ZipGait: Bridging Skeleton and Silhouette with Diffusion Model for Advancing Gait Recognition

Fanxu Min, Qing Cai, Shaoxiang Guo, Yang Yu, Fan Hao, Junyu Dong

School of Information Science and Engineering, Ocean University of China,

Abstract

Current gait recognition research predominantly focuses on extracting appearance features effectively, but the performance is severely compromised by the vulnerability of silhouettes under unconstrained scenes. Consequently, numerous studies have explored how to harness information from various models, particularly by sufficiently utilizing the intrinsic information of skeleton sequences. While these model-based methods have achieved significant performance, there is still a huge gap compared to appearance-based methods, which implies the potential value of bridging silhouettes and skeletons. In this work, we make the first attempt to reconstruct dense body shapes from discrete skeleton distributions via the diffusion model, demonstrating a new approach that connects cross-modal features rather than focusing solely on intrinsic features to improve model-based methods. To realize this idea, we propose a novel gait diffusion model named DiffGait, which has been designed with four specific adaptations suitable for gait recognition. Furthermore, to effectively utilize the reconstructed silhouettes and skeletons, we introduce Perception Gait Integration (PGI) to integrate different gait features through a two-stage process. Incorporating those modifications leads to an efficient model-based gait recognition framework called **ZipGait**. Through extensive experiments on four public benchmarks, ZipGait demonstrates superior performance, outperforming the state-of-the-art methods by a large margin under both cross-domain and intra-domain settings, while achieving significant plug-and-play performance improvements.

1 Introduction

Gait recognition, as a biometric technology, facilitates remote identification in uncontrolled settings without necessitating subject participation, discerning individuals based on their gait patterns (Wu et al. 2016; Wang et al. 2003). Current research focuses on utilizing shape information of silhouettes for recognition, known as appearance-based methods (Dou et al. 2024, 2023; Wang et al. 2023a; Lin, Zhang, and Yu 2021; Zheng et al. 2022a). However, these methods are vulnerable to complex backgrounds, severe occlusion, arbitrary viewpoints, and diverse clothing changes. This gives rise to a surging interest in model-based methods that extract gait features through various models, including 2D/3D Skeleton, SMPL model, and Point cloud (Loper et al. 2023; Yam, Nixon, and Carter 2004; Shen et al. 2023).

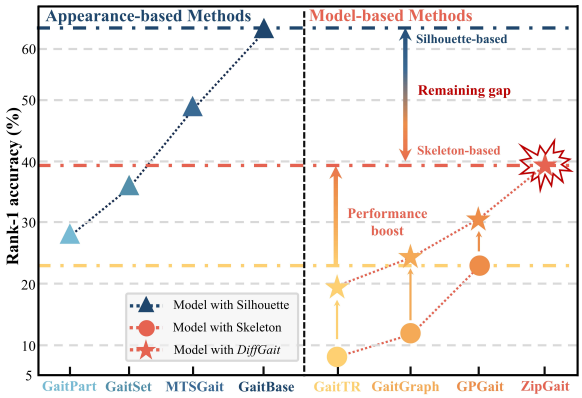

Figure 1: Comparison with alternative methods on the Gait3D test set elucidates the performance disparities between two types of methods in real-world scenarios. DiffGait enhances model-based methods through a plug-and-play approach, reducing the performance gap. Meanwhile, ZipGait achieves the best-performing compared to methods using skeletons.

Among them, 2D skeleton is particularly attractive since it achieves higher pose estimation accuracy due to low spatial complexity which demonstrates simplicity and effectiveness (Liao et al. 2020; Li et al. 2020; Teepe et al. 2021). Most existing model-based works, using the skeleton as input, focus on extracting intrinsic information from skeleton sequences to provide sufficient features, broadly categorized as follows: utilization of skeleton structural features (i.e. position, angle, length, local-global (Liao et al. 2020)); computation of physical information from skeleton sequences (i.e. motion velocity, gait periodicity (Frank, Mannor, and Precup 2010; Liu et al. 2022)); generation of complementary skeleton information (i.e. multi-view, frame-level correlations (Gao et al. 2022; Wang, Chen, and Liu 2022)).

Although existing model-based methods have shown progressive improvements in performance, they continue to lag behind appearance-based approaches, suggesting that shape features can offer more gait imformation, as indicated in Fig 1. This observation prompts us to consider whether merely utilizing intrinsic information of skeletons could bridge this huge gap between these two types of methods. We propose a novel concept: establishing a natural correla-

Preprint

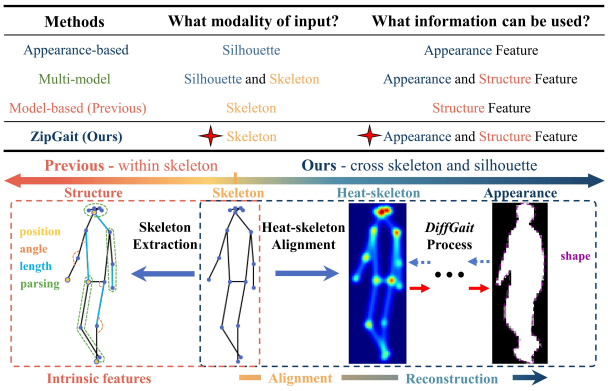

Figure 2: Differences between our approach and related gait recognition methods. Comparison of our cross-modal insight with previous works that primarily focused on skeleton extraction.

tion between silhouettes and skeletons to reconstruct dense body shapes from sparse skeleton structures, inspired by the progress of diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020). As shown in Fig. 2, instead of solely leveraging information from intra-modality gait features, we shift towards connecting cross-modality gait features for a more effective gait representation, thereby enhancing the performance of model-based methods.

To realize the aforementioned idea, we introduce a novel Gait Diffusion Model named **DiffGait**, which denoises the skeleton features mixed with standard Gaussian noise using a diffusion model to retrieve matching silhouettes. Intuitively, several limitations hinder the diffusion model's direct application in gait recognition, which DiffGait addresses with four aspects: **1)** Given that silhouettes and skeletons represent two unalignable modalities that inherently resist direct transformation via the diffusion process, we employ Heat-skeleton Alignment to transform 2D skeleton joints into a unified mesh grid, bridging skeletons and silhouettes across spatial-temporal dimensions; **2)** To address the cumbersome algorithms of the standard diffusion process for gait recognition, the DiffGait Forward Process has been developed to convert the diffusion object from the silhouette to the Hybrid Gait Volume (HGV), streamlining the condition and noise encoding stages; **3)** Recognizing that existing diffusion models neglect the intrinsic correspondence between modalities during the denoising process, we implement the DiffGait Reverse Process to produce multi-level silhouettes, which improves the distinction in denoising outcomes; **4)** Mainstream diffusion models, predominantly based on U-Net (Ronneberger, Fischer, and Brox 2015), are unsuitable for gait recognition due to their large scale. In response, our DiffGait Architecture adopts a decoder-only configuration to achieve a surprisingly efficient model with just 1.9M parameters and a rapid inference speed of 3628 FPS.

Furthermore, to efficiently leverage appearance and structure features, we propose a gait feature fusion module termed **Perceptual Gait Integration (PGI)**, which operates through a two-stage process that refines silhouettes and extracts hybrid gait features respectively. By incorporating these two designs into our baseline, we create a simple-but-effective model-based gait recognition architecture, named **ZipGait**, which seamlessly bridges skeleton and silhouette like a zipper. Fig. 2 summarizes the distinctions between our proposed method and existing gait recognition approaches, which are further discussed in detail in the Appendix. C. Extensive experiments on four dominant datasets demonstrate that our method outperforms state-of-the-art approaches in both cross-domain and intra-domain settings while achieving significant plug-and-play performance improvements.

2 Related Work

Model-based Gait Recognition. These methods utilize the underlying structure of the human body as input. In particular, skeleton representation is enhanced by extracting intrinsic information from skeletal sequences. GaitGraph2 (Teepe et al. 2022) pre-computes skeletons to obtain joint positions, motion velocities, and length of bones as gait features. GPGait (Fu et al. 2023) transforms the arbitrary human pose into a unified representation and allows efficient partitions of the human graph. (Fan et al. 2024; Min et al. 2024) represents the coordinates of human joints as a heatmap to provide explicit structural features. (Li and Zhao 2022; Choi et al. 2019) utilize gait periodicity priors and frame-level discriminative power, respectively. (Huang et al. 2023) dynamically adapts to the specific attributes of each skeleton sequence and the corresponding view angle. (Pan et al. 2023) generate multi-view pose sequences for each single-view sample to reduce the cross-view variance. Existing studies have significantly improved the extraction of efficient gait features from skeletons. However, none has explored the connection between structure and appearance to improve the performance of model-based methods.

Appearance-based Gait Recognition. These methods attempt to learn gait features directly from silhouette sequences, which has a big performance gap compared with model-based methods. GaitSet (Chao et al. 2019) innovatively regarded the gait sequence as a set and compressed frame-level spatial features. GaitPart (Fan et al. 2020) carefully explored the local details of the input silhouette and modeled the temporal dependencies. GaitBase (Fan et al. 2023) is an efficient baseline model that demonstrates applicability across various frameworks and gait modalities. MTSGait (Zheng et al. 2022a) learns spatial features and multi-scale temporal features simultaneously. This demonstrates the feasibility of bridging the gap between skeletons and silhouettes to improve the model-based methods.

