GAITWM: GENERATIVE WORLD MODELS FOR SPARSE-VIEW GAIT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gait recognition has achieved remarkable progress with the advancement of deep learning, showing promising performance in both controlled environments and unconstrained real-world scenarios. Existing cross-view gait recognition methods typically rely on densely collected multi-view gait sequences to learn view-consistent representations, assuming sufficient observations across predefined viewpoints. However, such assumptions rarely hold in practical applications, where gait data are often collected from sparse and unpredictable viewpoints due to the high cost and difficulty of multi-view acquisition and annotation. This limitation significantly degrades recognition performance when encountering unseen viewpoints. To address this challenge, we propose **GaitWM**, utilizing a ***generative world model*** for label-efficient ***sparse-view gait recognition***. Unlike conventional methods that learn view-invariant representations solely from existing observations, GaitWM learns the underlying gait dynamics and view transformation knowledge from generative video world, enabling the generation of plausible gait sequences from unobserved viewpoints. By synthesizing complementary views conditioned on sparse observations, GaitWM effectively expands viewpoint coverage and provides richer motion cues for recognition without requiring densely collected multi-view gait datasets. Extensive experiments on CASIA-B and SUSTech1K demonstrate that GaitWM consistently outperforms existing methods under sparse-view settings, achieving robust generalization to unseen viewpoints. The results validate the effectiveness of generative video models as a powerful paradigm for reducing reliance on exhaustive multi-view gait data collection and advancing practical gait recognition systems. *The code will be released publicly upon publication.*

1 INTRODUCTION

Gait recognition aims to facilitate human retrieval in challenging scenarios, enabling the identification of pedestrians at long distances, under low-resolution conditions, and without requiring their cooperation (Shen et al., 2022; Sepas-Moghaddam & Etemad, 2022). This capability allows gait recognition to be applied in diverse applications, including public security, healthcare, and robotics.

Cross-view gait recognition has achieved remarkable success on both laboratory (Yu et al., 2006a; Takemura et al., 2018) and in-the-wild benchmarks (Shen et al., 2023; Li et al., 2023; Zheng et al., 2022; Zhu et al., 2021). However, these impressive results are largely built upon an idealized assumption of *dense-view training*, where each identity is observed from sufficient viewpoints. Under this setting, the representative RGB-based method BigGait (Ye et al., 2024) can make over 95.3% and 95.9% Rank-1 accuracy on laboratory (Yu et al., 2006a) and in-the-wild datasets (Shen et al., 2023), respectively. We argue that such an assumption is difficult to satisfy in practice, as collecting and annotating densely multi-view gait sequences is both costly and difficult to scale (Fan et al., 2023a; Wang et al., 2023). More importantly, dense-view coverage and identity scalability (Zou et al., 2024) are inherently conflicting.

In this paper, we are interested in the problem of *sparse-view gait recognition*, aiming to achieve robust cross-view recognition when each identity is observed from only a few viewpoints with sparse-view training. However, we find that this setting is substantially more challenging than the conventional dense-view gait recognition. When training data are limited to sparse-view observations, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rank-1 accuracy of BigGait drops dramatically to around 45.2% and 49.6% on laboratory and in-the-wild datasets, respectively. Such observations have also been reported in previous literature (Wang et al., 2023; Fan et al., 2023a). The fundamental challenge of sparse-view gait recognition lies in the absence of cross-view supervision, which prevents deep learning-based methods from extracting view-consistent gait representations. To the best of our knowledge, there is still no effective paradigm for addressing sparse-view gait recognition.

To this end, we propose **GaitWM**, a new paradigm that leverages a generative video world model for label-efficient sparse-view gait recognition, as illustrated in Figure 1. GaitWM distills gait dynamics and viewpoint transformation knowledge from large-scale video generation models (Oquab et al., 2024; Wan Team, 2025; Lin et al., 2026) to synthesize gait sequences in novel view, providing complementary cross-view supervision beyond sparse observations. To further bridge the gap between generated and real samples, we futher propose an *Generative Identity Regularization* that adaptively relaxes the identity supervision on generated samples, enabling more effective utilization of synthetic data. Moreover, GaitWM can be seamlessly integrated into existing gait recognition frameworks.

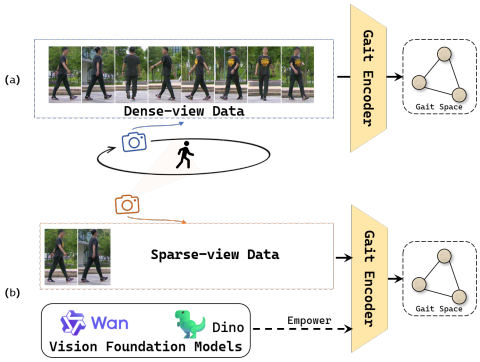


Figure 1: Comparison between dense-view and sparse-view gait recognition. (a) Dense-view training relies on multiple captured viewpoints to construct a comprehensive gait space. (b) Sparse-view training observes only limited viewpoints. To address this limitation, we introduce a generative paradigm that transfers priors from large-scale vision foundation models to recover a view-invariant gait space.

Extensive experiments on both laboratory and in-the-wild benchmarks under diverse sparse-view protocols demonstrate that GaitWM consistently improves state-of-the-art gait recognition methods. In particular, GaitWM achieves substantial gains under extremely sparse-view settings while exhibiting strong generalization to unseen viewpoints and significantly reducing the reliance on densely annotated multi-view gait datasets. The main contributions are summarized:

- We propose **GaitWM**, a novel generative paradigm that leverages a video generative model to synthesize novel view gait sequences, effectively enriching viewpoint diversity and providing complementary cross-view supervision from sparse observations.
- We propose an *Generative Identity Regularization* that adaptively relaxes the identity supervision imposed on generated samples, enabling more effective utilization of synthetic gait sequences for recognition.
- Extensive experiments on both laboratory and in-the-wild benchmarks validate the effectiveness and generalizability of GaitWM, consistently improving existing gait recognition frameworks across diverse sparse-view settings.

2 Related Work

2.1 Gait Recognition

Cross-view gait recognition aims to alleviate the significant appearance variations caused by viewpoint changes. Existing methods mainly address this challenge by constructing multi-view gait pairs (Ye et al., 2025a;b; Wang et al., 2025a) and learning view-consistent representations through contrastive learning (Fan et al., 2023a; Zheng et al., 2022; Li et al., 2023), view transformation (Zheng et al., 2011; Xu et al., 2020), feature disentanglement (Zhang et al., 2020), or 3D-aware modeling (Shen et al., 2025; Ahn et al., 2022; Han et al., 2024). Benefiting from densely collected multi-view datasets such as CASIA-B (Yu et al., 2006a) and OU-MVLP (Takemura et al., 2018), recent methods (Liang et al., 2022; Ma et al., 2023; Lin et al., 2021) have achieved remarkable recognition accuracy under both laboratory and unconstrained settings. Nevertheless, these approaches implicitly assume that sufficient cross-view observations are available for each identity

during training. Such dense-view supervision becomes prohibitively expensive to collect in real-world scenarios, limiting their scalability to practical applications.

