Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

D^{3}F: Diffusion-Driven Dual-Stream Framework for Occluded Person Re-Identification

Xiaohao Xie | **Wei Meng** | **Wenhua Jiao**

Abstract

Person Re-Identification (ReID) under severe occlusion remains a formidable challenge, primarily because existing discriminative models lack the capacity to infer missing topological structures, inevitably leading to fragmented feature representations and compromised metric spaces. To break this perceptual limitation, we propose a novel end-to-end Diffusion-Driven Dual-stream Framework (D^{3}F), which seamlessly integrates generative structural priors from Diffusion Transformers (DiT) into vision-language ReID. To overcome the high computational overhead and semantic gaps inherent in diffusion models, we first design a truncated prior extraction mechanism alongside a Cascaded Inverted Modality Shuffle Network (CIMSN) to achieve lightweight and deep cross-modal interaction. Furthermore, recognizing that divergent generative features can severely contaminate the strict distance metric space, we propose an Asymmetric Decoupled Feature Guidance (ADFG) strategy. ADFG strictly utilizes the fused generative prior as a contextual lens to update text prompts, while reverting to pure discriminative features for mask localization and final metric pooling. This restrained decoupling strategy strikes an optimal balance between occlusion-resistant inference and metric space purity. Extensive experiments demonstrate that the proposed D^{3}F significantly outperforms existing methods, achieving state-of-the-art (SOTA) performance on both occluded and holistic ReID benchmark datasets.

KEYWORDS

Occluded person re-identification, Diffusion models, Generative structural prior, Vision-language models, Feature disentanglement.

1 | INTRODUCTION

Person Re-Identification (ReID) aims to retrieve specific individuals across non-overlapping camera networks. It has significant application value in intelligent security and urban computing [1,2]. Recently, the rise of vision-language foundation models, such as CLIP[3], has brought a major shift to the ReID paradigm. Existing methods like CLIP-ReID[4] successfully utilize the rich semantics of open vocabularies by aligning image features with text descriptions. To further address complex occlusion scenarios, recent studies have introduced prompt learning mechanisms for local feature disentanglement. For instance, pioneering works like ProFD[5] use learnable text prompts as semantic anchors to interact with visual features. This generates fine-grained part masks, which helps strip away background noise and occlusions to some extent.

Despite the notable progress of vision-language disentanglement models, fundamental bottlenecks remain when dealing with severe occlusion or drastic pose variation. The root cause is that conventional architectures—whether CNNs, ViTs, or CLIP—are essentially discriminative models. Discriminative models rely heavily on visible visual cues in the image. When a large area of the human topological structure is obscured, the discriminative network often extracts only fragmented local features. It lacks the ability to infer or reconstruct the missing parts. As a result, the generated part masks are prone to severe drifting in the occluded regions, which degrades the intra-class compactness of the metric space.

To break the perceptual limits of discriminative feature extraction, diffusion models offer a highly promising solution. Recent studies have revealed that generative diffusion models can capture rich low-level textures and, more importantly, internalize deep

structural priors about the geometric topology and structural integrity of objects during their forward and reverse processes[6]. Building upon this foundation, Transformer-based diffusion models (DiT[7]) have recently scaled these generative capabilities by learning from massive real-world datasets. Intuitively, introducing the powerful structural priors of DiT into ReID would equip the network with the ability to infer the complete human structure through occlusions. However, applying generative priors to ReID tasks faces two major challenges. First is the computational overhead. The complete forward generation process of DiT is highly expensive and cannot meet the efficient feature extraction requirements of ReID. Second is the semantic gap in the feature space. Generative feature distributions are highly divergent and contain a large amount of random reconstructive information. Directly concatenating or adding them to discriminative features would cause semantic confusion. It would also contaminate the clean metric space that the ReID task heavily relies on. To address these challenges, this paper proposes a novel Diffusion-Driven Dual-stream Framework (D^3F). To solve the computational overhead issue, we design a truncated generative prior extraction mechanism. By injecting structured noise at a specific time step and applying physical truncation at the intermediate layers of DiT, we extract the optimal prior representation at the boundary of semantics and textures with minimal cost. To overcome modality barriers, we propose a Cascaded Inverted Modality Shuffle Network (CIMSN). It uses inverted dimensional expansion and 2D spatial depthwise convolutions to achieve deep channel interaction between CLIP discriminative features and DiT generative features, while preserving the local inductive bias of the 2D space. Furthermore, to prevent feature space contamination, this paper proposes an Asymmetric Decoupled Feature Guidance (ADFG) mechanism. Instead of the traditional direct feature fusion approach, we use the features fused with DiT priors strictly as a "contextual lens." Through cross-attention, it guides the semantic update of text prompts, allowing the text domain to preemptively perceive occlusion-resistant topological information. Meanwhile, when using text to calculate part masks and perform final feature pooling, the network strictly reverts to using pure CLIP visual features. This decoupled design equips the text domain with generative inference capabilities while maintaining a highly clean identity metric space. The main contributions of this paper are summarized as follows:

- We propose D^3F, a novel end-to-end dual-stream framework. It introduces the generative structural priors of DiT into the vision-language ReID task at a very low computational cost, compensating for the lack of topological inference capability in discriminative models under severe occlusion.
- We design the lightweight Cascaded Inverted Modality Shuffle Network (CIMSN). Through cross-modal channel shuffling and 2D spatial mixing, it enables deep interaction between heterogeneous features and mitigates the feature over-smoothing issue common in standard attention mechanisms.
- We propose an asymmetric decoupled feature guidance strategy. This strategy uses fusion priors to update the text context while retaining pure visual features for distance metric learning. It achieves an effective balance between absorbing generative occlusion-resistant knowledge and maintaining metric space stability. Extensive experiments show that D^3F achieves state-of-the-art (SOTA) performance on multiple mainstream occluded and holistic ReID datasets.

2 | RELATED WORK

In this section, we review literature related to our proposed method from three key areas: occluded person re-identification, the application of Vision Transformers and CLIP in ReID, and the recent use of diffusion priors in discriminative tasks.

2.1 | Occluded Person Re-identification

Person re-identification aims to match a specific pedestrian across non-overlapping camera views. Although holistic ReID has made substantial progress, occluded ReID remains challenging due to background interference and the loss of discriminative body parts. Existing semantic-based methods generally fall into three categories:

- External model-driven methods: These utilize off-the-shelf pose estimators[8] or human parsing models[9] to locate visible body parts. However, these methods usually require heavy computation and are vulnerable to cascading errors from the external models.
- Occlusion simulation methods: These generate synthetic occlusions during training to improve model robustness. Typical approaches include random erasing[10] or batch-level key-retained occlusion augmentation[11]. While useful, they do not explicitly recover the missing information caused by occlusion.

- Self-attention-based methods: These leverage the inherent feature grouping capabilities of Transformers [12,13] to implicitly discover visible regions. Our method belongs to the semantic-based category. The core difference is that we do not merely rely on the weak cues provided by corrupted visual features. Instead, we introduce strong generative priors to assist the discriminative network in topological inference and structural reconstruction of occluded areas.

2.2 | Vision Transformers and CLIP in ReID

Since its introduction, the Vision Transformer (ViT) [14] has quickly become dominant in the ReID field due to its strong ability to capture long-range dependencies. Works like TransReID [15] first introduced ViT to ReID with specific modifications for side information embedding. Building on this, the success of Contrastive Language-Image Pre-training (CLIP) [3] inspired researchers to utilize its strong vision-text alignment representations. CLIP-ReID [4] and its variants focus on fine-tuning the frozen CLIP visual encoder and using text descriptions to regularize visual features. Recently, methods like ProFD [5] introduced prompt learning, using text prompts as semantic anchors to generate part masks and disentangle local features. Despite achieving impressive results, these methods primarily treat the image encoder as a pure discriminative feature extractor. Under extreme occlusion, it is difficult to generate precise part-aware semantic masks relying only on visual features from the contrastive learning space.

