Gait Recognition via Collaborating Discriminative and Generative Diffusion Models

Haijun Xiong, Bin Feng*, Bang Wang, Xinggang Wang, Wenyu Liu

School of Electronic Information and Communications, Huazhong University of Science and Technology
{xionghj, fengbin}@hust.edu.cn

Abstract

Gait recognition offers a non-intrusive biometric solution by identifying individuals through their walking patterns. Although discriminative models have achieved notable success in this domain, the full potential of generative models remains largely underexplored. In this paper, we introduce **CoD^2**, a novel framework that combines the data distribution modeling capabilities of diffusion models with the semantic representation learning strengths of discriminative models to extract robust gait features. We propose a Multi-level Conditional Control strategy that incorporates both high-level identity-aware semantic conditions and low-level visual details. Specifically, the high-level condition, extracted by the discriminative extractor, guides the generation of identity-consistent gait sequences, whereas low-level visual details, such as appearance and motion, are preserved to enhance consistency. Furthermore, the generated sequences facilitate the discriminative extractor's learning, enabling it to capture more comprehensive high-level semantic features. Extensive experiments on four datasets (SUSTech1K, CCPG, GREW, and Gait3D) demonstrate that CoD^2 achieves state-of-the-art performance and can be seamlessly integrated with existing discriminative methods, yielding consistent improvements.

Introduction

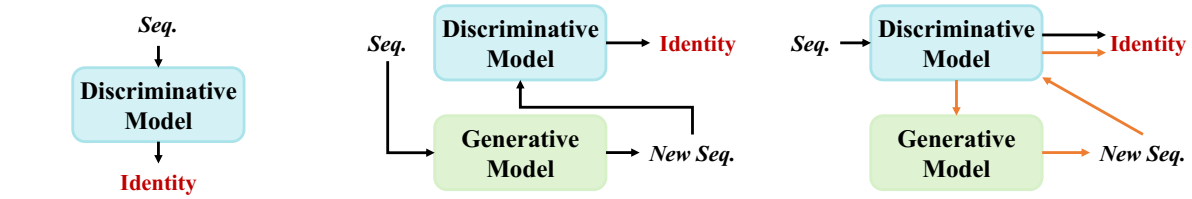
Gait recognition is a biometric technology that distinguishes individuals by unique walking patterns. Unlike other biometric modalities, such as face, iris, and fingerprint recognition, gait can be captured from a distance without requiring subject cooperation, making it particularly suitable for applications in crime prevention, sports science, and healthcare (Venkat and De Wilde 2011; Sepas-Moghaddam and Etemad 2022). Despite significant progress in gait recognition, existing discriminative methods (Wang et al. 2023c; Ye et al. 2024; Xiong et al. 2025) (Figure 1 (a)) continue to struggle in complex scenarios involving variations in clothing, viewpoints, occlusions, and carried objects, which complicate the extraction of robust discriminative features.

Generative models, particularly diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021), have recently gained significant attention for their remarkable capability to generate high-quality, visually compelling images. These models excel at capturing complex data distributions and generate realistic samples by iteratively reversing a noise injection process. Beyond image synthesis, the potential of diffusion models has been increasingly explored in video generation (Ho et al. 2022), where they effectively capture temporal coherence and high-level structural dynamics. Such characteristics make them especially suitable for tasks that demand both realistic visual generation and consistent motion evolution, including video synthesis and dynamic scene modeling (Yu et al. 2024; Wu et al. 2025). Furthermore, due to their powerful representational capacity, recent works have leveraged pre-trained diffusion models for a variety of downstream applications, achieving promising results in pose estimation (Feng et al. 2023), mesh recovery (Zhu et al. 2024; Foo et al. 2023), and action recognition (Wu et al. 2024a; Li, Huang, and Mao 2023).

Previous studies (Jin et al. 2025) (Figure 1 (b)) have employed diffusion models to denoise RGB gait sequences and generate clean gait representations. However, such methods do not fully exploit the intrinsic relationship between generative and discriminative models, thereby limiting the potential of the generative model. While discriminative models emphasize inter-class separability, generative models focus on modeling the underlying data distribution. These two paradigms provide complementary perspectives on the data, and their integration can yield a more holistic understanding of gait patterns. Consequently, combining discriminative models with generative diffusion models is essential for enhancing the feature extraction capabilities of both, ultimately leading to more effective gait recognition.

To address these aforementioned challenges, we propose a novel gait recognition framework, **CoD^2**. As illustrated in Figure 1 (c), CoD^2 differs fundamentally from prior works by integrating the data distribution modeling capability of diffusion models with the semantic representation learning strength of discriminative models, thereby extracting more robust gait features. We further present a Multi-level Conditional Control strategy that combines both high-level and low-level conditions to guide the generative learning process of the diffusion model. Specifically, the high-level condition, derived from the discriminative feature extractor, provides identity-aware semantic information to generate identity-consistent gait sequences. In contrast, the low-level condi-

*Corresponding author



(a) Naive Discriminative methods (b) Generative-assisted methods (c) Our method

Figure 1: **Comparison of different methods for gait recognition.** (a) Naive discriminative methods, such as GaitSet (Chao et al. 2019); (b) Generative-assisted methods, such as DenoisingGait (Jin et al. 2025); (c) Our proposed CoD^2, which integrates collaborating discriminative and generative models.

tion preserves essential visual details, such as appearance and motion information, which are critical for maintaining identity consistency in the generated sequences. Moreover, the generated sequences in turn promote the training of the discriminative extractor, enabling it to capture richer and more comprehensive semantic representations. We evaluate CoD^2 through extensive experiments on four datasets (Shen et al. 2023; Li et al. 2023; Zhu et al. 2021; Zheng et al. 2022b), achieving state-of-the-art Rank-1 performance. Furthermore, integrating CoD^2 with four representative discriminative methods (Chao et al. 2019; Lin, Zhang, and Yu 2021; Fan et al. 2023, 2025) consistently improves performance across all datasets, demonstrating its strong versatility. Notably, CoD^2 introduces only a marginal increase in training consumption, with no impact on testing efficiency. In summary, the main contributions are as follows:

- We introduce CoD^2, a novel gait recognition framework that integrates the data distribution modeling capacity of generative diffusion models with the semantic representation learning ability of discriminative models, enhancing gait feature extraction through their complementary strengths.
- We propose a Multi-level Conditional Control strategy that jointly leverages high-level identity-aware semantic features with low-level visual details to guide the diffusion model's generative process. The generated sequences facilitate the discriminative model's learning, further improving feature robustness.
- Extensive experiments demonstrate that CoD^2 achieves state-of-the-art performance and can be seamlessly integrated with existing discriminative methods, consistently improving performance with minimal impact on training consumption and no effect on testing efficiency.

