GEAM: Graph-Enhanced Attribute-Aware Modeling for Cloth-Changing Person Re-Identification

Yongkang Ding[1], **Zi Ye**[1], **Tiantian Gong**[2], **Liyan Zhang**[1*]
[1]College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[2]School of Information and Artificial Intelligence
Yangzhou University, Yangzhou, China
{ykding,yuuuileaf,zhangliyan}@nuaa.edu.cn
gongtiantian@yzu.edu.cn

Abstract

Cloth-changing person re-identification aims to retrieve images of the same person across cameras when clothing may vary over time. Unlike conventional Re-ID, appearance cues such as clothing color and texture become highly unreliable in cloth-changing scenarios. Existing methods either rely on auxiliary clothing-invariant biometric cues that may introduce estimation noise, or model attribute semantics as isolated labels, making it difficult to simultaneously achieve robustness to attribute noise and fine-grained visual–semantic interaction. To address these issues, we propose a Graph-Enhanced Attribute-Aware Modeling framework, termed GEAM. Specifically, we first preprocess the attribute representation to explicitly suppress clothing-related semantics. We then introduce an adaptive attribute graph to model the latent dependencies among identity-related attributes, thereby alleviating the interference caused by attribute prediction errors. Based on the refined attribute representation, we further map the enhanced attribute semantics into compact semantic tokens and inject them into the visual backbone in a hierarchical manner, enabling local visual features to adaptively absorb identity-relevant yet clothing-irrelevant discriminative cues. This design effectively improves robustness to cloth-changing scenarios while preserving the stability of pretrained visual representations as much as possible. Extensive experiments on multiple mainstream CC-ReID benchmarks demonstrate that GEAM consistently outperforms a variety of state-of-the-art methods under the challenging clothes-changing setting. The source code will be available at `https://github.com/ykding666/GEAM`.

1 Introduction

Person re-identification (Re-ID)[1, 2, 3, 4, 5] aims to retrieve images of the same individual across non-overlapping camera networks and is one of the fundamental tasks in intelligent video analysis. Most existing Re-ID methods are developed under a short-term observation assumption, i.e., they implicitly assume that a person wears the same clothes during cross-camera matching. However, in real-world long-term surveillance scenarios, pedestrians often reappear after days or even weeks, usually accompanied by noticeable clothing changes[6, 7, 8]. Under such circumstances, conventional appearance cues based on color, texture, and clothing style become highly unreliable. As illustrated in Fig. 1(a), the same identity may exhibit drastically different appearances after changing clothes, while different identities may be easily confused due to similar clothing, making cloth-changing person re-identification (CC-ReID) a significantly more challenging task.

*Corresponding author: Liyan Zhang.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

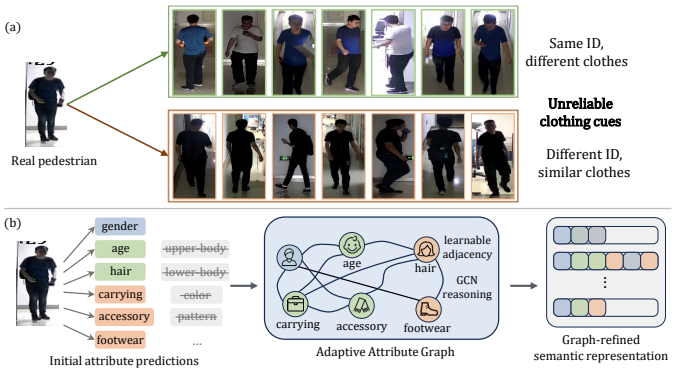


Figure 1: Motivation of the proposed GEAM. (a) Clothing cues are unreliable in CC-ReID, where the same identity may wear different clothes and different identities may share similar clothing. (b) GEAM suppresses clothing-related attributes and refines identity-related semantics through an adaptive attribute graph, providing more robust cues for cloth-changing identity matching.

Existing CC-ReID methods can be broadly divided into two categories. The first category attempts to exploit clothing-invariant biometric cues, such as body shape, pose structure, skeletal topology, or gait information, to improve robustness against clothing variation[9, 10, 11, 12, 13, 14]. However, such methods usually rely on additional structural estimators and may discard fine-grained visual details that are still beneficial for identity discrimination while suppressing clothing-related interference. The second category introduces textual descriptions or attribute semantics to mitigate the influence of clothing changes through semantic supervision[15, 16, 17, 18, 19]. Compared with purely visual de-clothing modeling, this line of research can preserve the representational capability of the visual backbone while providing more interpretable auxiliary information for identity modeling. Nevertheless, existing semantic-based methods typically treat attributes as independent discrete labels and fuse them with visual features via simple concatenation or global alignment. As a result, they fail to capture the intrinsic dependencies among attributes and lack the ability to interact with local visual regions in a fine-grained manner. Meanwhile, attribute prediction is inevitably subject to false positives, false negatives, and noise, which further limits the stability of the semantic branch in complex scenarios.

Motivated by the above observations, we propose a Graph-Enhanced Attribute-Aware Modeling framework, namely **GEAM**, for CC-ReID. The core idea is straightforward: instead of directly using raw attributes and naively injecting them into visual features, we first filter, refine, and structurally model the attribute semantics, and then integrate them into the visual backbone in a more stable manner. Specifically, on the one hand, we explicitly suppress attributes that are strongly correlated with clothing variation and retain only more stable semantics that are beneficial for identity discrimination. We further employ an adaptive attribute graph to model the latent correlations among attributes, enabling the model to compensate for and refine unreliable predictions through attribute co-occurrence relationships, thereby producing more robust semantic representations. On the other hand, as shown in Fig. 1(b), we map the enhanced attribute semantics into compact semantic tokens and inject them into the visual backbone hierarchically, allowing local visual regions to selectively absorb identity-relevant and clothing-irrelevant discriminative cues on demand. This design avoids the representational disturbance caused by directly concatenating semantic tokens with the visual sequence, while preserving the original spatial modeling mechanism of the visual backbone. Consequently, more effective visual–semantic collaborative modeling can be achieved without disrupting the distribution of pretrained visual features.

The main contributions of this work are summarized as follows:

1. We propose a graph-enhanced attribute modeling framework that improves the robustness of attribute semantics by explicitly suppressing clothing-related attributes and modeling the dependencies among identity-related attributes.
2. We propose a hierarchical semantic injection mechanism that enables enhanced attribute semantics to interact with visual features in a more fine-grained manner, while minimizing disturbance to the pretrained visual backbone as much as possible.