Sparse-View Gait Recognition. In practice, acquiring densely paired gait sequences from multiple viewpoints is often infeasible (Ye et al., 2025b; Wang et al., 2023; 2025a), making sparse-view gait recognition a more realistic yet significantly more challenging setting. The lack of sufficient cross-view supervision prevents existing methods from learning reliable view-consistent representations, resulting in dramatic performance degradation even with large identity sets. To alleviate this issue, GaitCSV (Wang et al., 2023) formulates sparse-view gait recognition and introduces causal intervention to reduce viewpoint bias during representation learning. GaitSSB (Fan et al., 2023a) explores large-scale self-supervised pretraining to improve representation transferability under limited annotations. However, these methods (Jin et al., 2025; Kim et al., 2025; Huang et al., 2026) still rely solely on the observed gait sequences and cannot recover the missing viewpoints that are never captured during training. Consequently, the intrinsic sparsity of viewpoint supervision remains unresolved. Different from existing representation learning approaches, GaitWM explicitly generate the missing-view distribution by leveraging a generative video world model, providing complementary cross-view supervision beyond the observed gait sequences.

2.2 Video Generation Models

Recent video generation models (Wan Team, 2025; Kong et al., 2024; Zheng et al., 2024) have demonstrated impressive capabilities in modeling temporal dynamics, camera motion, and object appearance from large-scale videos. Beyond video synthesis, these models have recently been adopted in various vision tasks such as video prediction (Ding et al., 2025), robotic world modeling (Qiu et al., 2026), and data augmentation (Yuan et al., 2025), demonstrating that generative models can learn transferable motion priors from massive unlabeled videos. Nevertheless, existing studies mainly exploit generated videos for visual generation (DeAndres-Tame et al., 2025) or data augmentation (Yu et al., 2019; Ma et al., 2024a;b). Their potential for biometric recognition remains largely unexplored. Moreover, directly using generated samples for identity recognition is non-trivial because synthesized sequences inevitably introduce identity uncertainty. Our method addresses this issue through an Generative Identity Regularization, enabling reliable utilization of generated gait sequences.

3 GaitWM

3.1 Problem Formulation

We first formulate the problem of cross-view gait recognition. Let $\mathcal{D} = \{(\mathbf{X}_i, y_i, v_i)\}_{i=1}^{N}$ denote the training set, where $\mathbf{X}_i = \{\mathbf{x}_{i,t}\}_{t=1}^{T_i}$ is the gait sequence of the i-th sample, and $y_i \in \mathcal{Y}$ and $v_i \in \mathcal{V}$ denote its identity label and observed viewpoint, respectively. For each identity y, $\mathcal{V}_y$ denotes the set of its observed viewpoints.

Prior cross-view gait recognition methods are typically developed under the ***dense-view*** setting, where each identity is observed from sufficient viewpoints. Such multi-view observations provide abundant cross-view positive pairs, empowering the recognition model to learn identity-discriminative representations that are robust to viewpoint variations. In other words, sequences of the same identity are encouraged to remain closed in the feature space despite viewpoint changes.

In this paper, we focus on cross-view gait recognition under the ***sparse-view*** training setting, where each identity is observed from only a small number of viewpoints. Under this setting, most cross-view positive pairs are unavailable during training. Consequently, the model can only distinguish identities based on the limited observed views, without sufficient supervision to associate the same identity across different viewpoints. As a result, the learned representations are more sensitive to viewpoint variations and generalize poorly to unseen viewpoints. To solve this problem, we proposed GaitWM that takes an observed gait sequence $\mathbf{X}_i$ as input and generates a complementary sequence $\widehat{\mathbf{X}}_i^{v_t}$ from a novel target viewpoint v_t, as illustrated in Figure 2. The generated sequence inherits the identity label y_i from its source sequence and is jointly used with real samples for training. In this way, GaitWM enriches the viewpoint diversity of each identity and provides additional cross-view supervision without collecting multi-view data.

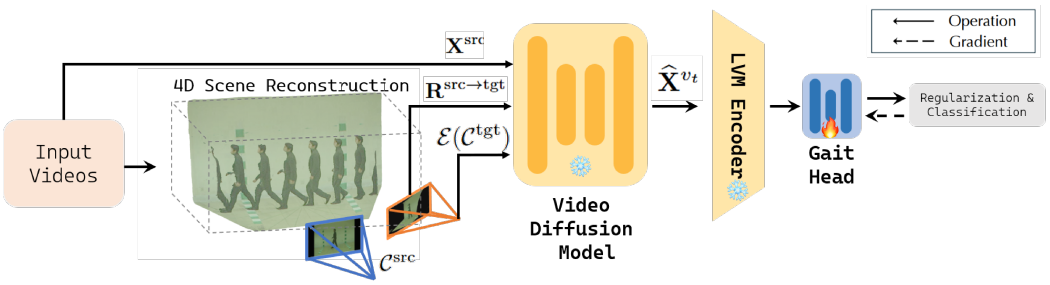


Figure 2: Overview of the proposed GaitWM framework. GaitWM first reconstructs a 4D scene representation and then employs a video diffusion model conditioned on the source observation and target viewpoint to generate identity-preserving novel-view gait sequences. The generated sequences are subsequently processed by a frozen LVM encoder and a trainable gait recognition network, while generative identity regularization improves the reliability of identity supervision for generated samples.

3.2 Novel View Synthesis for Dynamic Gait

Our key idea is to leverage the geometric and generative priors learned by foundation video models to synthesize identity-preserving gait sequences from novel viewpoints. Existing human video generation methods (Yuan et al., 2025; Ma et al., 2025) commonly crop pedestrians into fixed bounding boxes and use skeletons or SMPL models as generation conditions. Such template-based guidance is suitable for person re-identification, where texture and appearance provide strong identity cues. However, it is less suitable for gait recognition, where subtle body dynamics and temporal motion patterns are essential for identifying different subjects. In particular, fixed bounding boxes and predefined human templates may lose the original spatial context or restrict fine-grained motion variations.

Recent progress in monocular depth estimation (Lin et al., 2025; Wang et al., 2025b) and large-scale video generation (Wan Team, 2025) makes it possible to reconstruct a dynamic scene and reshoot it from a controllable camera viewpoint. Based on this observation, we synthesize gait sequences from novel viewpoints to provide additional cross-view supervision. Given a source gait video $\mathbf{X}^{\text{src}} = \{\mathbf{x}_t^{\text{src}}\}_{t=1}^{T}$, we adopt the pretrained Vista4D (Lin et al., 2026) to reshoot the walking subject under a target camera. The synthetic examples are shown in Figure 3. Different from frame-level view transformation, Vista4D first represents the source video as a dynamic 4D point cloud, allowing the original appearance and walking dynamics to be better preserved under large viewpoint changes.

Dynamic scene reconstruction. We first reconstruct the source video into a dynamic 4D scene, obtaining the frame-wise depth $\mathbf{D}^{\text{src}}$, camera intrinsics $\mathbf{K}^{\text{src}}$, and camera extrinsics $\mathbf{T}^{\text{src}}$. Meanwhile, SAM 2 (Ravi et al., 2025) is used to segment the walking subject. The remaining background regions are treated as static regions and represented by a static mask $\mathbf{M}^{\text{stc}}$. We then lift the RGB frames and their estimated depths into a world-coordinate point cloud:

$$\mathbf{P} = \Omega\left(\Phi^{-1}\left([\mathbf{X}^{\text{src}}, \mathbf{D}^{\text{src}}], \mathbf{K}^{\text{src}}\right), \mathbf{T}^{\text{src}}\right), \tag{1}$$

where Φ^{-1} denotes inverse perspective projection, and Ω transforms the reconstructed points from the camera coordinate system to the world coordinate system.