2.3 | Diffusion Priors for Discriminative Tasks

Diffusion models [16], especially Transformer-based diffusion models (DiT) [7], have reshaped generative modeling and demonstrated strong capabilities in capturing complex real-world data distributions. Recently, researchers have started exploring how to use the rich representations in pre-trained diffusion models for downstream perception and discriminative tasks. Early attempts explored using the intermediate features of diffusion models for semantic segmentation [6] or depth estimation [17].

In the ReID field, there have been some preliminary explorations, such as using diffusion models for data expansion [18] or sampling local features with intrinsic semantic structures via spatial diffusion processes [19]. However, these methods are mostly limited to the data generation level or auxiliary pseudo-label optimization. Deeply integrating diffusion models into person ReID—a highly fine-grained retrieval task constrained by strict metric spaces—is still largely unexplored. Our work bridges this gap. We adopt a truncated DiT as the generative prior branch. More importantly, we do not use it for image synthesis, nor do we rigidly concatenate it with visual features. Instead, we construct an asymmetric decoupled feature guidance mechanism. This allows us to use DiT's rich contextual inference ability to guide the discriminative CLIP branch in generating precise semantic masks and robust local features, avoiding metric space contamination.

3 | METHODOLOGY

3.1 | Overall Architecture

Existing ReID methods based on vision-language pre-training (e.g., CLIP) rely primarily on discriminative features obtained through contrastive learning. However, their robustness is often limited under severe occlusion or complex background interference because they lack deep inference capabilities for human topology and fine-grained high-frequency textures. To address this, we propose the Diffusion-Driven Dual-stream Framework (D^3F). This framework (Figure 1) is an end-to-end joint training architecture containing two independent models (a CLIP encoder and a DiT module). It is designed to utilize the generative prior of diffusion models to enhance the structural consistency of discriminative features.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width, the forward data flow of D^3F proceeds along two parallel branches:

- Discriminative visual representation extraction: Image I is fed into the largely frozen CLIP visual encoder to extract global and local spatial features with strong generalization capabilities, denoted as $F_{vis} \in \mathbb{R}^{L\times D_{vis}}$. Here, $L = \frac{H}{P} \times \frac{W}{P}$ represents the number of spatial patches (P is the patch size), and D_{vis} is the channel dimension of the visual features.
- Generative structural prior extraction: The same image I is mapped to a latent space via a Variational Autoencoder (VAE). Structured noise at a specific time step is injected, and the variable is fed into a truncated Diffusion Transformer (DiT)

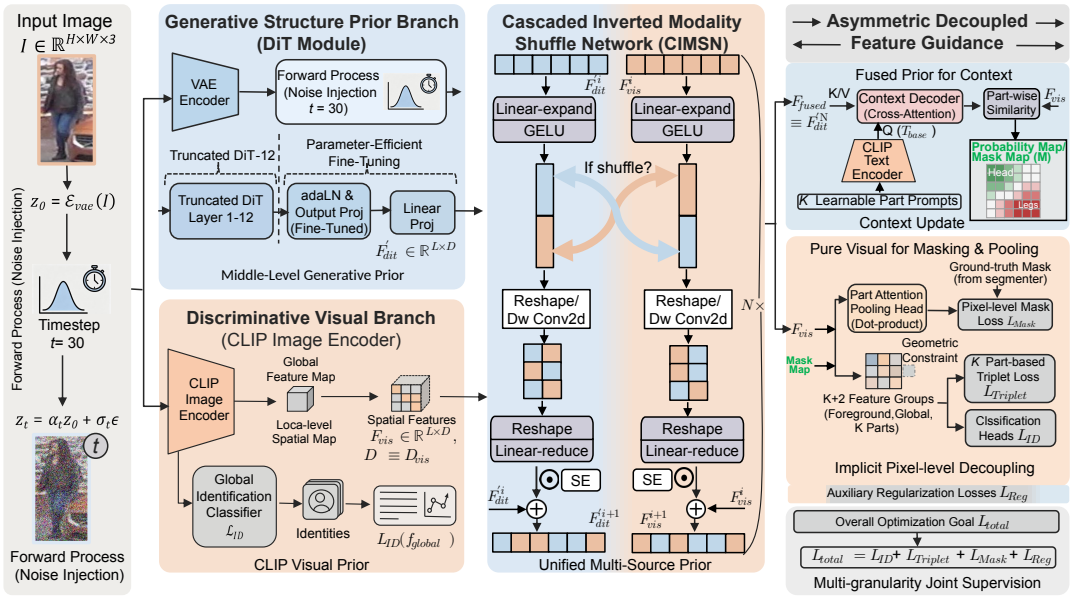


FIGURE 1 Overall architecture of the proposed Diffusion-Driven Dual-stream Framework (D^{3}F). Given an input image, the framework processes the data flow via two parallel branches. **Top:** The Generative Structural Prior branch injects structured noise at a specific timestep (t) and extracts a middle-level generative prior (F'_{dit}) via a truncated Diffusion Transformer (DiT) equipped with adaptive Parameter-Efficient Fine-Tuning (PEFT). **Bottom:** The Discriminative Visual branch extracts purely discriminative spatial features (F_{vis}) via a frozen CLIP image encoder. Subsequently, the Cascaded Inverted Modality Shuffle Network (CIMSN) performs lightweight cross-modal interaction to yield the fused feature F_{fused}. Finally, the Asymmetric Decoupled Feature Guidance (ADFG) mechanism strictly utilizes F_{fused} solely for updating text context prompts, while reverting to the pure F_{vis} for mask generation and feature pooling, thereby achieving a perfect balance between topological inference and metric space purity.

module. This branch aims to extract diffusion prior features containing rich human structural and texture information, denoted as $F_{dit} \in \mathbb{R}^{L \times D_{dit}}$, where D_{dit} is the output channel dimension of the specific DiT layer.

After feature extraction and dimensional alignment, the framework introduces the Cascaded Inverted Modality Shuffle Network (CIMSN) to perform deep, lightweight cross-modal interaction between F_{vis} and the projected generative prior F'_{dit}, outputting the fused feature F_{fused}. Finally, to prevent generative features from disrupting the metric space, our Asymmetric Decoupled Feature Guidance (ADFG) mechanism is integrated into the prompt-guided feature disentanglement pipeline. Crucially, this decoupling operates within a single end-to-end backpropagation flow, rather than as a post-processing step. By using the fused feature F_{fused} solely to update text prompts and reverting to the pure visual feature F_{vis} for mask generation and pooling, ADFG acts as a smart gradient gate. It explicitly cuts off the gradient interference from divergent generative features to the metric space, while preserving their ability to optimize the text context. This end-to-end gradient control successfully absorbs the macroscopic topological priors of DiT while maintaining the absolute purity of the identity metric space.

3.2 | Preliminary: Prompt-Guided Feature Disentangling

Our vision-language ReID baseline builds upon the prompt-guided feature disentanglement paradigm (e.g., ProFD[5]). Given an input image I, a pre-trained CLIP visual encoder extracts spatial feature maps $F_{vis} \in \mathbb{R}^{L \times D_{vis}}$. Simultaneously, we define a set of learnable text prompts containing K local body semantics (e.g., "head," "torso," "legs"). These prompts are initially processed by the CLIP text encoder to obtain the base textual embeddings $T_{base} \in \mathbb{R}^{K \times D_{vis}}$.

Instance-Aware Context Enhancement: A purely static text prompt struggles to align with highly variable human appearances in camera networks. To address this, the baseline utilizes a cross-modal Context Decoder. Specifically, it employs a cross-attention mechanism where the base textual embeddings T_{base} act as queries and the visual spatial features F_{vis} act as keys

and values. This injects instance-specific visual context into the text domain to yield dynamically enhanced prompt embeddings $F_{txt} \in \mathbb{R}^{K \times D_{vis}}$:

$$F_{txt} = \text{ContextDecoder}(T_{base}, F_{vis}) \tag{1}$$

For detailed architectural specifics of the context decoder, readers are referred to[5].

