Text-guided Feature Disentanglement for Cross-modal Gait Recognition

Supplementary Material

1. Data Augmentation

1.1. Patch Exchange

Inspired by previous VI-ReID approaches [2, 3], we propose a patch exchange data augmentation strategy for gait recognition tasks to enhance the model's generalization ability. Specifically, for the 2D Silhouettes and 3D Depth maps with input sizes of 64×64, their data shapes are $(s, h, w, c = 1)$ and $(s, h, w, c = 3)$, respectively. First, we adjust the initial number of channels of both by replicating the 2D Silhouettes along the dimension, so that both data have the same shape of $(s, h, w, 3)$. Then, paired inputs are divided into non-overlapping patches of size $p \times p$, resulting in $\frac{h}{p} \times \frac{w}{p}$ patches in total. For a patch at position (i, j), we define an exchange probability ξ to exchange paired 2D and 3D data. The patch exchange data augmentation process is illustrated in Figure 1.

2. Text Encoder

We leverage the CLIP text encoder to extract semantic features from the textual descriptions in the GMTD. These features are subsequently transformed via a lightweight Adapter to align with the target embedding space. Formally,

$$v_j^m = \text{Dropout}\left(W_2\left(\sigma\left(\text{LN}\left(W_1\left(\text{CLIP}_t\left(t_j^m\right)\right)\right)\right)\right)\right), \tag{1}$$

whereLN represents Layer Normalization, σ denotes the LeakyReLU activation, $W_1 \in \mathbb{R}^{d\times 4d}$ and $W_2 \in \mathbb{R}^{4d\times d}$ are projection matrices, and the dropout layer is applied with rate of 0.1.

Table 1. LiDAR → Camera accuracy of ablation study on patch size $p \in \{1, 2, 4, 8, 16, 32\}$ under a fixed swapping probability $\xi = 0.1$.

p	1	2	4	8	16	32
Rank-1 (%)	55.4	58.9	60.7	61.2	**61.7**	60.8
Rank-5 (%)	70.8	73.6	81.9	82.0	**82.5**	81.1

Table 2. Ablation study on patch exchange probability $\xi \in \{0.1, 0.2, \ldots, 0.6\}$ under a fixed patch size $p = 16$.

ξ	0.0	0.1	0.2	0.3	0.4	0.5	0.6
Rank-1 (%)	61.2	**61.7**	60.7	60.1	54.5	48.7	39.2
Rank-5 (%)	82.1	**82.5**	81.2	80.9	70.5	61.9	51.7

Table 3. Ablation study of the LiDAR → Camera accuracy within the proposed disentanglement loss functions.**MA**: Modality Alignment Loss; **MO**: Modality Orthogonality Loss; **HSIC**: HSIC-based Independence Loss.

MA	MO	HSIC	Rank-1(%)	Rank-5(%)
✓			56.4	72.6
✓	✓		58.6	76.8
✓		✓	59.7	78.1
✓	✓	✓	**61.7**	**82.5**

Table 4. **Ablation of Text Factorization** on SUSTech1K (L→C). We compare different prompt settings to evaluate the contribution of modality and viewpoint semantics.

Text Setting	Semantic Components	Rank-1 (%)
Full text (Ours)	**Modality + Viewpoint**	**61.7%**
Modality-only text	Modality semantics only	60.6%
Viewpoint-only text	Viewpoint strings only	57.4%
Learnable Prompts	Non-linguistic tokens	58.3%
No Text (Baseline)	Visual-only features	56.2%

3. More Ablation Study

3.1. Patch Exchange and Loss Functions

We conduct comprehensive ablation studies to investigate the influence of patch exchange hyperparameters, including patch size p and exchange probability ξ, as summarized in Table 1 and 2, respectively. Furthermore, we evaluate the effectiveness of the proposed feature disentanglement loss functions through additional ablation experiments. As shown in Table 3, these losses play a critical role in driving the disentanglement process and enhancing representation quality.

3.2. Analysis of Text Prompt Components

Beyond the architectural hyperparameters, we further investigate the impact of the text-based prompt dictionary on the cross-modal retrieval performance. Specifically, we analyze the contribution of different semantic components within our Gait Modality Text Dictionary (GMTD).

As summarized in Table 4, our Full text configuration, which integrates both modality and viewpoint semantics, achieves the superior Rank-1 accuracy of 61.7%. Removing either modality or viewpoint information leads to a performance drop, particularly when viewpoint strings are excluded (decreasing from 61.7% to 60.6%). This underscores the importance of spatial awareness in cross-modal alignment. No-

