Silhouette-based Gait Foundation Model

Dingqiang Ye[1,2], Chao Fan[2,⋆], Kartik Narayan[1],
Bingzhe Wu[2], Chengwen Luo[2], Jianqiang Li[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, USA
[2] Shenzhen University, China
dye6@jh.edu, chaofan996@szu.edu.cn, knaraya4@jh.edu,
wubingzheagent@gmail.com, {chengwen, lijq}@szu.edu.cn, vpatel36@jhu.edu

Abstract. Gait patterns play a critical role in human identification and healthcare analytics, yet current progress remains constrained by small, narrowly designed models that fail to scale or generalize. Building a unified gait foundation model requires addressing two longstanding barriers: (a) Scalability – Why have gait models historically failed to follow empirical scaling trends? (b) Generalization – Can one model serve the diverse gait tasks that have traditionally been studied in isolation? We introduce **FoundationGait**, the first scalable, self-supervised pretraining framework for vision-based gait understanding. Its largest version has nearly 0.13 billion parameters and is pretrained on 12 public gait datasets comprising over 2 million walking sequences. Extensive experiments demonstrate that FoundationGait, with or without fine-tuning, performs robustly across a wide spectrum of gait datasets, conditions, tasks (e.g., human identification, scoliosis screening, depression prediction, and attribute estimation), and even input modality. Notably, it achieves 48.0% **self-supervised** rank-1 accuracy on the challenging in-the-wild Gait3D dataset (1,000 test subjects) and 64.5% on the largest in-the-lab OU-MVLP dataset (5,000+ test subjects), setting a new milestone in robust gait recognition. These results establish FoundationGait as a strong and versatile foundation for future gait research. All code and models: https://github.com/ShiqiYu/OpenGait.

Keywords: Gait Recognition · Gait Healthcare · Vision Foundation Model · Human Biometrics

1 Introduction

Vision-based gait offer a compelling solution for large-scale, low-cost, and non-invasive applications—including human identification and healthcare [39, 47, 48, 65, 67, 97]—as illustrated in Fig. 1. Their ability to capture motion remotely and unobtrusively in unconstrained environments makes gait uniquely suitable for real-world deployment. Yet, despite this promise, a dedicated vision-based Gait Foundation Model has not emerged, even as large vision models (LVMs) [32, 49, 50] have reshaped many related domains.

⋆ Corresponding Author.

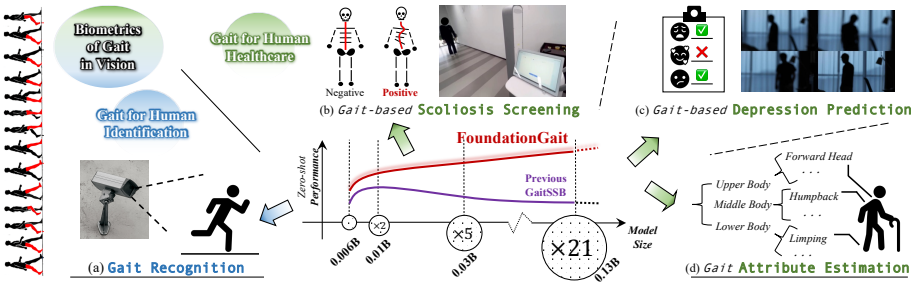

Fig. 1: FoundationGait: A scalable and unified gait foundation model for diverse gait-related tasks, including person identification, scoliosis screening, gait attribute estimation, and depression detection.

Recent progress in domain-specific LVMs [17, 19, 30, 31, 59], which balance specialization with broad generalization, has demonstrated their practical value and attracted significant attention from both academia and industry [1]. Motivated by these developments, this paper takes a step toward establishing such a foundation model for gait.

Recent efforts toward building a gait foundation model generally follow two directions. The first, exemplified by BigGait [85] and subsequent studies [29, 84], focuses on transferring capacity from existing LVMs—such as DINOv2 [49] and Stable Diffusion [54]—into the gait domain. The second, represented by GaitSSB [10] and related work [44, 67], instead trains gait models from scratch using vision-only or vision–language contrastive pretraining strategies [4, 53].

Both directions demonstrate that rich and discriminative gait representations can indeed be learned from walking videos under supervised or self-supervised regimes. Despite the significant progress, they—similar to much of the long-standing gait literature [9, 39, 67, 97]—remain task-specific, typically optimizing models for individual datasets rather than pursuing unified, scalable solutions.