Mask Localization and Feature Pooling: With the enhanced part-aware prompts F_{txt}, the network locates the body parts by calculating the channel-wise dot-product similarity between the visual and textual features to generate fine-grained part masks $M \in \mathbb{R}^{K \times L}$:

$$M = \text{Softmax}(\alpha F_{txt} F_{vis}^{T}) \tag{2}$$

where α is a scaling factor, and the Softmax operation is applied along the spatial dimension L to normalize the attention distribution for each semantic part. To extract pure, occlusion-free local representations, the generated part masks M are utilized as spatial attention weights. Instead of generic global average pooling, we aggregate the visual spatial features F_{vis} through mask-guided weighted pooling. Specifically, the decoupled embedding $f_{part}^{k} \in \mathbb{R}^{D_{vis}}$ for the k-th semantic part is formulated as a spatial weighted sum:

$$f_{part}^{k} = \sum_{i=1}^{L} M_{k,i} F_{vis,i} \tag{3}$$

where $M_{k,i}$ denotes the predicted probability of the i-th spatial patch belonging to the k-th body part, and $F_{vis,i} \in \mathbb{R}^{D_{vis}}$ is the feature vector of the corresponding patch. This operation strictly filters out background noise and occlusions by assigning near-zero weights to corrupted regions. Finally, a linear dimensionality reduction head is applied to obtain the compact part embeddings.

Limitation in Occluded Scenarios: Under severe occlusion, relying solely on CLIP's contrastive learning features makes it difficult to precisely align local semantics. Specifically, the corrupted F_{vis} propagates errors during the context enhancement stage, causing the text prompts to deviate and resulting in inaccurate mask generation. Therefore, we introduce generative diffusion priors into this baseline (detailed in Section 3.3 and 3.5) to enhance the structural robustness of local feature disentanglement.

3.3 | Truncated Generative Prior Extraction

Traditional contrastive learning visual encoders perform well in capturing global semantics, but often fail to infer and reconstruct local human structures and high-frequency textures in severe occlusion scenarios. In contrast, diffusion models learn broad real-world data distributions and contain deep priors regarding object geometric topology and structural integrity. To introduce this generative prior into the ReID task while controlling computational cost, we design a truncated generative prior extraction module.

- Latent Mapping and Noise Injection: Given an input image I, we project it into a low-dimensional compact latent space using a frozen VAE encoder $\mathcal{E}_{vae}$, obtaining the initial latent variable $z_0 = \mathcal{E}_{vae}(I)$. Unlike standard diffusion denoising, we use the forward process to inject structured noise at a specific time step t into z_0 to acquire structurally discriminative information:

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

  where α_t and σ_t are scaling coefficients determined by a pre-defined linear noise schedule. The time step t (default $t = 25$ in this paper) provides a moderate smoothing mechanism. It effectively suppresses hard edge noise caused by occlusions while preserving the key human topological backbone, helping the network focus on more robust macroscopic structures.
- Physical Truncation for Prior Extraction: The noisy latent variable z_t is then fed into a pre-trained Diffusion Transformer (DiT). To avoid the high computational cost of a complete forward generation pass and to extract intermediate semantic features suitable for visual understanding, we perform physical truncation on the DiT. Specifically, the network retains the first l layers (set to $l = 12$ in experiments) and discards the subsequent computations. The generative structural prior feature $F_{dit} \in \mathbb{R}^{L \times D_{dit}}$ extracted after truncation is formulated as:

$$F_{dit} = \text{DiT}^{1 \to l}(z_t, t) \tag{5}$$

  This truncation strategy significantly reduces memory and computational burdens while extracting prior representations with both high-dimensional semantics and low-level texture attributes.

- Adaptive Parameter-Efficient Fine-Tuning: There is an inherent semantic gap between the feature distribution of generative tasks and the strict distance metric space required for ReID. To align the feature spaces effectively, we propose a localized Parameter-Efficient Fine-Tuning (PEFT) strategy. During training, most of DiT's weights are strictly frozen to preserve pre-trained generative knowledge. We only unfreeze the last N Transformer blocks before truncation (e.g., $N = 3$) and open the gradients of the adaptive layer normalization (adaLN_modulation) modules and output projection layers. Finally, a linear projection layer $\mathcal{P}(\cdot)$ maps the features to the target dimension (consistent with the D_{vis} channel size of CLIP visual features): $F'_{dit} = \mathcal{P}(F_{dit})$. By locally unfreezing adaLN, the network can adaptively modulate the feature space while maintaining the stability of DiT's pre-trained self-attention maps, successfully aligning generative features with the discriminative task.

3.4 | Cascaded Inverted Modality Shuffle Network

After extracting the discriminative visual feature F_{vis} and the generative structural prior F'_{dit}, effectively fusing these heterogeneous features without disrupting their spatial topologies is the core challenge. Traditional direct addition or global cross-attention often introduces heavy computational overhead and can lead to semantic over-smoothing of the feature space. Therefore, we propose a lightweight and efficient Cascaded Inverted Modality Shuffle Network (CIMSN). CIMSN consists of N stacked Inverted Modality Shuffle Blocks (IMSB). For the i-th block, with inputs $F^{(i)}_{vis}$ and $F'^{(i)}_{dit}$, the internal forward computation includes four key stages:

· Inverted Expansion: Inspired by lightweight mobile network designs, we first map the features to a higher-dimensional hidden space via Layer Normalization and linear projection to increase feature representation capacity:

$$\begin{aligned} X_v &= \sigma\left(W^v_{exp}\mathrm{LN}(F^{(i)}_{vis})\right), \\ X_d &= \sigma\left(W^d_{exp}\mathrm{LN}(F'^{(i)}_{dit})\right) \end{aligned} \tag{6}$$

where W_{exp} is the expansion projection matrix and $\sigma(\cdot)$ denotes the GELU activation function.

· Cross-Modal Channel Shuffle: To enable cross-modal interaction without expensive attention computation, we design a channel-level shuffle mechanism. Specifically, the high-dimensional features X_v and X_d are equally divided into two parts along the channel dimension: $X_v = [X_{v,1}, X_{v,2}]$ and $X_d = [X_{d,1}, X_{d,2}]$. The latter halves of the channels of these two modalities are then swapped for feature recombination:

$$\hat{X}_v = [X_{v,1}, X_{d,2}], \quad \hat{X}_d = [X_{d,1}, X_{v,2}] \tag{7}$$

This operation forces the discriminative features and generative priors to interact macroscopically at the channel group level, breaking the modality barrier.

· 2D Spatial Mixing and Inductive Bias: Since the ViT architecture lacks inductive biases for processing local spatial details, we reshape the recombined 1D sequence feature $\hat{X}_v$ back to a 2D spatial topology $\mathbb{R}^{B\times C\times H\times W}$. A 2D depthwise convolution is then applied to model local contexts among adjacent pixels:

$$Y_v = \mathrm{Flatten}\left(\mathrm{DWConv}_{3\times 3}(\mathrm{Reshape}(\hat{X}_v))\right) \tag{8}$$

This operation recovers the 2D spatial adjacency disrupted by the flatten operation and effectively extracts local texture features useful for handling occlusions. A symmetric 2D spatial mixing operation is also performed on $\hat{X}_d$ to obtain Y_d.