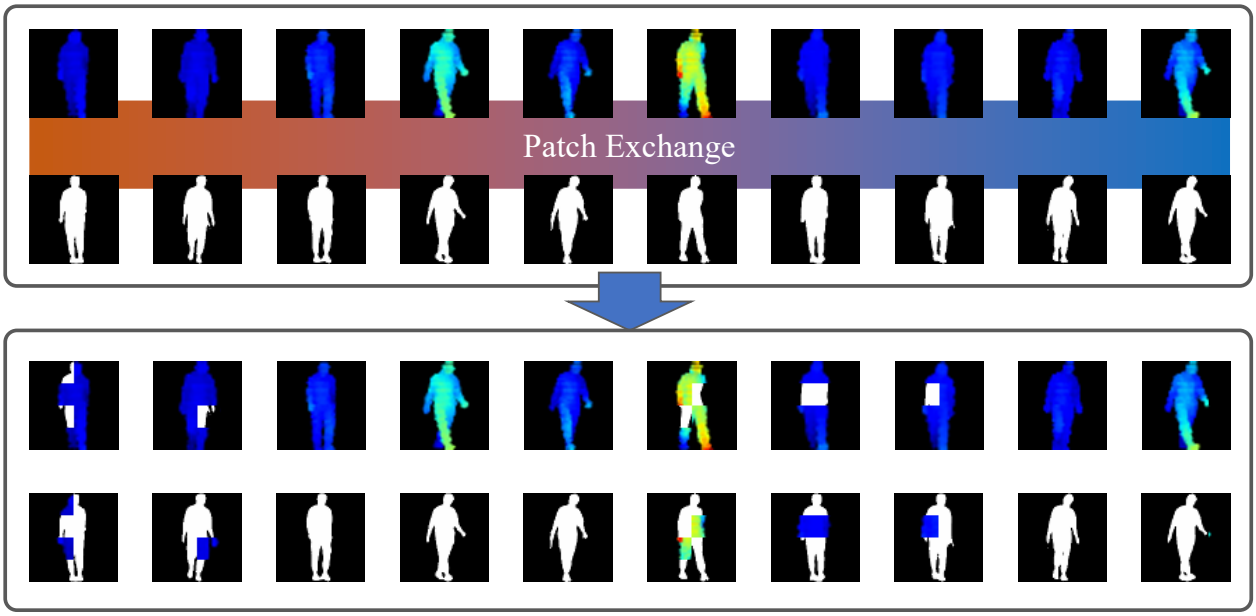
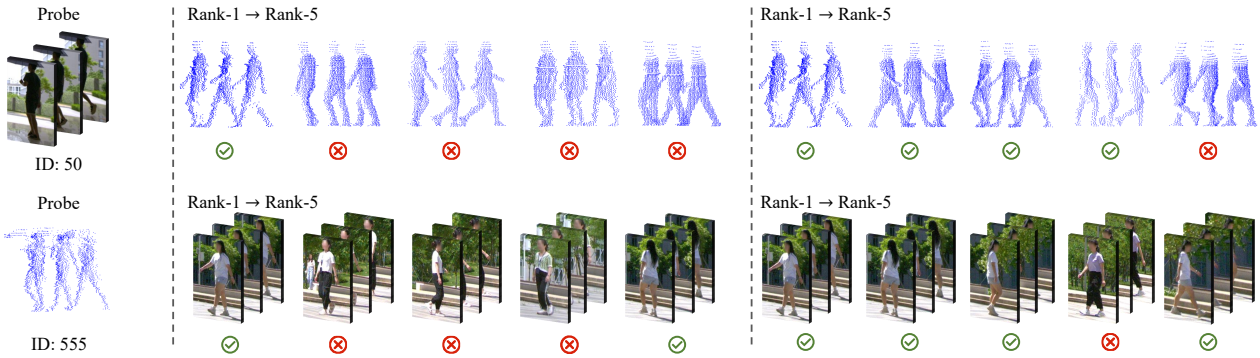


Figure 1. Visualization of the patch exchange data augmentation strategy.



(a) CrossGait (b) TCFDNet(Ours)

Figure 2. The visualization comparison of cross-modal gait retrieval results includes outcomes from Rank-1 to Rank-5, where a green circle with a checkmark indicates a correct retrieval, and a red circle with a cross denotes an incorrect result.

tably, our linguistic-based approach outperforms the Learnable Prompts baseline (58.3%), suggesting that explicit semantic constraints provided by natural language offer better generalization than purely data-driven tokens.

Fairness and Practicality of Viewpoint Information. A key design consideration of our method is the practical deployment without requiring auxiliary sensors. We emphasize that our framework does not utilize ground-truth viewpoint labels or external pose estimators during the inference phase. Instead, the GMTD serves as a comprehensive, pre-defined dictionary covering eight discrete viewpoints (from 0° to 315°). The model dynamically retrieves the most relevant viewpoint description from this fixed dictionary via the $top\text{-}k$ mechanism, based solely on the inherent visual features of the input (e.g., images or point clouds). Consequently, our method maintains a strictly fair comparison with existing baselines by utilizing the same input data, while providing a more robust semantic guidance through the automated retrieval of viewpoint-aware prompts.

Table 5. **Backbone Generalization Analysis** on SUSTech1K (L→C). We compare the default CLIP backbone with a stronger VLM (SigLIP) and a unimodal self-supervised model (DINOv2) to validate scalability and independence.