Additionally, recent works aim to effectively integrate gait modalities to enhance recognition performance (Peng et al. 2024; Cui and Kang 2023; Dong et al. 2024; Zheng et al. 2022b), and some research is devoted to obtaining richer gait representations (Guo et al. 2023; Pinyoanuntapong et al. 2023; Han and Bhanu 2005; Wang et al. 2023b; Zheng et al. 2023b; Zou et al. 2024; Wang et al. 2024). Our proposed ZipGait diverges from these multimodal methods that use both silhouettes and skeletons as inputs, we only use skeletons. For a detailed comparison, see the Appendix. C.

Diffusion Model. Diffusion models, also widely known as DDPM, are a series of models generated through Markov chains trained via variational inference (Ho, Jain, and Abbeel 2020). (Song, Meng, and Ermon 2020) replaced the

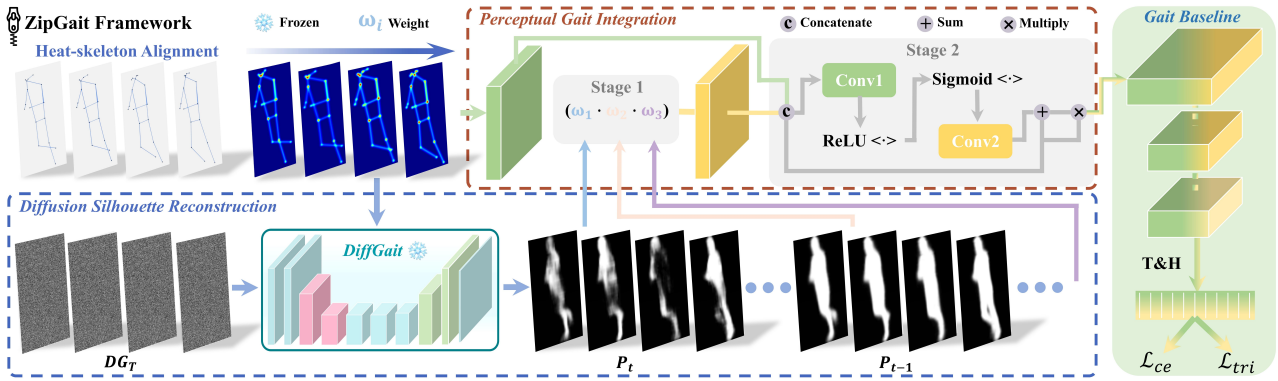

Figure 3: Overall pipeline of the proposed ZipGait framework. It consists of two fundamental improvements: DiffGait and Perceptual Gait Integration. The entire denoising process of DiffGait is summarized as Diffusion Silhouette Reconstruction. **T&H** represents the horizontal mapping (Fu et al. 2019) and temporal aggregation.

Markovian forward process used in earlier studies with a non-Markovian process, introducing the well-known DDIM. LDM (Rombach et al. 2022) has achieved image generation guided by textual inputs, advancing research in cross-modal generation using diffusion models. Diffusion-based methods have achieved impressive results in generation and segmentation tasks (Chen et al. 2023; Feng et al. 2023; Holmquist and Wandt 2023; Gong et al. 2023). Inspired by these studies, we have introduced diffusion models to generate dense body shapes from sparse skeleton distributions, thereby establishing a direct relational model between gait modalities for the first time.

3 Method

3.1 Problem Formulation

Denote a human skeleton sequence with total N frames as $\mathcal{V}_s = \{\mathcal{S}_t\}_{t=1}^{N}$, where $\mathcal{S}_t \in \mathbb{R}^{kp\times 3}$ refers to the t-th frame, previous methods identify individuals based on skeleton structure features $\mathcal{F}_s$ within the video sequence $\mathcal{V}_s$. Our work is the first attempt to establish the natural correlation between appearance and structure features to eliminate discrepancies between silhouette and skeleton, which can easily reconstruct silhouette appearance features $\mathcal{F}_a$ from $\mathcal{F}_s$ and achieve excellent performance.

3.2 ZipGait Workflow

Inspired by (Min et al. 2024; Duan et al. 2022; Fan et al. 2024), our work employs heatmaps to represent skeletons, termed 'Heat-skeleton'. Leveraging the OpenGait (Fan et al. 2023) framework, we have optimized its structure for enhanced fine-grained gait fusion, as shown in Fig. 3.

Given a skeleton frame $\mathcal{S}_t$, it is first transformed into Heat-skeletons $\mathcal{I}_t$. Subsequently, $\mathcal{I}_t$ is input into *DiffGait* to reconstruct the corresponding silhouettes. DiffGait Reverse Process could generate multi-level silhouettes under different sampling steps. This process undergoes M rounds of sampling to predict $\{P_i\}_{i=1}^{M}$, which we regard as Diffusion Silhouette Reconstruction.

$\mathcal{I}_t$, $\{P_i\}_{i=1}^{M}$ serve as inputs to Perceptual Gait Integration to obtain a comprehensive gait representation. In stage one, $\{P_i\}_{i=1}^{M}$ are then refined through dynamic weight allocation to yield the final composite silhouette $\mathcal{P}_t$. In stage two, after convolutional initialization, $\mathcal{P}_t$ and $\mathcal{I}_t$ are blended through gait fusion layer to obtain the hybrid gait feature $\mathcal{H}_t$.

Finally, $\mathcal{H}_t$ is processed through GaitBase (He et al. 2016; Fan et al. 2023) to generate the predicted identity as a one-hot vector. During the training phase, the model's learning process is supervised by calculating both the triplet loss (Hermans, Beyer, and Leibe 2017) and the cross-entropy loss, which can be formulated as:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T F_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T F_i + b_j}}, \tag{1}$$

$$\mathcal{L}_{tri} = \varphi\left[\mathcal{D}\left(F_i, F_k\right) - \mathcal{D}\left(F_i, F_j\right) + m\right], \tag{2}$$

F_i, F_k are the features of sample i and sample k, φ is equal to $max\,(\alpha, 0)$, $\mathcal{D}$ represents the Euclidean distance, m denotes the margin for the triplet loss.

3.3 DiffGait

We designed a pioneering diffusion model, termed DiffGait, aimed at reconstructing dense body shapes from sparse skeleton structures. However, directly applying diffusion models for this task inevitably encounters limitations due to modality inconsistency, non-specific diffusion process and model scale. To handle these issues, we initially employ Heat-skeleton Alignment to unify the modalities; subsequently, we design the DiffGait Process consisting of two opposite processes: the DiffGait Forward Process and the DiffGait Reverse Process; ultimately, we develop a highly streamlined DiffGait Architecture, as shown in Fig. 4.

Heat-skeleton Alignment. This improvement focuses on two aspects: **1)** Heat-skeletons explicitly provide spatial structure features; **2)** Heat-skeletons and silhouettes share modality consistency.

According to the method proposed by (Duan et al. 2022), the joint-based heatmap $\mathcal{J}$ centered on each skeleton point is created using the coordinate triplets (x_k, y_k, c_k):

$$\mathcal{J}_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k, \tag{3}$$

σ regulates the variance of the Gaussian maps, while (x_k, y_k) represents the spatial location of the k-th joint, and

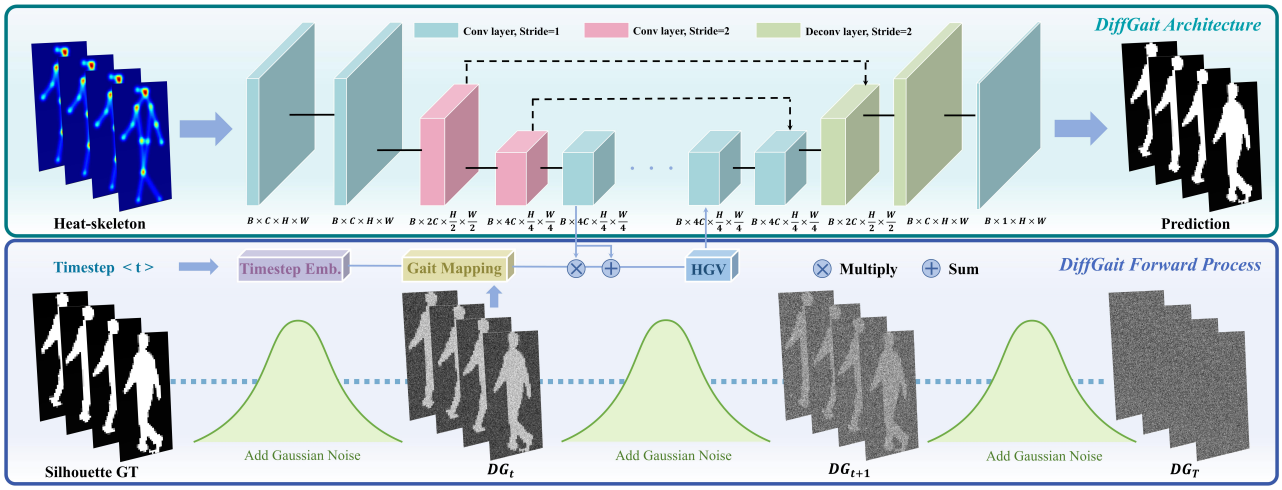

Figure 4: The upper half of the diagram delineates the specific network architecture and modules of DiffGait, while the lower half illustrates the forward process of DiffGait. The entire figure provides a detailed explanation of the feature flow process within DiffGait.

c_k represents the corresponding confidence score. We can also create the limb-based heatmap $\mathcal{L}$:

$$\mathcal{L}_{kij} = e^{-\frac{\mathcal{D}((i,j),seg[a_k,b_k])^2}{2*\sigma^2}} * \min\left(\mathcal{C}_{a_k}, \mathcal{C}_{b_k}\right). \quad (4)$$

The limb indexed as k connects two joints, a_k and b_k. The function $\mathcal{D}$ calculates the distance from the point (i, j) to the segment $[(x_{a_k}, y_{a_k}), (x_{b_k}, y_{b_k})]$.

