological traits) and another for clothing (apparel, colors, and patterns). We apply syntax checks and validation to ensure the outputs remain grammatical and structured.

C. Bidirectional Decoupled Alignment Module

Adopting CLIP for global embedding alignment presents two key limitations: first, its limited capacity for fine-grained semantics hinders the separation of identity from clothing, and second, its holistic encoding lacks token-level cross-modal modeling, reducing robustness in complex scenes.

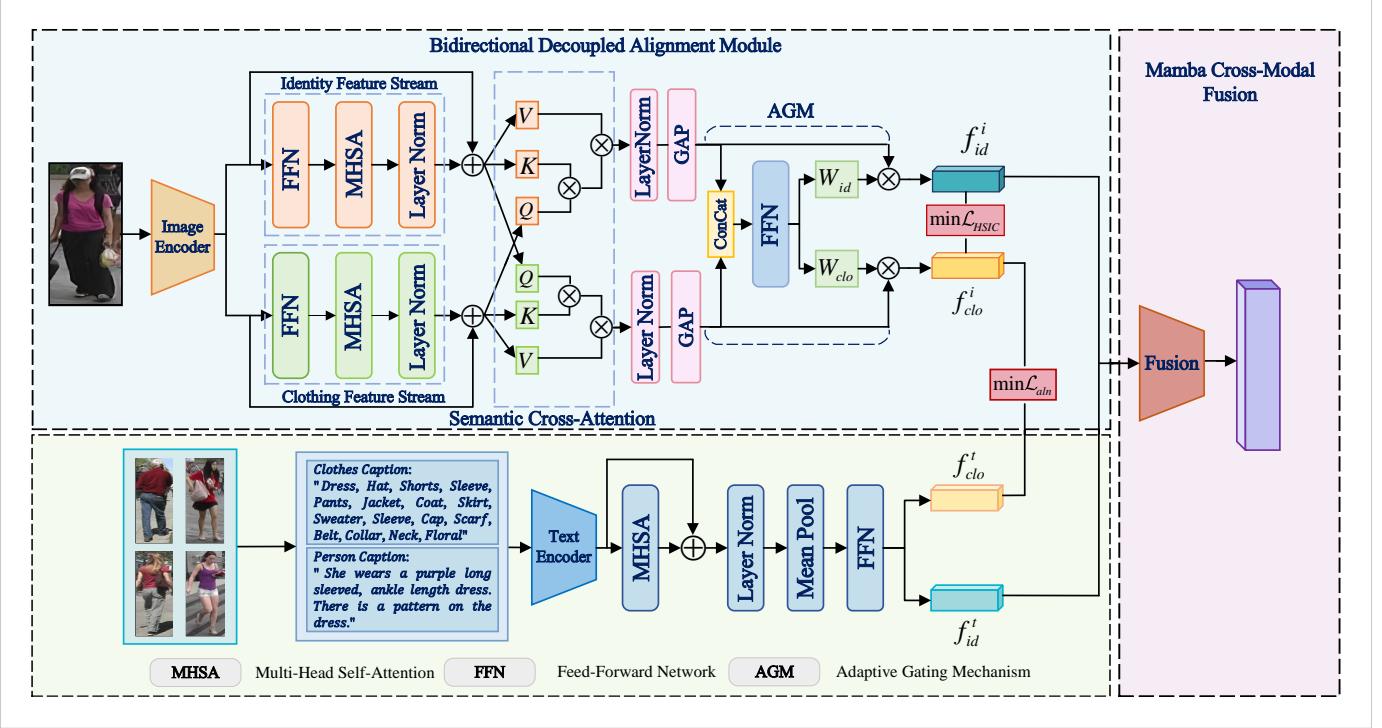


Fig. 2. Overview of the proposed framework. An MLLM generates identity (f_{id}^t) and clothing (f_{clo}^t) descriptions to supervise the BDAM. The BDAM module disentangles the input visual feature f_i into identity (f_{id}) and clothing (f_{clo}) features, which are optimized via contrastive and decoupling losses. Finally, the fusion module integrates the visual and textual identity features (f_{id} and f_{id}^t).

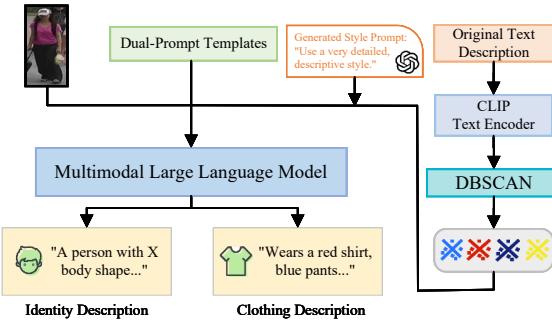


Fig. 3. Overview of the offline pipeline using an MLLM to generate decoupled identity and clothing descriptions. Style prompts are derived via CLIP and DBSCAN to enhance diversity.