3. We validate the effectiveness of the proposed method on multiple mainstream CC-ReID benchmarks, and the results demonstrate that it achieves consistent and significant performance gains, particularly under the more challenging cloth-changing retrieval settings.

2 Related Work

Cloth-Changing Person Re-Identification. Existing studies on cloth-changing person re-identification (CC-ReID) mainly mitigate clothing variations through visual representation debiasing, structural cue exploitation, and human part modeling[20, 21, 22, 23, 24, 25, 26]. RGB-based methods avoid relying on extra modalities and instead reduce the dependence on clothing color and texture by feature disentanglement, causal debiasing, or adversarial learning[27, 28, 21, 20]. For example, CAL[27] introduces a clothing classifier for adversarial training to suppress clothing-discriminative information, while AIM models clothing bias from the perspective of causal intervention and weakens its influence on identity representations with a dual-branch architecture. Beyond RGB debiasing, structure-assisted methods exploit relatively stable biometric cues such as pose, silhouette, gait, or 3D body shape[9, 10, 11, 12, 14], where 3DSL[11] reconstructs 3D human shape to obtain representations more robust to texture changes. Human parsing and part-based modeling methods further introduce local region priors to separate clothing-related and clothing-irrelevant features, as in DCR-ReID[29]. These methods have advanced CC-ReID by reducing the dominance of clothing appearance from different perspectives. However, RGB-based debiasing is still easily affected by clothing-dominated visual signals, while structure- and parsing-based methods may suffer from errors introduced by additional estimators. Moreover, excessively suppressing clothing appearance may discard fine-grained visual cues that remain useful for identity discrimination. In contrast, this work introduces structured attribute semantics as complementary information while preserving the discriminative capability of the visual backbone, thereby improving identity robustness in cloth-changing scenarios.

Semantic Attribute Modeling for Re-Identification. Different from directly removing clothing appearance or relying on additional structural estimation, semantic attributes provide lightweight and interpretable auxiliary information for CC-ReID[15, 16, 17, 18, 19]. Early attribute-assisted Re-ID methods use attributes such as gender, age, hairstyle, body shape, and carried objects as additional supervision, or fuse them with visual features to enhance identity representations. With the development of vision–language pretrained models, semantic prompts, textual descriptions, and cross-modal alignment have also been introduced into Re-ID[30, 31, 32]. For example, CLIP-ReID[30] employs textual prompts to provide semantic constraints, MADE[15] converts pedestrian attributes or anatomical semantics into semantic tokens, and MTAT[16] learns clothing-invariant features by masking clothing-related words and performing cross-modal adversarial learning. Despite their effectiveness, most existing methods still model attributes or textual semantics as independent prompts, without explicitly characterizing dependencies among attributes, making them vulnerable to noisy predictions such as false positives and false negatives. Meanwhile, common fusion strategies such as feature concatenation, token concatenation, or single-layer cross-modal alignment are insufficient for fine-grained interactions between hierarchical visual regions and identity-related semantics. To address these limitations, GEAM models inter-attribute relationships through an adaptive attribute graph and integrates enhanced semantic tokens into the visual backbone via hierarchical semantic injection, enabling more stable visual–semantic collaborative modeling.

3 Method

3.1 Overall Framework

We propose a **Graph-Enhanced Attribute-Aware Modeling** framework, termed **GEAM**, to improve the robustness of identity representations in cloth-changing scenarios by jointly leveraging attribute enhancement and hierarchical semantic injection. The overall pipeline is illustrated in Fig. 2. Given an input pedestrian image $x \in \mathbb{R}^{H\times W\times 3}$, the model adopts a dual-branch architecture consisting of a visual branch and an attribute branch.

In the visual branch, the input image is first transformed by Patch Embedding into a visual sequence composed of one `[CLS]` token and multiple patch tokens, which is then fed into a visual Transformer backbone. To enable hierarchical interaction, we divide the backbone into three consecutive stages, namely *Stage 1*, *Stage 2*, and *Stage 3*, and insert two attribute injectors between Stage 1 and Stage 2

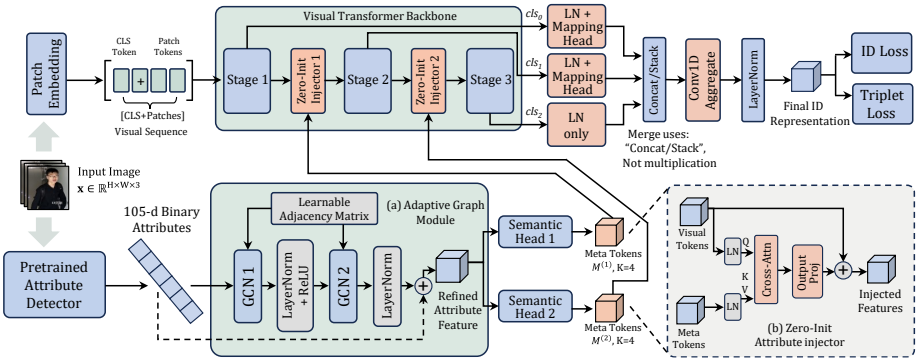


Figure 2: The refined semantics are mapped into semantic tokens and injected into different visual stages via gated residual attribute injectors.

and between Stage 2 and Stage 3, respectively. Semantic tokens are not directly concatenated with the visual sequence; instead, they are integrated into visual features in a residual manner through the injectors, thereby preserving the original representation distribution of the pretrained backbone as much as possible.

In the attribute branch, the input image is first processed by a pretrained attribute detector to obtain a 105-dimensional binary attribute vector. Since this vector contains many clothing-related attributes that are highly variable over time, we first perform explicit filtering and random perturbation on the raw attributes, and then model the dependencies among attributes through an adaptive attribute graph to obtain more robust semantic representations. Subsequently, two independent semantic mapping heads are employed to project the enhanced attribute representation into two groups of semantic tokens, which are used for the two injection stages, respectively.

At the end of the backbone, the model extracts the `[CLS]` representations from the three stages, denoted as cls_0, cls_1, and cls_2, respectively. The outputs from the first two stages are first passed through lightweight mapping modules, and are then concatenated and aggregated together with the last-stage representation. A 1D convolution followed by layer normalization is applied to obtain the final identity representation f, which is used for identity classification and retrieval learning. Overall, the attribute branch is responsible for producing stable identity-aware semantics, while the visual branch models local appearance structures, and the two branches are coupled through hierarchical injection for fine-grained collaboration.