Since the walking subject changes continuously over time, the human points are associated with their corresponding frames. In contrast, the points belonging to static regions are retained across the whole sequence according to $\mathbf{M}^{\text{src}}$. By combining the frame-wise dynamic human points with the persistent static points, we obtain a temporally persistent 4D point cloud $\overline{\mathbf{P}}$. It preserves both the walking dynamics of the subject and the spatial context of the source scene.

Target camera construction. We then construct a target camera according to the desired novel viewpoint. Specifically, the center of the walking subject is first estimated from the reconstructed dynamic point cloud. Given the source viewpoint and a target viewpoint v_t, we apply a relative azimuth rotation $\Delta\theta$ to the reconstructed source camera around the vertical axis passing through the

subject center. The same relative transformation is applied to the camera poses of all frames, while the camera intrinsics, camera height, and camera-to-subject distance are kept unchanged.

This produces a target camera trajectory $\mathcal{C}^{\text{tgt}} = \{(\mathbf{K}_t^{\text{tgt}}, \mathbf{T}_t^{\text{tgt}})\}_{t=1}^{T}$, which observes the original walking process from the target viewpoint while maintaining temporal consistency across frames. We then render the temporally persistent point cloud $\overline{\mathbf{P}}$ using the target camera trajectory, obtaining a target-view point-map sequence $\mathbf{R}^{\text{src}\to\text{tgt}}$ and its corresponding alpha mask $\mathbf{M}^{\text{src}\to\text{tgt}}$. The rendered point map provides an explicit geometric condition for the target viewpoint, while the alpha mask indicates the visible regions and the regions that need to be completed by the video generation model.

Dynamic gait generation. We finally apply Vista4D (Lin et al., 2026) built upon the Wan2.1-T2V-14B video diffusion model (Wan Team, 2025) to generate dynamic gait sequences. The generation process is jointly conditioned on the source gait video, the target-view point map, the alpha mask, and the target camera trajectory:

$$\widehat{\mathbf{X}}^{v_t} = G_\psi\left(\mathbf{X}^{\text{src}}, \mathbf{R}^{\text{src}\to\text{tgt}}, \mathbf{M}^{\text{src}\to\text{tgt}}, \mathcal{E}(\mathcal{C}^{\text{tgt}})\right), \tag{2}$$

where G_ψ denotes the pretrained Vista4D model, and $\mathcal{E}(\mathcal{C}^{\text{tgt}})$ denotes the Plücker embedding of the target camera rays. The source video provides the appearance and walking dynamics of the subject, while the point-map and camera conditions determine how the source content should be observed from the target viewpoint. The generated sequence $\widehat{\mathbf{X}}^{v_t}$ therefore preserves the original walking process while presenting it from a novel viewpoint v_t. It is subsequently used as a complementary training sample for cross-view gait recognition.

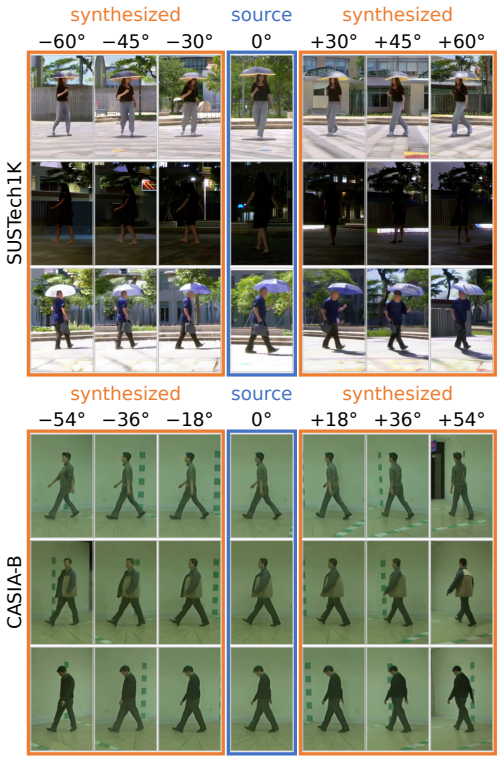
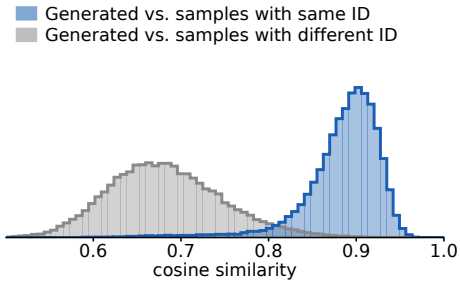
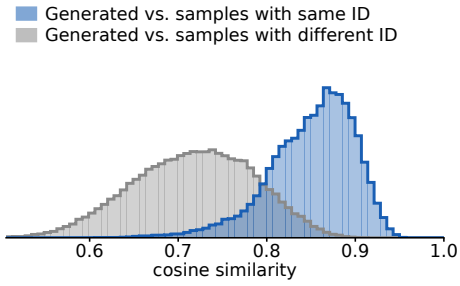


Figure 3: Visualization of novel-view gait sequences synthesized by GaitWM. Each row shows one sequence, with the center column representing the captured view and the others showing synthesized views at a relative azimuth rotation $\Delta\theta$.



(a) SUSTech1K



(b) CASIA-B

Figure 4: Similarity distribution between generated samples and real samples. The cosine similarity between generated samples and samples from the same identity is significantly higher than that with different identities on both SUSTech1K and CASIA-B, demonstrating that GaitWM generates identity-preserving novel views.

3.3 Generative Identity Regularization

The generated gait sequences inherit identity labels from their source sequences, but identity information may not be perfectly preserved during generation, especially under large viewpoint changes. Directly applying the same identity supervision to all generated samples may therefore introduce noisy constraints. To alleviate this issue, we slightly adapt the commonly used cross-entropy and triplet losses according to the reliability of generated sequences.

For cross-entropy loss, we assign a larger label-smoothing (Fan et al., 2023b) factor ϵ_{gen} to generated samples, where $\epsilon_{\text{gen}} > \epsilon_{\text{real}}$. This reduces the confidence of their inherited identity labels and prevents the recognition model from overfitting inaccurate identity details.

For triplet loss, we adapt the original margin m according to the relative viewpoint rotation $\Delta\theta_i$:

$$m_i = m\left[1 - \lambda_m \min\left(\frac{|\Delta\theta_i|}{\Delta\theta_{\max}}, 1\right)\right], \tag{3}$$

where $\Delta\theta_{\max}$ denotes the maximum viewpoint rotation considered during generation, and λ_m controls the maximum reduction of the triplet margin. A larger viewpoint rotation leads to a smaller margin and thus a weaker positive-pair constraint. For each anchor-positive pair, we use the smaller margin of the two samples, while keeping the negative-pair constraint unchanged.

When $\epsilon_{\text{gen}} = \epsilon_{\text{real}}$ and $\lambda_m = 0$, the proposed adaptation reduces to the original cross-entropy and triplet losses. Therefore, Generative Identity Regularization can be directly applied to existing gait recognition frameworks without modifying their network architectures.