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

Instead of using the ViT [CLS] token, we exploit the full patch sequence and introduce cross-attention between the

branches to exchange information and reinforce semantic distinctions. Each stream applies global average pooling (GAP) to produce \hat{f}_{id} and \hat{f}_{clo} . To enable adaptive soft disentanglement, we design a gating mechanism: the concatenated global vectors are fed to a lightweight linear network with a Sigmoid output, producing a gate $g \in \mathbb{R}^{B \times D}$. The final features are $f_{id}^i = g \odot \hat{f}_{id}$ and $f_{clo}^i = (1 - g) \odot \hat{f}_{clo}$, where \odot denotes element-wise multiplication and \hat{f}_{id} is sent to the fusion module.

To train BDAM and enforce separation, we introduce two specialized loss functions: a clothing alignment loss (\mathcal{L}_{aln}) and an HSIC-based decoupling loss ($\mathcal{L}_{Decouple}$).

The clothing alignment loss supervises the visual clothing features with the MLLM-generated clothing descriptions:

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

where $s_{ij} = \hat{f}_{clo}^i \cdot (f_{clo}^t)^j$ is the dot-product similarity between the visual clothing feature of sample i and the textual clothing feature of sample j , and τ is a temperature parameter. In practice, clothing features are linearly projected to the text dimension and L2 normalized for stable similarity estimation. This alignment objective explicitly ensures the clothing stream learns accurate semantic representations, which indirectly enhances the purity of the identity features by providing clear guidance on what constitutes clothing information. To further

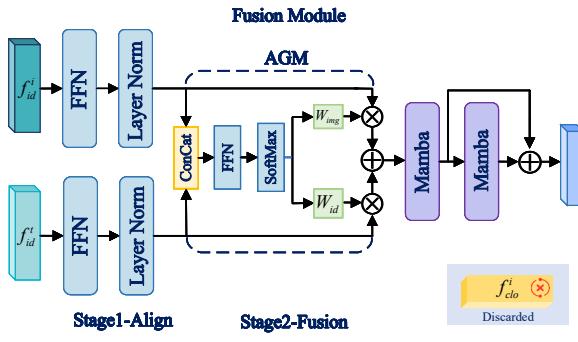


Fig. 4. Architecture of the Mamba Fusion Module, which fuses visual (f_{id}^i) and textual (f_{id}^t) identity features while discarding clothing features (f_{clo}^i).

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

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

Here, $f_{id}^i, f_{clo}^i \in \mathbb{R}^{B \times D}$ are the gated features; $K_{id} = f_{id}^i(f_{id}^i)^T$ and $K_{clo} = f_{clo}^i(f_{clo}^i)^T$ are their kernel matrices; and $H = I_N - (1/N)\mathbf{1}_N\mathbf{1}_N^T$ is the centering matrix. HSIC measures the statistical dependence between features; minimizing this value encourages the features to be statistically independent.

D. Feature Fusion

We introduce the Mamba SSM for efficient, semantic-sensitive feature fusion. 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 isolated by the BDAM. As illustrated in Fig. 4, the process begins with an FFN performing dimensional alignment on the decoupled visual features f_{id}^i and the textual features f_{id}^t to generate 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: since the BDAM (supervised by \mathcal{L}_{aln} and $\mathcal{L}_{\text{Decouple}}$) is tasked with purging irrelevant information into f_{clo}^i , excluding this feature forces the model 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 $g_{\text{fus}} \in \mathbb{R}^{B \times 2}$, which is normalized via a SoftMax layer to produce image W_{img} and text W_{txt} weights ($W_{img} + W_{txt} = 1$). The resulting fusion is computed as: $f_{\text{fusion}} = W_{img} \cdot f_{img} + W_{txt} \cdot f_{txt}$.

The resulting f_{fusion} features are then processed by the Mamba SSM to enhance cross-modal interaction. Leveraging its capability to model long-range dependencies, Mamba effectively captures complex sequential relationships. We employ a stack of Mamba layers, where each layer updates its input $f_{\text{fusion}}^{(l)}$ using a residual connection: $f_{\text{fusion}}^{(l+1)} =$

$\text{Mamba}(f_{\text{fusion}}^{(l)}) + f_{\text{fusion}}^{(l)}$. This structure improves information flow. Finally, the output from the last Mamba layer is projected to produce the final representation, $f_{\text{final}} \in \mathbb{R}^{B \times D_{out}}$.