Related Work

Gait Recognition

Current gait recognition methods can be broadly categorized into model-based and appearance-based methods, depending on the input modality.

Model-based methods (Teepe et al. 2021, 2022; Li and Zhao 2022; Fu et al. 2023) exploit structural human priors, such as skeletons and 3D meshes. For example, PoseGait (Liao et al. 2020) integrates multiple skeleton-based features with human prior knowledge to enhance recognition performance, while CAG (Huang et al. 2023) uses adaptive conditional networks to extract fine-grained representations. Other studies (Pinyoanuntapong et al. 2023; Zhang et al. 2023) adopt transformer architectures to capture long-range spatial dependencies, and SMPLGait (Zheng et al. 2022b) further improves recognition by utilizing dense 3D mesh representations reconstructed from RGB images.

Appearance-based methods (Fan et al. 2023; Wang et al. 2024; Ma et al. 2023; Peng et al. 2024a; Zheng et al. 2022a; Wang et al. 2023b; Zheng et al. 2023; Xiong et al. 2024a; Zheng et al. 2024) directly learn spatial-temporal representations from gait silhouettes or RGB sequences. GaitSet (Chao et al. 2019) is the first to treat gait sequences as unordered frame sets. Subsequent methods (Fan et al. 2020; Huang et al. 2021; Lin, Zhang, and Yu 2021) adopt 1D or 3D CNNs to model local motion patterns across frames, while deeper architectures (Ma et al. 2024; Fan et al. 2025) have been developed to extract richer identity-discriminative features. Recent studies (Dou et al. 2023; Wang et al. 2023a; Xiong et al. 2024b) revisit gait recognition from a causal inference perspective, and DenoisingGait (Jin et al. 2025) employs diffusion models to generate noise-free gait representations. Moreover, alternative modalities, such as point clouds and RGB videos, have recently been incorporated into gait recognition frameworks (Shen et al. 2023; Ye et al. 2024), broadening the scope of this research field.

Diffusion Models for Representation Learning

Diffusion models have emerged as a powerful paradigm for generative modeling, particularly in image and video synthesis (Ho, Jain, and Abbeel 2020; Ho et al. 2022). These models generate high-quality visual content by progressively refining Gaussian noise through an iterative denoising process. Building on their remarkable success, recent studies have extended diffusion models to a wide range of downstream tasks (Xu, Guo, and Peng 2024; Wu et al. 2024a; Feng et al. 2023; Chen et al. 2023; Vogel et al. 2024; Toker et al. 2024; Kara et al. 2024; Wu et al. 2024b). For example, DPMesh (Zhu et al. 2024) leverages spatial structural priors from pre-trained diffusion models to reconstruct occluded human meshes, while HOIAnimator (Song et al. 2024) introduces Perceptive Diffusion Models to enhance the realism of human–object interactions in animations. Moreover, ControlNet (Zhang, Rao, and Agrawala 2023) integrates spatial conditioning mechanisms into pre-trained diffusion models

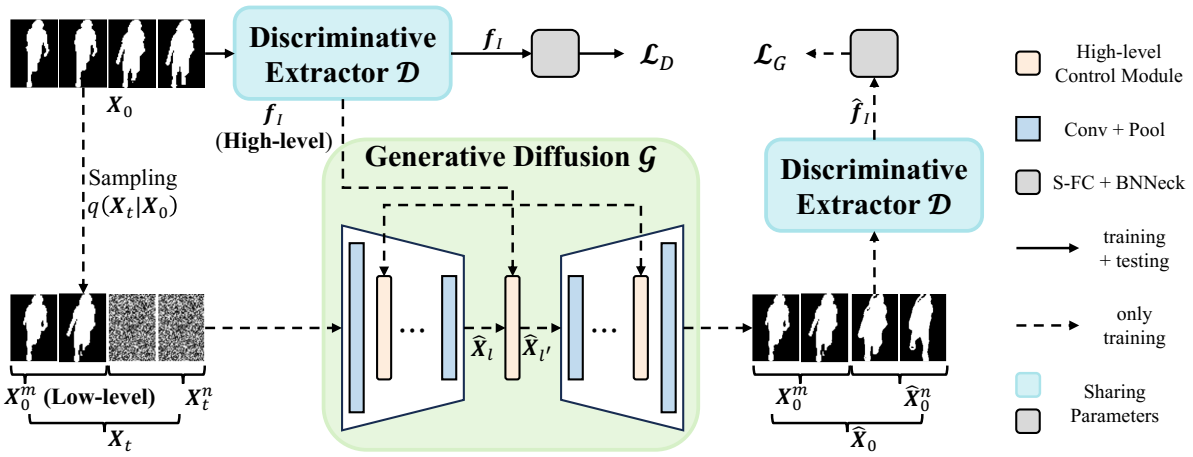


Figure 2: **Overview of our proposed method.** The discriminative extractor $\mathcal{D}$ (*e.g.*, GaitSet, GaitGL, GaitBase, or DeepGaitV2) first extracts the identity feature $\boldsymbol{f}_I$ from the input gait sequence $\boldsymbol{X}_0$. This feature serves as a high-level semantic condition to guide the generative diffusion model $\mathcal{G}$ during sequence generation. The noise sequence $\boldsymbol{X}_t$ is composed of Gaussian noise $\boldsymbol{X}_t^n \sim \mathcal{N}(0, I)$ and low-level visual information $\boldsymbol{X}_0^m$ sampled from $\boldsymbol{X}_0$. The generated gait sequence $\hat{\boldsymbol{X}}_0$ is then fed into $\mathcal{D}$ to extract the corresponding identity feature $\hat{\boldsymbol{f}}_I$. Finally, $\mathcal{D}$ and $\mathcal{G}$ are jointly optimized with the loss $\mathcal{L}_D$ and $\mathcal{L}_G$, where $\mathcal{G}$ is employed only for training, while $\mathcal{D}$ is used for both training and inference.

for precise detail manipulation, and AYG (Ling et al. 2024) combines Gaussian Splatting with diffusion models to enable text-to-4D generation.