Two core properties required for a true gait foundation model remain insufficiently explored. (a) **Size Scalability.** Although empirical scaling trends are a known driver of progress in modern foundation models, systematic investigations of scalability for video-based gait modeling are rare. In practice, we find that many gait-specific architectures fail to benefit from increasing parameters beyond relatively small sizes (≈10M); performance often saturates or even drops sharply as model capacity grows (Sec. 3.1). Historically, this limitation is mainly attributed to the simplicity of gait inputs—low resolution (64×44) sparse binary silhouettes—together with limited training data, leading to overfitting and hindering large-scale models. However, we observe that large-scale models suffering from fine-grained detail loss is also a key factor to the unscalable issue. (b) **Cross-task Generalization.** A defining feature of foundation models is their ability to transfer across diverse downstream tasks. Yet, gait research has traditionally treated recognition-oriented tasks and healthcare-oriented analyses

as separate problem spaces, despite both relying on subtle, fine-grained motion dynamics.

Motivated by these gaps, we investigate the design principles of a true gait foundation model and introduce FoundationGait, a scalable self-supervised pretraining framework that generalizes effectively across a broad spectrum of gait datasets and downstream tasks. Thanks to its frequent usage (62.8% in recent top venues according to OpenGait [9]), FoundationGait takes the popular silhouette sequence as input. Our exploration centers on two key questions:

- *How can we scale gait models from 0.01B parameters to the 1B range*? In Fig. 3, even with a million-level diverse training data, current gait models still fail to scale. We find that preserving fine-grained local body-part cues is essential for scaling. Building on this insight, we develop a novel *part-aware pretraining* strategy that enables FoundationGait to scale by **20**× over widely used baselines (e.g., DeepGaitV2's ≈0.006B backbone [9]), while achieving consistent performance gains across datasets and tasks.
- *How can we build an efficiently large gait corpus for pretraining?* Although the existing GaitLU-1M [10] dataset contains nearly one million silhouette sequences, it has been criticized for its limited photographic diversity [44]. To address this limitation, we curate 12 publicly available gait datasets spanning both recognition tasks [10, 35, 55, 57, 62, 87, 94, 99, 100] and healthcare applications [39, 67, 97], collectively forming WebGait-2M[3]. This collection comprises over 2.35 million walking sequences, covering a broad spectrum of gait-related challenges (Table 1) and providing a more representative characterization of the real-world distribution of open-source gait data studied by the research community.

In extensive experiments, FoundationGait delivers strong and consistent performance across both recognition and healthcare settings. On challenging gait recognition benchmarks—Gait3D [94] and OU-MVLP [62], containing 1,000 in-the-wild and 5,000+ in-the-lab test subjects, respectively—FoundationGait achieves 48.0% and 64.5% self-supervised rank-1 accuracy, establishing new milestones for robust cross-condition recognition. In healthcare-oriented tasks, such as scoliosis screening on the Scoliosis1K dataset (1,000+ participants), FoundationGait reaches 97.0% accuracy after fine-tuning. Detailed evaluations and ablations are provided in Sec. 4.

Overall, this paper makes three major contributions:

- We demonstrate—systematically and for the first time—that vision-based gait pretraining can benefit from empirical scaling trends.
- We unify human identification and healthcare tasks within a single, scalable gait foundation model.
- Extensive experiments, with and without fine-tuning, verify the strong performance and broad cross-task generalization of **FoundationGait**.

[3] We neither process nor redistribute any data; "WebGait-2M" is a convenient name.

2 Related Works

Walking patterns, as a fundamental and natural behavior, are widely recognized for their potential to reveal both human identity and health status.

Gait Recognition. Gait recognition plays a crucial role in surveillance, particularly in unconstrained and long-distance scenarios. Recent research can be broadly categorized into five main areas [9]: (a) Mitigating gait-unrelated factors, such as the influence of RGB-encoded clothing and backgrounds in walking videos. Approaches in this direction include human parsing techniques [73, 95, 100], SMPL models [75, 76, 79, 94], multi-modal learning [12, 28], end-to-end training with inductive biases [29, 36, 84, 85], and the use of advanced sensors such as LiDAR [55, 56]. (b) Enhancing networks with gait-specific priors, such as spatio-temporal and local-global feature learning, combined with attention refinements [13, 20, 27, 38, 42, 51, 70, 71]. (c) Improving the learning process through novel frameworks, such as auto-encoder regression [18], generative modeling [25, 29, 43, 86], causal intervention [7, 68, 80], and contrastive learning [10, 44, 67]. (d) Addressing real-world challenges such as occlusion [21–23, 26, 81], clothing variation [35], and low-quality data [74]. (e) Developing improved loss functions for better optimization [60, 88, 92].

