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

Anonymous ICME submission

Abstract—**Text-to-Image Person Re-Identification is critically hampered by clothing-induced interference and a persistent modality gap. We propose a novel framework centered on feature decoupling guided by a Multimodal Large Language Model. We first employ this model to automatically generate separate identity and clothing descriptions for fine-grained supervision. These descriptions guide our core component, the Bidirectional Decoupled Alignment Module, which disentangles visual features into identity and clothing subspaces. A multi-task loss strategy, comprising an alignment loss and a kernel-based orthogonal constraint, enforces this separation. Furthermore, we pioneer the integration of the Mamba state space model as an efficient fusion module, leveraging its linear-time complexity to model long-range dependencies. Comprehensive experiments on benchmark datasets demonstrate our method's superior performance and robustness.**

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]. It 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.

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 [4], but struggled with high intra-class and low inter-class variance. To overcome this, subsequent studies introduced feature disentanglement [5]. These are broadly explicit, using auxiliary modules for part-alignment [6], or implicit, using regularizers to associate noun phrases with regions [7]. This progression highlights that distinguishing identity-relevant from irrelevant semantics is essential for advancing T2I-ReID.

This pursuit of disentanglement has been propelled by powerful backbones. Models employing ViT [8] capture fine-grained details, while methods leveraging CLIP [9] learn a well-aligned joint space. Despite these advances, a critical limitation persists: both lines of work commonly treat the textual description holistically. This overlooks the semantic
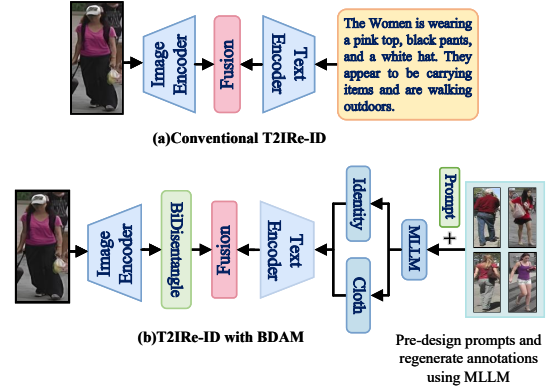

Fig. 1. Comparison of ReID methods. (a) Traditional: Direct fusion of image and text features without distinguishing identity from non-identity information limits alignment. (b) Proposed (BDAM): Introduces a decoupling module that aligns separated identity/clothing features with MLLM-generated descriptions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

distinction between content relevant to identity (e.g., gender, body shape) and content irrelevant to identity (e.g., clothing, hairstyle). 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. Inspired by the style-clustering paradigm [10], our approach employs Multimodal Large Language Model (MLLM) to guide feature decoupling by automatically generating 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-task loss strategy, including an alignment loss and an HSIC-based orthogonality constraint. Furthermore, we pioneer the integration of the Mamba state space model (SSM) as an efficient fusion module, adept at capturing long-range cross-modal dependencies with linear complexity.

Our main contributions are threefold: (1) An automatic prompt construction pipeline that combines style clustering with an MLLM to produce fine-grained, decoupled identity

and clothing descriptions; (2) The BDAM, which achieves precise decoupling and alignment reinforced by a multi-task loss strategy combining an alignment loss and an orthogonality constraint based on HSIC; and (3) The novel integration of a Mamba SSM fusion module that models long-range cross-modal dependencies with linear complexity.

II. RELATED WORK

Feature Disentanglement. Prior work in ReID [11] often relies on implicit regularizers or external detectors, which limits scalability and can lead to spurious separations. Our framework addresses this by pairing model-level disentanglement with explicit semantic supervision from MLLM-generated descriptions, ensuring controllable separation.

Feature Fusion. Existing fusion approaches are often hindered by the quadratic complexity of Transformers [12] or the holistic, cue-blurring nature of CLIP [13]. This motivates our design: our BDAM supplies factor-aware representations, which are then efficiently modeled by a Mamba SSM fusion module that captures long-range dependencies with linear complexity.

III. METHOD

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 ($\mathcal{L}_{\text{aln}}$) to supervise the visual clothing features using clothing descriptions, and an HSIC-based loss to ensure orthogonality. Finally, the Mamba SSM efficiently fuses the disentangled visual identity and textual semantics, enhancing the model's overall representation.