DiffGait Forward Process. To simplify the diffusion process, we convert the diffusion object from silhouettes to Hybrid Gait Volume (HGV). Given a Heat-skeleton sequence with total N frames as $\mathcal{V}_h = \{\mathcal{I}_t\}_{t=1}^{N}$, where $\mathcal{I}_t \in \mathbb{R}^{h\times w\times 2}$ refers to the t-th frame.

To obtain DG_t, we designate the silhouettes aligned with Heat-skeletons as the diffusion target and define them as the initial state DG_0 of the diffusion process. We introduce Gaussian noise at various timesteps to facilitate the forward diffusion of the silhouettes,

$$DG_t = \sqrt{\overline{\alpha}_t} DG_0 + \sqrt{1-\overline{\alpha}_t}\epsilon. \quad (5)$$

Subsequently, $G_{ske} \in \mathbb{R}^{C\times H/4\times W/4}$ is extracted by $\mathcal{E}$. For dimension matching, we apply *Gait Mapping* to reduce DG_t by a factor of four via a two-layer convolution and increase feature channels to C. This approach retains more details compared to the direct bilinear interpolation used in DiffuVolume (Zheng et al. 2023a).

Furthermore, timestep embedding of dimension C corresponding to t is generated by *Timestep Embedding* and is added to DG_t, then we match the features of DG_t and G_{ske} by element-wise multiplication and introduce skip connections to ensure a smooth transition of the features,

$$HGV_t = G_{ske} \odot (GM(DG_t) + TE(t)) + G_{ske}, \quad (6)$$

where $\odot$ is the element-wise multiplication, HGV_t means the Hybrid Gait Volume, GM means the Gait Mapping, t is the selected timestep and TE means Timestep Embedding.

DiffGait Reverse Process. We have modified the DDIM (Song, Meng, and Ermon 2020) sampling method to generate multi-level silhouettes under different timesteps and reduce Gaussian uncertainty, as shown in Fig. 5.

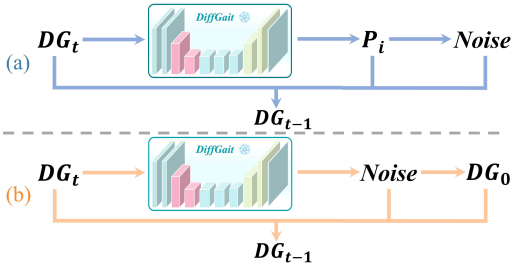

Figure 5: Visualization of the reverse process of (a) ours and (b) previous method.

Standard Gaussian noise DG_T serves as the initial input. Then, HGV_T is calculated based on Equation 6, and P_T is predicted by $\mathcal{DE}$. We consider P_T as the maximum likelihood of DG_0, denoted as DG_0^T. Then we calculate the noise for the current timestep via the following formula:

$$\epsilon = \frac{1}{\sqrt{1-\overline{\alpha}_T}}(DG_T - \sqrt{\overline{\alpha}_T} DG_T^0). \quad (7)$$

Furthermore, ϵ and DG_0^l are used to recover the DG_{T-1}. The formulation is expressed as,

$$DG_{T-1} = \sqrt{\overline{\alpha}_{T-1}} DG_T^0 + \sigma\epsilon^* + \sqrt{1-\overline{\alpha}_{T-1}-\sigma^2}\epsilon, \quad (8)$$

where $\sigma = \eta\sqrt{(1-\frac{\overline{\alpha}_T}{\overline{\alpha}_{T-1}})\cdot\frac{1-\overline{\alpha}_{T-1}}{1-\overline{\alpha}_T}}$, η means sampling coefficient and ϵ^* means the Standard Gaussian Noise. DG_{T-1} will continue to participate in the reverse process, following the method mentioned above, until the end of the sampling process.

DiffGait Architecture. The inference speed and model parameter size exhibited by existing UNet-based diffusion models render their application to gait recognition entirely impractical. Considering gait modalities are 'clean', DiffGait employs an extremely streamlined architecture with a surprisingly low model parameter size of 1.9M and an impressive inference speed of 3628 FPS.

Compared with previous models, DiffGait incorporates two key improvements: **1)** DiffGait adopts a Decoder-only

Table 1: Comparison of different datasets with their implementation details and setting.

DataSet	Batch Size	Milestones	Total Steps	Train Set		Test Set		Scenario	Modalities
				#ID	#Seq	#ID	#Seq		
CASIA-B	(8, 16)	(20k, 40k, 50k)	60k	74	8,140	50	5,500	Constrained	Sil., RGB
OU-MVLP	(32, 8)	(60k, 80k, 100k)	120k	5,153	144,284	5,154	144,412	Constrained	Sil., Ske.
Gait3D	(32, 4)	(20k, 40k, 50k)	60k	3,000	18,940	1,000	6,369	Real-world	Sil., Ske., Mesh
GREW	(32, 4)	(80k, 120k, 150k)	180k	20,000	102,887	6,000	24,000	Real-world	Sil., Ske.

Table 2: Quantitative evaluation. Comparison with other SOTA gait recognition methods across four authoritative datasets. The best performances are in **blod**, and the second best methods are underlined.

Input	Method	Testing Datasets									
		Gait3D				GREW	OU-MVLP	CASIA-B			
								NM	BG	CL	Mean
		Rank-1	Rank-5	mAP	mINP	Rank-1	Rank-1	Rank-1			
Silhouette	GaitSet(AAAI19)	36.7	59.3	30.0	17.3	46.3	87.1	95.0	87.2	70.4	84.2
	GaitPart(CVPR20)	28.2	47.6	21.6	12.4	44.0	88.5	96.2	91.5	78.7	88.8
	MTSGait(ACM MM22)	48.7	67.1	37.6	22.0	55.3	-	-	-	-	-
	GaitBase(CVPR23)	64.6	-	-	-	60.1	90.8	97.6	94.0	77.4	89.7
Skeleton	GaitGraph(ICIP21)	12.6	28.7	11.0	6.5	10.2	4.2	86.3	76.5	65.2	76.0
	GaitGraph2(CVPRW22)	7.2	15.9	5.2	3.0	34.8	70.7	80.3	71.4	63.8	71.8
	GPGait(ICCV23)	22.4	-	-	-	57.0	59.1	93.6	80.2	69.3	81.0
	SkeletonGait(AAAI24)	38.1	56.7	28.9	16.1	**77.4**	67.4	-	-	-	-
	Ours	**39.5**	**60.1**	**30.4**	**17.1**	72.5	**68.2**	**94.1**	**81.3**	**74.9**	**83.4**

architecture. **2)** DiffGait eliminates unnecessary convolutional layers and attention modules, operating within a low-dimensional feature space. To be specific, a five-layer ResNet-like encoder $\mathcal{E}$ is designed to extract skeleton structure features G_{ske}. *Gait Mapping* is adopted for downsampling silhouette-based Gaussian noise DG_t. A similarly sized decoder $\mathcal{DE}$, denoises silhouettes from Hybrid Gait Volume (HGV) integrated by G_{ske} and DG_t.

3.4 Perceptual Gait Integration (PGI)

Through the above designs, DiffGait produces multi-level silhouettes $\{P_i\}_{i=1}^{M}$, each emphasizing various appearance feature levels. Effective integration of structure and appearance features to construct a comprehensive gait representation significantly influences the overall recognition performance, so we propose Perceptual Gait Integration (PGI) consisting of two stages:

In stage one, our strategy is to allocate varying weights to different timesteps to generate the well-defined silhouette $\mathcal{P}_t$, acknowledging that the reconstruction process progresses from local to global features, with later steps yielding enhanced outcomes. We set the weights ω as (0, 0, 0.2, 0.3, 0.5) and discuss them in Tab. 9.

In stage two, our goal is to mitigate noise arising from gait modalities fusion and effectively integrate structure and appearance features. The initialized $\mathcal{P}_t$ and $\mathcal{I}_t$ are concatenated to create a fused feature, which serves as the input for subsequent processing. Following (Min et al. 2024), by employing a perceptual structure centered on ReLU and Sigmoid activations, we assign higher weights to features that are similar across skeletons and silhouettes, eliminate the impact of redundant features, and derive $\mathcal{H}_t$ through an attention-like approach (Vaswani et al. 2017).