E. Loss Function

To achieve fine-grained alignment between modalities, we adopt the InfoNCE loss ($\mathcal{L}_{\text{info}}$), which maximizes similarity for positive image-text pairs while separating negatives:

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

Here, v_i is the final L2-normalized fused representation, t_i is the text feature for the i -th identity, and τ controls distribution sharpness.

To enhance intra-modality identity discrimination, we include a triplet loss ($\mathcal{L}_{\text{triplet}}$) to enforce class compactness and separation:

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

Here, f_a, f_p, f_n are the decoupled visual identity features for the anchor, positive, and negative samples, respectively, and m is the margin.

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:

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

Here, $\nabla_\theta(w_k \mathcal{L}_k)$ is the gradient of the weighted loss of task k w.r.t. shared parameters θ . $\tilde{r}_k = (\mathcal{L}_k / \mathcal{L}_k^0) / \bar{r}$ is the normalized loss ratio, where \mathcal{L}_k^0 is the initial loss, and G_{ref} is a reference gradient norm.

Finally, the overall loss function is defined as follows, including a regularization term to prevent instability:

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

Here, \mathcal{L}_k represents the loss term for task k (which includes $\mathcal{L}_{\text{info}}$, $\mathcal{L}_{\text{triplet}}$, \mathcal{L}_{aln} , and $\mathcal{L}_{\text{Decouple}}$), and $w_k = \exp(s_k)$ is the learnable task weight. The hyperparameters α and λ control the GradNorm strength and regularization, respectively.

IV. EXPERIMENTS

A. Implementation Details

1) *Datasets and Metrics*: We evaluate our method on three standard benchmarks: CUHK-PEDES [21], ICFG-PEDES [24], and RSTPReid [1]. We follow their official identity-based splits and report mean Average Precision (mAP) and Rank-k accuracy (R-1, R-5, R-10).

TABLE I
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
<i>Methods with CLIP backbone:</i>													
IRRA [7]	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 [26]	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 [4]	CLIP-ViT	74.46	90.19	94.01	-	64.72	81.35	86.31	-	61.49	82.26	89.23	-
Propot [27]	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 [28]	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 [10]	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
<i>Methods with ViT backbone:</i>													
CPCL [29]	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 [30]	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 [1]	ViT	61.37	80.15	86.73	-	54.23	72.63	79.53	-	43.50	67.80	77.15	-
CFine [31]	ViT	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
IVT [25]	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

2) *Model and Training:* Our model employs a pre-trained bert-base-uncased as the text encoder and a vit-base-patch16-224 as the visual encoder. All images are resized to 224×224 pixels. The BDAM module disentangles visual tokens into 768-dimensional identity and clothing features. A 2-layer Mamba fusion module (256-dimensional input/output, 16-dimensional state, 4-convolution kernel) integrates the representations. Dropout is 0.1 throughout. We use the Adam optimizer (LR 1×10^{-4} , WD 1×10^{-3}) with cosine annealing. The total loss combines InfoNCE, triplet, clothing alignment, and HSIC decoupling terms, which are dynamically balanced using GradNorm ($\alpha = 1.5$).

3) *Data Augmentation and Reporting:* We use clip-vit-base-patch32 and DBSCAN to find style clusters, then use ChatGPT-4 to generate decoupled identity and clothing descriptions. All experiments are repeated with three random seeds (0, 1, 2), and we report the mean results.

B. Parameter Analysis

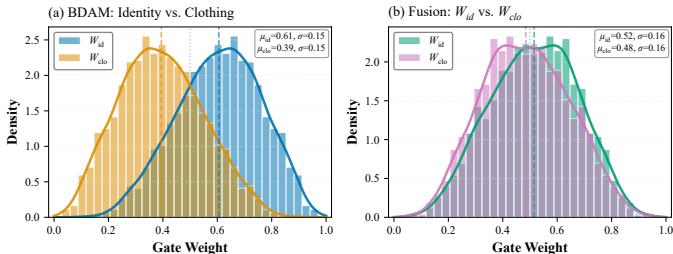


Fig. 5. Learned gate weight distributions on CUHK-PEDES. (a) BDAM's identity-centric bias and (b) Fusion's balanced modality contributions.