B. Bidirectional Decoupled Alignment Module

Some studies directly adopt CLIP [14] as a feature extractor for both modalities, aligning global embeddings for retrieval or discrimination. However, this approach presents two key limitations. First, its limited capacity for fine-grained semantics hinders the separation of identity from clothing. Second, its holistic encoding of images and text lacks the modeling of cross-modal structure at the token level, which reduces robustness in complex scenes.

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

are followed by multi-layer self-attention in each branch to enhance local consistency and contextual awareness.

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

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

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

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

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

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

Here, $f_{\text{id}}^i \in \mathbb{R}^{B \times D}$ and $f_{\text{clo}}^i \in \mathbb{R}{B \times D}$ are the gated identity and clothing features, respectively. $K_{\text{id}} = f_{\text{id}}^i (f_{\text{id}}^i)^T$ and $K_{\text{clo}} = f_{\text{clo}}^i (f_{\text{clo}}^i)^T$ are their respective kernel matrices. $H = I_N - (1/N)\mathbf{1}_N \mathbf{1}_N^T$ is the centering matrix, where I_N is the N-dimensional identity matrix and $\mathbf{1}_N$ is a column vector of all ones. HSIC measures the statistical dependence between features by calculating the mean trace of the product of their kernel matrices and the centering matrix. By minimizing
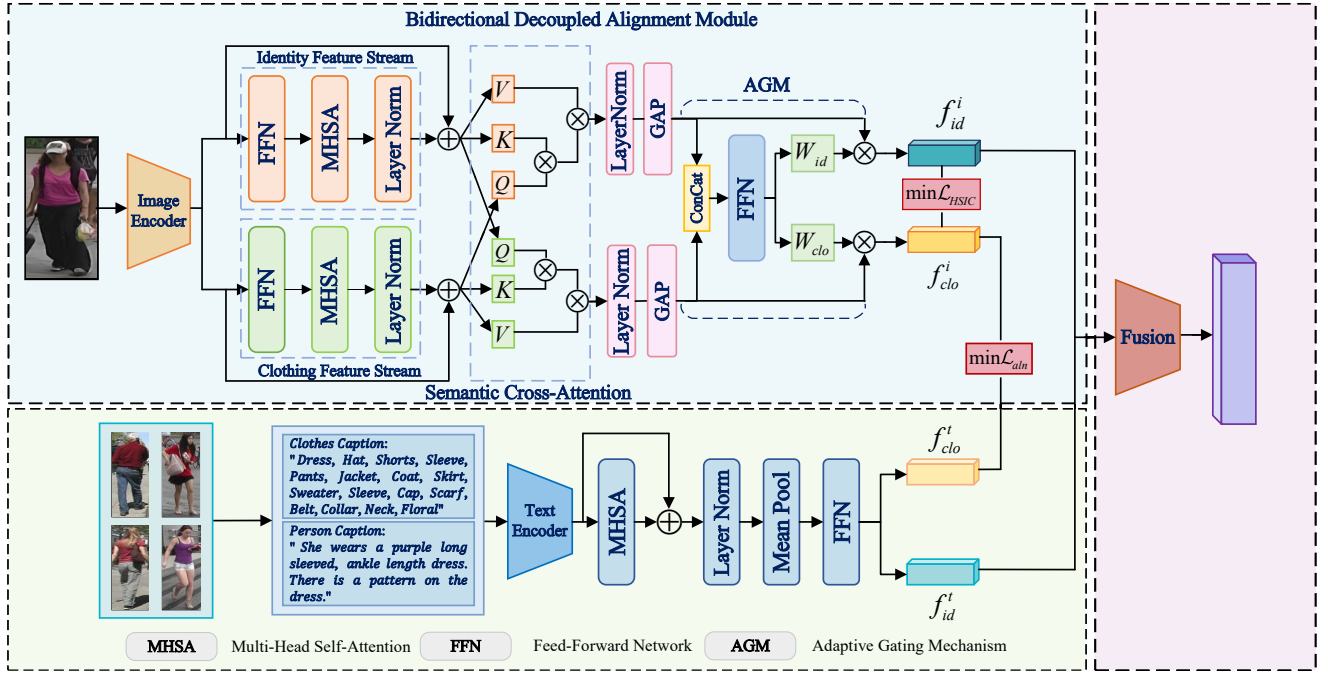
Fig. 2. First, we use an MLLM with predefined prompts to generate identity-related and clothing-related descriptions from the pedestrian image, which are encoded as f_{clo}^t and f_{id}^t, respectively. Then, the input pedestrian image is encoded into a visual feature f_i, which still resides in an entangled feature space. Subsequently, the BDAM module, composed of a dual-branch attention mechanism, decouples f_i into identity feature f_{id} and non-identity feature f_{clo}. The decoupling process is supervised and optimized through disentanglement loss and corresponding textual descriptions via contrastive learning. Finally, the fusion module integrates f_{id} and f_{id}^t to generate the final fused feature representation.

this value, the loss encourages the features to be statistically independent.