4 Experiment

4.1 Experimental Settings

Datasets & Metrics. We evaluated our proposed method on four mainstream datasets, including two outdoor datasets: Gait3D (Zheng et al. 2022b) GREW (Zhu et al. 2021), and two indoor datasets: CASIA-B (Yu, Tan, and Tan 2006) OU-MVLP (Takemura et al. 2018). The key statistics of these gait datasets are listed in Tab. 1. We use the following metrics to evaluate model performance quantitatively: rank retrieval (Rank-1, Rank-5), mean Average Precision(mAP) and mean Inverse Negative Penalty (mINP), and Rank-1 accuracy is considered the primary metric.

Implementation details. During the training and inference stages, we use PyTorch as the framework to conduct all experiments on two RTX 3090. For the full four datasets, we used silhouettes with a resolution of 64×44, and skeletons are the 2D coordinates of joints that conform to COCO 17 and transformed into Heat-skeletons with a resolution of $2 \times 64 \times 44$. Tab. 1 displays the main hyper-parameters of our experiments.

When training DiffGait, we choose Adam as the optimizer, with the initial learning rate set to 0.01 and batch size (16, 4). We set the timestep as 1000. A cosine schedule is adopted to set the noise coefficient β_t, α_t. When training ZipGait, the SGD optimizer with an initial learning rate of 0.1 and weight decay of 0.0005 is utilized.

4.2 Comparison with State-of-the-art Methods

4.2.1 Quantitative comparison We initiate our analysis with an *Intra-domain comparison*, where we systematically evaluate our approach against the current state-of-the-art model-based methods, with comparative results presented in Tab. 2. Our method improves the performance by 10.2% compared to GPGait and 1.8% compared to SkeletonGait in

Figure 6: Comparison with other methods on four authoritative Datasets. The proposed method achieves state-of-the-art generalization performance across in all scenarios.

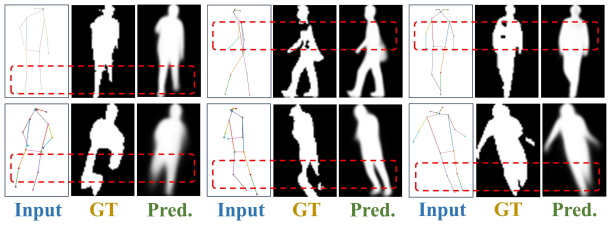

Figure 7: Qualitative evaluation. Visualization of the silhouette reconstruction results of DiffGait in diverse environments. Red boxes highlight the regions where features are missing.

Table 3: Efficiency comparison with different metrics. All models are evaluated on a single 3090 GPU. Params represents the number of model parameters.

Model	Inference speed (FPS) ↑	Params (M) ↓
DDPM	345	45.09
DDIM	1754	45.09
DiffGait (ours)	**3628**	**1.9**

Gait3D. Similarly, our method also holds a significant advantage over other methods in the indoor datasets.

We further perform a *Cross-domain comparison* to assess our method's efficacy across different testing scenarios, and detailed results are shown in Fig. 6. We compare two representative skeleton-based methods, GaitGraph2 and GPGait, with the latter demonstrating superior generalization capabilities. We can observe from the experimental results: 1) Skeletons are susceptible to data distribution; 2) Reconstructed appearance features can enhance representation robustness; 3) Training on real datasets leads to improved generalization capabilities.

4.2.2 Qualitative comparison As illustrated in Fig. 7, we randomly selected over a dozen test skeletons from four datasets. DiffGait can efficiently reconstruct occluded body parts by learning the a priori distribution of the human body. It can be observed that, compared to GT, various gait noises, such as occlusions, have been eliminated.

We further examine the ability of our model in dealing with challenging scenarios. We depict in Fig. 8 side-by-side comparisons of (a) our DiffGait against standard (b) DDIM and (c) DDPM, GT stands for ground truth. It is observed that our DiffGait consistently reconstructs silhouettes for various challenging scenes.

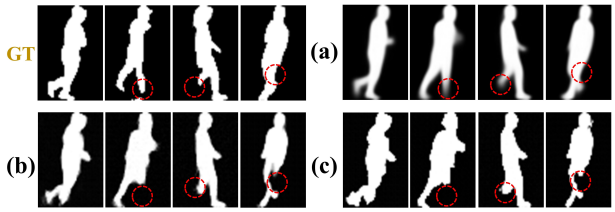

Figure 8: Qualitative evaluation. Visualization of the silhouette reconstruction results of our DiffGait (a), DDIM (b), DDPM (c) in diverse environments. Red circles mean challenging regions.

Table 4: Evaluation of Plug-and-Play Performance of DiffGait.

Methods \ Metrics	Rank-1 (%)	mAP (%)
GaitGraph(ICIP21)	12.6	11.0
GaitGraph + DiffGait	23.2 (10.6 ↑)	17.1 (6.1 ↑)
GaitTR(ES23)	7.8	6.6
GaitTR + DiffGait	19.6 (11.8 ↑)	15.7 (9.1 ↑)
GaitGraph2(CVPRW22)	7.2	5.2
GaitGraph2 + DiffGait	18.9 (11.7 ↑)	15.9 (10.7 ↑)
GPGait(ICCV23)	22.3	16.5
GPGait + DiffGait	31.3 (9.0 ↑)	22.9 (6.4 ↑)
SkeletonGait(AAAI24)	38.1	28.9
SkeletonGait + DiffGait	39.2 (1.1 ↑)	29.5 (0.6 ↑)

4.2.3 Efficiency comparison. In addition, to demonstrate the efficiency of our proposed method, we compare the number of model parameters and inference speed between our DiffGait and two typical methods DDPM and DDIM in Tab. 3. DiffGait achieves significant performance while only requiring approximately 1.9M model parameters. It outperforms them by achieving a speedup of about 10 times and 2 times faster in terms of Frames Per Second (FPS).

4.3 Plug-and-Play Performance Evaluation.

we demonstrate the plug-and-play performance of DiffGait by applying it to state-of-the-art model-based methods (Teepe et al. 2021, 2022; Zhang et al. 2023; Fu et al. 2023; Fan et al. 2024). The experimental outcomes, detailed in Tab. 4, reveal that DiffGait significantly enhances the performance of all existing model-based methods across various metrics. The effectiveness stems from reconstructed silhouettes, which provides richer gait features. Contrasting with ZipGait, we integrate gait features by padding and concate-

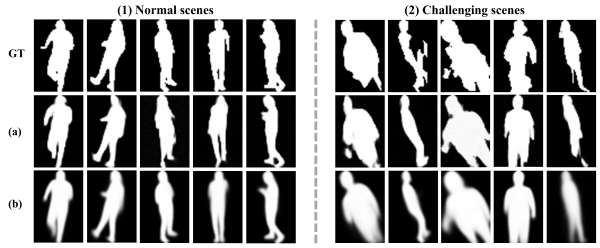

Figure 9: Visualization of the reconstructed results in normal and challenging scenarios. (a) donates DDIM, (b) donates DiffGait.

nating before passing them into the head layer.

4.4 Ablation Study

Study on components of ZipGait. We empirically evaluate the effectiveness of each proposed component, and report the results in Tab. 5 (a) is our baseline, using Heat-skeletons as input; (b) solely uses silhouettes predicted by DiffGait; (c) illustrates a simple sum fusion of shape and structure features; (d) demonstrates that our proposed PGI improves performance from 38.7% to 39.5%.

Table 5: Ablation of different components in ZipGait on Gait3D. Heat. denotes Heat-skeleton; Silh. donates Silhouette; PGI donates Perceptual Gait Integration.

Structure	Heat.	Silh.	PGI	Rank-1	mAP
(a) Baseline	✓	✗	✗	33.6	23.4
(b)	✗	✓	✗	24.7	18.1
(c)	✓	✓	✗	38.7	27.2
(d) ZipGait	✓	✓	✓	**39.5**	**30.4**

Table 6: Ablation of various designs in PGI on Gait3D.

Structure	Stage 1	Stage 2	Rank-1	mAP
(a)	✓	✗	39.0	28.9
(b)	✗	✓	39.2	29.5
(c) Full	✓	✓	**39.5**	**30.4**

Study on Perceptual Gait Integration. We further explore the influence of two stages within PGI, and tabulate the results in Tab. 6. (a) only utilizes stage 1, refining the predicted multi-level silhouettes through dynamic weight allocation; (b) employs a gait fusion layer to generate mixed gait features, representing stage 2; (c) demonstrates the complete two-stage integration. Each stage independently enhances performance, and their integration results in even greater improvements, thereby proving the efficacy of the modules.

Table 7: Ablation of different diffusion process.

Method	Rank-1	Rank-5	mAP	mINP
DDIM Process	38.4	58.1	29.2	16.5
DiffGait Process	**39.5**	**60.1**	**30.4**	**17.1**

Study on DiffGait Process. To validate the effect of the DiffGait Process, we consider two alternative ways to train our DiffGait. Fig. 10 visualizes the denoising process in detail, highlighting the distinctions between the two methods. Further experiments under identical conditions demonstrated superior performance of the DiffGait Process, providing a clear validation of its effectiveness, shown in Tab. 7. Model training and inference of DDIM process can be found in Appendix, along with more detailed comparisons.