· Gated Residual and Dimensionality Reduction: After spatial mixing, the features are reduced back to their original dimensions via a linear bottleneck. Concurrently, to suppress irrelevant noise potentially introduced during shuffling, we use a Squeeze-and-Excitation (SE) gating mechanism, calculating adaptive channel weights w_{se} via global average pooling. Finally, the residual update is completed using the DropPath mechanism:

$$F^{(i+1)}_{vis} = F^{(i)}_{vis} + \mathrm{DropPath}\left(w^v_{se} \odot W^v_{proj}(Y_v)\right) \tag{9}$$

· Late Fusion Alternating Strategy: Considering that F_{vis} and F'_{dit} reside in different feature spaces, dense cross-modal shuffling can easily cause semantic over-smoothing. Thus, CIMSN adopts an alternating communication mechanism. First is late fusion: in the initial shallow layers of the cascaded network (default $k = 2$ in this paper), "cross-modal channel

shuffling" is disabled, allowing the two branches to perform independent high-dimensional spatial feature extraction until the deeper layers activate shuffle interactions to ensure basic feature semantic stability. Second is alternating shuffle: channel swapping is skipped in the even-numbered layers among the deep network layers where interaction is active. This series of strategies acts as a natural regularization, ensuring the semantic sharpness and stability of heterogeneous features during fusion. The output of the final layer is defined as F_{fused}.

3.5 | Asymmetric Decoupled Feature Guidance

Our baseline model is built upon the prompt-guided feature disentanglement pipeline proposed by ProFD[5]. In the original architecture, the interaction between vision and text strictly follows three core stages: context injection, mask generation, and mask-weighted pooling. By introducing CIMSN, our network simultaneously holds the pure discriminative feature F_{vis} and the enhanced feature F_{fused} incorporating diffusion priors. To investigate at which stage of the pipeline the generative knowledge should be injected to achieve a balance between inference and metric space purity, we designed three micro-ablation switches corresponding to the three interaction stages:

- Stage One: Context Micro-Enhancement ($\mathcal{S}_{ctx}$). In the context decoder, visual features are used as memory to provide instance-level cues to enhance text prompt expression. This switch determines the input feature for this stage:

$$F_{ctx\_memory} = \begin{cases} F_{fused}, & \mathcal{S}_{ctx} = \text{ON} \\ F_{vis}, & \mathcal{S}_{ctx} = \text{OFF} \end{cases} \tag{10}$$

  This selected F_{ctx_memory} acts as a robust substitute for the original corrupted F_{vis} as the key and value input in the cross-modal Context Decoder (Equation (1)). This option verifies whether the fused feature can act as a "contextual lens," transferring its DiT-derived occlusion-resistant topological inference capability to the text domain cross-modally.
- Stage Two: Local Mask Localization ($\mathcal{S}_{mask}$). During part mask generation, the network locates body parts by calculating the inner product between text prompts and visual spatial features. This switch controls the feature source for the base map:

$$\begin{aligned} M &= \text{Softmax}(\alpha F_{txt}(F_{vis\_mask})^T), \\ F_{vis\_mask} &= \begin{cases} F_{fused}, & \mathcal{S}_{mask} = \text{ON} \\ F_{vis}, & \mathcal{S}_{mask} = \text{OFF} \end{cases} \end{aligned} \tag{11}$$

  This option explores whether fusion features with inference capabilities can produce clearer boundary pixel-level localization scores under severe occlusion compared to pure contrastive features.
- Stage Three: Final Feature Pooling ($\mathcal{S}_{pool}$). This is the core stage determining the success of the decoupling design. When using masks to perform weighted pooling to extract the collective local embeddings (i.e., $F_{pool} = [f^1_{part}, \ldots, f^K_{part}]$):

$$\begin{aligned} F_{pool} &= \text{Pooling}(M, F_{vis\_pool}), \\ F_{vis\_pool} &= \begin{cases} F_{fused}, & \mathcal{S}_{pool} = \text{ON} \\ F_{vis}, & \mathcal{S}_{pool} = \text{OFF} \end{cases} \end{aligned} \tag{12}$$

  Here, the Pooling$(\cdot)$ operation strictly follows the spatial weighted sum defined in Equation (3). ReID fundamentally relies on a strict distance metric in the feature space. Through this switch, we verify our core hypothesis: directly introducing generative features into final matching leads to semantic jitter; conversely, strictly reverting to pure discriminative features is necessary to defend physical space purity.

Asymmetric Decoupled Optimal Strategy and Texture Residual Gating. Through in-depth ablation of the above variable combinations (detailed in Section 4.4), we established (ON, OFF, OFF) as the default optimal strategy. This means we introduce priors only during the context enhancement stage and strictly revert to pure visual features during mask localization and metric pooling. Although $\mathcal{S}_{pool}$ must remain off to ensure the macroscopic purity of the metric space, the fine-grained high-frequency structural textures in the generative feature F'_{dit} are still highly valuable for retrieving hard samples. To balance preventing contamination and utilizing textures, we introduce a Learnable Texture Residual Gating (LTRG) mechanism after

the final feature extraction. Specifically, we use the generated precise part mask M to extract the local texture descriptor $F_{pool}^{dit} = \text{Pooling}(M, F_{dit}^{'})$ from the pure DiT generative prior. Then, through a learnable scalar weight γ (initialized to 0), it is smoothly injected into the backbone discriminative feature as a residual:

$$F_{final} = F_{pool} + \gamma F_{pool}^{dit} \tag{13}$$

Because γ is initialized to zero, this gate maintains the absolute stability of the discriminative metric space in the early training stage. In the later training stages, the model adaptively opens the gate, drawing critical high-frequency texture details from the generative priors (with occlusions filtered out), thereby achieving further performance breakthroughs.

3.6 | Optimization and Training Strategy

To enable the discriminative visual branch and the generative prior branch to converge synergistically, we designed a multi-granular joint optimization scheme. This scheme supervises the network from multiple dimensions, including pixel-level semantic alignment, global identity classification, and feature space metric learning. The total loss function $\mathcal{L}_{total}$ consists of the base ReID loss $\mathcal{L}_{ReID}$ and the auxiliary regularization loss $\mathcal{L}_{Reg}$:

$$\mathcal{L}_{total} = \mathcal{L}_{ReID} + \mathcal{L}_{Reg} \tag{14}$$

Base ReID Losses. The base ReID loss drives the core retrieval and classification capabilities of the model. It includes the identity classification loss $\mathcal{L}_{ID}$, the part-based triplet metric loss $\mathcal{L}_{Triplet}$, and the pixel-level mask loss $\mathcal{L}_{Mask}$:

$$\mathcal{L}_{ReID} = \mathcal{L}_{ID} + \mathcal{L}_{Triplet} + \lambda_{mask}\mathcal{L}_{Mask} \tag{15}$$

where λ_{mask} is a hyperparameter for pixel-level supervision strength, set to 5.0 in our experiments.

- Identity Classification Loss: This loss constrains the model via Cross-Entropy Loss to accurately map image features to their corresponding true identity labels. To extract robust identity representations across multiple scales, we apply this loss jointly to global features, foreground features, and concatenated multi-part local features:

$$\mathcal{L}_{ID} = \mathcal{L}_{CE}(f_{global}) + \mathcal{L}_{CE}(f_{foreg}) + \mathcal{L}_{CE}(f_{concat}) \tag{16}$$

  This multi-dimensional classification supervision forces the network to focus on the human macroscopic global contour while accounting for local fine-grained identity details.
- Part-based Triplet Loss: Since ReID is a feature retrieval problem, the model must pull same-identity samples closer and push different-identity samples apart in a continuous geometric feature space. We introduce a triplet metric loss to the K disentangled local part features. We strictly use Euclidean Distance for this measurement:

$$\mathcal{L}_{Triplet} = \sum_{k=1}^{K} \left[d_E(f_a^k, f_p^k) - d_E(f_a^k, f_n^k) + m \right]_+ \tag{17}$$

  where f_a^k, f_p^k, f_n^k represent the specific decoupled embeddings (f_{part}^k) extracted from the anchor, positive, and negative samples for the k-th local part; $d_E(\cdot,\cdot)$ denotes standard Euclidean distance; m is the margin hyperparameter; and $[\cdot]_+$ denotes $\max(0,\cdot)$. Using Euclidean distance imposes absolute spatial geometric constraints on the visual vectors, effectively avoiding metric space distortions that could be caused by generative priors.
- Pixel-level Mask Loss: To drive the text prompts to accurately locate human anatomical parts and strip away noise, the network computes the dot-product similarity between normalized text prompts and visual spatial features, aligning it with ground-truth semantic segmentation masks via cross-entropy loss.