C. Semantic Enhancement

We employ an MLLM to automatically generate fine-grained identity and clothing descriptions for pedestrians. This approach reduces the burden of manual annotation and enriches the available supervision. Fig. 3 illustrates this generation pipeline. Prior work, notably HAM [10], shows that modeling annotator styles can steer an MLLM to produce diverse texts. Adapting this core insight, we extend the pipeline to meet our model's design goals.

We first use the CLIP text encoder to embed the original descriptions into vectors of a fixed dimension. Using prompts, an MLLM generalizes and substitutes entity attributes to emphasize expression style rather than content. We then cluster these style embeddings with DBSCAN, which adaptively discovers dense regions without predefining the cluster count. To stabilize the clusters, we reassign noise points and merge small clusters. These discovered style categories are then used to formulate textual prompts, such as "Use a very detailed, descriptive style," which guide the MLLM's generation tone. This setup aligns the learned style categories with the identity and clothing disentanglement expected by BDAM.

A dual prompt generator, using content-specific templates (e.g., "Describe the person's identity" and "Describe the person's clothing"), guides the MLLM to output two distinct texts
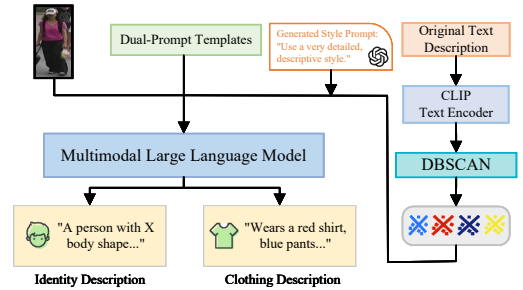

Fig. 3. Overview of the offline pipeline using an MLLM to generate decoupled identity and clothing descriptions. Style prompts are derived via CLIP and DBSCAN to enhance diversity.

per image: one description for identity, covering biological traits, and another for clothing, detailing apparel, colors, and patterns. We control the generation process with length and temperature constraints. We also apply syntax checks and validation for attribute coverage to ensure the outputs remain grammatical, structured, and parsable.

D. Feature Fusion

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

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

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

E. Loss Function

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

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

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

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

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

Here, f_a, f_p, f_n are decoupled visual identity features of the anchor, positive, and negative samples, respectively; $\|\cdot\|_2$ denotes the L2 norm; and m is the margin parameter used to enforce a minimum distance gap between positive and negative pairs.

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

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

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

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

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

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

IV. EXPERIMENTS

A. Implementation Details

1) Datasets and Metrics: We evaluate our method on three standard benchmarks: CUHK-PEDES [23], 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).

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 [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 [15]	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 [16]	CLIP-ViT	74.46	90.19	94.01	-	64.72	81.35	86.31	-	61.49	82.26	89.23	-
Propot [17]	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 [18]	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
Methods with ViT backbone:													
CPCL [19]	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 [20]	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 [21]	ViT	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
IVT [22]	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 and Efficiency Analysis

Gate Weight Analysis. We analyze the learned gate weights to validate our design. The BDAM identity weights (W_{id}) exhibit a mean of 0.61 ($\sigma = 0.17$), significantly exceeding clothing weights ($W_{clothing} = 0.38$). This asymmetry confirms that identity-relevant features dominate the representation. Conversely, a fusion module maintains balanced weights (means of 0.52 and 0.48) indicating stable cross-modal alignment without modality collapse. Detailed distributions are provided in the supplementary material.

C. Ablation Study

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

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 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	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

2) 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.

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

Table I compares our method against state-of-the-art approaches. On CUHK-PEDES and RSTPReid, 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. Notably, this high performance is achieved without leveraging large-scale pre-trained vision-language models, validating that our explicit identity-clothing disentanglement and efficient Mamba fusion provide superior cross-modal alignment through task-specific design.

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 RSTPReid. 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, 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*, vol. 132, pp. 104658, 2023.

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

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

[5] 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.

[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, 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.

[9] 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.

[10] 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.

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

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

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

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

[22] 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.

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

[24] 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.