We empirically validate our gating mechanisms by analyzing the learned gate weights (Fig. 5). The BDAM's dimension-level gate ($g_{\text{dis}} \in \mathbb{R}^{B \times D}$), guided by asymmetric losses (e.g., InfoNCE, HSIC) to force stream specialization, exhibits a clear identity-centric design: identity weights (mean 0.61) substantially exceed clothing weights (mean 0.38). This asymmetry confirms its ability to discriminate cues.

In contrast, the instance-level fusion gate ($g_{\text{fus}} \in \mathbb{R}^{B \times 2}$) maintains nearly equal modality contributions (means 0.52 and 0.48), demonstrating stable alignment without modality collapse. These empirical observations corroborate our theoretical design: BDAM enforces semantic disentanglement, while the fusion gate achieves dynamic equilibrium.

C. Ablation Study

We conduct systematic ablation studies on CUHK-PEDES to validate each component.

TABLE II
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

1) *Disentanglement Module:* Table II shows that BDAM yields substantial gains. Removing cross-attention, ablating the gating mechanism, or reducing depth all degrade performance, confirming that semantic interaction and adaptive control are essential for robust disentanglement.

TABLE III
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	72.61	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

2) *Fusion Module:* Table III confirms the synergy of our fusion components. The Mamba SSM is critical; its removal

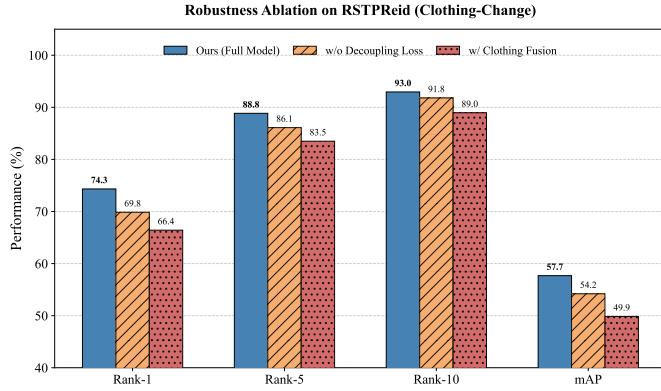


Fig. 6. **Robustness analysis on RSTPR Reid.** The performance degradation in variants lacking decoupling or retaining clothing features validates the efficacy of our interference-removal strategy.

causes the largest drop, highlighting the value of long-range dependency modeling. The gating and alignment layers provide further essential gains.

TABLE IV
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

3) *Loss and Prompting:* Table IV validates our multi-objective loss. InfoNCE and triplet losses are foundational. Crucially, the degradation without alignment or HSIC decoupling losses proves that explicit constraints are necessary to enforce identity-clothing separation. For prompt generation, we found density-based clustering (DBSCAN) significantly outperforms K-Me and random sampling, as it adaptively handles irregular style distributions. Full clustering results and t-SNE visualizations are in the supplementary material.

D. Robustness Mechanism Analysis

To verify robustness against clothing variations, we conduct an ablation on the cloth-changing RSTPR Reid dataset, focusing on two key designs: *Explicit Decoupling* via HSIC and *Explicit Exclusion* of clothing features. We compare the full model against two variants: *w/o Decoupling Loss* and *w/ Clothing Fusion* (intentionally re-introducing clothing features into the fusion module).

Results in Fig. 6 reveal two key insights. First, regarding *clothing features as interference*, the *w/ Clothing Fusion* variant suffers significant degradation across all metrics. This empirically confirms that in cloth-changing scenarios, visual clothing features act as confounding noise. By structurally discarding them, our method ensures reliance on stable identity semantics.

Second, regarding how *decoupling ensures precision*, the *w/o Decoupling Loss* variant shows a notable drop in Rank-1 accuracy, despite relatively stable Rank-10 performance. This disparity indicates that while entangled features suffice for coarse retrieval, the HSIC constraint is indispensable for the semantic purity required for precise top-rank matching. Thus, our robustness stems from the synergy of mathematically enforcing independence and structurally blocking interference.

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

Table I compares our method against state-of-the-art approaches. On CUHK-PEDES and RSTPR Reid, our method establishes new benchmarks, substantially surpassing both ViT-based methods and the strong CLIP-based HAM baseline. On ICFG-PEDES, we remain highly competitive. Remarkably, our ViT-based method outperforms strong CLIP-based competitors, such as HAM and IRRA, by a significant margin. This result challenges the prevailing dominance of large-scale vision-language pre-training in T2I-ReID. We attribute this success to two key factors:

Explicit Decoupling vs. Implicit Alignment: While CLIP models implicitly learn potential correlations, our BDAM forces the model to mathematically separate clothing from identity, preventing the network from overfitting to clothing shortcuts.