Vision Backbone / Text Encoder	Rank-1 (%)
CLIP (ViT-B/16) / CLIP (Default)	**61.7%**
SigLIP (ViT-B/16) / SigLIP	62.1%
DINOv2 (ViT-B/14) / CLIP	60.5%

3.3. Generalization Across Visual Backbones

To verify that the effectiveness of our framework is not solely dependent on the specific pre-aligned feature space of CLIP, we evaluate the scalability and independence of TCFDNet by employing different visual backbone architectures. The results are summarized in Table 5.

Scalability and Independence. First, we investigate the scalability of our method by adopting SigLIP [4], a more advanced Vision-Language Model. As shown in the table, replacing the default encoder with SigLIP boosts the Rank-1 accuracy to 62.1%, demonstrating that our architecture naturally benefits from stronger visual representations.

More importantly, we isolate the contribution of our framework by utilizing DINOv2 [1] as the visual encoder. DINOv2 is a unimodal self-supervised model that lacks the pre-aligned vision-language space inherent to CLIP. Despite this, TCFDNet maintains a high accuracy of 60.5%. This result serves as strong evidence that our proposed *Text-guided Disentanglement* mechanism actively acts as a bridge between disparate feature spaces, rather than passively relying on the initialization of contrastive pre-training.

Remark on Generative MLLMs. While we acknowledge the emerging capabilities of Generative Multimodal Large Language Models (MLLMs) such as InternVL and Qwen-VL, they fundamentally differ in architecture (autoregressive generation) compared to the dual-encoder retrieval paradigm used in this work. Consequently, we prioritize SigLIP for a direct comparison within the contrastive learning framework and leave the adaptation of our disentanglement strategy to generative MLLMs for future exploration.

4. HSIC Independence Loss Analysis

In our proposed framework, we disentangle the visual features into modality-specific features and modality-shared features, denoted as $\widetilde{F}_{(mod)_i}^{\ m}$ and $\widetilde{F}_{(shared)_i}^{\ m}$, respectively. Here, $m \in \{2d, 3d\}$ indicates the modality, and i indexes the i-th sample in a mini-batch of size N. While the modality-specific features preserve modality-dependent cues such as contours or geometric patterns, the modality-shared features are expected to encode identity-consistent and modality-invariant representations suitable for cross-modal matching. To enforce strict disentanglement, we incorporate the Hilbert-Schmidt Independence Criterion (HSIC), which quantifies the statistical dependence between two random variables in a kernel space. Unlike geometric orthogonality losses that operate only on directions, HSIC captures higher-order nonlinear correlations in a distribution-free manner.

Feature Representation. We first pool both $\widetilde{F}_{(mod)_i}^{\ m}$ and $\widetilde{F}_{(shared)_i}^{\ m}$ via global average pooling:

$$f^m_{(mod),i} = \text{Avgpool}(\widetilde{F}_{(mod)_i}^{\ m}) \in \mathbb{R}^C \tag{2}$$

$$f^m_{(shared),i} = \text{Avgpool}(\widetilde{F}_{(shared)_i}^{\ m}) \in \mathbb{R}^C. \tag{3}$$

We then stack the pooled features across the batch to obtain:

$$F^m_{(mod)} = [f^m_{(mod),1}, \ldots, f^m_{(mod),N}]^\top \in \mathbb{R}^{N\times C}, \tag{4}$$

$$F^m_{(shared)} = [f^m_{(shared),1}, \ldots, f^m_{(shared),N}]^\top \in \mathbb{R}^{N\times C}. \tag{5}$$

HSIC Computation. We define linear kernel Gram matrices:

$$K = F^m_{(mod)} \cdot (F^m_{(mod)})^\top, \quad L = F^m_{(shared)} \cdot (F^m_{(shared)})^\top. \tag{6}$$

Let $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ be the centering matrix. We compute:

$$K_c = HKH, \quad L_c = HLH. \tag{7}$$

Finally, the HSIC loss is given by:

$$\mathcal{L}_{\text{HSIC}} = \frac{1}{(N-1)^2} \cdot \text{tr}(K_c L_c), \tag{8}$$

where $\text{tr}(\cdot)$ denotes the matrix trace operator. A lower HSIC value indicates greater statistical independence between $F_{(mod)}$ and $F_{(shared)}$.

5. Cross-modal Retrieval Results

To provide a qualitative comparison between our method and the previous STOA approach, we visualized the cross-modal gait recognition results, as shown in Figure 2.

6. Limitation and Future Works

Existing cross-modal gait recognition datasets (e.g., SUSTech1K and FreeGait) collect data from individuals at short time intervals. However, in long-life scenarios, such data may become ineffective due to factors such as seasonal variations, differing weather conditions, or changes in clothing or movement patterns resulting from health conditions, all of which can lead to substantial domain shifts in the gait data. Exploring long-time interval cross-modal gait recognition in long-life settings presents both a compelling and practical research avenue.

References

[1] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[2] Zhihao Qian, Yutian Lin, and Bo Du. Visible–infrared person re-identification via patch-mixed cross-modality learning. *Pattern Recognition*, 157:110873, 2025. 1

[3] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 1

[4] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3