Disentangling Identity from Transient Attributes: 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.**

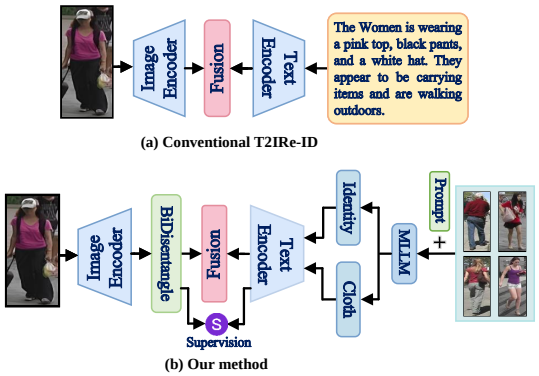

Fig. 1. Comparison of ReID methods. (a) Entangled global feature fusion. (b) *Our method*: MLLM-guided decoupling of identity and transient attributes (e.g., clothing and accessories).

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 (e.g., pose, viewpoint, illumination) obscuring identity cues and a persistent modality gap hampering fusion [4]. These issues are exacerbated at the fine-grained level, where semantic alignment is particularly difficult [5].

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 [6], 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 [7], and implicit approaches leveraging regularizers to associate noun phrases with image regions [8], [9]. 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 [10] capture fine-grained details, while methods leveraging CLIP [11] 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 attributes (e.g., clothing, accessories, and belongings; hereafter collectively referred to as "clothing" for brevity), leading to entangled representations where salient transient 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 transient 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 transient details. To bridge this gap, we introduce a Style-Guided Semantic Enhancement strategy.

II. Related Work

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

While these efforts improve semantic purity, critical limitations persist. Without an effective interaction mechanism, isolated factors may fail to support robust cross-modal matching [18]. Reliance on manual annotations or external detectors constrains scalability [17], and implicit regularizers can be underconstrained, yielding spurious separations on unseen data [8]. Consequently, recent work emphasizes coupling disentanglement with principled interaction and independence constraints [18], [19]. Our framework addresses this by pairing explicit supervision with model-level disentanglement to preserve cross-modal synergy.

III. METHODOLOGY

Drawing inspiration from the style-clustering paradigm [20], 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 [21]. Furthermore, we pioneer the integration of the Mamba as an efficient fusion module [22], 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.

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 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^t_{id} and f^t_{clo}. The BDAM leverages these textual features to guide the disentanglement of f_i. To enforce this separation, we employ two losses: an alignment loss ($\mathcal{L}_{\text{aln}}$) 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 (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 [20], 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

CLIP-based global alignment suffers from two limitations: insufficient fine-grained semantics for separating identity from clothing, and lack of token-level cross-modal interaction, which degrades robustness in complex scenes.

We adopt a pre-trained ViT (ViT-B/16) as the visual encoder E_v, producing entangled patch tokens $f_i \in \mathbb{R}^{B\times L\times D}$. These are projected into two branches yielding preliminary identity (f'_{id}) and clothing (f'_{clo}) features, which are refined via stream-wise self-attention. Unlike CLIP, we utilize the full patch sequence and introduce cross-attention between the two streams to enhance semantic distinction.

Each stream is globally averaged to obtain $\hat{f}_{\text{id}}$ and $\hat{f}_{\text{clo}}$. A lightweight gating network—comprising concatenation, a linear layer, and a Sigmoid—produces a soft mask $g \in \mathbb{R}^{B\times D}$. The final disentangled features are:

$$f^i_{\text{id}} = g \odot \hat{f}_{\text{id}}, \quad f^i_{\text{clo}} = (1-g) \odot \hat{f}_{\text{clo}},$$

where $\odot$ denotes element-wise multiplication. Only f^i_{id} is forwarded to the fusion module.