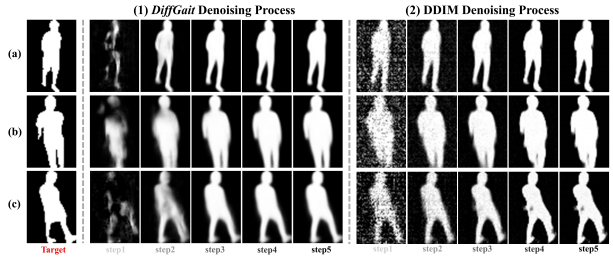

Figure 10: Visualization of the reconstructed results about whole denoising process under different algorithms.

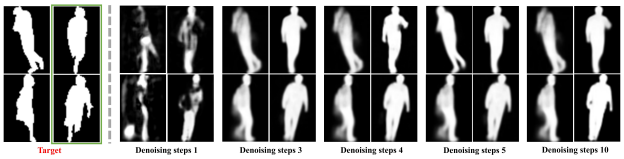

Figure 11: Visualization of the reconstructed results under different denoising steps of DiffGait.

4.5 Further Analysis

Denoising steps analysis. As shown in Fig. 11, setting the denoising steps to three or more can generate effective silhouettes. To select the most appropriate number of denoising steps, we analyzed several possible values, with performance and efficiency serving as the criteria. The experimental results in Tab. 8 reveal that the best performance is achieved with **5** denoising steps, striking a balance by enhancing performance without significant additional resource demands, with only a slight 0.2s increase per iteration.

Table 8: Parameter analysis. Comparison model performance on Gait3D under different sampling steps for DiffGait.

Sampling Steps	Training time	Rank-1	mAP
No DiffGait	37.2s	33.6	23.4
1 sampling steps	43.8s	6.8	1.7
3 sampling steps	53.4s	37.4	27.1
4 sampling steps	54.2s	38.4	28.3
5 sampling steps	59.6s	**38.7**	**29.2**
10 sampling steps	76.4s	37.9	28.8

Dynamic weight allocation analysis. As shown in Fig. 10, during the denoising process, the quality of predicted silhouettes varies at each step. We established various potential combinations of weights and conducted experimental validations. The results, displayed in Tab. 9, indicate that α_4 is the optimal choice for Stage 1 of PGI.

Table 9: Parameter analysis. Comparison model performance on Gait3D datasets under different weight combinations.

groups	ω_1	ω_2	ω_3	ω_4	ω_5	Rank-1	mAP
α_1	1	0	0	0	0	12.8	6.2
α_2	0	0	0	0	1	39.0	29.4
α_3	0	0	0	0.5	0.5	39.2	29.8
α_4	0	0	0.2	0.3	0.5	**39.5**	**30.4**
α_5	0.2	0.2	0.2	0.2	0.2	38.7	29.2

5 Conclusion

This paper establishes the connection between skeletons and silhouettes via the diffusion model, providing new insights into utilizing gait modalities. The proposed DiffGait is the first successful method to reconstruct dense body shapes

from sparse skeleton structures. Simultaneously, Perceptual Gait Integration is proposed for efficient fusion of gait modality. Ultimately, ZipGait demonstrates superior performance in both cross-domain and intra-domain settings and achieves plug-and-play improvements. This study highlights the potential of gait modality interaction in gait recognition.

References

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

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Choi, S.; Kim, J.; Kim, W.; and Kim, C. 2019. Skeleton-based gait recognition via robust frame-level matching. *IEEE Transactions on information forensics and security*, 14(10): 2577–2592.

Cui, Y.; and Kang, Y. 2023. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17949–17957.

Dong, Y.; Yu, C.; Ha, R.; Shi, Y.; Ma, Y.; Xu, L.; Fu, Y.; and Wang, J. 2024. HybridGait: A benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 1600–1608.

Dou, H.; Zhang, P.; Su, W.; Yu, Y.; Lin, Y.; and Li, X. 2023. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5578–5588.

Dou, H.; Zhang, P.; Zhao, Y.; Jin, L.; and Li, X. 2024. CLASH: Complementary Learning with Neural Architecture Search for Gait Recognition. *IEEE Transactions on image processing*.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2969–2978.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9707–9716.

Fan, C.; Ma, J.; Jin, D.; Shen, C.; and Yu, S. 2024. SkeletonGait: Gait Recognition Using Skeleton Maps. In *Proceedings of the AAAI conference on artificial intelligence*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 14861–14872.

Frank, J.; Mannor, S.; and Precup, D. 2010. Activity and gait recognition with time-delay embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, 1581–1586.

Fu, Y.; Meng, S.; Hou, S.; Hu, X.; and Huang, Y. 2023. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19595–19604.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8295–8302.

Gao, S.; Yun, J.; Zhao, Y.; and Liu, L. 2022. Gait-D: skeleton-based gait feature decomposition for gait recognition. *IET computer vision*, 16(2): 111–125.

Gong, J.; Foo, L. G.; Fan, Z.; Ke, Q.; Rahmani, H.; and Liu, J. 2023. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13041–13051.

Guo, Y.; Shah, A.; Liu, J.; Chellappa, R.; and Peng, C. 2023. GaitContour: Efficient Gait Recognition based on a Contour-Pose Representation. *arXiv preprint arXiv:2311.16497*.

Han, J.; and Bhanu, B. 2005. Individual recognition using gait energy image. *IEEE Transactions on pattern analysis and machine intelligence*, 28(2): 316–322.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Holmquist, K.; and Wandt, B. 2023. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15977–15987.

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 Transactions on image processing*.

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

Li, X.; Makihara, Y.; Xu, C.; Yagi, Y.; Yu, S.; and Ren, M. 2020. End-to-end model-based gait recognition. In *Proceedings of the asian conference on computer vision*.

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

Lin, B.; Zhang, S.; and Yu, X. 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proc5eedings of the IEEE/CVF international conference on computer vision*, 14648–14656.

Liu, X.; You, Z.; He, Y.; Bi, S.; and Wang, J. 2022. Symmetry-Driven hyper feature GCN for skeleton-based gait recognition. *Pattern recognition*, 125: 108520.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal graphics papers: pushing the boundaries, Volume 2*, 851–866.

Min, F.; Guo, S.; Hao, F.; and Dong, J. 2024. GaitMA: Pose-guided Multi-modal Feature Fusion for Gait Recognition. *arXiv preprint arXiv:2407.14812*.

Pan, H.; Chen, Y.; Xu, T.; He, Y.; and He, Z. 2023. Toward complete-view and high-level pose-based gait recognition. *IEEE Transactions on information forensics and security*, 18: 2104–2118.

Peng, Y.; Ma, K.; Zhang, Y.; and He, Z. 2024. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia tools and applications*, 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 2023-2023 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 1–5. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shen, C.; Fan, C.; Wu, W.; Wang, R.; Huang, G. Q.; and Yu, S. 2023. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1054–1063.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on computer vision and applications*, 10: 1–14.

Teepe, T.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2022. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 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 *2021 IEEE international conference on image processing (ICIP)*, 2314–2318. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Chen, J.; and Liu, Y. 2022. Frame-level refinement networks for skeleton-based gait recognition. *Computer cision and image understanding*, 222: 103500.

Wang, L.; Tan, T.; Ning, H.; and Hu, W. 2003. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on pattern analysis and machine intelligence*, 25(12): 1505–1518.

Wang, M.; Guo, X.; Lin, B.; Yang, T.; Zhu, Z.; Li, L.; Zhang, S.; and Yu, X. 2023a. DyGait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13424–13433.

Wang, Z.; Hou, S.; Zhang, M.; Liu, X.; Cao, C.; and Huang, Y. 2023b. GaitParsing: Human semantic parsing for gait recognition. *IEEE Transactions on multimedia*.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 5785–5793.

Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; and Tan, T. 2016. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on pattern analysis and machine intelligence*, 39(2): 209–226.

Yam, C.; Nixon, M. S.; and Carter, J. N. 2004. Automated person recognition by walking and running via model-based approaches. *Pattern recognition*, 37(5): 1057–1072.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International conference on pattern recognition (ICPR'06)*, volume 4, 441–444. IEEE.

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.

Zheng, D.; Wu, X.-M.; Liu, Z.; Meng, J.; and Zheng, W.-s. 2023a. Diffuvolume: Diffusion model for volume based stereo matching. *arXiv preprint arXiv:2308.15989*.

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 *Proceedings of the 30th ACM international conference on multimedia*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20228–20237.

Zheng, J.; Liu, X.; Wang, S.; Wang, L.; Yan, C.; and Liu, W. 2023b. Parsing is all you need for accurate gait recognition

in the wild. In *Proceedings of the 31st ACM international conference on multimedia*, 116–124.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 14789–14799.

Zou, S.; Fan, C.; Xiong, J.; Shen, C.; Yu, S.; and Tang, J. 2024. Cross-Covariate Gait Recognition: A Benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7855–7863.

Supplementary Material for ZipGait: Bridging Skeleton and Silhouette with Diffusion Model for Advancing Gait Recognition

In appendix, we provide background on diffusion models, DiffGait training and inference algorithms, comparison with other related gait recognition methods, analysis of the impact on generalization and more visualization details in Section A, Section B, Section C, Section D and Section E.

A. Background on Diffusion Models

Diffusion models typically consist of two basic processes: 1) a forward process that gradually adds Gaussian noise to sample data, and 2) a reverse process that learns to invert the forward diffusion.