4 Experiments

4.1 Experimental Setup

Dataset. SUSTech1K (Shen et al., 2023) and CASIA-B (Yu et al., 2006b) are used to evaluate GaitWM under the sparse-view gait recognition setting. We select these two datasets because they provide full-resolution gait videos, whereas most existing gait datasets contain only cropped pedestrian sequences. Specifically, **SUSTech1K** contains $25{,}239$ outdoor sequences of $1{,}050$ identities from 12 viewpoints under seven covariates, and **CASIA-B** contains 124 identities from 11 viewpoints under normal, bag and clothing-change conditions. We follow the official cross-view evaluation protocol and report rank-1 accuracy (%) excluding identical-view cases.

Training Details. (1) *Sparse-view training.* To simulate realistic sparse-view scenarios, each identity is only provided with gait sequences from a single viewpoint. We retain one 29-frame clip from this viewpoint and remove observations from all remaining viewpoints. Note that the selected viewpoint may still include multiple covariate conditions. The viewpoint is randomly sampled for each identity on SUSTech1K and fixed to 90° on CASIA-B. (2) *Novel view synthesis.* For each sparse-view sequence, we synthesize six novel-view sequences with different viewpoint changes. Specifically, the relative viewpoint variation $\Delta\theta$ is set to $\pm 30^\circ, \pm 45^\circ, \pm 60^\circ$ on SUSTech1K and $\pm 18^\circ, \pm 36^\circ, \pm 54^\circ$ on CASIA-B. The sequences are generated using the pretrained Vista4D (Lin et al., 2026) without additional fine-tuning. The generated sequences inherit the identity labels from their source sequences, while their silhouettes are extracted using the same procedure as real samples.

Setting. We implement four representative models spanning three input modalities: GaitBase (Fan et al., 2023b) and DeepGaitV2 (Fan et al., 2025) on silhouettes, BigGait (Ye et al., 2024) on RGB, and Gait-X (Wang et al., 2025c) on RGB and silhouettes, with BigGait as the default model in the ablations. All models are trained with SGD, an initial learning rate of 0.1 with milestone decay and a batch of 8×4 sequences of 30 frames, for 20k iterations except DeepGaitV2, which uses 30k; every baseline and its GaitWM counterpart share the same schedule. The Generative Identity Regularization is utilized, with hyper-parameters ϵ_{real}=0.1 and ϵ_{gen}=0.2. The triplet margin uses m=0.2, λ_m=0.5 and $\Delta\theta_{\max}$ set to the largest generated rotation, 60° on SUSTech1K and 54° on CASIA-B, so that the margin decreases from 0.2 on real pairs to 0.1 at the widest generated viewpoint, in proportion to how far a sample has been rotated away from its source.

Table 1: Comparison with existing gait recognition methods under dense-view and sparse-view training settings. The values in parentheses denote improvements over the corresponding sparse-view baselines.

Input	Method	Venue	SUSTech1K									CASIA-B			
			NM	BG	CL	CR	UB	UN	OC	NT	Overall	NM	BG	CL	Mean
Dense-view Training															
Sils	GaitBase	CVPR2023	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	96.5	91.5	78.0	88.7
	DeepGaitV2	TPAMI2025	87.4	84.1	53.4	81.3	86.1	84.8	88.5	28.8	82.3	94.3	90.0	78.6	87.6
RGB	BigGait	CVPR2024	95.1	96.8	66.2	96.6	95.5	92.1	99.2	83.3	95.3	99.8	99.3	88.6	95.9
RGB+Sils	Gait-X	ICCV2025	97.6	95.7	80.5	95.4	96.6	95.3	99.1	56.3	95.3	100.0	99.8	96.0	98.6
Sparse-view Training															
Sils	GaitBase	CVPR2023	7.6	6.6	4.0	6.5	5.4	4.7	5.0	8.5	6.4	20.8	16.6	10.8	16.1
	↪ + GaitWM	Ours	19.7	19.9	9.4	17.9	15.9	14.4	15.4	12.2	**17.9** (+11.5)	32.6	26.5	15.9	**25.0** (+8.9)
	DeepGaitV2	TPAMI2025	12.8	10.9	5.2	10.8	9.0	7.7	8.1	10.1	10.5	24.8	20.3	13.4	19.5
	↪ + GaitWM	Ours	15.1	13.8	7.9	13.2	11.5	10.9	13.1	10.5	**13.2** (+2.7)	34.4	25.9	12.9	**24.4** (+4.9)
RGB	BigGait	CVPR2024	45.5	46.6	28.4	46.0	46.5	35.9	36.9	53.4	45.2	56.9	53.6	38.3	49.6
	↪ + GaitWM	Ours	89.2	87.5	51.5	88.0	88.0	74.7	95.2	71.8	**87.1** (+41.9)	93.5	88.7	45.9	**76.0** (+26.4)
RGB+Sils	Gait-X	ICCV2025	31.6	29.8	17.1	29.4	27.2	22.3	25.1	17.2	28.9	47.7	42.8	30.2	40.3
	↪ + GaitWM	Ours	84.2	82.3	53.5	81.5	82.2	72.9	90.1	35.6	**81.4** (+52.5)	86.8	82.5	53.0	**74.1** (+33.8)

4.2 Main Results

Sparse-view v.s. Dense-view training. We first investigate how existing gait recognition methods behave when dense multi-view supervision is unavailable. Under the conventional ***dense-view training*** setting, where each identity is associated with multiple viewpoints, previous SoTA methods achieve strong cross-view recognition performance, as summarized in the upper half of Table 1. For example, the representative silhouette-based methods GaitBase and DeepGaitV2 obtain 76.1% and 82.3% mean rank-1 accuracy on SUSTech1K, respectively, while the RGB-based BigGait achieves over 95% rank-1 accuracy on both SUSTech1K and CASIA-B. However, when the training data are reduced to the ***sparse-view training*** setting and no multi-view gait sequences are available, these methods suffer from substantial performance degradation, as they can no longer explicitly learn view-consistent identity representations. For instance, BigGait drops from 95.3% to 45.2% mean rank-1 accuracy on SUSTech1K. In contrast, by generating complementary novel-view gait sequences, GaitWM effectively restores cross-view supervision and improves sparse-view recognition from 45.2% to 87.1% ($+41.9\%$) on SUSTech1K and from 49.6% to 76.0% ($+26.4\%$) on CASIA-B. Furthermore, GaitWM is not tied to a specific recognition architecture. When integrated with existing gait recognition methods, including GaitBase, DeepGaitV2, BigGait, and Gait-X, GaitWM consistently improves their performance under the sparse-view training protocol, demonstrating its effectiveness as a general data augmentation paradigm for label-efficient cross-view gait recognition.

4.3 Ablation Study

Effectiveness of synthetic novel views. As shown in Table 2, sparse-view training suffers from limited cross-view supervision, achieving only 45.2% and 49.6% accuracy on SUSTech1K and CASIA-B, respectively. Introducing only one synthesized novel view significantly improves performance under all settings, demonstrating that generated views effectively compensate for missing viewpoint information. Moreover, larger viewpoint variations generally bring greater improvements, as they provide more diverse cross-view supervision. For example, $\Delta\theta = -60^\circ$ improves the accuracy by 36.0% on SUSTech1K and 22.6% on CASIA-B. As further illustrated in Fig. 7, GaitWM substantially improves cross-view recognition consistency with synthetic novel views.

Effect of generative identity regularization. Table 3 verifies the effectiveness of the proposed generative identity regularization. Without either component, generated samples may introduce unreliable identity supervision. Combining label smoothing for generated samples and adaptive triplet margin achieves the best performance, reaching 87.1% and 76.0% accuracy on SUSTech1K and CASIA-B, respectively.