3.2 Graph-Enhanced Attribute Modeling

A key challenge in CC-ReID lies in the fact that raw attribute predictions not only contain substantial clothing-related interference, but also inevitably suffer from false positives and false negatives. To address this issue, we enhance the raw attributes through three steps, namely attribute preprocessing, graph-based relational reasoning, and semantic expansion.

Let the binary attribute vector predicted by the attribute detector be $a \in \{0,1\}^D$, where $D = 105$. We divide the attribute indices into a clothing-related set Ω_{cloth} and an identity-related set Ω_{id}, and explicitly suppress the clothing-related dimensions. For identity-related attributes, random masking is further applied during training to simulate the uncertainty of attribute prediction in complex scenarios. The preprocessed attribute representation is defined as

$$z_i = \begin{cases} 0, & i \in \Omega_{\mathrm{cloth}}, \\ a_i r_i, & i \in \Omega_{\mathrm{id}}, \end{cases} \tag{1}$$

where $r_i \in \{0,1\}$ denotes a random retention mask. When random perturbation is enabled, the selected identity-related dimensions satisfy

$$r_i \sim \mathrm{Bernoulli}(0.5). \tag{2}$$

This design suppresses clothing-related interference at the source and improves the robustness of the model to attribute noise.

Based on the preprocessed attributes, we further model the latent dependencies among attributes using a graph structure. For a batch of attribute representations, let the initial input be $Z^{(0)} \in \mathbb{R}^{B\times D}$. We treat each attribute dimension as a node in a graph and introduce a learnable adjacency matrix $A \in \mathbb{R}^{D\times D}$ to characterize the connection strength between attributes. Two graph-based attribute propagation layers are then applied for relational reasoning:

$$Z^{(l+1)} = \sigma\left(Z^{(l)} A W^{(l)}\right), \quad l = 0, 1, \tag{3}$$

where $W^{(l)}$ denotes a learnable transformation matrix in the l-th propagation layer , and $\sigma(\cdot)$ is a nonlinear activation function. Through graph propagation, each attribute node no longer depends solely on its own observation, but can absorb complementary information from correlated attributes, thereby alleviating the adverse effects caused by unreliable local attribute predictions.

To stabilize the training process, we adopt a residual graph enhancement strategy, in which the graph reasoning result is used as a progressive correction to the original attribute representation. Let ΔZ denote the output after the two graph convolution layers. The refined attribute representation is formulated as

$$Z_{\text{ref}} = Z^{(0)} + \gamma_{\text{gcn}} \Delta Z, \tag{4}$$

where γ_{gcn} is a learnable scaling coefficient initialized to zero at the beginning of training. This strategy enables the model to transition smoothly from the original attribute representation to the graph-enhanced one.

Although Z_{ref} contains more robust identity-aware semantics, it remains a compact global attribute vector and is therefore insufficient for direct interaction with a long visual sequence. To alleviate the semantic bandwidth mismatch, we use two independent semantic mapping heads to expand it into two groups of semantic tokens:

$$M^{(1)} = f^{(1)}_{\text{sem}}(Z_{\text{ref}}) \in \mathbb{R}^{B\times K\times C}, \tag{5}$$

$$M^{(2)} = f^{(2)}_{\text{sem}}(Z_{\text{ref}}) \in \mathbb{R}^{B\times K\times C}, \tag{6}$$

where $f^{(1)}_{\text{sem}}(\cdot)$ and $f^{(2)}_{\text{sem}}(\cdot)$ denote two independent semantic mapping functions, K is the number of semantic tokens, and C is the channel dimension of the visual backbone. In this work, we set $K = 4$. The two groups of semantic tokens are used for the subsequent two injection stages, respectively, so as to accommodate the semantic requirements of visual features at different depths.

3.3 Hierarchical Semantic Injection

In the cross-modal fusion stage, existing methods often directly merge semantic information into the visual sequence through token concatenation. However, this strategy not only changes the input structure of the pretrained visual backbone, but also fails to support fine-grained interaction between local visual regions and semantic information. To address this issue, we adopt a hierarchical semantic injection mechanism, in which the semantic branch is integrated into the visual backbone in a *visual-querying-semantic* manner.

Let the visual feature sequence at the current stage be $X_v \in \mathbb{R}^{B\times N\times C}$, and let the corresponding semantic tokens be $M^{(s)} \in \mathbb{R}^{B\times K\times C}$, where $s \in \{1, 2\}$ denotes the injection stage. We first apply layer normalization to both the visual features and the semantic tokens, and then project them into query, key, and value representations:

$$Q = \text{LN}(X_v)W_Q, \quad K = \text{LN}(M^{(s)})W_K, \quad V = \text{LN}(M^{(s)})W_V, \tag{7}$$

where W_Q, W_K, and W_V are learnable projection matrices. The visual–semantic interaction is then realized via scaled dot-product attention:

$$X_{\text{attn}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V, \tag{8}$$

where d is the dimension of each attention head. This design allows each local token in the visual sequence to adaptively select the most relevant identity cues from multiple semantic tokens according to its own context, thereby enabling more fine-grained cross-modal modeling.

To prevent the semantic branch from excessively perturbing pretrained visual features during the early stage of fine-tuning, we adopt a gated residual injection form:

$$X_{\text{out}} = X_v + \gamma_{\text{inj}}(X_{\text{attn}} W_O), \tag{9}$$

where W_O is the output projection matrix and γ_{inj} is a learnable gating coefficient. The injector is initialized to produce a near-zero residual response, so that it behaves approximately as an identity mapping at the beginning of training. In this way, the model can first optimize along the original visual manifold and then gradually incorporate semantic enhancement.

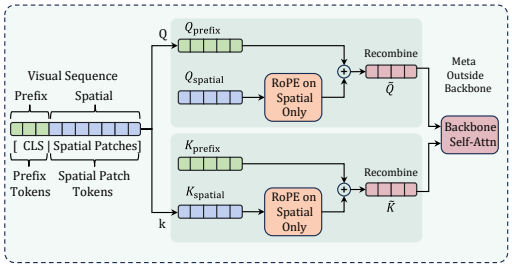


Figure 3: Visualizing Positional Decoupling.