In this paper, we propose CoD^2, the first framework that enhances feature extraction by unifying the semantic representation learning capability of discriminative models and the data distribution modeling power of generative models.

Methodology

Background

Before introducing our proposed method, we briefly review the key concepts of gait recognition and the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020).

Discriminative Gait Recognition. Given a gait sequence $\boldsymbol{X}_0 \in \mathbb{R}^{1\times T\times H\times W}$ with T frames, each of size (H, W), discriminative gait recognition methods typically process $\boldsymbol{X}_0$ through a discriminative feature extractor $\mathcal{D}$ to obtain the identity representation $\boldsymbol{f}_I \in \mathbb{R}^{C\times p}$, where C and p denote the number of channels and parts, respectively:

$$\boldsymbol{f}_I = \mathcal{D}(\boldsymbol{X}_0). \tag{1}$$

Subsequently, $\boldsymbol{f}_I$ is refined using a separate fully connected (S-FC) layer followed by BNNeck, and optimized with a combination of triplet and cross-entropy losses:

$$\mathcal{L}_D = \mathcal{L}_{tri} + \mathcal{L}_{ce}. \tag{2}$$

DDPM. DDPM generates high-quality visual content by iteratively denoising random Gaussian noise. It consists of two phases: a fixed forward diffusion process and a learnable reverse denoising process. In the forward phase, Gaussian noise is gradually added to the original image x_0 through a Markov chain, progressively transforming it into pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$. At each timestep t, the noised variable x_t depends only on its previous state x_{t-1}, as formulated by:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{3}$$

where β_t denotes a predefined variance schedule. The reverse process reconstructs x_0 from x_T through iterative denoising:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I), \tag{4}$$

where $\mu_\theta(x_t, t)$ is a parameterized function, typically implemented as a neural network, used to predict the mean $\hat{\mu}$ at each timestep. Recent methods, such as ControlNet (Zhang, Rao, and Agrawala 2023), extend diffusion models to controllable generation by incorporating conditional input. Given a condition c, the training objective can be formulated as:

$$\min_\theta \mathbb{E}_{x_0,c,t,\mu}\left[\|\mu - \mu_\theta(x_t, c, t)\|_2^2\right], \tag{5}$$

which enables the generation of realistic samples from Gaussian noise.

Pipeline

The overall framework of CoD^2 is shown in Figure 2. It comprises two discriminative extractors with shared parameters, and a generative diffusion module. Similar to previous methods, the first discriminative extractor $\mathcal{D}$ processes the input gait sequence $\boldsymbol{X}_0$ to obtain the identity feature $\boldsymbol{f}_I$. Meanwhile, a noise sequence $\boldsymbol{X}_t$ is constructed by combining the low-level condition $\boldsymbol{X}_0^m$ (a part of $\boldsymbol{X}_0$) with Gaussian noise $\boldsymbol{X}_t^n$. The identity feature, serving as a high-level condition, guides the denoising process of the generative diffusion module $\mathcal{G}$ by embedding identity-aware semantic information, resulting in a generated gait sequence $\hat{\boldsymbol{X}}_0$. The

second extractor $\mathcal{D}$ is then reapplied to extract the identity feature $\hat{\boldsymbol{f}}_I$ from $\hat{\boldsymbol{X}}_0$, ensuring identity consistency. This bidirectional interaction between $\mathcal{D}$ and $\mathcal{G}$ not only reinforces the generative module but also enhances the discriminative extractor's ability to capture more effective gait features.

Discriminative Extractor and Generative Diffusion Module

The discriminative extractor $\mathcal{D}$ serves as the core backbone and can be instantiated with various existing gait recognition models, such as GaitSet, GaitGL, GaitBase, and DeepGaitV2-P3D (abbreviated as DeepGaitV2). The versatility of CoD^2 is further validated in Table 5. Considering that binary silhouette are substantially simpler than RGB inputs and that directly predicting noise from noisy sequences provides limited discriminative information (Wu et al. 2024a; Guo et al. 2024), we adopt a lightweight generative diffusion module $\mathcal{G}$ (details in **Supplementary Materials**) to generate new sequences from noise.

Multi-level Conditional Control

The generative diffusion module $\mathcal{G}$ takes the noise sequence $\boldsymbol{X}_t$ and the identity feature $\boldsymbol{f}_I$ as input. Here, $\boldsymbol{f}_I$ serves as a high-level control condition, encapsulating identity-aware semantic information. Meanwhile, $\boldsymbol{X}_0^m$ in $\boldsymbol{X}_t$, derived from the original sequence $\boldsymbol{X}_0$, preserves low-level visual cues (such as appearance and motion), acting as a low-level control condition during the denoising process.

Low-level conditional control. Unlike text-to-video generation, gait sequence generation requires preserving visual details from original sequences, such as appearance and motion information. To achieve this, we introduce a sampling strategy that randomly selects continuous m frames from $\boldsymbol{X}_0$ as a reference, denoted as $\boldsymbol{X}_0^m \in \mathbb{R}^{1\times m\times H\times W}$. This reference is concatenated with Gaussian noise $\boldsymbol{X}_t^n \in \mathbb{R}^{1\times(T-m)\times H\times W}$ along the temporal dimension to construct the noise sequence $\boldsymbol{X}_t \in \mathbb{R}^{1\times T\times H\times W}$, formulated as:

$$\begin{aligned} \boldsymbol{X}_0^m &= \boldsymbol{X}_0[k:k+m], k\in[0,T-m], \\ \boldsymbol{X}_t^n &= \text{Sample}(\mathcal{N}(0,I)), \\ \boldsymbol{X}_t &= \text{Cat}(\boldsymbol{X}_0^m, \boldsymbol{X}_t^n), \end{aligned} \tag{6}$$

where $\text{Cat}(\cdot)$ denotes the concatenation operation. During denoising, the spatial-temporal modeling process transfers low-level visual cues from $\boldsymbol{X}_0^m$ to $\boldsymbol{X}_t^n$, ensuring that the generated sequences retain essential appearance and motion details. Inspired by LAMP (Wu et al. 2023), we keep the reference frames $\boldsymbol{X}_0^m$ noise-free during training, meaning that $\boldsymbol{X}_0^m$ remains unchanged after passing through a 3D convolutional layer in $\mathcal{G}$, *i.e.*,

$$\left[\hat{\boldsymbol{X}}_0^m, \hat{\boldsymbol{X}}_i^n\right] = \text{Conv}\left([\boldsymbol{X}_0^m, \boldsymbol{X}_i^n]\right), \quad \hat{\boldsymbol{X}}_0^m = \boldsymbol{X}_0^m,$$

which ensures both temporal identity consistency and the integrity of low-level visual details during denoising. By preserving low-level details, this strategy enhances control effectiveness and improves the overall sequence generation.