Our work extends the framework of Denoising Diffusion Probabilistic Models (DDPMs), a class of deep generative models that approximate the distribution of natural images using the terminal state of a Markov chain that originates from a standard Gaussian distribution. DDPMs are trained to reverse a diffusion process that gradually adds Gaussian noise to the training data x_0 over T steps until it becomes pure noise at x_T.

To be specific, the forward process may be formalized as sampling from a conditional distribution q, denoted as $\mathrm{q}\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$, which mainly involves adding noise to the data without depending on any parameterized distribution,

$$\mathrm{q}\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{9}$$

where $\mathbf{I}$ is the identity matrix and the rate at which the original data is diffused into noise is controlled by a variance scheduling given by $\beta_1, \ldots, \beta_T$. Importantly, the forward process and its schedule may be reparameterized as:

$$\alpha_t=\prod_{t=1}^{T}\left(1-\beta_t\right), \tag{10}$$

$$\mathrm{q}\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{\alpha_t} \mathbf{x}_{t-1},\left(1-\alpha_t\right) \mathbf{I}\right). \tag{11}$$

In contrast, the reverse process reconstructs the data distribution from noise, represented as $\mathrm{p}\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$. This reverse conditional distribution is arbitrarily complex, but we might approximate it via a denoising deep neural network,

$$\mathrm{p}\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \approx \mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_t\right). \tag{12}$$

where $\mu_\theta\left(\mathbf{x}_t, t\right)$ is a learned deep neural network that has as input both the noisy data x_t and its step t (usually position encoded); Σ_t depends on the variance schedule but is not otherwise learned.

The conditional diffusion models, which add an extra term in the denoising process, provide the opportunity for the cross-modal generation tasks:

$$\mathrm{p}\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c\right) \approx \mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, c, t\right), \Sigma_t\right), \tag{13}$$

where the conditioning inputs c added to ϵ_θ may be derived from the original modality.

B. DiffGait Training and Inference Algorithms

Training. During the training phase, we perform the diffusion process that corrupts ground truth silhouette DG_0 to noisy silhouette DG_t, then G_{ske} is integrated with DG_t to HGV_t. We learn the reverse denoising process by continuously using $\mathcal{D}$ to recover HGV_t into DG_0, thus optimizing *DiffGait* to gradually remove the uncertainty of DG_t and redefine the source distribution DG_0 to P_t, which can be formulated as:

$$P_t=\mathcal{D}(HGV_t). \tag{14}$$

We employ MSE to supervise the model training:

$$\mathcal{L}_{dg}=\left\|DG_0-P_t\right\|_2^2. \tag{15}$$

Algorithm A provides the overall training procedure.

Inference. Algorithm B summarizes the detailed inference procedure of *DiffGait*, which can be seen as iteratively generating more complete silhouettes. We have redesigned the sampling method based on DDIM to generate effective silhouettes at each sampling step, observing that later-generated features are more complete yet tend to lack some details. Specifically, for each sampling step, $\mathcal{D}$ takes the initial HGV_T or the HGV_t from the previous step as input and outputs the estimated silhouette for the current step. Then, our new sampling method is used to update the silhouette for the next step.

Algorithm A: DiffGait Training

Require: Heat-skeleton : $\mathcal{I}_t$, GT_Silhouette : DG_0
repeat
  $G_{ske}=\mathcal{E}(\mathcal{I}_t)$
  $t \sim$ Uniform$(\{1,\ldots,T\})$
  $\epsilon \sim \mathcal{N}(0,1)$
  $DG_t=\sqrt{\alpha_t}DG_0+\sqrt{1-\alpha_t}\epsilon$
  $HGV_t=G_{ske}\odot(GM(DG_t)+TE(t))+G_{ske}$
  Take gradient descent step on $\Delta\theta\|\mathcal{D}(HGV_t)-DG_0\|^2$
until converged

Algorithm B: DiffGait Inference

Require: Heat-skeleton : $\mathcal{I}_t$, steps : T
Ensure: Predicted_Silhouette : P_i
$DG_T \sim \mathcal{N}(0,1)$
$G_{ske}=\mathcal{E}(\mathcal{I}_t)$
$times=$ Reversed(Linspace$(-1,T,steps)$)
$time\_pairs=$ List(Zip$(times[:-1], times[1:])$)
$HGV_T=G_{ske}\odot(GM(DG_T)+TE(T))+G_{ske}$
for $(t_{\text{now}}, t_{\text{next}})$ in time_pairs **do**
  $P_i=\mathcal{D}(HGV_T, t_{\text{now}})$
  $HGV_t=$ *DiffGait*$(HGV_T, P_i i, t_{\text{now}}, t_{\text{next}})$
end for
return P_t

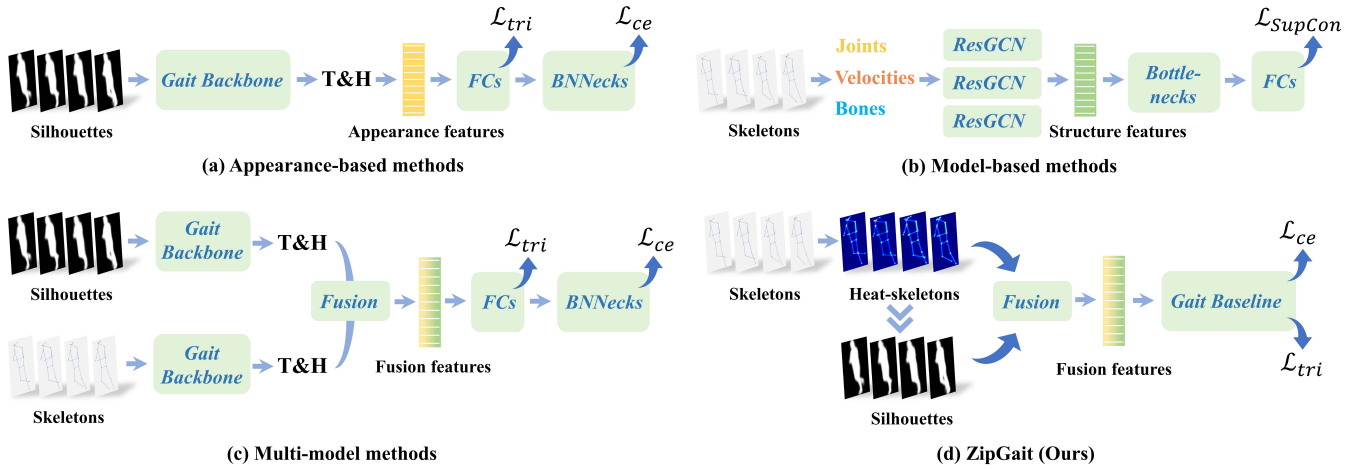

Figure A: Brief frameworks of related gait recognition methods, and our proposed ZipGait model. By providing an overall description of the differences in the procedures of each method. **T&H** represents the horizontal mapping and temporal aggregation.

C. Comparison with Other Related Gait Recognition Methods

In the aforementioned article, we briefly delineate the differences between our ZipGait approach and other common gait recognition methods in terms of input modalities and feature utilization. To concretely highlight the distinctions of our architecture, we have illustrated four pipelines in Fig A, each representing the overall workflow of our method compared to three other approaches in gait recognition.

As shown in Fig A(a), appearance-based methods are currently the mainstream algorithms in the field of gait recognition. This paradigm, followed by most studies, typically exhibits the best performance under general conditions. However, these methods inevitably suffer from the quality of silhouettes; in environments with complex backgrounds or significant occlusions, performance is limited, as low-quality silhouettes do not enable the models to learn robust gait representations, which is a very critical factor.

Fig A(b) illustrates model-based approaches, which primarily extract gait features from human body models. The skeleton is the only model that can be used as a standalone input for gait recognition, often referred to as skeleton-based methods. SMPL models and point clouds are usually employed as supplementary modalities to provide additional information. The main limitation of model-based methods is the accuracy of the models, especially three-dimensional ones, which currently cannot effectively estimate the absolute posture of the human body in three-dimensional space, making gait recognition using these models unreliable.

The basic workflow of multimodal methods is depicted in Fig A(c), which extracts features from different modalities through two branches and then merges them for feature classification. This approach does offer a novel strategy for obtaining more effective gait representations, but it requires inputs from multiple modalities and the processing of features from these modalities, which is particularly demanding in terms of computational and storage resources. Although this can achieve high performance, further efforts are needed to make the structure more lightweight.

Fig A(d) presents our ZipGait method, which borrows the advantages of the three aforementioned approaches while addressing some of their shortcomings. Our method solely uses the skeleton as an input but establishes a relationship with the silhouette through our DiffGait model. In terms of feature fusion, we have refined the dual-branch structure of current multimodal methods by merging features at a lower dimension, effectively simplifying the model structure. Overall, ZipGait employs a highly streamlined structure, using only one modality as input, yet achieving the effects of multi-modality.

It must be noted that our proposed method still lags behind appearance-based and multimodal methods in performance. This is primarily due to the predicted silhouettes not fully restoring all external details, tending instead to reconstruct a more averaged distribution.

Table A: Ablation study on the effect of *DiffGait* reconstructed silhouettes on generalization capability. Heat. denotes Heat-skeleton; Silh. donates Silhouette; The best performances are in **blod**.