As illustrated in Fig. 2, we introduce two independently parameterized attribute injectors between Stage 1 and Stage 2 and between Stage 2 and Stage 3, respectively, using $M^{(1)}$ and $M^{(2)}$ for the two hierarchical injection steps. Compared with single-stage fusion, hierarchical injection not only supplements shallow features with semantic information related to local patterns, but also further strengthens identity discrimination at deeper layers, which is more consistent with the hierarchical abstraction process of visual representations from local details to global semantics.

In addition, to reduce the interference of semantic injection with the spatial modeling mechanism of the visual backbone, we always keep the backbone input as a purely visual sequence and never directly concatenate semantic tokens into the backbone. Moreover, rotary positional embedding (RoPE) is applied only to spatial patch tokens with explicit 2D coordinates, while non-spatial tokens such as the `[CLS]` token are excluded from RoPE. As shown in Fig. 3, we denote this position-handling strategy as *Positional Decoupling (PD)*, which avoids assigning artificial spatial positions to non-spatial tokens and helps maintain stable spatial modeling in the backbone.

At the end of the backbone, we extract the `[CLS]` representations from the three stages, denoted as cls_0, cls_1, and cls_2, respectively. The outputs from the first two stages are first processed by layer normalization and lightweight mapping heads, and are then concatenated with the last-stage representation for aggregation. The final identity representation is formulated as

$$f = \text{LN}\Big(\text{Conv1D}\big([\phi_0(\text{cls}_0), \phi_1(\text{cls}_1), \phi_2(\text{cls}_2)]\big)\Big), \tag{10}$$

where $\phi_0(\cdot)$ and $\phi_1(\cdot)$ denote *layer normalization* + *mapping head*, and $\phi_2(\cdot)$ denotes layer normalization. This aggregation strategy integrates visual and semantic information from different stages, thereby facilitating a more robust final identity representation.

3.4 Training Loss

We optimize GEAM with an identity classification loss and a triplet loss:

$$L = \lambda_1 L_{\text{id}} + \lambda_2 L_{\text{tri}}, \tag{11}$$

where λ_1 and λ_2 are the loss weights. The identity classification loss is defined as

$$L_{\text{id}} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{i,y_i}, \tag{12}$$

where N is the batch size, y_i denotes the identity label of the i-th sample, and p_{i,y_i} is the predicted probability for the ground-truth class. To improve intra-class compactness and inter-class separability, we further adopt the triplet loss:

$$L_{\text{tri}} = \frac{1}{N} \sum_{i=1}^{N} \left[d(f_i, f_i^+) - d(f_i, f_i^-) + \alpha \right]_+, \tag{13}$$

where f_i, f_i^+, and f_i^- denote the anchor, positive, and negative features, respectively; $d(\cdot, \cdot)$ is the distance metric; α is the margin; and $[\cdot]_+ = \max(\cdot, 0)$.

Table 1: Comparison With State-of-the-art Methods On PRCC, LTCC and LaST.

	Methods	Venue	PRCC				LTCC				LaST	
			SC		CC		General		CC			
			R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
General	PCB[38]	ECCV18	99.8	97.0	41.8	38.7	65.1	30.6	23.5	10.0	-	-
	IANet[39]	CVPR19	99.4	98.3	46.3	45.9	63.7	31.0	25.0	12.6	-	-
	ISP[40]	ECCV20	92.8	-	36.6	-	-	-	27.8	11.0	-	-
	TransReID[2]	ICCV21	97.3	95.9	47.1	49.3	73.8	39.4	35.7	18.8	31.3	18.6
CC-ReID	CAL[27]	CVPR22	100	99.8	55.2	55.8	74.2	40.8	40.1	18.0	73.7	28.2
	AIM[28]	CVPR23	100	**99.9**	57.9	58.3	76.3	41.1	40.6	19.1	-	-
	DCR-ReID[29]	TCSVT23	100	99.7	57.2	57.4	76.1	42.3	41.1	20.4	-	-
	CLIP3DReID[13]	CVPR24	-	-	60.6	59.3	-	-	42.1	21.7	-	-
	CCPG[22]	CVPR24	100	99.6	61.8	58.3	77.2	42.9	46.2	22.9	-	-
	MCSC[24]	TIP24	99.8	99.8	57.8	57.3	73.9	40.2	42.2	19.4	75.4	31.0
	MADE[15]	TMM24	100	98.6	64.3	59.1	84.2	48.2	47.4	24.4	79.0	40.9
	PFAC[23]	MM25	-	-	55.6	56.7	-	-	49.8	26.2	-	-
	CaPu[18]	IJCV25	99.9	98.7	57.9	57.2	76.5	43.1	41.3	20.5	-	-
	MTAT[16]	TIFS25	-	-	-	-	81.5	45.0	48.0	22.3	-	-
	CTVNet[17]	TITS25	-	-	57.1	57.3	-	-	42.1	18.4	-	-
	HPRNET[7]	TCSVT25	100	99.6	62.3	60.1	77.1	42.6	45.9	19.2	-	-
Ours	GEAM	-	**100**	99.5	**66.0**	**61.6**	**83.8**	**51.4**	**51.5**	**28.7**	**81.8**	**45.2**

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed method on five mainstream CC-ReID benchmarks, including PRCC[33], LTCC[34], LaST[35], Celeb-reID-light[36], and Celeb-reID[37]. PRCC and LTCC are the most commonly used cloth-changing person re-identification datasets, where PRCC focuses on cross-view and cloth-changing retrieval in relatively controlled scenarios, while LTCC features a longer temporal span and more frequent clothing changes. LaST is a large-scale long-term spatio-temporal person re-identification benchmark that more closely reflects real-world long-term surveillance scenarios. Celeb-reID-light and Celeb-reID further cover richer appearance variations in open environments, and are thus suitable for evaluating the cross-scene generalization ability of the model.

Evaluation Protocols. We follow the official evaluation protocols of each dataset and adopt Rank-1 and mAP as the evaluation metrics throughout. According to the settings of different datasets, we mainly report results under three evaluation settings, namely the *general setting*, the *same-clothes setting (SC)*, and the *clothes-changing setting (CC)*, in order to assess the performance of the proposed method under different retrieval difficulties.