High-level conditional control. The high-level condition embeds identity-aware semantic information into the generative diffusion module $\mathcal{G}$, providing effective guidance during the generation process. While ControlNet (Zhang, Rao, and Agrawala 2023) performs element-wise addition for condition fusion after convolutional layers, we find this operation too coarse for gait sequence generation, leading to degraded performance (as shown in Table 7). To address this limitation, we propose a refined High-level Control Module that seamlessly integrates $\boldsymbol{f}_I$ into $\mathcal{G}$, facilitating identity-aware guidance and improving the generated sequences of generated sequences (as illustrated in Figure 3).

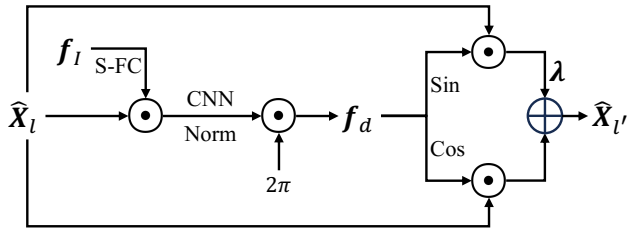


Figure 3: **Details of High-level Control Module**. The S-FC represents a separate fully connected layer, and $\boldsymbol{\lambda} \in \mathbb{R}^{C'}$ is a learnable channel-wise control vector that regulates the adjustment intensity across different feature channels.

We draw inspiration from Euler's formula:

$$e^{i\theta} = \cos(\theta) + i\sin(\theta), \tag{7}$$

which represents a signal as a rotation in the complex plane, thereby encoding both amplitude and phase information. Motivated by this, we design a phase modulation module based on sinusoidal projection to effectively embed high-level identity semantics into the generative process.

Specifically, given an intermediate noisy sequence $\hat{\boldsymbol{X}}_l \in \mathbb{R}^{C'\times T\times H\times W}$ from the reverse diffusion process, we compute a spatially varying phase feature $\boldsymbol{f}_d$ conditioned on the identity feature $\boldsymbol{f}_I$:

$$\boldsymbol{f}_d = 2\pi \cdot \text{Norm}(\text{Conv}(\hat{\boldsymbol{X}}_l \cdot \text{S-Fc}(\boldsymbol{f}_I))). \tag{8}$$

Here, S-FC($\boldsymbol{f}_I$) denotes a spatially broadcasted identity embedding, and $\text{Norm}(\boldsymbol{x}) = \frac{\boldsymbol{x}-\boldsymbol{x}_{\min}}{\boldsymbol{x}_{\max}-\boldsymbol{x}_{\min}}$ normalizes the values to the range $[0, 2\pi]$ via min-max normalization. We then apply sinusoidal modulation to inject identity-aware semantics into the sequence:

$$\hat{\boldsymbol{X}}_{l'} = \hat{\boldsymbol{X}}_l \cdot \cos(\boldsymbol{f}_d) + \boldsymbol{\lambda} \cdot \hat{\boldsymbol{X}}_l \cdot \sin(\boldsymbol{f}_d), \tag{9}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{C'}$ is a learnable channel-wise scaling vector. This formulation, grounded in Euler's identity (Equation 7), effectively modulates the intermediate representation $\hat{\boldsymbol{X}}_l$ with a phase shift parameterized by $\boldsymbol{f}_I$.

This identity-conditioned phase modulation enables the network to impose global semantic control in a spatially adaptive manner. As shown in Figure 3, the sinusoidal components allow smooth and differentiable injection of identity semantics, facilitating the generation of identity-consistent gait sequences.

By jointly incorporating high-level semantic and low-level visual conditions, our method ensures that the generated sequence $\hat{\boldsymbol{X}}_0$ preserves appearance and motion details while maintaining strong identity consistency, thereby enhancing discriminative effectiveness.

Modality	Method	Venue	Probe Sequence (R-1)								Overall	
			NM	BG	CL	CR	UB	UN	OC	NT	R-1	R-5
Silhouette	GaitSet	AAAI19	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
	GaitPart	CVPR19	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
	GaitGL	ICCV21	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
	GaitBase	CVPR23	81.5	77.5	<u>49.6</u>	75.8	75.5	76.7	81.4	25.9	76.1	89.4
	DeepGaitV2	TPAMI25	83.5	79.5	46.3	76.8	79.1	78.5	81.1	27.3	77.4	90.2
Silhouette + Skeleton	BiFusion	MTAP24	69.8	62.3	45.4	60.9	54.3	63.5	77.8	33.7	62.1	83.4
	SkeletonGait++	AAAI24	<u>85.1</u>	<u>82.9</u>	46.6	<u>81.9</u>	<u>80.8</u>	<u>82.5</u>	<u>86.2</u>	**47.5**	<u>81.3</u>	<u>95.5</u>
Silhouette	**Ours**	-	**87.9**	**84.5**	**55.4**	**82.8**	**87.2**	**85.1**	**88.7**	<u>38.6</u>	**83.8**	**95.8**

Table 1: Performance comparisons on SUSTech1K. The **best** and <u>second-best</u> results are highlighted in bold and underlined, respectively.

Dataset	Batch Size	Optimizer	Steps
SUSTech1K	(8, 4)	Adam (dr=0.1)	50K
CCPG	(8, 8)	lr=1e−4	60K
GREW	(32, 2)	wd=5e−4	180K
Gait3D	(32, 2)		60K

Table 2: **Implementation details**. The batch size (P, K) denotes P subjects and K sequences per subject. The parameters dr, lr, and wd refer to the decay rate, learning rate, and weight decay, respectively.