Anomaly Gait Detection. For instance, scoliosis, which affects 0.47%–5.2% of adolescents [33], was addressed by Zhou et al. [97] with the Scoliosis1K dataset, comprising over 1,000 adolescent patients, and the multi-task learning model ScoNet. Similarly, depression impacts 7.1% of U.S. adults annually [45]. Liu et al. [39] released a video-based depression dataset with 100+ participants and evaluated various gait recognition models on it. Wang et al. [67] developed a gait attribute benchmark involving 15 attributes (e.g., toe-out, head forward, humpback) from 533 subjects and over 120,000 sequences. They also created a CLIP-like [53] model for potential use in security and healthcare.

As noted, gait analysis [77, 78] for both identification and healthcare sometimes benefits from each other. For example, Zhou [97] incorporated identification signals to enhance scoliosis classification, while Wang et al. [67] considered the potential of motion attribute estimation for identification tasks. Despite sharing a common focus on human gait, no framework has yet integrated gait recognition and anomaly detection across diverse tasks. Our FoundationGait addresses this gap, enabling a scalable and unified solution that leverages the power of large-scale pretraining.

Domain-Specific Foundation Models. Beyond general-purpose LVMs [4, 32, 49, 82], domain-specific LVMs offer a more practical solution, as they are better suited to understand field data, particularly when it differs significantly from everyday internet frames, such as satellite or industrial images [17, 19]. In related research, Khirodkar et al. [30] introduced a comprehensive foundation model tailored for human-centric vision tasks, including 2D pose estimation, part segmentation, depth estimation, and normal prediction. Similarly, Kim et al. [31] developed a foundation model focused on human face [46, 98] and body recognition [61]. In this paper, FoundationGait takes a step further by focusing on human gait for both recognition and healthcare applications.

Non-Vision-based Gait Foundation Models. Beyond vision-based methods, physical signal–driven gait foundation models (e.g., ground reaction forces and joint kinematics) have recently attracted growing interest. GaitFM [83] pretrains on physics-based synthetic gaits for data-efficient adaptation to dementia-related tasks. GaitDynamics [63] employs a diffusion transformer for flexible kinematic modeling and experiment-free prediction of gait modification outcomes. PathoFM [5] focuses on pathological gait, learning transferable representations from spinal cord injury data through self-supervised objectives. In contrast, FoundationGait unifies video-based gait tasks, namely, human recognition and healthcare, within a single scalable framework. Unless otherwise specified, the remainder of this paper focuses exclusively on the vision-based paradigm.

3 Method

Why has the gait research community not yet benefited from empirical scaling trends? Sec. 3.1 investigates this issue and provides a detailed analysis, while Sec. 3.2 introduces our proposed solution. For clarity, experimental settings are presented separately in Sec. 4.

3.1 Unscalable Issue in Gait Models

Scalable size is often considered a defining characteristic of foundation models. However, existing representative gait models remain relatively small, typically containing fewer than 5M parameters [2, 11] in their backbone. To explore the potential of scaling in this domain, we first enlarge the model size.

Scaling the Supervised DeepGaitV2. As shown in Fig. 2 (left, purple curve), when scaling up DeepGaitV2 [8, 9] on the Gait3D dataset [94], one of the largest and most challenging gait recognition benchmarks (6K+ seq.), the rank-1 accuracy quickly saturates and even degrades as model size increases. In contrast, the part-aware strategy proposed by FoundationGait (in Sec. 3.2) effectively alleviates this issue, exhibiting considerable performance gains as the model scales up (Fig. 2, left, red curve).

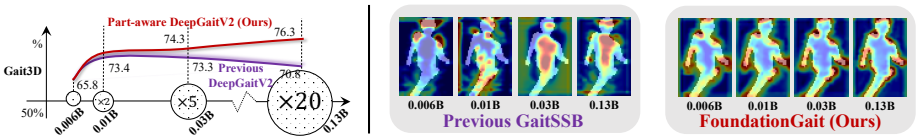

Fig. 2: Left: Unscalable Issue in Supervised DeepGaitV2 [9]. Our part-aware strategy effectively alleviates this issue. **Right:** FoundationGait presents a scale-invariant ability to capture fine-grained full-body details (*e.g.,* hands, contours), whereas GaitSSB prefers brittle shortcuts (*i.e.,* head, chest). The second layer GradCAM activation maps for FoundationGait and GaitSSB.