Auxiliary Regularization Losses. To further prevent local disentangled features from semantic collapse, improve feature compactness in the clustering space, and enhance model robustness to severe occlusion distributions, we adopt the auxiliary

regularization mechanisms from the baseline model ProFD[5]:

$$\mathcal{L}_{Reg} = \mathcal{L}_{Attn} + \mathcal{L}_{PCL} + \mathcal{L}_{Dis} + \lambda_{vis}\mathcal{L}_{Vis} \tag{18}$$

Specifically: (1) Attention alignment loss $\mathcal{L}_{Attn}$ forces the predicted attention distribution generated by text prompts to approximate the true human body part distribution via cross-entropy constraint; (2) Cluster contrastive memory loss $\mathcal{L}_{PCL}$ maintains a momentum-updated Cluster Memory Bank, applying InfoNCE-like constraints on global and concatenated part features to strengthen intra-class compactness; (3) Feature dissimilarity loss $\mathcal{L}_{Dis}$ forces different local feature vectors (e.g., "head" and "legs") to remain distinct, avoiding homogenization; (4) Part visibility prediction loss $\mathcal{L}_{Vis}$ uses Focal Loss to supervise part visibility prediction, alleviating the extreme imbalance between occluded and unoccluded samples. λ_{vis} is its scaling weight, set to 10.0. For detailed derivations of these auxiliary losses, please refer to[5].

Layered Learning Rate Fine-tuning Strategy. Since this framework involves joint training of two large pre-trained foundation models (CLIP and DiT) along with multiple heterogeneous new modules, we implement a strict layered adaptive learning rate strategy to prevent catastrophic forgetting and promote fast convergence. Let the base learning rate be lr_0 (7.5×10^{-6} in experiments). The optimizer divides weights into three parameter groups with different fine-tuning steps:

- Base pre-trained layers group: Includes the CLIP visual encoder (image_encoder), the partially unfrozen DiT module (dit_encoder), and the projection layer (dit_proj). The learning rate for weights is set to lr_0; to activate bias expression sensitivity, the learning rate for biases is $2 \times lr_0$, with a lightweight weight decay of 1×10^{-4}.
- Learnable context prompt group: Targets the continuous conditional context vectors (ctx) in text prompts. Since this space is difficult to align cross-modally directly, its learning rate is increased to $70 \times lr_0$ to accelerate semantic bridging.
- Adaptive convergence group: Includes the context decoder, identity classification heads, and CIMSN, which are initialized randomly. These layers are assigned a learning rate of $10 \times lr_0$ to quickly establish cross-modal interaction pathways. The full framework uses the AdamW optimizer for end-to-end fine-tuning. The total training duration is 60 epochs, with step-wise learning rate decay applied at epochs 30 and 50.

4 | EXPERIMENTS

To evaluate the effectiveness of D^{3}F, we conducted extensive experiments on multiple mainstream ReID benchmark datasets. This section introduces the datasets, evaluation metrics, implementation details, and provides quantitative and qualitative comparisons with current state-of-the-art methods.

4.1 | Datasets and Evaluation Metrics

We selected four widely used datasets: two specifically for occluded scenarios (Occluded-Duke[8] and Occluded-REID[20]), and two mainstream holistic datasets (Market-1501[21] and DukeMTMC-reID[22]).

- Occluded-Duke: A challenging dataset selected from DukeMTMC-reID. It contains 15,618 training images (702 identities), 2,210 severely occluded query images (519 identities), and 17,661 gallery images. It is an ideal testbed for evaluating occlusion-resistant inference capabilities.
- Occluded-REID: An occluded dataset collected via mobile devices in a campus environment. It contains 2,000 images (200 identities) and is used solely for testing.
- Market-1501: A widely used holistic benchmark with 32,668 images of 1,501 identities.
- DukeMTMC-reID: A large-scale holistic dataset containing 36,411 images of 1,404 identities.

Evaluation Metrics: Following existing works[5,4], we use Rank-1 accuracy from the Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) as the primary metrics. No post-processing techniques (e.g., Re-ranking) are used during testing.

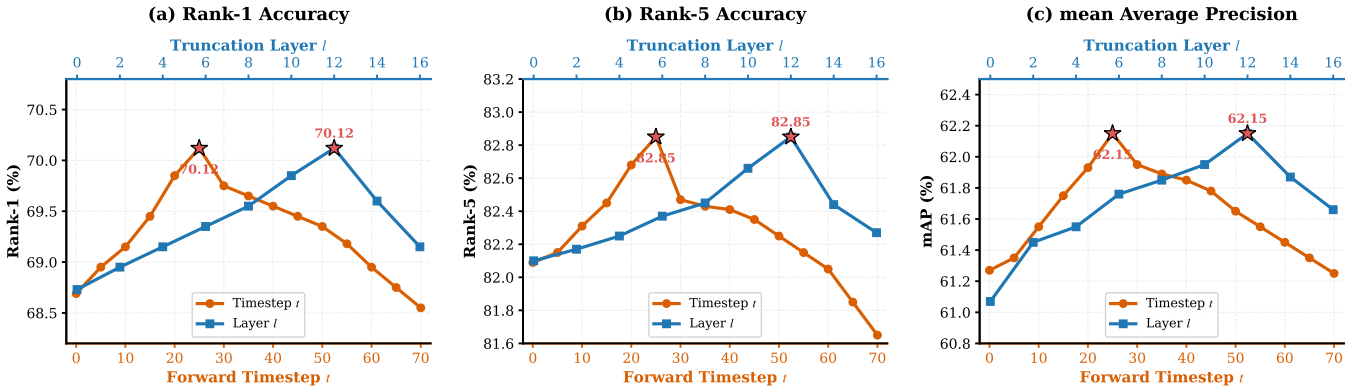


FIGURE 2 **Ablation study of generative prior extraction strategies.** We evaluate the impact of the forward noise time step t and the physical truncation layer index l on Rank-1, Rank-5, and mAP. Both parameters exhibit a clear inverted-U trend, indicating that moderate structured noise ($t = 25$) and intermediate-layer features ($l = 12$) achieve the optimal balance between macroscopic topological inference and microscopic texture preservation.

4.2 | Implementation Details

We implemented D^{3}F in PyTorch. Network initialization utilizes the pre-trained ViT-B/16 (from OpenAI's CLIP) as the discriminative visual backbone. The generative prior branch uses the pre-trained DiT-XL/2 model[7] with its VAE encoder frozen. To reduce VRAM usage, the VAE is set to half-precision (FP16) and slicing is enabled. All input images are resized to 256×128 pixels. Random Erasing is applied during training with a probability of 0.5. Weight decay is slightly lowered to 1×10^{-4}. The batch size is 64, training for 60 epochs. The base learning rate is 7.5×10^{-6} with Cosine Annealing scheduling. In CIMSN, the cascade depth N is 7, and the inverted expansion ratio E is 1.5. The ADFG mechanism adopts the (ON, OFF, OFF) configuration. The margin m for the triplet loss is set to 0.3.

4.3 | Comparison with State-of-the-Art Methods

We compared our performance with representative SOTA methods across four datasets. These include traditional CNN-based local/global alignment methods, Vision Transformer or attention-based methods, and recent VLM-based disentanglement frameworks (e.g., CLIP-ReID, ProFD). Quantitative comparison results are shown in Tables 1 and 2, Specifically, on the challenging occluded datasets (Table 1), our proposed D^{3}F demonstrates significant superiority over existing state-of-the-art methods, notably outperforming recent vision-language models such as ProFD. This substantial gain confirms that incorporating DiT-based generative structural priors effectively infers missing topologies under severe occlusions. Furthermore, on holistic datasets (Table 2) like Market-1501 and DukeMTMC-reID, D^{3}F also achieves state-of-the-art (SOTA) performance, particularly setting new records in the critical mAP metric (91.2% on Market-1501 and 85.8% on DukeMTMC-reID). This indicates that our Asymmetric Decoupled Feature Guidance (ADFG) mechanism successfully prevents generative feature divergence from contaminating the metric space, thereby maintaining exceptional discriminative capabilities even in unoccluded scenarios.