Training Objective

After obtaining the identity features $\boldsymbol{f}_I$ and $\hat{\boldsymbol{f}}_I$, we adopt a joint loss $\mathcal{L}$ to simultaneously optimize the discriminative extractor and the generative diffusion module. The overall objective is formulated as:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_G, \tag{10}$$

where $\mathcal{L}_D$ (defined in Equation 2) supervises $\boldsymbol{f}_I$, while $\mathcal{L}_G = \mathcal{L}_{tri} + \mathcal{L}_{ce}$ is applied to supervise the identity feature $\hat{\boldsymbol{f}}_I$ of the generated sequence to enforce identity consistency.

Experiments

In this section, we first describe the datasets used and implementation details. We then conduct extensive experiments to evaluate CoD^2, including both quantitative and qualitative analyses. Finally, comprehensive ablation studies on four datasets are performed to assess the contribution of each component within CoD^2. More experiments are provided in the **Supplementary Materials**.

Datasets and Evaluation Metrics

Datasets: We evaluate our method on four widely used datasets: SUSTech1K (Shen et al. 2023), CCPG (Li et al. 2023), GREW (Zhu et al. 2021), and Gait3D (Zheng et al. 2022b). SUSTech1K, collected in laboratory, includes conditions such as normal, clothing changes, night, and occlusion. CCPG is designed for cross-domain evaluation, comprising four clothing-change scenarios (*i.e.*, full-body, upper-body, lower-body, and backpacks changes). GREW and Gait3D are large-scale real-world datasets with significant challenges due to diverse environmental conditions. All training and testing splits strictly follow the official dataset protocols.

Metrics: Following prior work (Xiong et al. 2024b), we use Rank-k accuracy (R-k) and mean Average Precision (mAP) to evaluate the performance of CoD^2.

Implementation Details

(1) All images are resized to 64×44, and an ordered sampling strategy with a fixed sequence length of 30 frames is adopted during training. (2) We primarily employ DeepGaitV2 (Fan et al. 2025) as the discriminative extractor to validate CoD^2, and further assess its versatility with other baselines, including GaitSet (Chao et al. 2019), GaitGL (Lin, Zhang, and Yu 2021), and GaitBase (Fan et al. 2023). (3) Dataset-specific configurations are provided in Table 2. To ensure fairness, the batch size is halved due to the reuse of the discriminative extractor. (4) Further architectural detail of the generative diffusion module are presented in the **Supplementary Materials**. (5) The number of continuous frames m in Equation 6 is fixed to 5. (6) All experiments are conducted on Nvidia GeForce RTX 3090 GPUs.

Quantitative Results

Evaluation on SUSTech1K and CCPG. We compare CoD^2 with several recent methods (Chao et al. 2019; Fan et al. 2020; Lin, Zhang, and Yu 2021; Fan et al. 2023, 2025; Peng et al. 2024b; Fan et al. 2024) on the SUSTech1K and CCPG datasets. These results underscore the superiority of CoD^2.

Key observations from Table 1 are as follows: (1) Silhouette-based methods perform poorly under low-light conditions, achieving a maximum accuracy of only 27.3%, primarily due to degraded image quality caused by insufficient lighting. Despite this, CoD^2 consistently outperforms these methods across all conditions, with a notable improvement of +11.3% under the night condition compared to DeepGaitV2, which achieves the second-highest accuracy (silhouette-based methods) at 27.3%. This highlights CoD^2's enhanced ability to extract discriminative features,

Method	Venue	Gait Evaluation Protocol				
		CL	UP	DN	BG	Mean
GaitSet	AAAI19	60.2	65.2	65.1	68.5	64.8
GaitPart	CVPR20	64.3	67.8	68.6	71.7	68.1
GaitBase	CVPR23	71.6	75.0	76.8	78.6	75.5
DeepGaitV2	TPAMI25	78.6	84.8	80.7	89.2	83.3
Ours	-	**80.1**	**86.9**	**81.6**	**90.9**	**84.8**

Table 3: Performance comparisons on CCPG.

Method	Venue	GREW		Gait3D	
		R-1	R-5	R-1	mAP
GaitSet	AAAI19	46.3	63.6	36.7	30.0
GaitPart	CVPR19	44.0	60.7	28.2	21.6
GaitGL	ICCV21	47.3	63.6	29.7	22.3
SMPLGait	CVPR22	-	-	46.3	37.2
DANet	CVPR23	-	-	48.0	-
GaitBase	CVPR23	60.1	-	64.6	-
GaitGCI	CVPR23	68.5	80.8	50.3	39.5
HSTL	ICCV23	62.7	76.6	61.3	55.5
DyGait	ICCV23	71.4	83.2	66.3	56.4
DeepGaitV2	TPAMI25	77.7	87.9	74.4	65.8
QAGait	AAAI24	59.1	74.0	67.0	56.5
VPNet	CVPR24	<u>80.0</u>	<u>89.4</u>	75.4	-
CLTD	ECCV24	78.0	87.8	69.7	-
WaveLoss	AAAI25	-	-	<u>75.6</u>	<u>66.5</u>
Ours	-	**81.2**	**90.8**	**78.3**	**71.2**

Table 4: Performance comparisons on GREW and Gait3D.

especially in challenging low-quality silhouette scenarios, such as those encountered at night. (2) CoD^2 achieves state-of-the-art results in most conditions (seven out of eight), outperforming SkeletonGait++ (a multimodal-based method), demonstrating that our method effectively leverages silhouette data alone without relying on additional modalities.

In Table 3, CoD^2 achieves SOTA results across all scenarios, with an average Rank-1 accuracy of 84.8%. This demonstrates that CoD^2 effectively combines the strengths of discriminative and generative models, significantly improving the discriminative model under various clothing conditions.