Table 2: Ablation of synthesized novel view under sparse-view training. Adding only one novel view substantially improves cross-view recognition significantly on both datasets. Larger viewpoint rotations generally provide greater gains.

Training setting	Rotation $\Delta\theta$		Overall accuracy (%)	
	SUSTech1K	CASIA-B	SUSTech1K	CASIA-B
w/o novel view	–	–	45.2	49.6
w/ novel view	$-30°$	$-18°$	72.7 (Δ+27.5)	57.1 (Δ+7.5)
	$+30°$	$+18°$	76.2 (Δ+31.0)	58.8 (Δ+9.2)
	$-45°$	$-36°$	79.9 (Δ+34.7)	66.4 (Δ+16.8)
	$+45°$	$+36°$	80.9 (Δ+35.7)	62.9 (Δ+13.3)
	$-60°$	$-54°$	**81.2** (Δ+36.0)	**72.2** (Δ+22.6)
	$+60°$	$+54°$	<u>80.2</u> (Δ+35.0)	<u>68.5</u> (Δ+18.9)

Table 3: Ablation of Generative Identity Regularization. The full regularization consistently improves sparse-view gait recognition on both datasets.

Generative Identity Regularization		Overall accuracy (%)	
$\epsilon_{\text{gen}} > \epsilon_{\text{real}}$	$\lambda_m > 0$	SUSTech1K	CASIA-B
✗	✗	86.0	75.2
✗	✓	<u>86.8</u>	75.2
✓	✗	85.5	<u>75.8</u>
✓	✓	**87.1**	**76.0**

Table 4: Identity consistency evaluation of generated gait sequences. Real and Synthesized denote captured and generated probes, respectively. The comparable recognition accuracy demonstrates that GaitWM preserves identity information during novel-view synthesis.

SUSTech1K		CASIA-B	
Real	Synthesized	Real	Synthesized
76.0	61.7	71.7	68.7

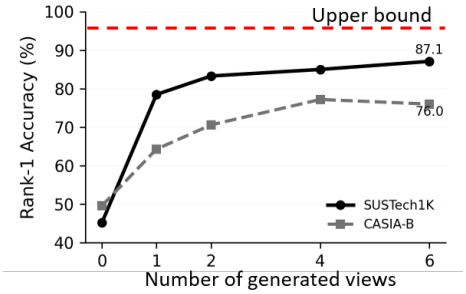


Figure 5: Ablation study on the number of generated views.

Effect of the Number of Generated Views. We analyze how the number of generated views affects sparse-view gait recognition. As shown in Fig. 5, increasing generated viewpoint coverage consistently improves performance on both benchmarks. Without generated views, the model achieves only 45.2% and 49.6% Rank-1 accuracy on SUSTech1K and CASIA-B due to severe cross-view ambiguity. With six synthesized views, the performance increases to 87.1% and 76.0%, approaching the upper-bound results obtained with fully collected multi-view data (95.3% and 95.9%). This demonstrates that GaitWM can effectively recover cross-view supervision from sparse observations. These results validate that viewpoint diversity is critical for sparse-view gait recognition. By leveraging large-scale video priors, GaitWM learns view transformations and generates identity-preserving novel views, substantially expanding viewpoint coverage without expensive multi-view data collection.

4.4 Qualitative Results

Identity consistency of generated views. We evaluate whether synthesized views preserve identity consistency using both recognition accuracy and feature similarity. As shown in Table 4, generated sequences achieve 61.7% and 68.7% probe accuracy on SUSTech1K and CASIA-B, compared with 76.0% and 71.7% for real probes. Figure 4 further shows that generated sequences remain substantially closer to real sequences of the same identity than to those of different identities. The average cosine similarities are 0.884 vs. 0.683 on SUSTech1K and 0.850 vs. 0.717 on CASIA-B. These results indicate that GaitWM changes the viewpoint while largely preserving identity information, making the generated sequences effective cross-view training samples.

Embedding visualization. Figure 6 visualizes the test embeddings learned under sparse-view training and with GaitWM-generated views. Compared with the sparse-view baseline, embeddings learned with GaitWM exhibit more compact intra-identity clusters and clearer separation between different identities on both SUSTech1K and CASIA-B. This indicates that the synthesized

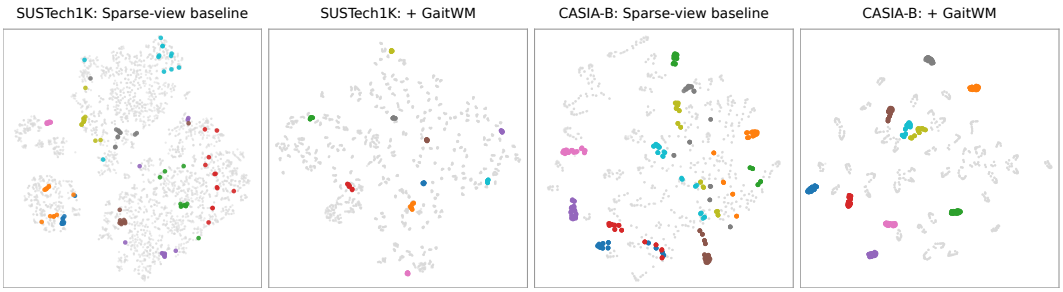
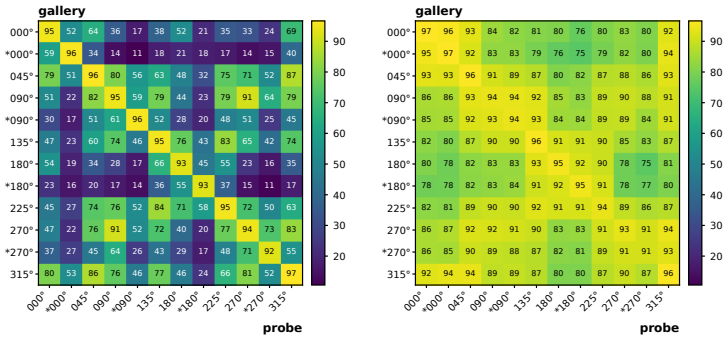


Figure 6: t-SNE visualization. Different colors denote ten identities, while remaining samples are shown in gray. Compared with sparse-view training (left), adding generated views (right) produces more compact and identity-discriminative embeddings. Best viewed in color.



(a) Without Generated Views (b) With Generated Views

Figure 7: Cross-view rank-1 accuracy (%) on SUSTech1K. Compared with the sparse-view baseline, GaitWM substantially improves cross-view recognition consistency by generating complementary novel views, especially under large viewpoint variations.

novel views provide effective cross-view supervision, enabling model to learn view-consistent and identity-discriminative gait representations.

4.5 Cost Analysis

We compare the cost of obtaining multi-view gait data through real-world collection and video generation. Taking SUSTech1K as an example, dense-view data collection requires approximately \$50K and 500 hours of human recording, whereas GaitWM costs about \$32K for 3,840 GPU hours. Although generation is computationally intensive, it can be parallelized and reused across identities and viewpoint configurations. These advantages in time, cost, and scalability suggest that generative models may play an important role in enabling large-scale gait recognition.