Implementation Details. All experiments are implemented in PyTorch and conducted on two NVIDIA RTX 3090 GPUs. We employ a Transformer-based visual extractor [41] for pedestrian image representation, and resize all input images to 224×224. In the attribute branch, following SOLIDER [42], each image is assigned a frozen 105-dimensional binary attribute vector, which serves as the fixed input to the graph-enhanced attribute modeling module. During training, we use the SGD optimizer and train the model for 20 epochs. The initial learning rate is set to 2×10^{-5}, with a linear warmup strategy applied in the early training stage, and the weight decay is set to 0.05. As described in Section 3, the model is jointly supervised by the identity classification loss and the triplet loss, where the margin of the triplet loss is set to $\alpha = 0.3$. Additional implementation details are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

We further compare GEAM with existing general Re-ID methods and advanced CC-ReID methods. The corresponding results are summarized in Tables 1 and 2. Overall, GEAM achieves either the best or highly competitive performance on multiple datasets, with its advantage being particularly

evident under the more challenging clothes-changing setting. These results indicate that the proposed method can more effectively suppress the interference caused by clothing changes and learn more robust identity representations.

On the two representative cloth-changing benchmarks, namely PRCC and LTCC, GEAM achieves the best results under the clothes-changing setting. Specifically, on PRCC-CC, GEAM achieves 66.0% Rank-1 and 61.6% mAP, outperforming the second-best result in Table 1 by 1.7 and 1.5 percentage points, respectively. On the more challenging LTCC-CC benchmark, GEAM obtains 51.5% Rank-1 and 28.7% mAP, surpassing the second-best method by 1.7 and 2.5 percentage points, respectively. Considering that LTCC involves more frequent clothing changes, longer time intervals, and more complex background interference, these results demonstrate that GEAM exhibits stronger robustness in long-term cloth-changing scenarios. Meanwhile, GEAM also remains competitive under more conventional evaluation settings, such as the same-clothes and general settings, indicating that the performance gains of the proposed method are not achieved at the expense of retrieval performance in standard scenarios.

Table 2: Comparison With SOTA Methods.

Method	Venue	Celeb-reID-light		Celeb-reID	
		rank-1	mAP	rank-1	mAP
CLIP3DReID[13]	CVPR24	39.4	26.3	63.1	26.3
MCSC[24]	TIP24	29.2	16.2	57.8	13.0
MADE[15]	TMM24	72.0	52.3	-	-
ZUT-ResNet[43]	AAAI25	-	-	53.3	-
CSFRNet[25]	IJCV25	48.1	26.2	68.0	19.7
MLDC[44]	PR25	40.6	24.2	61.4	16.9
MTAT[16]	TIFS25	71.0	51.4	74.5	36.8
GEAM	-	**74.3**	**55.4**	**79.1**	**40.2**

GEAM also demonstrates consistent advantages on larger-scale or more open benchmarks. As shown in Table 1, GEAM achieves 81.8% Rank-1 and 45.2% mAP on LaST, outperforming the second-best method by 2.8 and 4.3 percentage points, respectively. Table 2 further shows that GEAM achieves $74.3\%/55.4\%$ and $79.1\%/40.2\%$ Rank-1/mAP on Celeb-reID-light and Celeb-reID, respectively, both of which are clearly superior to existing methods. These results suggest that the performance improvement brought by GEAM is not limited to a specific dataset, but remains consistent across different scene complexities and appearance variation conditions, demonstrating favorable generalization capability.

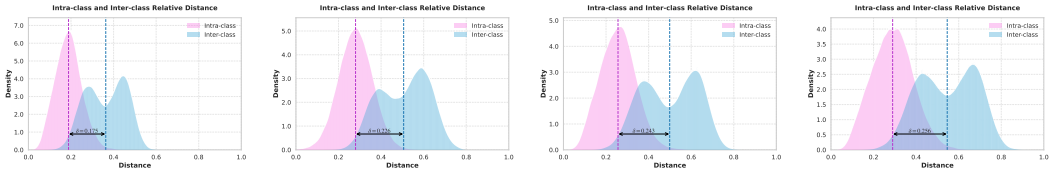
Table 3: Ablation of major components in GEAM on representative cloth-changing benchmarks.

No.	Method	AP	AAG	SI	PD	PRCC-CC		LTCC-CC	
						R-1	mAP	R-1	mAP
A1	Baseline					61.6	57.3	43.1	23.0
A2	+ PD				✓	62.0	57.8	43.9	23.6
A3	+ AP	✓			✓	63.0	58.7	46.5	24.8
A4	+ AP + AAG (w/o PD)	✓	✓			64.1	59.9	48.1	26.0
A5	+ AP + AAG	✓	✓		✓	64.4	60.3	48.9	26.5
A6	+ AP + AAG + SI (w/o PD)	✓	✓	✓		65.6	61.1	50.9	28.0
A7	Full GEAM	✓	✓	✓	✓	**66.0**	**61.6**	**51.5**	**28.7**

4.3 Ablation Study

To further analyze the contribution of each component, we conduct ablation studies on two representative cloth-changing retrieval benchmarks, namely PRCC-CC and LTCC-CC. The results are reported in Table 3. Overall, the full model achieves the best performance on both datasets. Compared with the baseline, Full GEAM improves the performance on PRCC-CC from $61.6\%/57.3\%$ to $66.0\%/61.6\%$, and on LTCC-CC from $43.1\%/23.0\%$ to $51.5\%/28.7\%$. The gain on LTCC-CC is particularly pronounced, indicating that the proposed design yields stronger overall benefits in more complex and longer-term cloth-changing scenarios.

From the perspective of the source of performance gains, the enhancement design in the attribute branch serves as the primary contributor. First, introducing only Positional Decoupling (PD) already leads to a certain improvement, suggesting that this design helps maintain the consistency of spatial modeling and provides a more stable foundation for stacking subsequent modules. Building on this, Attribute Preprocessing (AP) brings more evident gains. For example, after introducing AP on LTCC-CC, the performance improves from $43.9\%/23.6\%$ to $46.5\%/24.8\%$, demonstrating that explicitly



(a) A1 (Baseline) (b) A4 (c) A6 (d) A7 (GEAM)

Figure 4: Visualization of intra-class and inter-class relative distance distributions on LTCC-CC.

suppressing clothing-related attributes is necessary for identity modeling in cloth-changing scenarios. After further introducing the Adaptive Attribute Graph (AAG), the results on LTCC-CC continue to improve to $48.9\%/26.5\%$, indicating that modeling the latent dependencies among attributes can effectively enhance the robustness of semantic representations.