Beyond quantitative metrics, we provide intuitive qualitative evaluations to further demonstrate the superiority of our approach. As visualized in the t-SNE feature space distributions (Figure 3), the pure baseline model struggles with severe feature entanglement, as occlusions severely disrupt identity representations. In contrast, by seamlessly injecting DiT generative structural priors, D^{3}F significantly improves intra-class compactness and inter-class separability, forming highly distinct identity clusters. Furthermore, the retrieval results on hard samples (Figure 4) visually confirm this advantage. While the baseline easily fails when target pedestrians are heavily obscured, our framework successfully infers the missing topologies and accurately retrieves the correct matches. This visually proves its exceptional robustness against complex real-world occlusions.

TABLE 1 Comparison with State-of-the-Art (SOTA) Methods on Occluded-ReID/Occluded-Duke Datasets: * denotes models with Transformer backbone, † indicates adding sie_camera, and ‡ indicates encoders with a small-step sliding-window setting (Stride13*13). Bold values = best performance, underlined values = second-best.

Methods	Occluded-ReID		Occluded-Duke	
	R1	mAP	R1	mAP
PVPM [9]	66.8	59.5	-	-
CAAO [23]	65.2	61.1	67.8	55.8
CAAO*	87.1	83.4	68.5	59.5
PFD* [24]	79.8	81.3	67.7	60.1
TransReID* [15]	70.2	67.3	64.2	55.7
TransReID*†	-	-	66.4	59.2
PAT* [12]	81.6	72.1	64.5	53.6
FED* [25]	86.3	79.3	68.1	56.4
DPM* [26]	80.2	75.2	66.7	57.2
SGFNet* [27]	<u>93</u>	<u>90.3</u>	**76.5**	**67.2**
THCB-Net* [28]	87.3	84.5	72.3	62.6
RGANet* [29]	86.4	80	71.6	62.4
ProFD* [5]	91	88.3	68.9	61
SPT* [30]	87.8	81.1	<u>74.7</u>	63
MLRAT* [31]	88.9	83.6	73.3	63.1
D^3F(Ours)	**93.4**	91.1	70.8	62.8
D^3F†	-	-	71.1	63
D^3F†‡	93.3	**91.7**	72.2	<u>64.4</u>

TABLE 2 Comparison with State-of-the-Art (SOTA) Methods on Market1501/DukeMTMCReID

Methods	Market1501		DukeMTMC	
	R1	mAP	R1	mAP
CAAO [23]	95.1	87.3	88.9	77.5
CAAO*	95.3	88	89.8	80.9
PFD* [24]	95.5	89.6	90.6	82.2
TransReID* [15]	95	88.2	89.6	80.6
PAT* [12]	95.4	88	88.8	78.2
FED* [25]	95	86.3	89.4	78
SGFNet* [27]	**96.2**	<u>91.1</u>	**92.8**	<u>84.1</u>
THCB-Net* [28]	**96.2**	90.6	91.7	83.5
RGANet* [29]	95.5	89.8	-	-
ProFD* [5]	95.1	90.0	91.7	83.2
SPT* [30]	95.5	89.4	91.1	82.4
MLRAT* [31]	95.8	89.9	-	-
D^3F(Ours)	95.3	90.4	91.9	84.4
D^3F†	95.3	90.5	<u>92.6</u>	85.3
D^3F†‡	<u>95.9</u>	**91.2**	92.5	**85.8**

4.4 | Ablation Study

To verify that the strong performance of D^3F is not merely from stacking modules, we conducted a systematic drop-one-out ablation analysis on the Occluded-Duke dataset. To ensure rigor, all other modules were fixed at the default optimal configuration when probing a specific hyperparameter (i.e., $t = 25, l = 12$, CIMSN with 7 layers / alternating communication / 1.5 expand ratio, Context enhancement only, and the final Learnable Texture Residual Gating is disabled by default).

4.4.1 | Generative Prior Extraction

We investigated the impact of the forward noise time step (t) and the physical truncation layer index (l).

TABLE 3 Micro-ablation analysis of Asymmetric Decoupled Feature Guidance (ADFG). $\mathcal{S}_{ctx}$, $\mathcal{S}_{mask}$, and $\mathcal{S}_{pool}$ represent injecting fusion priors at the context enhancement, mask localization, and final feature pooling stages, respectively. Results indicate that injecting priors exclusively at the context stage (Config 4) avoids metric space contamination and cross-modal semantic conflicts, yielding the optimal performance.

Config	$\mathcal{S}_{ctx}$	$\mathcal{S}_{mask}$	$\mathcal{S}_{pool}$	Rank-1 (%)	Rank-5 (%)	mAP (%)
1 (Baseline)	×	×	×	68.87	82.08	60.99
2	×	×	✓	68.95	81.85	60.85
3	×	✓	×	69.55	82.45	61.65
4 (Optimal)	✓	×	×	**70.12**	**82.85**	**62.15**
5	×	✓	✓	69.10	82.15	61.45
6	✓	×	✓	69.45	82.25	61.55
7	✓	✓	×	68.45	81.75	60.85

TABLE 4 Ablation of the interaction depth N in CIMSN. Shallow networks fail to align heterogeneous features adequately, while overly deep networks cause optimization difficulties. $N = 7$ achieves the optimal feature fusion capacity.

Interaction Depth (N)	Rank-1 (%)	Rank-5 (%)	mAP (%)
1	68.11	81.27	61.34
2	68.45	81.45	61.52
3	68.95	81.75	61.77
4	69.55	82.40	62.10
5	69.85	82.62	**62.25**
6	69.95	82.74	62.11
7	**70.12**	**82.85**	62.15
8	69.85	82.67	62.05
9	69.65	82.42	61.94
10	69.25	82.12	61.75
11	68.37	81.99	61.88

TABLE 5 Ablation of the inverted expansion ratio E. An expansion ratio of 1.5 strikes the best balance between parameter growth and high-dimensional feature representation space.

Expand Ratio (E)	Rank-1 (%)	Rank-5 (%)	mAP (%)
0.5	68.65	81.95	61.61
1.0	69.45	82.44	61.86
1.5	**70.12**	**82.85**	**62.15**
2.0	69.75	82.59	61.94
2.5	69.41	82.35	61.84
3.0	69.22	82.31	61.77
4.0	69.05	82.14	61.55

- Time step t: As shown in Figure 2, when $t = 0$ (no noise), the network focuses on micro details with limited macro topology extraction capacity (Rank-1 68.69%). As t increases, hard edge noise from occlusions is smoothed, and the main human structure dominates, reaching a peak Rank-1 of 70.12% at $t = 25$. However, when $t > 25$, excessive noise begins to gradually destroy fine-grained identity information, causing a smooth performance drop. This validates the critical role of moderate structured noise in stimulating macroscopic inference.
- Layer index l: Fixing $t = 25$, we tested the truncation depth (Figure 2). Intuitively, shallow features (e.g., $l < 6$) remain at the base texture level, lacking high-level semantics. Deep features (e.g., $l = 16$) lean heavily toward pure image reconstruction, creating a semantic gap with ReID requirements. Features extracted at the intermediate level ($l = 12$) achieved the optimal balance of semantics and detail, yielding 70.12% Rank-1.