Heat.	Silh.	Trained on CASIA-B		Trained on Gait3D	
		CASIA-B	Gait3D	CASIA-B	Gait3D
✓	✗	77.2	10.2	27.4	33.6
✗	✓	67.1	9.8	34.5	24.7
✓	✓	**83.4**	**12.8**	**44.2**	**39.5**

D. Analysis of the Impact on Generalization

Our proposed ZipGait outperforms other methods in cross-domain evaluations. To ascertain whether the enhanced generalization ability is due to the use of DiffGait in reconstructing silhouettes, we specifically designed experiments to separately assess the intra-domain and cross-domain evaluation effects of Heat-skeleton and predicted Silhouette. The experiments demonstrate that incorporating DiffGait is a key factor in improving the model's generalization capability.

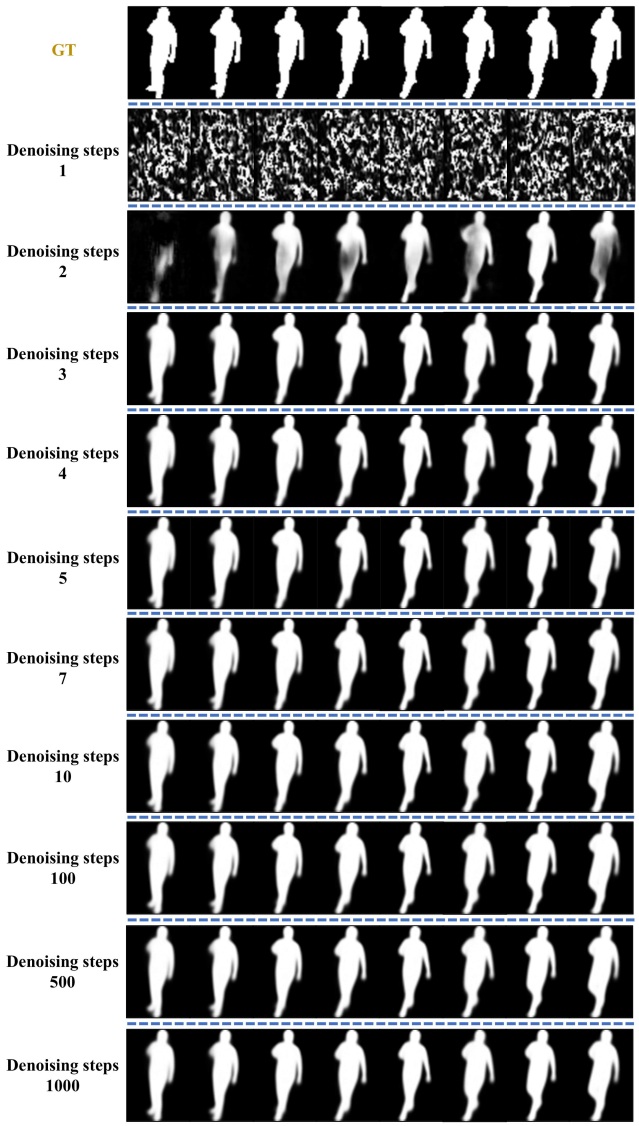

Figure B: Visualization of the reconstructed results under different denoising steps of DiffGait on Gait3D.

E. More Visualization Details.

In this section, we provide more comprehensive visualizations to demonstrate the effectiveness of our proposed DiffGait model. Fig. B displays the denoising results at different denoising steps, showing that visually effective silhouettes can be achieved when the denoising step is set to three or more,. Fig. C compares the performance with other diffusion models, revealing that our model does not significantly differ in denoising quality from the DDIM and DDPM models, and that our DiffGait is far more efficient than these methods. Fig. D visualizes the intermediate modalities used in ZipGait: skeleton, Heat-skeleton, and reconstructed silhouette, along with the ground truth silhouette. Fig. E compares the predicted silhouettes of two long sequences (each with 60 frames) derived from our DiffGait. Fig. F presents visualizations under normal conditions, where there is enough shape

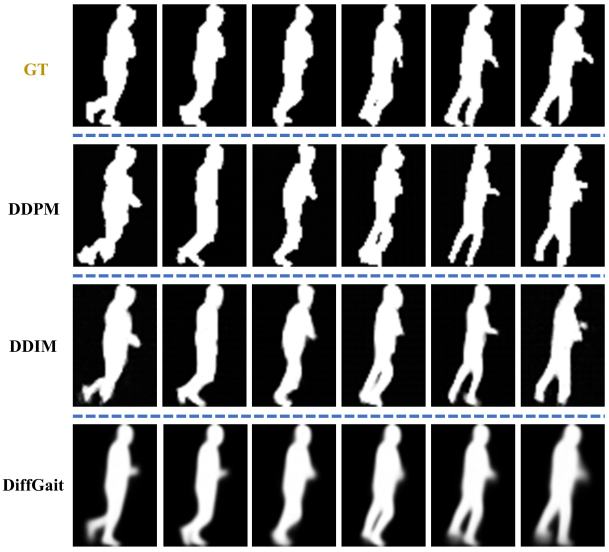

Figure C: Visualization of the silhouette reconstruction results of our DiffGait, DDIM, DDPM in one gait sequence.

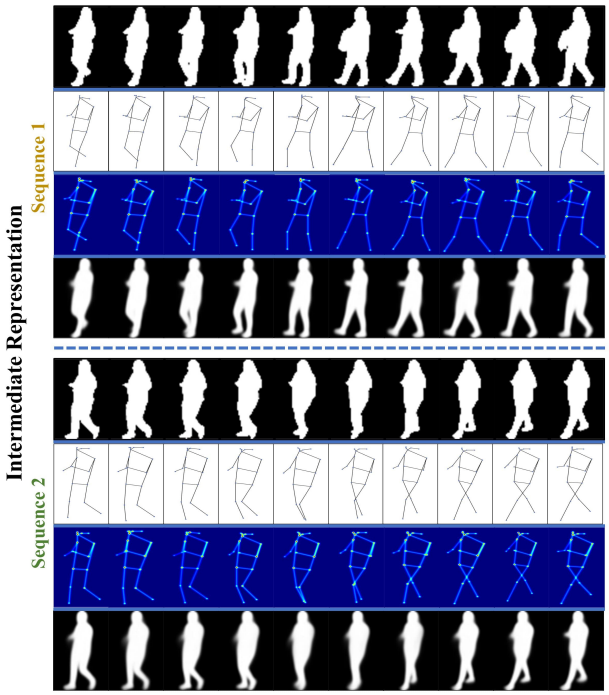

Figure D: Visual results of our DiffGait on different intermediate representations in the whole process.

information to reconstruct high-quality contours. Fig. G shows visualizations under challenging conditions, where the silhouettes reconstructed by DiffGait provide more effective features, as the robustness displayed by the skeleton allows it to address the impacts of occlusions.

These sufficient visualizations evaluate the quality of the DiffGait reconstructed silhouettes from different aspects and can fully demonstrate the effectiveness of our method.