5 Conclusion

In this work, we introduce GaitWM, a generative world model for sparse-view gait recognition. Different from existing approaches that rely on densely collected multi-view gait data, GaitWM learns view-invariant priors from large-scale video generative model and generates identity-preserving novel viewpoints from sparse observations. With synthesized novel views, GaitWM effectively expands viewpoint coverage and provides significant cross-view supervision. Extensive experiments demonstrate that GaitWM substantially improves sparse-view recognition performance and approaches the upper-bound performance achieved with fully collected multi-view data. Beyond gait recognition, our study reveals the potential of generative models as a scalable alternative to expensive real-world data collection. By transforming limited observations into richer and more generalizable training distributions, GaitWM offers a promising direction for data-efficient gait recognition.

AI USE STATEMENT

We used the pretrained Vista4D generative video model to synthesize novel-view gait sequences for training, as described in the paper. We also used ChatGPT to assist with drafting and editing portions of the manuscript and with LaTeX formatting of figures and tables. The authors reviewed and revised all AI-assisted material, checked the reported results and scientific claims against the underlying experiments and cited sources, and take full responsibility for the final content of this work.

REFERENCES

Jeongho Ahn, Kazuto Nakashima, Koki Yoshino, Yumi Iwashita, and Ryo Kurazume. 2v-gait: Gait recognition using 3d lidar robust to changes in walking direction and measurement distance. In *IEEE/SICE SII*, pp. 602–607, 2022.

Ivan DeAndres-Tame, Chengwei Ye, Ruben Tolosana, Ruben Vera-Rodriguez, and Shiqi Yu. Is visual realism enough? evaluating gait biometric fidelity in generative ai human animation, 2025.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14920–14937, 2023a.

Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9707–9716, 2023b.

Chao Fan, Saihui Hou, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, Yongzhen Huang, and Shiqi Yu. Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xiao Han, Yiming Ren, Peishan Cong, Yujing Sun, Jingya Wang, Lan Xu, and Yuexin Ma. Gait recognition in large-scale free environment via single lidar. In *ACM International Conference on Multimedia*, pp. 380–389, 2024.

Zhanbo Huang, Dingqiang Ye, Xiaoming Liu, and Yu Kong. Unlocking motion from large vision models with a semantic and kinematic duality for gait recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28379–28390, 2026.

Dongyang Jin, Chao Fan, Jingzhe Ma, Jingkai Zhou, Weihua Chen, and Shiqi Yu. On denoising walking videos for gait recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12347–12357, 2025.

Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13937–13947, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2024.

Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2023.

Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *European Conference on Computer Vision*, pp. 375–390, 2022.

Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pp. 14648–14656, 2021.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4d: Video reshooting with 4d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Changxiao Ma, Chao Yuan, Xincheng Shi, Yuzhuo Ma, Yongfei Zhang, Longkun Zhou, Yujia Zhang, Shangze Li, and Yifan Xu. Omniperson: Unified identity-preserving pedestrian generation, 2025.

Jingzhe Ma, Dingqiang Ye, Chao Fan, and Shiqi Yu. Pedestrian attribute editing for gait recognition and anonymization, 2023.

Jingzhe Ma, Haoyu Luo, Zixu Huang, Dongyang Jin, Rui Wang, Johann A Briffa, Norman Poh, and Shiqi Yu. Passersby-anonymizer: Safeguard the privacy of passersby in social videos. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2024a.

Jingzhe Ma, Xiaoqing Zhang, and Shiqi Yu. An identity-preserved framework for human motion transfer. *IEEE Transactions on Information Forensics and Security*, 19:3495–3509, 2024b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Boxiang Qiu, Liliang Chen, Yue Liao, Nan Wang, Lintao Wang, Jiayi Luo, Wenzhi Zhao, Shengcong Chen, Di Chen, Ye Li, et al. Ge-sim 2.0: A roadmap towards comprehensive closed-loop video world simulators for robotic manipulation, 2026.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):264–284, 2022.

Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q Huang, and Liang Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1054–1063, 2023.

Chuanfu Shen, Rui Wang, Lixin Duan, and Shiqi Yu. Lidargait++: Learning local features and size awareness from lidar point clouds for 3d gait recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6627–6636, 2025.

Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. on Computer Vision and Applications*, 10(4):1–14, 2018.

Wan Team. Wan: Open and advanced large-scale video generative models, 2025.

Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Liang Wang. Causal intervention for sparse-view gait recognition. In *ACM International Conference on Multimedia*, pp. 77–85, 2023.

Jilong Wang, Saihui Hou, Xianda Guo, Yan Huang, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Gaitc^{3}i: Robust cross-covariate gait recognition via causal intervention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Permutation-equivariant visual geometry learning, 2025b.

Zengbin Wang, Saihui Hou, Junjie Li, Xu Liu, Chunshui Cao, Yongzhen Huang, Siye Wang, and Man Zhang. Gait-x: Exploring x modality for generalized gait recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 13259–13269, 2025c.

Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Gait recognition from a single image using a phase-aware gait cycle reconstruction network. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2020.

Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 200–210, 2024.

Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Dingqiang Ye, Chao Fan, Kartik Narayan, Bingzhe Wu, Chengwen Luo, Jianqiang Li, and Vishal M Patel. Silhouette-based gait foundation model, 2025b.

Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition*, pp. 441–444, 2006a.

Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition*, volume 4, pp. 441–444, 2006b.

Shiqi Yu, Rijun Liao, Weizhi An, Haifeng Chen, Edel B Garcia, Yongzhen Huang, and Norman Poh. Gaitganv2: Invariant gait feature extraction using generative adversarial networks. *Pattern Recognition*, 87:179–189, 2019.

Chao Yuan, Guiwei Zhang, Changxiao Ma, Tianyi Zhang, and Guanglin Niu. From poses to identity: Training-free person re-identification via feature centralization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24409–24418, 2025.

Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 345–360, 2020.

Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20228–20237, 2022.

Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He, and Tieniu Tan. Robust view transformation model for gait recognition. In *2011 18th IEEE international conference on image processing*, pp. 2073–2076. IEEE, 2011.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024.

Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *IEEE/CVF International Conference on Computer Vision*, pp. 14789–14799, 2021.

Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-covariate gait recognition: A benchmark. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 7855–7863, 2024.

A Datasets, Splits and Evaluation Protocol

A.1 Datasets

We use the public versions of SUSTech1K and CASIA-B. SUSTech1K contains 25,239 outdoor sequences from 1,050 identities, captured across 12 viewpoints. CASIA-B contains 124 identities observed from 11 viewpoints under normal walking, bag-carrying, and clothing-change conditions. We evaluate on these benchmarks because they provide **full-resolution gait videos**, whereas most existing datasets release only cropped pedestrian sequences that lack sufficient scene context for dynamic scene reconstruction.

A.2 Construction of the Sparse-View Split

We retain the official train/test partitions and modify only the viewpoints available during training.

SUSTech1K. We use the official split of 250 training and 800 test identities. Each training identity is restricted to **one of eight** canonical viewpoints, $000°, 045°, \ldots, 315°$. All sequences and covariates under the selected viewpoint are retained. This deterministic assignment requires no random seed, and the resulting split is released with the code. Test identities remain unchanged with all viewpoints.

CASIA-B. We follow the standard 74/50 train/test split. All training identities are restricted to the $090°$ viewpoint, while test identities retain all 11 viewpoints.

B Novel-View Generation Pipeline

We provide the implementation details of the three-stage pipeline described in the main paper.