After fixing the attribute modeling part, Hierarchical Semantic Injection (SI) further brings significant improvements. As shown in Table 3, by introducing SI on top of AP and AAG, the model achieves $66.0\%/61.6\%$ and $51.5\%/28.7\%$ on PRCC-CC and LTCC-CC, respectively, indicating that injection-based fusion can more effectively translate attribute semantics into visual discriminative capability. Meanwhile, comparing Rows 4 and 5 as well as Rows 6 and 7 reveals that further incorporating PD after introducing AAG or SI consistently improves performance. For instance, on LTCC-CC, enabling PD after AAG improves the result from $48.1\%/26.0\%$ to $48.9\%/26.5\%$, while when both AAG and SI are enabled, PD further boosts the performance from $50.9\%/28.0\%$ to $51.5\%/28.7\%$. These observations indicate that Positional Decoupling not only stabilizes the spatial modeling process, but also works synergistically with attribute enhancement and semantic injection. Overall, graph-enhanced attribute modeling and hierarchical semantic injection jointly constitute the major source of performance gains, while Positional Decoupling further improves the stability of the overall framework in cloth-changing scenarios.

4.4 Visualization Analysis

To further provide an intuitive analysis of how each module affects feature discriminability, we visualize the distributions of intra-class and inter-class relative distances for four representative model variants on LTCC-CC, as shown in Fig. 4. Specifically, A1 denotes the baseline, A4 incorporates Attribute Preprocessing (AP) and the Adaptive Attribute Graph (AAG), A6 further introduces Hierarchical Semantic Injection (SI), and A7 corresponds to the full GEAM model. It can be observed that the baseline exhibits a noticeable overlap between the intra-class and inter-class distributions, indicating that identity features in cloth-changing scenarios remain highly susceptible to appearance variations. With the progressive introduction of graph-enhanced attribute modeling and semantic injection, the intra-class distribution gradually becomes more concentrated at smaller distances, while the inter-class distribution shifts overall toward larger distances. As a result, the margin δ between the two distributions consistently increases from 0.175 to 0.256. This trend is consistent with the ablation results reported in Table 3, demonstrating that GEAM effectively improves intra-class compactness and inter-class separability, thereby learning more robust identity representations for cloth-changing person re-identification.

5 Conclusion

In this paper, we addressed the challenges of unstable clothing appearance, insufficient structural modeling of attribute semantics, and inadequate visual–semantic interaction in cloth-changing person re-identification, and proposed a Graph-Enhanced Attribute-Aware Modeling framework, termed GEAM. GEAM suppresses clothing-related semantics through attribute preprocessing, models latent dependencies among identity-related attributes with an adaptive attribute graph, and injects the enhanced semantic tokens into the visual backbone hierarchically for effective visual–semantic collaboration. Extensive experiments on multiple mainstream CC-ReID benchmarks demonstrate that GEAM achieves stable and significant improvements, especially under the challenging clothes-changing setting, and ablation studies further verify the effectiveness of the proposed components. Nevertheless, GEAM relies on a pretrained attribute detector to provide initial semantic representa-

tions, and inaccurate attribute predictions under severe occlusion, low resolution, or unusual poses may affect the quality of semantic enhancement. In future work, we will explore more robust attribute estimation and noise-tolerant semantic modeling strategies to further improve the stability of GEAM in complex real-world scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62427819 and 62601699, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20230031, and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant 26CXJH1040.

References

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[2] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[3] Yongxiang Li, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Learning source-free domain adaptation for visible-infrared person re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[4] Zhiqi Pang, Lingling Zhao, Junjie Wang, and Chunyu Wang. Lvlm-driven attribute-aware modeling for visible-infrared person re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Zhongao Zhou, Bin Yang, Wenke Huang, Jun Chen, and Mang Ye. Unbiased prototype consistency learning for multi-modal and multi-task object re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[6] Yongkang Ding, Zi Ye, and Liyan Zhang. Cloth-changing person re-identification: A survey on challenges, methods, and frontier advances. 2026.

[7] Mingfu Xiong, Longlong Ge, Ruimin Hu, Khan Muhammad, Sambit Bakshi, Javier Del Ser, Xiaokang Yang, and Bin Sheng. Hprnet: Human parsing reconstruction with non-local multi-scale perception network for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[8] Mengmeng Liu, Zhi Ma, Tao Li, Yanfeng Jiang, and Kai Wang. Long-term person re-identification with dramatic appearance change: Algorithm and benchmark. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6406–6415, 2022.

[9] Yu-Jhe Li, Xinshuo Weng, and Kris M Kitani. Learning shape representations for person re-identification under clothing change. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2432–2441, 2021.

[10] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021.

[11] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8146–8155, 2021.

[12] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2022.

[13] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 256–266, 2024.

[14] Vuong D Nguyen, Pranav Mantini, and Shishir K Shah. Temporal 3d shape modeling for video-based cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 173–182, 2024.

[15] Chunlei Peng, Boyu Wang, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Masked attribute description embedding for cloth-changing person re-identification. *IEEE Transactions on Multimedia*, 27:1475–1485, 2024.

[16] Chengrui Hao, Boyu Wang, Chunlei Peng, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Masked text adversarial training for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

[17] Xiejing Yin, Hua Han, Kaiyu Xu, Li Huang, and AAM Muzahid. Causality-driven explainable multimodal fusion with visual-text parallel computing for cloth-changing pedestrian re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 27(1):460–471, 2025.

[18] Zhengwei Yang, Huilin Zhu, Nan Lei, Basura Fernando, and Zheng Wang. Clothing purification with causality meets vision-language pretraining models. *International Journal of Computer Vision*, 133(11):7748–7771, 2025.

[19] Xiyu Han, Xian Zhong, Wenxin Huang, Xuemei Jia, Xiaohan Yu, and Alex Chichung Kot. See what you seek: Semantic contextual integration for cloth-changing person re-identification. *arXiv preprint arXiv:2412.01345*, 2024.

[20] Yuming Yan, Huimin Yu, Shuzhao Li, Zhaohui Lu, Jianfeng He, Haozhuo Zhang, and Runfa Wang. Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In *IJCAI*, pages 1523–1529, 2022.

[21] Zhengwei Yang, Xian Zhong, Zhun Zhong, Hong Liu, Zheng Wang, and Shin'Ichi Satoh. Win-win by competition: Auxiliary-free cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 32:2985–2999, 2023.

[22] Vuong D Nguyen, Pranav Mantini, and Shishir K Shah. Contrastive clothing and pose generation for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7541–7549, 2024.