Evaluation on GREW and Gait3D. The results on the challenging GREW and Gait3D datasets, summarized in Table 4, demonstrate that CoD^2 outperforms all previous methods. Specifically, on GREW, CoD^2 surpasses VPNet and CLTD by +1.2% and +3.2%, respectively, achieving a Rank-1 accuracy of 81.2%. On Gait3D, CoD^2 improves upon VPNet by +2.9% and WaveLoss (Wang and Wu 2025) by +2.7%, reaching a Rank-1 accuracy of 78.3%. Importantly, CoD^2 significantly outperforms its baseline, DeepGaitV2, with improvements of +3.5% on GREW (81.2% *vs.* 77.7%) and +3.9% on Gait3D (78.3% *vs.* 74.4%). These results further validate the effectiveness of CoD^2 in extracting discriminative gait features under real-world conditions.

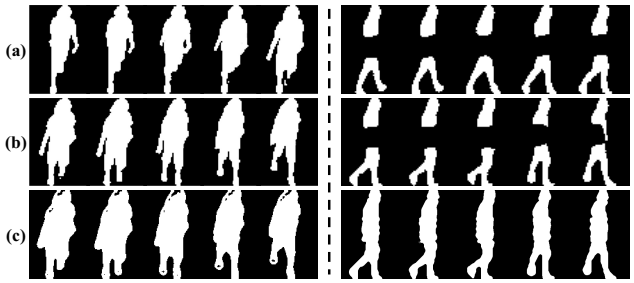


Figure 4: **Visualization of the generated sequence.** From top to bottom, sequences represent $\boldsymbol{X}_0^m$, the ground truth of $\boldsymbol{X}_t^n$, and $\hat{\boldsymbol{X}}_0^n$ in Figure 2, respectively.

Method	SUSTech1K	CCPG	GREW	Gait3D
GaitSet	65.0	64.8	46.3	36.7
+ CoD^2	71.3$^{+6.3\%}$	68.9$^{+4.1\%}$	54.1$^{+7.8\%}$	44.3$^{+7.6\%}$
GaitGL	63.1	66.2	47.3	29.7
+ CoD^2	69.9$^{+6.8\%}$	68.9$^{+2.7\%}$	51.5$^{+4.2\%}$	35.7$^{+6.0\%}$
GaitBase	76.1	75.5	60.1	64.6
+ CoD^2	84.2$^{+8.1\%}$	79.4$^{+3.9\%}$	71.1$^{+11.0\%}$	72.6$^{+8.0\%}$
DeepGaitV2	77.4	83.3	77.7	74.4
+ CoD^2	83.8$^{+6.4\%}$	84.8$^{+1.5\%}$	81.2$^{+3.5\%}$	78.3$^{+3.9\%}$

Table 5: Performance improvements (Rank-1 accuracy) of CoD^2 across different baselines on four datasets.

Qualitative Results

Figure 4 illustrates that the generated sequences closely resemble the original ones, demonstrating the effectiveness of our generative diffusion module in synthesizing realistic gait sequences. This success is attributed to the integration of visual details (*e.g.*, appearance and motion) with high-level identity-aware semantic information. The visualizations also highlight the discriminative extractor's ability to learn discriminative gait features, even when the generated sequences deviate from the originals. Notably, as shown on the right side of Figure 4, the goal of the generative diffusion model is not merely to replicate the ground truth, but to capture and enhance discriminative gait information, thereby improving recognition robustness.

Ablation Studies

Versatility of CoD^2. Table 5 demonstrates that our method significantly improves performance across four discriminative extractors on four datasets, highlighting the effectiveness and versatility of collaboratively integrating discriminative and generative diffusion models for gait recognition. Notably, we observe that incorporating CoD^2 with non-temporal modeling methods (*e.g.*, GaitSet and GaitBase) yields greater performance improvements compared to temporal modeling methods (*e.g.*, GaitGL, and DeepGaitV2). This is due to the generative diffusion model's ability to introduce rich temporal dynamics, which particularly benefits non-temporal modeling methods.

High-level	Low-level	SUSTech1K	CCPG	GREW	Gait3D
✗	✗	77.4	83.3	77.7	74.4
✓	✗	81.9	84.0	80.4	77.4
✗	✓	81.3	83.7	79.9	77.2
✓	✓	**83.8**	**84.8**	**81.2**	**78.3**

Table 6: The ablation study of Multi-level Conditional Control strategy.

Method	SUSTech1K	CCPG	GREW	Gait3D
Baseline	81.3	83.7	79.9	77.2
w/ addition	83.0	84.4	80.7	77.6
w/ Ours	**83.8**	**84.8**	**81.2**	**78.3**

Table 7: The ablation study of High-level Control Module.

m	SUSTech1K	CCPG	GREW	Gait3D
1	81.2	83.7	79.6	76.9
3	83.3	84.5	80.5	77.6
5	83.8	**84.8**	**81.2**	**78.3**
7	**84.0**	**84.8**	80.9	78.0
9	83.4	84.2	80.6	77.8

Table 8: The ablation study of the number of continuous frames m in Equation 6.

Effectiveness of Multi-level Conditional Control strategy. Table 6 investigates the impact of the Multi-level Conditional Control strategy. The results indicate that both conditions independently improve recognition accuracy, confirming that the extractor $\mathcal{D}$ effectively learns discriminative identity features. Moreover, the combination of both conditions leads to even greater performance, underscoring their complementary properties.

Effectiveness of High-level Control Module. Table 7 evaluates the effectiveness of High-level Control Module by comparing it with the baseline (low-level conditional control only) and an element-wise addition strategy. The results demonstrate that our control strategy outperforms direct element-wise addition, highlighting the advantages of High-level Control Module in improving the identity consistency and quality of generated sequences.

Impact of the number of continuous frames m. The number of continuous frames, m, plays a crucial role in balancing the low-level visual and high-level semantic conditions. As listed in Table 8, experiments with different values of m (*i.e.*, $m \in \{1, 3, 5, 7, 9\}$) reveal the following trends: (1) When low-level visual information is highly limited (*i.e.*, $m = 1$), the generative diffusion model struggles to learn discriminative features, resulting in suboptimal performance compared to using a single condition. (2) When m is too large, the generative diffusion model does not need to learn sufficiently, limiting its ability to guide the discriminative extractor and degrading performance. Based on a comprehensive evaluation across four datasets, we set m to 5.