Input Cropping. For each sequence, we select the center 29 frames and compute the union of their subject bounding boxes. A fixed crop centered on this union is applied to all frames, preserving a static background. We use a tight subject-centered crop as the source view and a second crop enlarged by $2\times$ for reconstruction. The wider crop retains additional scene geometry and improves background stability under large viewpoint rotations.

Target Camera. The target camera is obtained by rotating the reconstructed source camera around the vertical axis through the subject centroid by the desired angle. It is then translated backward along its optical axis until the projected subject points from all frames occupy at most 85% of the image dimensions. Before computing the centroid, points beyond six median absolute deviations from the median are removed to suppress depth outliers. The camera intrinsics and height remain unchanged.

Generation. Each sequence contains 29 frames at a resolution of 384×672, with the diffusion seed fixed to 10027. We generate views at $\pm 30°$, $\pm 45°$, and $\pm 60°$ for SUSTech1K, and $\pm 18°$, $\pm 36°$, and $\pm 54°$ for CASIA-B. This produces six synthesized sequences per source sequence. Each sequence inherits the identity label of its source, requiring no additional annotation.

C Additional Results

Within-domain and Cross-domain Generalization. We further evaluate whether GaitWM can benefit a strong gait recognition encoder under both within-domain and cross-domain settings. As shown in Table 5, augmenting dense-view SUSTech1K with synthetic views substantially improves generalization to unseen datasets. On CCPG, adding four synthetic views improves the BG accuracy from 45.5% to 62.9%, while consistent gains are also observed under CL, UP, and DN. Improvements on CASIA-B are relatively moderate but remain evident, particularly under the challenging CL condition, where the accuracy increases from 18.7% to 22.7%. In contrast, the within-domain performance on SUSTech1K decreases slightly from 95.3% to around 95.0% when one to four synthetic views are introduced. This small degradation may arise because the generated viewpoints

Table 5: Within-domain and cross-domain evaluation of GaitWM using BigGait as the gait recognition encoder. Synthetic views are progressively added to the dense-view SUSTech1K training set. Rank-1 accuracy (%) is reported.

Training Data	#Sequence	#ID	Test Set															
			SUSTech1K									CASIA-B			CCPG			
			NM	BG	CL	CR	UB	UN	OC	NT	MEAN	NM	BG	CL	CL	UP	DN	BG
Dense-view data	6011	250	95.4	96.2	69.4	96.5	95.3	92.1	98.8	82.8	**95.3**	91.1	85.8	18.7	4.5	11.5	11.9	45.5
+ 1 synthetic view	$6011 + 250 \times 1$	250	95.0	95.4	66.3	**96.3**	95.1	91.1	99.2	82.7	95.0	91.1	84.0	18.8	6.0	13.0	15.7	57.1
+ 2 synthetic views	$6011 + 250 \times 2$	250	**95.3**	95.3	**67.6**	96.0	95.0	**91.3**	**99.8**	**83.7**	95.0	92.2	**86.9**	**22.7**	5.0	12.9	15.2	60.0
+ 4 synthetic views	$6011 + 250 \times 4$	250	94.6	**95.7**	65.3	96.2	**95.4**	90.7	99.5	82.7	94.9	**93.0**	86.3	20.5	**6.7**	**14.2**	**17.5**	**62.9**
+ 6 synthetic views	$6011 + 250 \times 6$	250	89.2	87.5	51.5	88.0	88.0	74.7	95.2	71.9	87.1	88.4	80.7	20.4	6.2	11.8	17.5	58.1

alter the original viewpoint distribution of the training set, introducing a distribution shift relative to the captured SUSTech1K test data. When six synthetic views are added, this shift becomes more pronounced and leads to a larger within-domain decrease. Nevertheless, the same viewpoint diversification reduces reliance on source-domain-specific view distributions and improves robustness to unseen domains. These results indicate that GaitWM trades a small amount of source-domain downdegradation for stronger cross-domain generalization.

Table 6: Cross-domain evaluation under sparse-view training. Each model is trained on the indicated source dataset with sparse-view training data and directly evaluated on unseen test sets without adaptation. Rank-1 accuracy (%) is reported with identical-viewpoint pairs excluded. Bold indicates the better result.

Source Dataset: SUSTech1K								
Input	Method	Test Set						
		CASIA-B			CCPG			
		NM	BG	CL	CL	UP	DN	BG
Sils	GaitBase	18.7	11.6	6.4	5.4	7.5	9.4	11.1
	+ GaitWM	**27.2**	**20.4**	**10.0**	**8.7**	**12.0**	**14.6**	**20.6**
	DeepGaitV2	23.9	17.3	8.8	7.0	10.3	11.7	16.5
	+ GaitWM	**31.4**	**20.9**	**11.3**	**8.2**	**13.2**	**14.6**	**21.7**
RGB	BigGait	55.9	49.3	15.9	4.8	7.8	11.6	20.7
	+ GaitWM	**88.4**	**80.7**	**20.4**	**6.2**	**11.8**	**17.5**	**58.1**
RGB+Sils	Gait-X	46.5	40.6	19.4	**5.7**	6.6	13.1	16.1
	+ GaitWM	**87.3**	**80.4**	**32.6**	5.6	**10.1**	**21.8**	**40.6**
Source Dataset: CASIA-B								
Input	Method	Test Set						
		SUSTech1K			CCPG			
		CL	NT	Overall	CL	UP	DN	BG
Sils	GaitBase	3.8	6.7	4.9	6.3	8.2	8.6	9.6
	+ GaitWM	**6.6**	**8.7**	**8.8**	**8.4**	**11.0**	**11.3**	**13.6**
	DeepGaitV2	3.8	7.8	4.8	7.0	9.1	8.9	10.2
	+ GaitWM	**4.1**	**8.8**	**6.8**	**7.6**	**10.7**	**13.0**	**14.3**
RGB	BigGait	18.8	31.0	27.8	5.4	9.1	8.1	20.4
	+ GaitWM	**30.3**	**60.9**	**66.9**	5.4	**15.5**	**14.4**	**47.9**
RGB+Sils	Gait-X	8.8	14.3	18.0	4.1	5.8	8.0	11.4
	+ GaitWM	**19.0**	**30.0**	**36.7**	**5.3**	**7.9**	**14.7**	**21.9**

Cross-domain Generalization. Table 6 evaluates whether the benefits of GaitWM transfer beyond the source dataset. Across both training directions, GaitWM consistently improves the cross-

domain performance of different gait encoders and input modalities. When trained on SUSTech1K, GaitWM improves BigGait from 55.9% to 88.4% on CASIA-B NM and from 20.7% to 58.1% on CCPG BG. Similarly, when models are trainined on CASIA-B, the overall accuracy of BigGait on SUSTech1K increases from 27.8% to 66.9%, while its CCPG BG accuracy improves from 20.4% to 47.9%. Consistent gains are also observed for GaitBase, DeepGaitV2, and Gait-X, showing that GaitWM is not tied to a particular architecture or input representation. These improvements indicate that synthesized viewpoints provide complementary view variations that reduce reliance on source-specific viewpoint distributions, thereby improving generalization to unseen datasets without target-domain adaptation.