[23] Jiajun Zhang, Xin Li, Si Wu, Yong Xu, and Yaowei Wang. Prior-free augmentation for cloth-changing person re-identification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10612–10621, 2025.

[24] Yan Huang, Qiang Wu, Zhang Zhang, Caifeng Shan, Yi Zhong, and Liang Wang. Meta clothing status calibration for long-term person re-identification. *IEEE Transactions on Image Processing*, 2024.

[25] Yan Huang, Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Csfrnet: Integrating clothing status awareness for long-term person re-identification. *International Journal of Computer Vision*, 133(6):3180–3202, 2025.

[26] Yongkang Ding, Jiechen Li, Hao Wang, Ziang Liu, and Anqi Wang. Attention-enhanced multimodal feature fusion network for clothes-changing person re-identification. *Complex & Intelligent Systems*, 11(1):1–15, 2025.

[27] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[28] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1472–1481, 2023.

[29] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4415–4428, 2023.

[30] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

[31] Min Jiang, Xinyu Liu, and Jun Kong. Text-guided prototype generation for occluded person re-identification. *IEEE Signal Processing Letters*, 2024.

[32] Yongkang Ding, Yuqing Wu, Chenwei Wu, Meina Qu, and Liyan Zhang. Person parsing-driven and text-guided for cloth-changing person re-identification. *IEEE Internet of Things Journal*, 2025.

[33] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019.

[34] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4390–4403, 2021.

[36] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[37] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3459–3471, 2019.

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[39] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-And-Aggregation Network for Person Re-Identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9309–9318, Long Beach, CA, USA, June 2019. IEEE.

[40] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020.

[41] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[42] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[43] Wenjie Pan, Jianqing Zhu, and Huanqiang Zeng. Fair training with zero inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6317–6325, 2025.

[44] Guoqing Zhang, Jieqiong Zhou, Yuhui Zheng, Gaven Martin, and Ruili Wang. Adaptive transformer with pyramid fusion for cloth-changing person re-identification. *Pattern Recognition*, 163:111443, 2025.

A Dataset Details

We conduct experiments on five mainstream CC-ReID benchmarks, including PRCC, LTCC, Celeb-reID, Celeb-reID-light, and LaST, to comprehensively evaluate the effectiveness of the proposed GEAM under controlled cloth-changing, long-term cloth-changing, and open complex scenarios. The statistics of these datasets are summarized in Table 4.

PRCC and LTCC are two of the most commonly used datasets for cloth-changing person re-identification. PRCC contains 221 identities and 33,698 images, where each identity has both same-clothes and clothes-changing images. It is mainly used to evaluate cross-camera and cloth-changing retrieval ability in controlled surveillance scenarios. LTCC contains 152 identities and 17,138 images, where each identity is associated with 2 to 14 outfits. Since LTCC involves more frequent clothing changes and a longer temporal span, it is more suitable for evaluating the robustness of a model in long-term cloth-changing scenarios.

Celeb-reID and Celeb-reID-light are collected from celebrity images in open scenarios, containing richer variations in pose, illumination, background, and imaging conditions. Specifically, Celeb-reID contains 1,052 identities and 34,186 images, with approximately 70% of the identities undergoing clothing changes. Celeb-reID-light contains 590 identities and 10,842 images, where samples of the same identity do not share the same clothing, thereby posing a greater challenge to clothing-invariant identity representation learning.

LaST is a large-scale long-term spatio-temporal person re-identification dataset, containing 10,862 identities and 228,156 images. It involves a larger identity scale and more complex spatio-temporal variations, making it closer to real-world long-term retrieval applications. By evaluating GEAM on the above datasets, we can verify its effectiveness and robustness from multiple perspectives, including typical cloth-changing retrieval, long-term cloth-changing retrieval, and generalization to open scenarios.

Table 4: Statistics of the datasets used in the experiments.

Dataset	#IDs	#Cameras	#Samples	#Clothes for each identity	Scene
PRCC	221	3	33,698	2	Indoor
LTCC	152	12	17,138	2-14	Indoor
Celeb-reID	1052	vary	34186	70%	Outdoor
Celeb-reID-light	590	vary	10842	100%	Outdoor
LaST	10,862	vary	228,156	76%	Movie

B Loss Weight Hyperparameter Analysis

To further analyze the impact of loss weights on model performance, we investigate the variation trends of the identity classification loss weight λ_1 and the triplet loss weight λ_2 on the cloth-changing settings of LTCC and Celeb-reID-light, respectively. The results are shown in Fig. 5. Overall, the model performance fluctuates to some extent as the hyperparameters vary, but no clear monotonic trend is observed, indicating that GEAM is not extremely sensitive to the loss weights. Among the two datasets, LTCC is relatively more sensitive to hyperparameter changes. In particular, when λ_1 is fixed, the model performance generally improves as λ_2 increases, suggesting that the metric constraint provided by the triplet loss is more important in the more challenging long-term cloth-changing scenario. In contrast, the performance variation on Celeb-reID-light is much smoother, where both Rank-1 and mAP remain within a relatively small fluctuation range under different parameter settings. This indicates that the performance on this dataset is less dependent on the loss weights.

On the other hand, compared with the optimal parameter combinations, the performance degradation under the default setting $(\lambda_1, \lambda_2) = (1.0, 1.0)$ is limited, with an overall gain of less than 1% after additional tuning. Specifically, under the default setting, the model achieves 51.5% Rank-1 and 28.7% mAP on LTCC, and 74.3% Rank-1 and 55.4% mAP on Celeb-reID-light. Although additional dataset-specific hyperparameter tuning can bring slight performance improvements, the gain is relatively marginal. Considering the extra computational cost introduced by further hyperparameter

search, as well as the need to ensure consistent experimental settings across different datasets and fair comparisons with existing methods, we adopt the default setting $(\lambda_1, \lambda_2) = (1.0, 1.0)$ as the final configuration in the main experiments.