$\boldsymbol{\lambda}$	SUSTech1K	CCPG	GREW	Gait3D
1	83.2	84.3	80.3	77.4
learnable scalar	83.5	84.4	80.8	77.8
learnable vector	**83.8**	**84.8**	**81.2**	**78.3**

Table 9: The ablation study of the learnable vector $\boldsymbol{\lambda}$ in Equation 9.

Method	Training (hour)	Testing (second)
GaitSet	0.95	41
+ CoD^2	1.08 (**+13.7%**)	41 (**+0%**)
GaitGL	2.91	43
+ CoD^2	3.35 (**+15.1%**)	43 (**+0%**)
GaitBase	5.83	77
+ CoD^2	6.29 (**+7.9%**)	77 (**+0%**)
DeepGaitV2	9.98	95
+ CoD^2	10.94 (**+9.6%**)	95 (**+0%**)

Table 10: Training and testing resource consumption on Gait3D. Training is calculated across four GPUs, while testing uses a single GPU.

Impact of the learnable vector $\boldsymbol{\lambda}$. Table 9 compares different strategies for $\boldsymbol{\lambda}$ (*i.e.*, a fixed value, a learnable scalar, and our learnable vector) in Equation 9. The results show that using a learnable weight to control the adjustment intensity generally enhances recognition performance. Furthermore, the learnable vector achieves the best results, as it allows for adaptive adjustments across different channels.

Training and Testing Resource Consumption. Table 10 analyzes the resource consumption of our method on Gait3D during both training and testing. While training demands increase by 7.9% to 15.1% compared to baselines (*i.e.*, GaitSet, GaitGL, GaitBase, and DeepGaitV2), the overhead remains acceptable by halving the batch size, despite the reuse of the discriminative extractor and the introduction of the generative diffusion model. Notably, testing resource requirements remain unchanged. The results in Table 5 and Table 10 demonstrate that CoD^2 achieves significant performance gains without compromising testing efficiency.

Conclusion

In this paper, we propose CoD^2, an novel gait recognition framework that collaboratively combines the data distribution modeling capabilities of diffusion models with the semantic representation learning strengths of discriminative models. We introduce a Multi-level Conditional Control strategy that integrates high-level identity-aware semantic information with low-level visual details to guide the generation process. Furthermore, ensuring identity consistency of generated sequences enhances the discriminative model's ability to learn robust gait features. We assess the effectiveness and versatility of CoD^2 on four datasets.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62376102).

References

Chao, H.; He, Y.; Zhang, J.; and Feng, J. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, 8126–8133.

Chen, L.-H.; Zhang, J.; Li, Y.; Pang, Y.; Xia, X.; and Liu, T. 2023. Humanmac: Masked motion completion for human motion prediction. In *CVPR*, 9544–9555.

Dou, H.; Zhang, P.; Su, W.; Yu, Y.; Lin, Y.; and Li, X. 2023. Gaitgci: Generative counterfactual intervention for gait recognition. In *CVPR*, 5578–5588.

Fan, C.; Hou, S.; Liang, J.; Shen, C.; Ma, J.; Jin, D.; Huang, Y.; and Yu, S. 2025. OpenGait: A Comprehensive Benchmark Study for Gait Recognition towards Better Practicality. *IEEE TPAMI*, 47(10): 8397–8414.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023. Opengait: Revisiting gait recognition towards better practicality. In *CVPR*, 9707–9716.

Fan, C.; Ma, J.; Jin, D.; Shen, C.; and Yu, S. 2024. SkeletonGait: Gait Recognition Using Skeleton Maps. In *AAAI*, volume 38, 1662–1669.

Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; and He, Z. 2020. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, 14225–14233.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023. DiffPose: SpatioTemporal diffusion model for video-based human pose estimation. In *ICCV*, 14861–14872.

Foo, L. G.; Gong, J.; Rahmani, H.; and Liu, J. 2023. Distribution-aligned diffusion for human mesh recovery. In *ICCV*, 9221–9232.

Fu, Y.; Meng, S.; Hou, S.; Hu, X.; and Huang, Y. 2023. Gpgait: Generalized pose-based gait recognition. In *ICCV*, 19595–19604.

Guo, X.; Zheng, M.; Hou, L.; Gao, Y.; Deng, Y.; Wan, P.; Zhang, D.; Liu, Y.; Hu, W.; Zha, Z.; et al. 2024. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *NeurIPS*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video Diffusion Models. *NeurIPS*, 35: 8633–8646.

Huang, X.; Wang, X.; Jin, Z.; Yang, B.; He, B.; Feng, B.; and Liu, W. 2023. Condition-adaptive graph convolution learning for skeleton-based gait recognition. *IEEE TIP*, 32: 4773–4784.

Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B.; He, B.; Liu, W.; and Feng, B. 2021. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, 12909–12918.

Jin, D.; Fan, C.; Ma, J.; Zhou, J.; Chen, W.; and Yu, S. 2025. On Denoising Walking Videos for Gait Recognition. In *CVPR*, 12347–12357.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, 6507–6516.

Li, C.; Huang, Q.; and Mao, Y. 2023. DD-GCN: Directed diffusion graph convolutional network for skeleton-based human action recognition. In *ICME*, 786–791. IEEE.

Li, N.; and Zhao, X. 2022. A strong and robust skeleton-based gait recognition method with gait periodicity priors. *IEEE TMM*, 25: 3046–3058.

Li, W.; Hou, S.; Zhang, C.; Cao, C.; Liu, X.; Huang, Y.; and Zhao, Y. 2023. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *CVPR*, 13824–13833.

Liao, R.; Yu, S.; An, W.; and Huang, Y. 2020. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 98: 107069.

Lin, B.; Zhang, S.; and Yu, X. 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, 14648–14656.

Ling, H.; Kim, S. W.; Torralba, A.; Fidler, S.; and Kreis, K. 2024. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *CVPR*, 8576–8588.

Ma, K.; Fu, Y.; Cao, C.; Hou, S.; Huang, Y.; and Zheng, D. 2024. Learning Visual Prompt for Gait Recognition. In *CVPR*, 593–603.

Ma, K.; Fu, Y.; Zheng, D.; Cao, C.; Hu, X.; and Huang, Y. 2023. Dynamic aggregated network for gait recognition. In *CVPR*, 22076–22085.