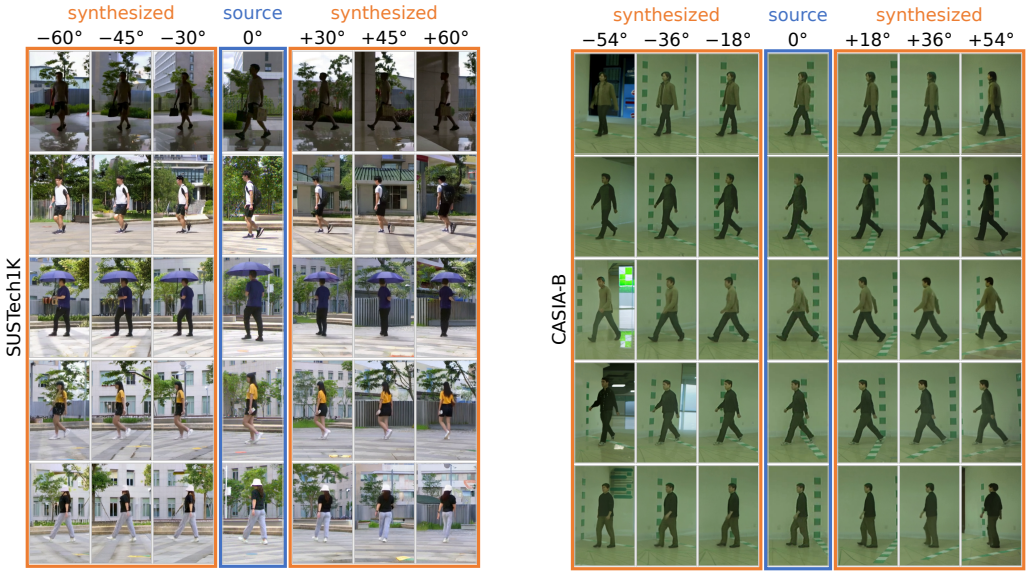


Figure 8: Synthesized views for five source sequences per benchmark.

Qualitative Analysis. Figures 8 and 9 show that GaitWM preserves visually recognizable subject characteristics while producing plausible viewpoint transformations across diverse walking conditions. The synthesized sequences maintain consistent appearance across viewpoints, suggesting that GaitWM preserves temporal gait dynamics rather than generating independently plausible frames. As the rotation magnitude increases, background distortions become more noticeable, whereas the foreground subject and walking motion remain generally stable within the adopted generation range. At extreme rotations of $\pm 90^\circ$, the subject may disappear or remain close to the source viewpoint, as illustrated in Figure 10. These observations motivate limiting the generation range to $\pm 60^\circ$ on SUSTech1K and $\pm 54^\circ$ on CASIA-B. They also support the use of generative identity regularization, which downweights less reliable synthesized samples while retaining the benefits of expanded viewpoint coverage.

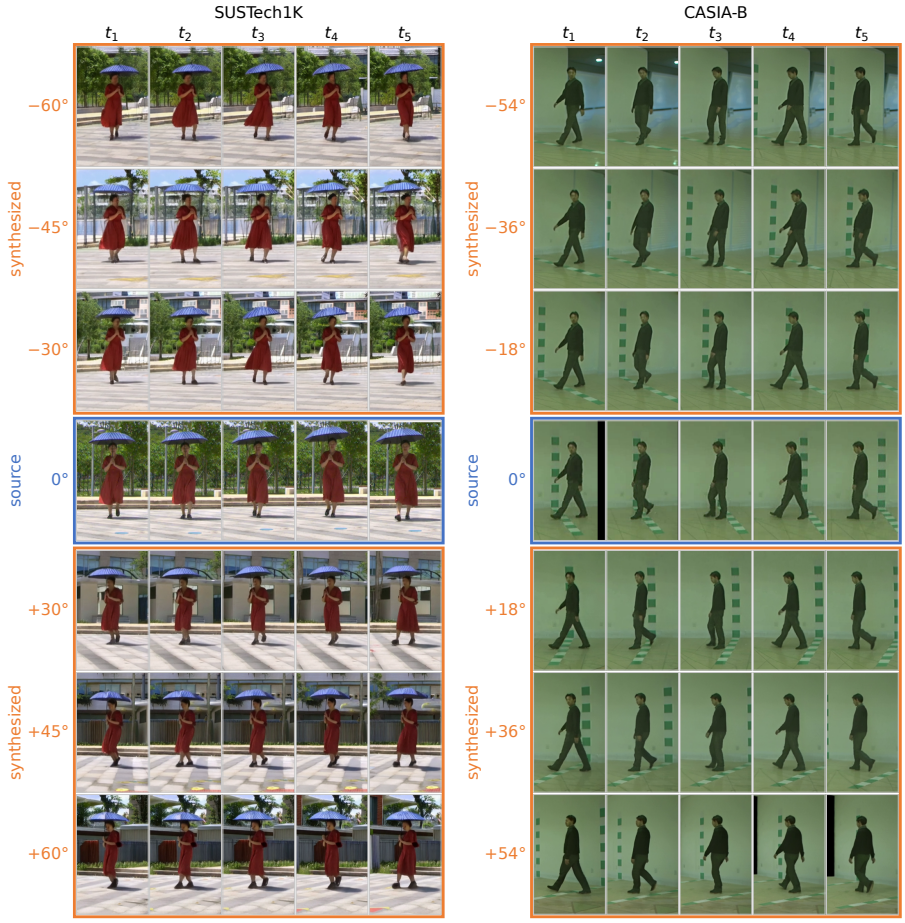


Figure 9: Temporal consistency of synthesized views for one sequence from each benchmark. The middle row shows the captured source view, while the rows above and below show views synthesized at negative and positive rotations, respectively. Each column displays the same frame index across all seven viewpoints, enabling direct comparison of gait phases at the same instant.

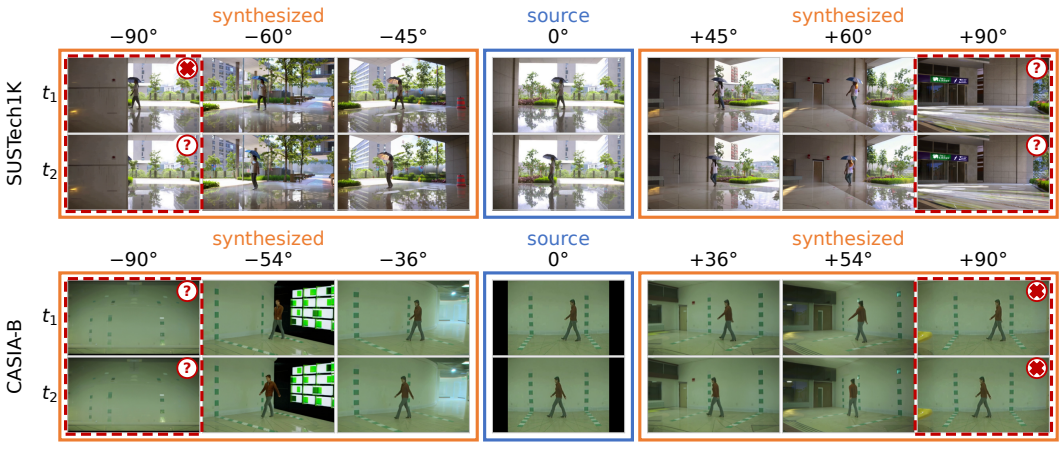


Figure 10: Generation behavior beyond the adopted rotation range. The outermost columns correspond to $\pm 90°$, are enclosed by dashed boxes, and are generated only for visualization rather than training. A circled question mark indicates a missing subject, whereas a circled cross indicates that the subject is rendered but the requested viewpoint rotation is not achieved.