Training is guided by two losses. First, a clothing alignment loss $\mathcal{L}_{\text{aln}}$ aligns visual clothing features with MLLM-generated textual descriptions:

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

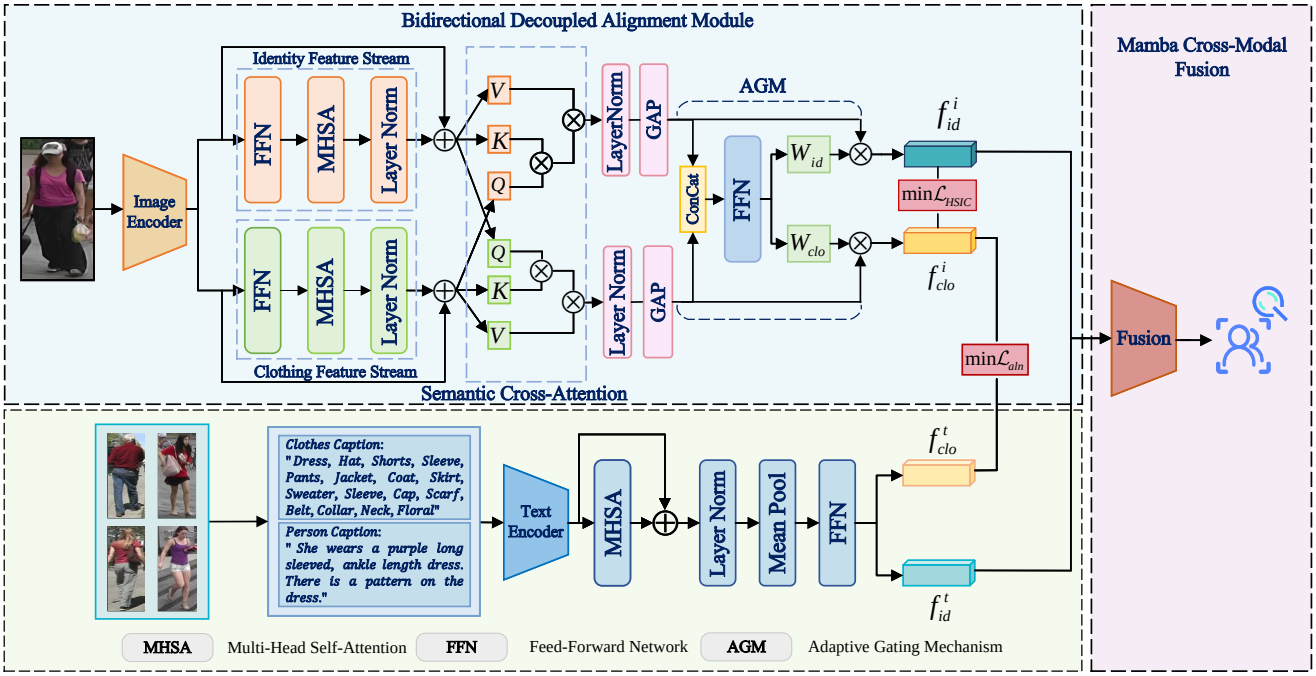

Fig. 2. Overview of the proposed framework. An MLLM generates identity (f^t_{id}) and clothing (f^t_{clo}) descriptions to supervise the BDAM. The BDAM module disentangles the input visual feature f_i into identity (f_{id}) and clothing (f_{clo}) features, which are optimized via contrastive and decoupling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

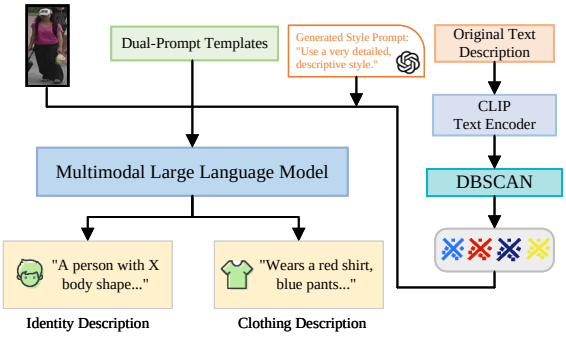

Fig. 3. Overview of the offline pipeline using an MLLM to generate decoupled identity and clothing descriptions. Style prompts are derived via CLIP and DBSCAN to enhance diversity.

where $s_{ij} = \hat{f}^i_{\text{clo}} \cdot (f^t_{\text{clo}})^j$ is the dot-product similarity (after L2 normalization and dimension alignment), and τ is a temperature. This explicitly defines clothing semantics, indirectly purifying the identity stream.