Peng, G.; Wang, Y.; Zhao, Y.; Zhang, S.; and Li, A. 2024a. Glgait: a global-local temporal receptive field network for gait recognition in the wild. In *ACM MM*, 826–835.

Peng, Y.; Ma, K.; Zhang, Y.; and He, Z. 2024b. Learning rich features for gait recognition by integrating skeletons and silhouettes. *MTAP*, 83(3): 7273–7294.

Pinyoanuntapong, E.; Ali, A.; Wang, P.; Lee, M.; and Chen, C. 2023. Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. In *ICASSP*, 1–5. IEEE.

Sepas-Moghaddam, A.; and Etemad, A. 2022. Deep Gait Recognition: A Survey. *IEEE TPAMI*, 45(1): 264–284.

Shen, C.; Fan, C.; Wu, W.; Wang, R.; Huang, G. Q.; and Yu, S. 2023. Lidargait: Benchmarking 3d gait recognition with point clouds. In *CVPR*, 1054–1063.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024. HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models. In *CVPR*, 811–820.

Teepe, T.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2022. Towards a deeper understanding of skeleton-based gait recognition. In *CVPRW*, 1569–1577.

Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2021. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, 2314–2318.

Toker, A.; Eisenberger, M.; Cremers, D.; and Leal-Taixé, L. 2024. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *CVPR*, 27695–27705.

Venkat, I.; and De Wilde, P. 2011. Robust gait recognition by learning and exploiting sub-gait characteristics. *IJCV*, 91: 7–23.

Vogel, M.; Tateno, K.; Pollefeys, M.; Tombari, F.; Rakotosaona, M.-J.; and Engelmann, F. 2024. P2P-Bridge: Diffusion Bridges for 3D Point Cloud Denoising. In *ECCV*, 184–201. Springer.

Wang, J.; Hou, S.; Huang, Y.; Cao, C.; Liu, X.; Huang, Y.; and Wang, L. 2023a. Causal intervention for sparse-view gait recognition. In *ACM MM*, 77–85.

Wang, L.; Liu, B.; Liang, F.; and Wang, B. 2023b. Hierarchical Spatio-Temporal Representation Learning for Gait Recognition. In *ICCV*, 19639–19649.

Wang, M.; Guo, X.; Lin, B.; Yang, T.; Zhu, Z.; Li, L.; Zhang, S.; and Yu, X. 2023c. DyGait: Exploiting Dynamic Representations for High-performance Gait Recognition. In *ICCV*, 13424–13433.

Wang, Z.; Hou, S.; Zhang, M.; Liu, X.; Cao, C.; Huang, Y.; Li, P.; and Xu, S. 2024. QAGait: Revisit Gait Recognition from a Quality Perspective. In *AAAI*, volume 38, 5785–5793.

Wang, Z.; and Wu, Q. 2025. WaveLoss: An Adaptive Dynamic Loss for Deep Gait Recognition. In *AAAI*, volume 39, 8259–8267.

Wu, L.; Lin, L.; Zhang, J.; Ma, Y.; and Liu, J. 2024a. MacDiff: Unified Skeleton Modeling with Masked Conditional Diffusion. In *ECCV*, 110–128. Springer.

Wu, R.; Chen, L.; Yang, T.; Guo, C.; Li, C.; and Zhang, X. 2023. LAMP: Learn A Motion Pattern for Few-Shot-Based Video Generation. *arXiv preprint arXiv:2310.10769*.

Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2025. Cat4d: Create anything in 4d with multi-view video diffusion models. In *CVPR*, 26057–26068.

Wu, W.; Fan, Q.; Qin, S.; Gu, H.; Zhao, R.; and Chan, A. B. 2024b. FreeDiff: Progressive Frequency Truncation for Image Editing with Diffusion Models. In *ECCV*, 194–209. Springer.

Xiong, H.; Deng, Y.; Feng, B.; Wang, X.; and Liu, W. 2024a. Gaitgs: Temporal feature learning in granularity and span dimension for gait recognition. In *2024 IEEE International Conference on Image Processing (ICIP)*, 2410–2416. IEEE.

Xiong, H.; Feng, B.; Wang, B.; Wang, X.; and Liu, W. 2025. MambaGait: Gait recognition approach combining explicit representation and implicit state space model. *IMAVIS*, 105597.

Xiong, H.; Feng, B.; Wang, X.; and Liu, W. 2024b. Causality-inspired Discriminative Feature Learning in Triple Domains for Gait Recognition. In *ECCV*, 251–270. Springer.

Xu, J.; Guo, Y.; and Peng, Y. 2024. FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Models. In *CVPR*, 561–570.

Ye, D.; Fan, C.; Ma, J.; Liu, X.; and Yu, S. 2024. BigGait: Learning Gait Representation You Want by Large Vision Models. In *CVPR*, 200–210.

Yu, H.; Wang, C.; Zhuang, P.; Menapace, W.; Siarohin, A.; Cao, J.; Jeni, L.; Tulyakov, S.; and Lee, H.-Y. 2024. 4real: Towards photorealistic 4d scene generation via video diffusion models. *NeurIPS*, 37: 45256–45280.

Zhang, C.; Chen, X.-P.; Han, G.-Q.; and Liu, X.-J. 2023. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6): e13244.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zheng, J.; Liu, X.; Gu, X.; Sun, Y.; Gan, C.; Zhang, J.; Liu, W.; and Yan, C. 2022a. Gait recognition in the wild with multi-hop temporal switch. In *ACM MM*, 6136–6145.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022b. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, 20228–20237.

Zheng, J.; Liu, X.; Wang, S.; Wang, L.; Yan, C.; and Liu, W. 2023. Parsing is all you need for accurate gait recognition in the wild. In *ACM MM*, 116–124.

Zheng, J.; Liu, X.; Zhang, B.; Yan, C.; Zhang, J.; Liu, W.; and Zhang, Y. 2024. It takes two: Accurate gait recognition in the wild via cross-granularity alignment. In *ACM MM*, 8786–8794.

Zhu, Y.; Li, A.; Tang, Y.; Zhao, W.; Zhou, J.; and Lu, J. 2024. DPMesh: Exploiting Diffusion Prior for Occluded Human Mesh Recovery. In *CVPR*, 1101–1110.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021. Gait recognition in the wild: A benchmark. In *ICCV*, 14789–14799.