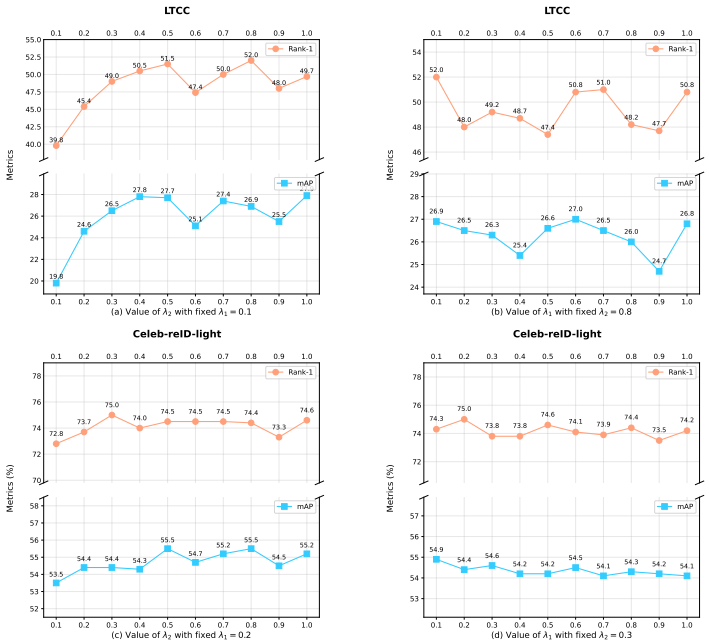


Figure 5: Hyperparameter analysis of loss weights on LTCC and Celeb-reID-light. We report Rank-1 and mAP under different values of λ_1 and λ_2. Subfigures (a) and (c) vary λ_2 with fixed λ_1, while subfigures (b) and (d) vary λ_1 with fixed λ_2.

C Supplementary Visualization Analysis

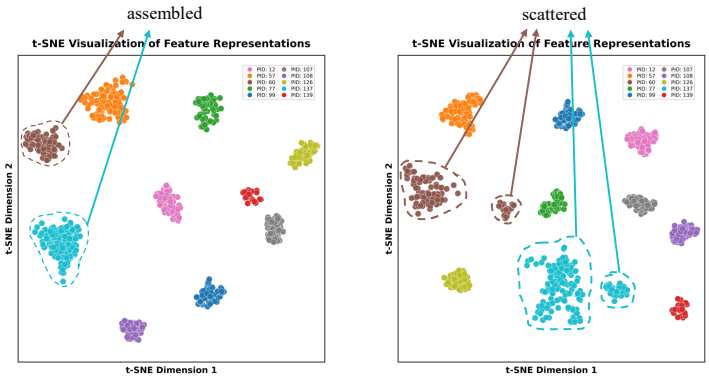
C.1 t-SNE Visualization of Feature Distributions

To further verify the ability of GEAM to learn identity-invariant representations in long-term cloth-changing scenarios, we compare the feature distributions of GEAM and MADE on the LTCC dataset using t-SNE visualization. The results are shown in Fig. 6. It should be noted that MADE is also a CC-ReID method that exploits pedestrian attribute information. Therefore, this comparison can more directly reflect the advantages of the proposed graph-enhanced attribute modeling and attribute injection mechanism over conventional attribute fusion strategies.

As shown in Fig. 6, MADE can also form relatively clear clusters for some identities, indicating that attribute information itself is beneficial for identity representation learning in cloth-changing scenarios. However, for some identities with large appearance variations or unstable attribute predictions, the sample clusters produced by MADE still exhibit certain dispersion, resulting in relatively large intra-class distances. In contrast, samples belonging to the same identity are more compactly clustered by GEAM, while different identities show clearer separation. This demonstrates that, while preserving the advantages of attribute semantics, the proposed method further enhances the robustness of attribute representations by suppressing clothing-related semantic interference through attribute preprocessing and modeling the latent dependencies among identity-related attributes via an adaptive attribute graph. Meanwhile, the staged attribute injection mechanism integrates the enhanced attribute semantics into visual features in a more fine-grained manner, thereby improving intra-class consistency and inter-class separability in the feature space under cloth-changing scenarios.

C.2 Retrieval Result Visualization

To further qualitatively validate the effectiveness of the proposed method in the actual retrieval process, we present the Rank-10 retrieval results of MADE and GEAM on the LTCC and Celeb-



(a) GEAM (Ours) (b) MADE

Figure 6: t-SNE visualization of feature distributions produced by MADE and GEAM. Different colors denote different identities, and the circled regions highlight representative identity clusters.



Figure 7: Visualization of Rank-10 retrieval results on LTCC and Celeb-reID-light. The leftmost image in each row denotes the query image, followed by the top-10 retrieved gallery images. Green boxes indicate correct matches with the same identity as the query, while red boxes indicate incorrect matches.

reID-light datasets, as shown in Fig. 7. Compared with feature distribution visualization, retrieval results can more directly reflect the model's ability to rank gallery images with the correct identity, and are particularly useful for observing whether the model can maintain stable identity matching performance under complex conditions such as clothing changes, viewpoint variations, pose variations, and background interference.

From the visualization results, it can be observed that although MADE also introduces pedestrian attribute information, it still produces obvious incorrect retrieval results for some challenging samples. In particular, when the query image only captures the side or back view of a pedestrian, identity cues such as frontal body structure, facial contour, and complete clothing details are relatively missing. In such cases, MADE tends to rank incorrect identities with similar clothing colors, garment silhouettes, hairstyles, or local poses at top positions. In contrast, GEAM retrieves more correct identity samples under these complex conditions, and the correct matches are usually ranked higher. Even when there exist significant clothing changes, viewpoint differences, and pose variations between the query and gallery images, GEAM can still maintain favorable matching stability. This indicates that the proposed method can more effectively suppress the interference caused by clothing appearance and local similarity, while focusing on more stable identity-discriminative cues.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the proposed method, experimental comparisons, ablation studies, and visualizations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper discusses a limitation in the Conclusion: GEAM relies on a pre-trained attribute detector, and inaccurate attribute predictions may affect semantic enhancement.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results, theorems, or formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the implementation details needed to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments are conducted on public benchmarks, and we will release the source code and pretrained weights upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, evaluation protocols, metrics, optimizer, training epochs, input resolution, backbone, losses, and major hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We fix the random seed for all experiments, and the reported results are reproducible under the same experimental setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted on two NVIDIA RTX 3090 GPUs with 24GB memory each.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses existing public benchmarks and follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work may benefit robust long-term person retrieval research, while potential privacy and surveillance concerns should be considered before real-world deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The released code and models are intended for research use, and users should follow the licenses and usage terms of the corresponding datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses public datasets and existing models, and the corresponding original works are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release the source code and pretrained weights for each dataset on GitHub with documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not collect new human-subject data, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.