Second, to enforce statistical independence, we minimize an HSIC-based decoupling loss:

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

where $K_{\text{id}} = f^i_{\text{id}}(f^i_{\text{id}})^\top$, $K_{\text{clo}} = f^i_{\text{clo}}(f^i_{\text{clo}})^\top$, and $H = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ is the centering matrix. Minimizing HSIC encourages orthogonality between identity and clothing subspaces.

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^{i}_{id} and the textual features f^t_{id} to generate f_{img} and f_{txt}.

Notably, the decoupled visual clothing feature $f^{\text{i}}_{\text{clo}}$ is *intentionally discarded* during fusion. This design is central to our goal: since the BDAM (supervised by $\mathcal{L}_{\text{aln}}$ and $\mathcal{L}_{\text{Decouple}}$) is tasked with purging irrelevant information into $f^{\text{i}}_{\text{clo}}$, 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: $f_{\text{fusion}} = W_{img} \cdot f_{\text{img}} + W_{txt} \cdot f_{\text{txt}}$.

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

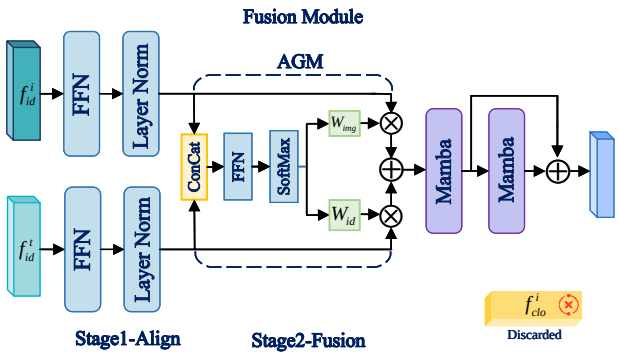

Fig. 4. Architecture of the Mamba Fusion Module, which fuses visual (f^{i}_{id}) and textual (f^{t}_{id}) identity features while discarding clothing features (f^{i}_{clo}).

its input $f^{(l)}_{fusion}$ using a residual connection: $f^{(l+1)}_{fusion} = \text{Mamba}(f^{(l)}_{fusion}) + f^{(l)}_{fusion}$. This structure improves information flow. Finally, the output from the last Mamba layer is projected to produce the final representation, $f_{final} \in \mathbb{R}^{B\times D_{out}}$.

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:

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

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:

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

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:

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

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

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

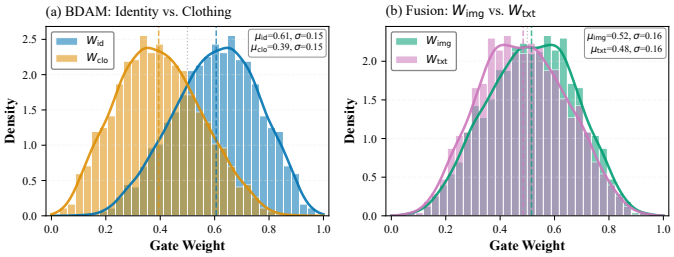

Fig. 5. Learned gate weight distributions on CUHK-PEDES. (a) BDAM's identity-centric bias and (b) Fusion's balanced modality contributions.

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 learnable task weight. The hyperparameters α and λ control the GradNorm strength and regularization, respectively. In Table IV, *w/o Weight Reg.* denotes removing the regularizer $\lambda\sum_k(\log w_k)^2$ in Eq. 6.