4.4.2 | Capacity of CIMSN

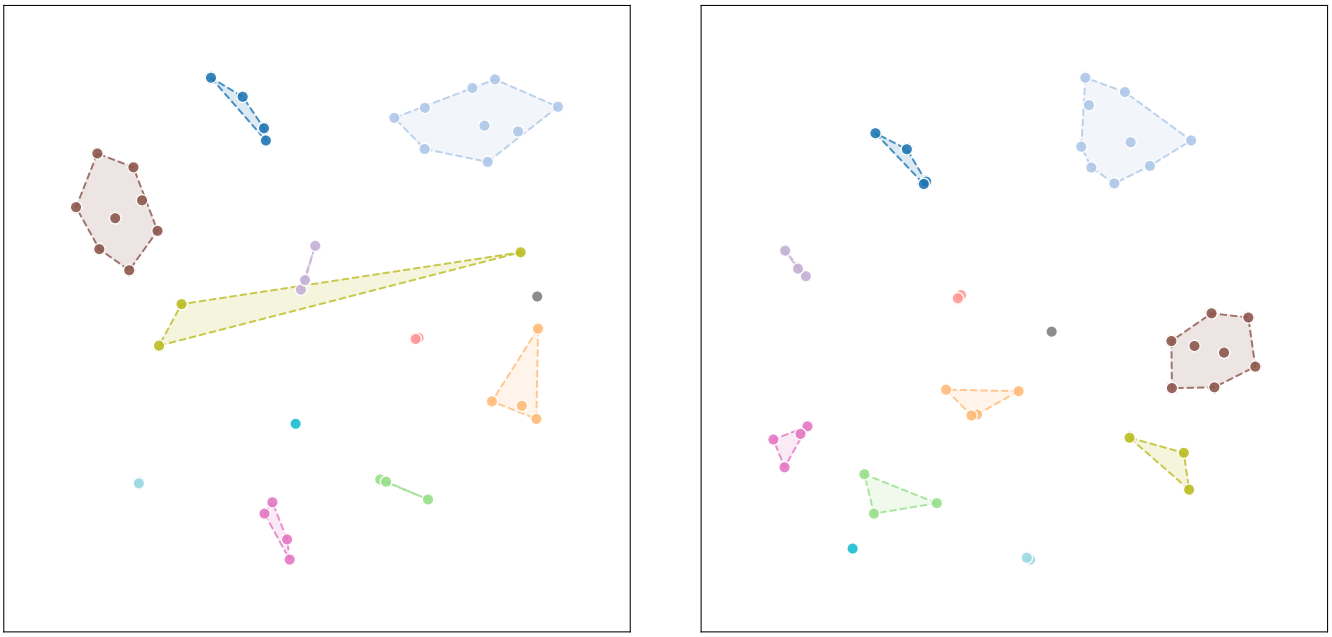
We analyzed the depth (N), expand ratio (E), and communication strategy of CIMSN.

TABLE 6 Comparison of cross-modal communication strategies. Sparse alternating communication effectively mitigates the "semantic over-smoothing" caused by full communication while breaking the modality barrier of no communication.

Communication Strategy	Rank-1 (%)	Rank-5 (%)	mAP (%)
Full Communication	68.95	82.30	61.75
No Communication	69.05	82.49	61.85
Alternating (Ours)	**70.12**	**82.85**	**62.15**

TABLE 7 Effectiveness of the generative texture residual injection. Here, "Pure Baseline" denotes the vision-language alignment without generative priors. "D^{3}F (w/o Gating)" relies solely on ADFG decoupled inference to avoid metric space contamination. Finally, "D^{3}F (w/ Gating)" introduces the Learnable Texture Residual Gating to inject crucial high-frequency details in a highly restrained manner, achieving further performance breakthroughs.

Model Configuration	Rank-1 (%)	Rank-5 (%)	mAP (%)
Pure Baseline	68.87	82.08	60.99
D^{3}F (w/o Gating)	70.12	82.85	62.15
D^{3}F (w/ Gating)	**70.82**	**83.17**	**62.79**



(a) Baseline (b) D^{3}F (Ours)

FIGURE 3 t-SNE visualization of the learned feature spaces. Different colors represent distinct pedestrian identities. (a) In the pure baseline model, features of different identities are entangled and overlapping due to severe occlusions. (b) By injecting DiT generative structural priors, our D^{3}F framework significantly improves intra-class compactness and inter-class separability, forming clear and distinct identity clusters.

- Depth N and Expand Ratio E: As clearly seen in Table 4, performance follows an inverted-U trend with depth N. Shallow interactions ($N \leq 3$) cannot fully fuse the distributions, while $N = 7$ reaches the peak (Rank-1 70.12%). Deeper networks ($N = 11$) result in performance degradation (68.37%) due to optimization difficulty. Regarding the expand ratio (Table 5), $E = 1.5$ provided an optimal trade-off between representation space and parameter size, achieving a distinct advantage in Rank-1 (70.12%).
- Cross-modal communication strategy: We compared different communication strategies (Table 6). The "full communication" strategy (shuffling at every layer) yielded 68.95% Rank-1, while "no communication" yielded 69.05%. In contrast, our

FIGURE 4 Visualization of retrieval results on hard samples. The leftmost column shows the severely occluded query images, followed by the top retrieved results from the gallery by the Baseline and our D^{3}F framework. Green and red bounding boxes denote correct and incorrect matches, respectively. Benefiting from the injection of generative structural information, our method successfully infers the occluded topologies, demonstrating a distinct advantage over the baseline in retrieving challenging occluded targets.

"alternating communication" strategy achieved 70.12%. This confirms that dense interactions cause cross-modal semantic over-smoothing, whereas sparse alternating communication acts as a natural regularization, maintaining original feature sharpness while allowing knowledge transfer.

4.4.3 | Micro-Ablation of ADFG

We systematically ablated the decoupling pipeline using the $\mathcal{S}_{ctx}, \mathcal{S}_{mask}, \mathcal{S}_{pool}$ control switches (Table 3). The pure baseline model without generative priors showed limited performance (Rank-1 68.87%). Over-introducing generative priors also failed to improve performance. The model reached an optimal Rank-1 of 70.12% (with mAP 62.15%) only when $\mathcal{S}_{ctx}$ = ON (Group 4). To deeply investigate this phenomenon, we compared two other typical configurations:

- Metric space contamination: Directly introducing DiT features into the final distance pooling layer (Group 2) caused a slight performance drop. This provides strong empirical evidence that divergent generative distributions directly contaminate the fragile metric space that identity retrieval depends on.
- Cross-modal semantic conflict: Enabling fusion features in both Context and Mask stages (Group 7) dropped performance to a global low of 68.45%. This indicates that modifying both the text and mask base maps simultaneously causes severe semantic misalignment.

In conclusion, the data validate our "leveraging without overstepping" philosophy. Using the generative prior solely as a "lens" to guide text updates is highly efficient, whereas the network must strictly revert to the pure discriminative space during mask localization and pooling to defend physical purity.

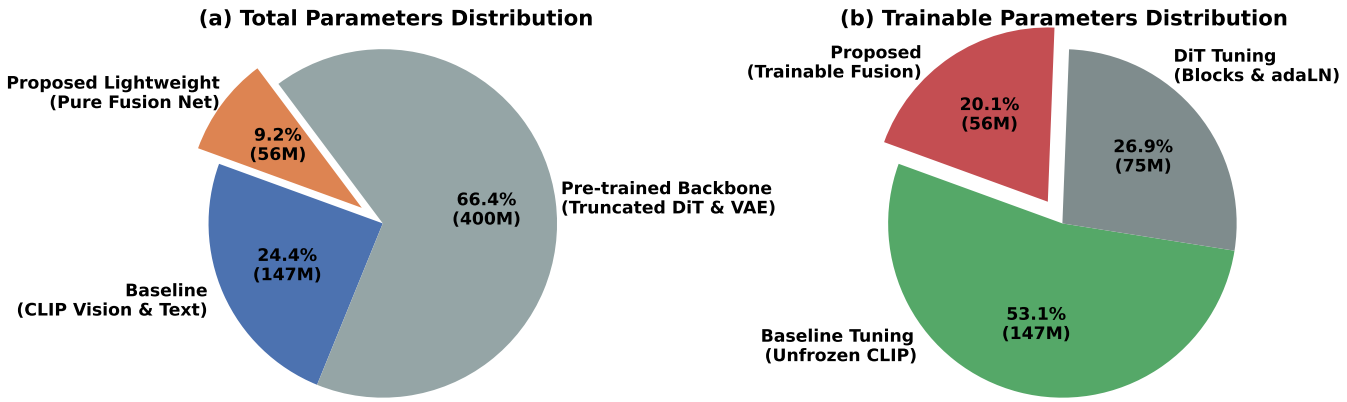


FIGURE 5 Parameter complexity analysis of the D^{3}F framework. (a) Total parameters distribution. The proposed lightweight fusion modules (CIMSN and projections) account for only 9.2% of the total parameters. (b) Trainable parameters distribution. By employing a parameter-efficient fine-tuning (PEFT) strategy on the DiT backbone (Blocks & adaLN), the trainable parameters of our proposed modules are effectively constrained to 20.1%, demonstrating the high efficiency of our cross-modal interaction design.