MLLM as a Superior Teacher: The rich, fine-grained descriptions generated by the MLLM provide high-quality semantic targets that are more informative than the noisy, coarse-grained alignment learned by CLIP. This validates that task-specific structural design combined with high-quality semantic guidance can effectively override the benefits of massive generic pre-training.

V. CONCLUSION AND LIMITATIONS

We proposed a framework to address clothing interference in T2I-ReID by using a Multimodal Large Language Model (MLLM) to guide feature decoupling. Our Bidirectional Decoupling Alignment Module (BDAM) leverages MLLM-generated descriptions to isolate identity from clothing features, enforced by an alignment and kernel-based orthogonal loss. A Mamba SSM provides efficient modality fusion. This method achieved new state-of-the-art results on CUHK-PEDES and RSTPR Reid. Limitations include potential MLLM-generated noise and large-scale deployment overhead. Future work will focus on improving description reliability, enhancing inference efficiency, and testing robustness on clothing-change scenarios.

REFERENCES

- [1] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao, “Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification,” 2021.
- [2] Maryam Bukhari et al., “Language and vision based Person Re-Identification for surveillance systems using deep learning with LIP layers,” *Image and Vision Computing*, vol. 132, pp. 104658, 2023.
- [3] Hiren Galiyawala and Mehul S. Raval, “Person Retrieval in Surveillance Using Textual Query: A Review,” *arXiv*, 2021.
- [4] Jialong Zuo et al., “UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity,” 2024.

- [5] Liwei Wang, Yin Li, and Svetlana Lazebnik, “Learning Deep Structure-Preserving Image-Text Embeddings,” 2016.
- [6] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang, “ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language,” 2020.
- [7] Ding Jiang and Mang Ye, “Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval,” 2023.
- [8] Alexey Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” 2021.
- [9] Lewei Yao et al., “FILIP: Fine-grained Interactive Language-Image Pre-Training,” 2021.
- [10] Jiayu Jiang et al., “Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification,” 2025.
- [11] Xin Wang, Hong Chen, Si’ao Tang, Zihao Wu, and Wenwu Zhu, “Disentangled Representation Learning,” 2024.
- [12] 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*, Salt Lake City, UT, USA, 2018, pp. 3743–3751.
- [13] Hao Cheng, Yufei Wang, Haoliang Li, Alex C. Kot, and Bihan Wen, “Disentangled Feature Representation for Few-shot Image Classification,” 2021.
- [14] Joanna Materzynska, Antonio Torralba, and David Bau, “Disentangling visual and written concepts in CLIP,” 2022.
- [15] Yubo Li et al., “Disentangling Identity Features from Interference Factors for Cloth-Changing Person Re-identification,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne VIC Australia, 2024, pp. 2252–2261.
- [16] Shehreen Azad and Yogesh Singh Rawat, “Activity-Biometrics: Person Identification from Daily Activities,” 2024.
- [17] Can Cui et al., “ProFD: Prompt-Guided Feature Disentangling for Occluded Person Re-Identification,” 2024.
- [18] Johann Schmidt and Sebastian Stober, “Robust Canonicalization through Bootstrapped Data Re-Alignment,” 2025.
- [19] Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings, “GraFT: Gradual Fusion Transformer for Multimodal Re-Identification,” 2023.
- [20] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak, “Extending CLIP’s Image-Text Alignment to Referring Image Segmentation,” 2024.
- [21] Shuang Li et al., “Person Search with Natural Language Description,” 2017.
- [22] Xinyu Feng et al., “Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis,” 2024.
- [23] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma, “Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion,” 2023.
- [24] Aichun Zhu et al., “DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval,” 2021.
- [25] Xiujun Shu et al., “See Finer, See More: Implicit Modality Alignment for Text-based Person Retrieval,” 2022.
- [26] 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*, vol. 38, pp. 14052–14060, 2024.
- [27] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang, “Prototypical Prompting for Text-to-image Person Re-identification,” 2024.
- [28] Yang Qin et al., “Noisy-Correspondence Learning for Text-to-Image Person Re-identification,” 2024.
- [29] Yanwei Zheng et al., “CPCL: Cross-Modal Prototypical Contrastive Learning for Weakly Supervised Text-based Person Re-Identification,” 2024.
- [30] Weihao Li, Lei Tan, Pingyang Dai, and Yan Zhang, “Prompt Decoupling for Text-to-Image Person Re-identification,” 2024.
- [31] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang, “CLIP-Driven Fine-grained Text-Image Person Re-identification,” 2022.