IV. Experiments

A. Implementation Details

1) Datasets and Metrics: We evaluate on three benchmarks: CUHK-PEDES [30], ICFG-PEDES [31], and RSTPReid [1], adhering to their official identity-based splits. Performance is reported using mean Average Precision (mAP) and Rank-k accuracy ($k = 1, 5, 10$).

2) Model and Training: We employ pre-trained *bert-base-uncased* and *vit-base-patch16-224* as text and visual encoders, respectively, with 224×224 inputs. BDAM outputs 768-dimensional features, which are integrated by a 2-layer Mamba fusion module (dim 256, state 16, kernel 4) with 0.1 dropout. Optimization uses Adam ($LR = 1 \times 10^{-4}$, $WD = 1 \times 10^{-3}$) with cosine annealing. The total loss (InfoNCE, triplet, alignment, HSIC) is dynamically balanced via GradNorm ($\alpha = 1.5$).

3) Data Augmentation and Reporting: We generate decoupled descriptions using ChatGPT-4, guided by style clusters derived from CLIP and DBSCAN. All experiments are repeated across three random seeds (0, 1, 2) to report mean results.

B. Parameter Analysis

We empirically validate our gating mechanisms by analyzing the learned weights in Fig. 5. Driven by asymmetric losses, the BDAM's dimension-level gate ($(g_{\text{dis}} \equiv g)$) exhibits a clear identity-centric bias, with identity weights ($\mu_{\text{id}} = 0.61$) significantly exceeding clothing weights ($\mu_{\text{clo}} = 0.39$). In contrast, the instance-level fusion gate (g_{fus}) maintains equilibrium between modalities, with comparable image and text contributions ($\mu_{\text{img}} = 0.52$ vs. $\mu_{\text{txt}} = 0.48$), preventing modality collapse. These statistics corroborate our design: BDAM enforces semantic disentanglement, while the fusion module achieves dynamic cross-modal balance.

C. Ablation Study

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

TABLE I

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THREE BENCHMARKS. "OURS" CORRESPONDS TO THE FULL MODEL WITH BDAM, MAMBA FUSION, AND ALL LOSS COMPONENTS ENABLED. UNLESS OTHERWISE SPECIFIED, THE SAME CONFIGURATION IS USED IN ALL ABLATION AND ROBUSTNESS STUDIES.

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 [8]	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 [23]	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 [5]	CLIP-ViT	74.46	90.19	94.01	-	64.72	81.35	86.31	-	61.49	82.26	89.23	-
Propot [24]	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 [25]	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 [20]	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 [26]	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 [27]	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 [28]	ViT	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
IVT [29]	ViT	65.59	83.11	89.21	-	56.04	73.60	80.22	-	46.70	70.00	78.80	-
Ours	ViT	**79.93**	**91.93**	**96.47**	**72.61**	**68.68**	**84.29**	**89.74**	**41.78**	**74.33**	**88.85**	**92.95**	**57.68**

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

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	79.93	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

TABLE IV

ABLATION STUDY ON INDIVIDUAL LOSS COMPONENTS.

Method	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Full Model	72.61	79.93	91.93	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 Weight Reg.	71.98	79.23	91.35	95.71

TABLE V

ROBUSTNESS ANALYSIS AGAINST SEMANTIC INTERFERENCE ON CUHK-PEDES.

Setting	mAP(%) ↑	R-1(%) ↑	R-5(%) ↑	R-10(%) ↑
Standard	72.61	79.93	91.93	96.47
Identity-Only	70.92	78.45	91.80	95.88
Clothes-Conflict	68.24	76.12	89.65	94.30

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.

Fusion module. Table III confirms the synergy of our fusion components. The Mamba SSM is critical; its removal causes the largest drop, highlighting the value of long-range dependency modeling. The gating and alignment layers provide further essential gains.