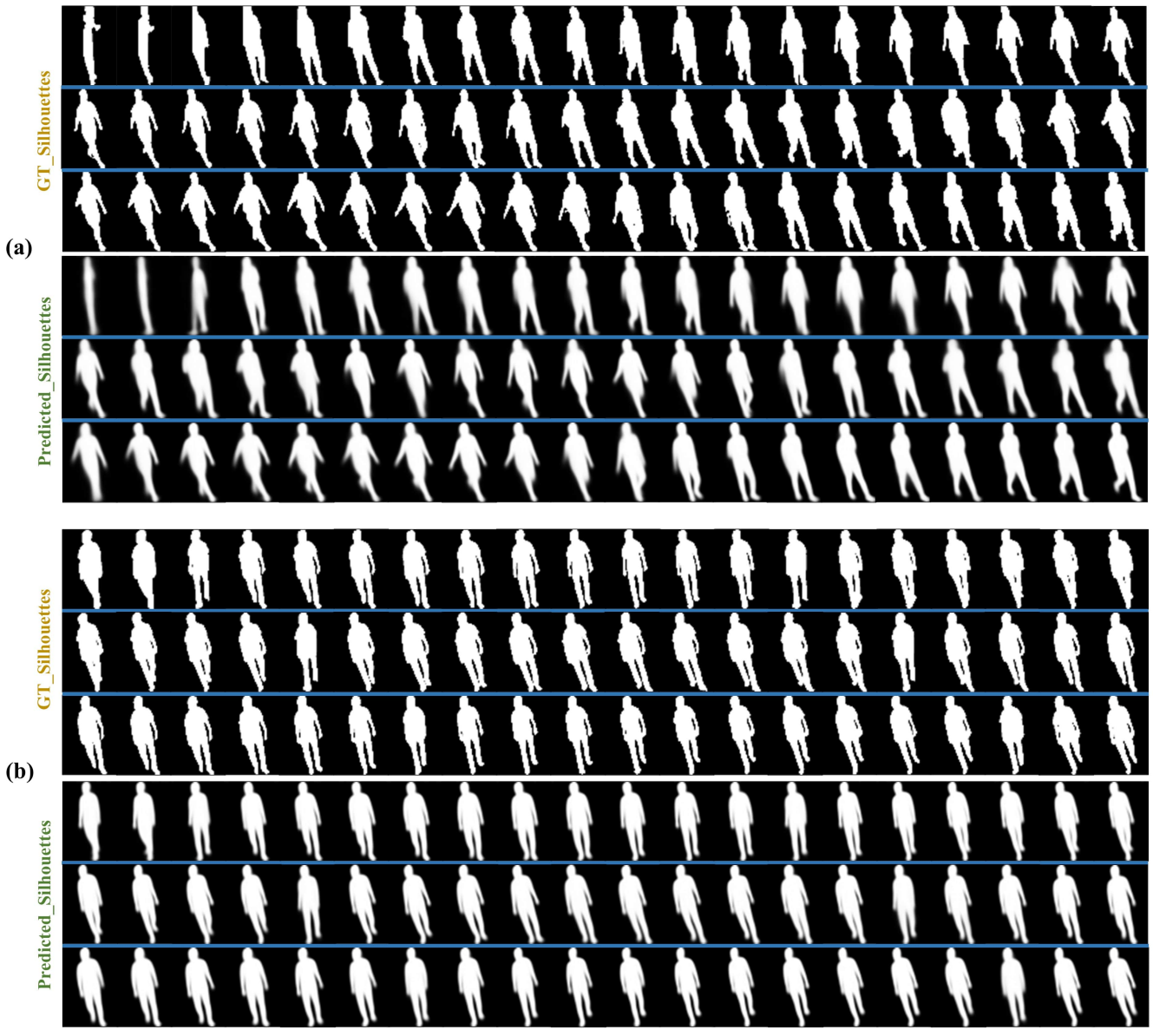

Figure E: Visual results of our DiffGait on different long sequences.

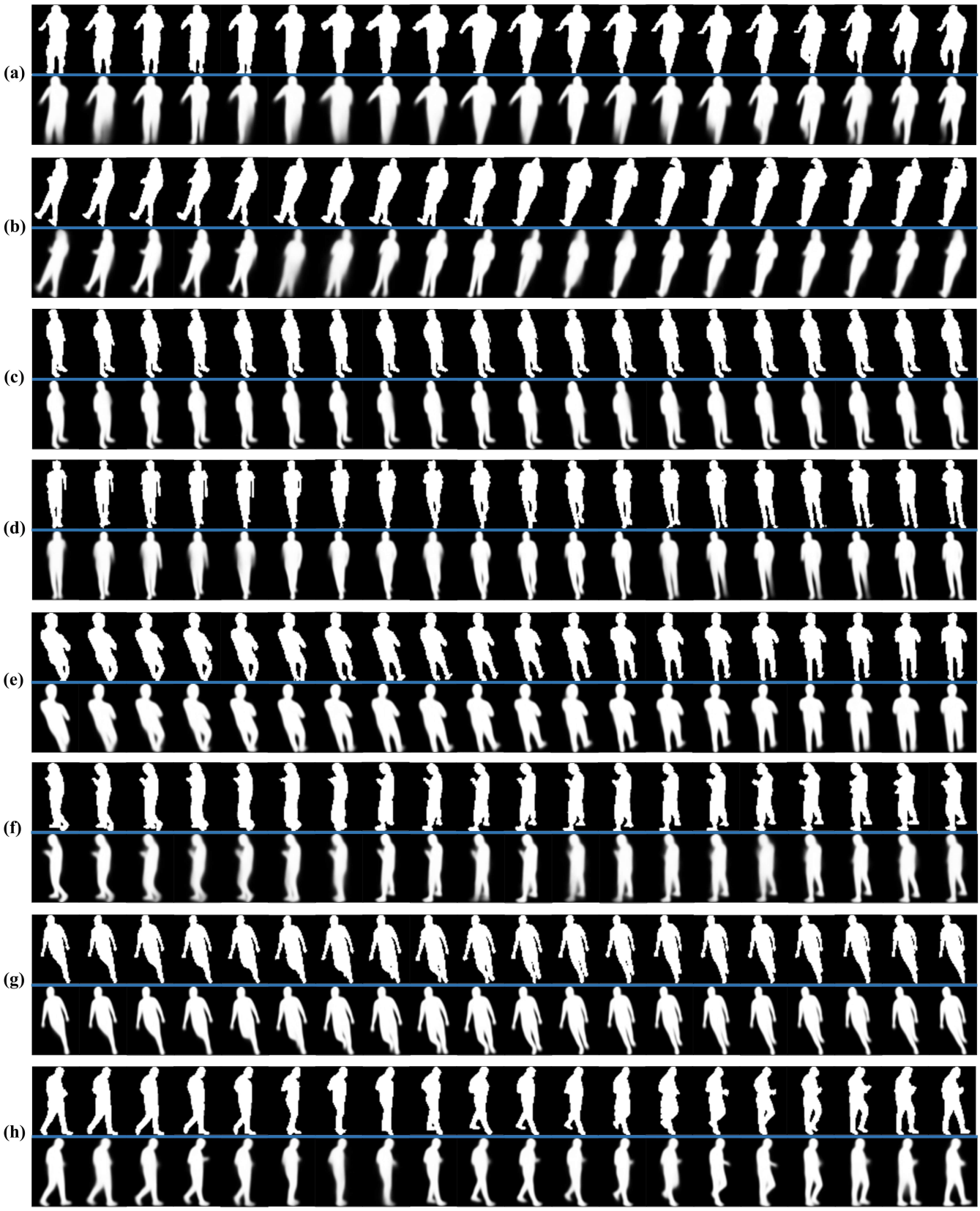

Figure F: Visual results of our DiffGait on Gait3D. Normal scenes with relatively complete shape information.

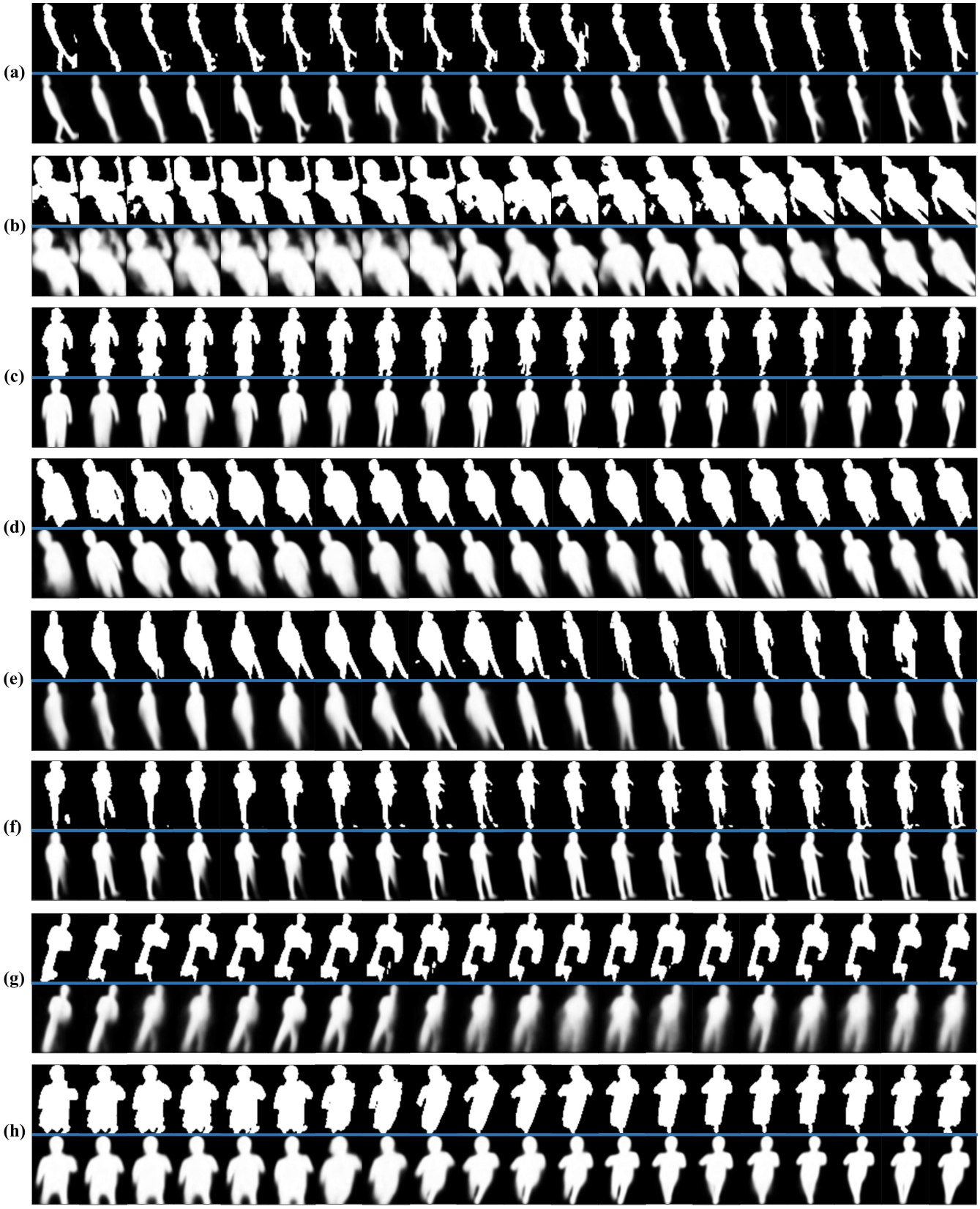

Figure G: Visual results of our DiffGait on Gait3D. Challenging scenes such as arbitrary views or occlusions are involved.