4.4.4 | Effectiveness of Learnable Texture Residual Gating

Experiments in Table 3 clearly indicate that directly placing divergent generative features into the pooling layer contaminates the metric space. However, the high-frequency structural textures in DiT priors remain valuable for fine-grained retrieval. To safely utilize this information, we introduced a Learnable Texture Residual Gating mechanism on top of the optimal asynchronous decoupling configuration. As shown in Table 7, when the model relies solely on ADFG for context inference, Rank-1 reaches 70.12%. After enabling residual gating, the network uses the mask to perform strict "occlusion filtering" on the generative prior, and then smoothly injects it into the final metric space with a tiny dynamic weight. This highly restrained strategy does not destroy the purity of the metric space; instead, it successfully supplements crucial high-frequency details, pushing Rank-1 to a final SOTA performance of 70.82%. This strongly demonstrates the perfect balance achieved by the residual gating mechanism between "preventing contamination" and "extracting textures."

4.5 | Parameter Complexity Analysis

To demonstrate the efficiency of the proposed D^{3}F, we analyze the parameter distribution of the entire framework, as illustrated in Fig. 5. Although the framework incorporates large-scale pre-trained foundation models (i.e., the CLIP discriminative baseline and the generative backbone comprising a truncated DiT and VAE), our proposed lightweight modules (i.e., the Cascaded Inverted Modality Shuffle Network (CIMSN) and projection layers) account for only 9.2% (56M) of the total parameters. In terms of trainable parameters, thanks to our Parameter-Efficient Fine-Tuning (PEFT) strategy, the tuning of the generative branch is strictly limited to the specific DiT blocks and adaptive layer normalization (adaLN) modules (accounting for 26.9% of trainable parameters). Consequently, the proposed trainable fusion modules occupy only 20.1% of the overall trainable parameters. This explicit distribution strongly indicates that the remarkable performance of D^{3}F does not rely on brute-force parameter stacking. Instead, it successfully leverages massive generative structural priors with minimal parameter cost through a highly efficient cross-modal interaction mechanism.

5 | CONCLUSION

This paper proposes a novel Diffusion-Driven Dual-stream Framework (D^{3}F) to address the perceptual limits of discriminative models in severe occlusion scenarios. By integrating a pre-trained Diffusion Transformer (DiT), we efficiently extract generative structural priors and seamlessly integrate them into vision-language ReID tasks.

Beyond lightweight cross-modal fusion via CIMSN, our core contribution lies in the elegant gradient flow control during end-to-end training. The proposed ADFG mechanism acts as a strict gradient gate. It uses generative priors solely as a "lens" to optimize text prompts, while strictly reverting to pure discriminative features for mask localization and metric matching. This design successfully cuts off divergent generative gradients from contaminating the metric space. Ultimately, D^3F proves that generative topological inference and strict discriminative metric learning can be harmoniously jointly optimized. Extensive experiments demonstrate that our framework significantly outperforms existing state-of-the-art methods.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: A benchmark. In: 2015:1116–1124.
2. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence.* 2021;44(6):2872–2893.
3. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: PmLR. 2021:8748–8763.
4. Li S, Sun L, Li Q. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: . 37. 2023:1405–1413.
5. Cui C, Huang S, Song W, Ding P, Zhang M, Wang D. Profd: Prompt-guided feature disentangling for occluded person re-identification. In: 2024:1583–1592.
6. Baranchuk D, Voynov A, Rubachev I, Khrulkov V, Babenko A. Label-Efficient Semantic Segmentation with Diffusion Models. In: .
7. Peebles W, Xie S. Scalable diffusion models with transformers. In: 2023:4195–4205.
8. Miao J, Wu Y, Liu P, Ding Y, Yang Y. Pose-guided feature alignment for occluded person re-identification. In: 2019:542–551.
9. Gao S, Wang J, Lu H, Liu Z. Pose-guided visible part matching for occluded person reid. In: 2020:11744–11752.
10. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. In: . 34. 2020:13001–13008.
11. Wang H, Ma Y, Chen X. KRE: A key-retained random erasing method for occluded person re-identification. In: 2023:467–473.
12. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F. Diverse part discovery: Occluded person re-identification with part-aware transformer. In: 2021:2898–2907.
13. Chen P, Liu W, Dai P, et al. Occlude them all: Occlusion-aware attention network for occluded person re-id. In: 2021:11833–11842.
14. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: 2021.
15. He S, Luo H, Wang P, Wang F, Li H, Jiang W. Transreid: Transformer-based object re-identification. In: 2021:15013–15022.
16. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Advances in neural information processing systems.* 2020;33:6840–6851.
17. Zhao W, Rao Y, Liu Z, Liu B, Zhou J, Lu J. Unleashing text-to-image diffusion models for visual perception. In: 2023:5729–5739.
18. Siddiqui N, Croitoru FA, Nayak GK, Ionescu RT, Shah M. DLCR: A generative data expansion framework via diffusion for clothes-changing person Re-ID. In: IEEE. 2025:1608–1617.
19. Tao X, Kong J, Jiang M, Lu M, Mian A. Unsupervised learning of intrinsic semantics with diffusion model for person re-identification. *IEEE Transactions on Image Processing.* 2024;33:6705–6719.
20. Zhuo J, Chen Z, Lai J, Wang G. Occluded person re-identification. In: IEEE. 2018:1–6.
21. Zheng L, Shen L, Tian L, Wang S, Bu J, Tian Q. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171.* 2015.
22. Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: 2017:3754–3762.
23. Zhao C, Qu Z, Jiang X, Tu Y, Bai X. Content-adaptive auto-occlusion network for occluded person re-identification. *IEEE Transactions on Image Processing.* 2023;32:4223-4236.
24. Wang T, Liu H, Song P, Guo T, Shi W. Pose-guided feature disentangling for occluded person re-identification based on transformer. In: . 36. 2022:2540–2549.
25. Wang Z, Zhu F, Tang S, Zhao R, He L, Song J. Feature erasing and diffusion network for occluded person re-identification. In: 2022:4754–4763.
26. Tan L, Dai P, Ji R, Wu Y. Dynamic prototype mask for occluded person re-identification. In: 2022:531–540.
27. Lin G, Yang S, Zheng WS, Li Z, Huang Z. A semantically guided and focused network for occluded person re-identification. *IEEE Transactions on Information Forensics and Security.* 2025.
28. Wang C, He S, Wu M, Lam SK, Tiwari P, Gao X. Looking Clearer with Text: A Hierarchical Context Blending Network for Occluded Person Re-Identification. *IEEE Transactions on Information Forensics and Security.* 2025.
29. He S, Chen W, Wang K, et al. Region generation and assessment network for occluded person re-identification. *IEEE transactions on information forensics and security.* 2023;19:120–132.
30. Tan L, Xia J, Liu W, Dai P, Wu Y, Cao L. Occluded person re-identification via saliency-guided patch transfer. In: . 38. 2024:5070–5078.
31. Lin G, Bao Z, Huang Z, Li Z, Zheng Ws, Chen Y. A multi-level relation-aware transformer model for occluded person re-identification. *Neural Networks.* 2024;177:106382.