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 Analysis

To rigorously evaluate the efficacy of the proposed decoupling strategy—specifically the architectural decision to structurally discard visual clothing features during fusion—we conducted a robustness analysis focusing on semantic interference. We devised three distinct inference configurations on the CUHK-PEDES dataset based on the MLLM-generated annotations. The first configuration, denoted as **Standard**, utilizes the original, correct identity and clothing descriptions to establish the performance upper bound. The second, **Identity-Only**, masks all clothing-related descriptions in the textual input to assess the model's dependency on transient attributes. The third and most challenging configuration is the **Clothes-Conflict** attack, where we retain correct identity descriptions but shuffle the clothing descriptions across the entire test set. This creates a severe semantic mismatch where the input text describes an outfit completely contradictory to the target image.

The quantitative results are presented in Table V. In the

Identity-Only setting, the model exhibits only a marginal performance decline compared to the Standard setting. This stability corroborates that the BDAM successfully isolates sufficient identity-related semantics into the identity subspace, allowing accurate retrieval without reliance on clothing context. More notably, in the Clothes-Conflict scenario, where conventional global-alignment methods typically suffer from drastic degradation due to misaligned semantics, our method maintains a highly competitive accuracy. This robustness validates our core design: by enforcing statistical independence via the HSIC loss and intentionally discarding the decoupled visual clothing feature (f^i_{clo}) within the Mamba fusion module, the framework effectively shields the identity retrieval process from misleading semantic noise.

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

Table I shows that our method establishes new benchmarks on CUHK-PEDES and RSTPReid while remaining highly competitive on ICFG-PEDES. Notably, our ViT-based framework outperforms strong CLIP-based competitors (e.g., HAM, IRRA), challenging the dominance of large-scale pre-training in T2I-ReID. We attribute this success to two factors:

Explicit Decoupling vs. Implicit Alignment. Unlike CLIP which implicitly learns potential correlations, our BDAM mathematically enforces the separation of clothing from identity, effectively preventing the network from overfitting to clothing shortcuts.

MLLM as a Teacher. The fine-grained descriptions generated by the MLLM provide superior semantic targets compared to the noisy, coarse-grained alignment of CLIP. This validates that task-specific structural design combined with high-quality semantic guidance can effectively surpass the benefits of massive generic pre-training.

V. Conclusion and Limitations

We proposed a framework to address clothing interference in T2I-ReID by using MLLM to guide feature decoupling. Our 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 RSTPReid. Limitations include potential MLLM-generated noise and large-scale deployment overhead.

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] Wenbo Dai, Lijing Lu, and Zhihang Li, "Diffusion-based Synthetic Data Generation for Visible-Infrared Person Re-Identification," 2025.

[5] Jialong Zuo et al., "UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity," 2024.

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

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

[8] Ding Jiang and Mang Ye, "Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval," 2023.

[9] Huadong Zhang, Shuli Cheng, and Anyu Du, "Multi-Stage Auxiliary Learning for Visible-Infrared Person Re-Identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, pp. 12032–12047, 2024.

[10] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

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

[13] 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.

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

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

[16] 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.

[17] Can Cui et al., "ProFD: Prompt-Guided Feature Disentangling for Occluded Person Re-Identification," 2024.

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

[19] Weikai Lu, Ziqian Zeng, Kehua Zhang, Haoran Li, Huiping Zhuang, Ruidong Wang, Cen Chen, and Hao Peng, "ARGUS: Defending Against Multimodal Indirect Prompt Injection via Steering Instruction-Following Behavior," 2025.

[20] Jiayu Jiang et al., "Modeling Thousands of Human Annotators for Generalizable Text-to-Image Person Re-identification," 2025.

[21] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *International conference on algorithmic learning theory*. Springer, 2005, pp. 63–77.

[22] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[23] 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.

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

[25] Yang Qin et al., "Noisy-Correspondence Learning for Text-to-Image Person Re-identification," 2024.

[26] Yanwei Zheng et al., "CPCL: Cross-Modal Prototypical Contrastive Learning for Weakly Supervised Text-based Person Re-Identification," 2024.

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

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

[29] Xiujun Shu et al., "See Finer, See More: Implicit Modality Alignment for Text-based Person Retrieval," 2022.

[30] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1970–1979.

[31] Aichun Zhu et al., "DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval," 2021.