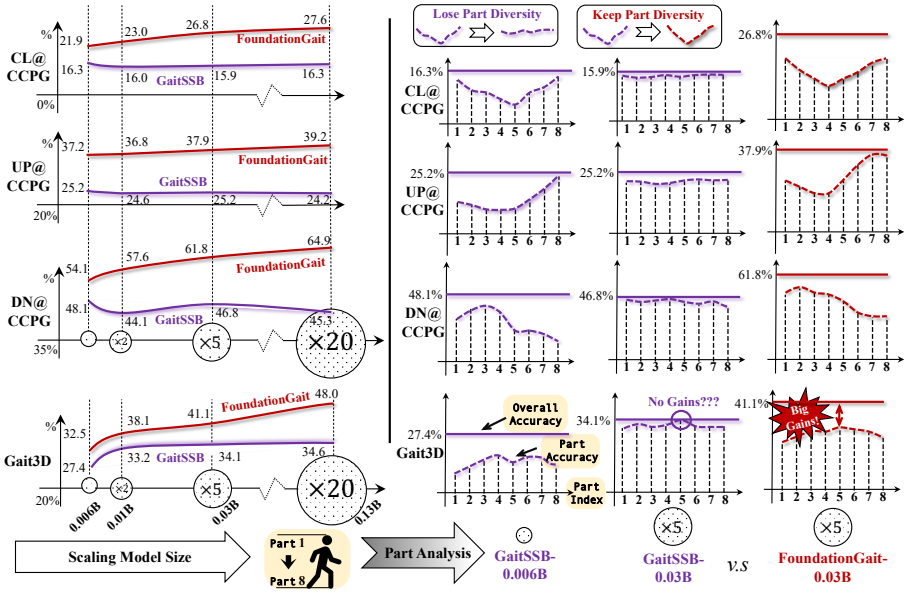

Fig. 3: Unscalable Issue in Self-supervised GaitSSB [10]. **Left:** GaitSSB scales unstably and saturates early, while FoundationGait scales steadily from 2× to 20×. **Right:** At 5×, GaitSSB loses part diversity (Overall ≈ Part Accuracy), whereas FoundationGait preserves diverse, complementary parts and achieves larger fusion gains. All models are pretrained on WebGait-2M.

Scaling the Self-supervised GaitSSB. We further scale up GaitSSB [10], a representative self-supervised gait pretraining framework, on the large-scale WebGait-2M dataset. Despite using extensive pretraining data (over 2M+ walking sequences), as shown in Fig. 3 (left, purple curves), rank-1 accuracy still saturates and even degrades on both CCPG [35], a challenging benchmark with 8K+ sequences spanning diverse clothing variations including full-clothing (CL@CCPG), upper-clothing (UP@CCPG), and lower-clothing (DN@CCPG) changes, and Gait3D [94], revealing a fundamental scalability bottleneck in existing approaches.

In contrast, our FoundationGait (red curves in Fig. 3, left) largely adheres to the expected scaling trends. We find that the devil lies in the details, specifically, in the body parts.

The gait community generally agrees that walking patterns require fine-grained descriptions. Therefore, most popular gait models [9] adopt a part-based pooling design, dividing the output feature map horizontally into several strips, each assumed to correspond to a specific body region. **However, even under this framework, we still observe that**, as shown in Fig. 3 (right), naively enlarging model size leads to diminished diversity among part features (GaitSSB-

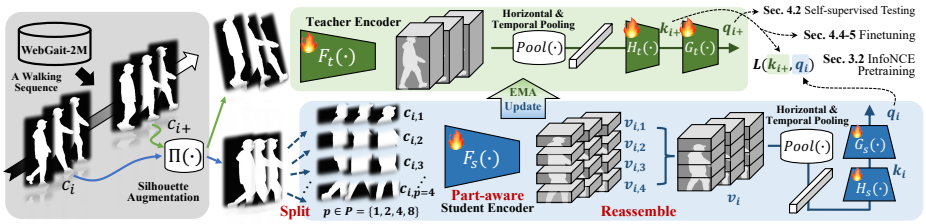

Fig. 4: Overview of FoundationGait. Two non-overlapping clips are sampled from a sequence and augmented as two views. The teacher encodes one clip to produce projected feature k_{i+} and predicted feature q_{i+}. The student applies a *Part-aware Strategy* to the other clip: each clip is horizontally split into p parts, encoded independently, and reassembled into a global representation v_i that inherits part-level diversity. An InfoNCE loss is computed between student predictions q_i and teacher projections k_{i+}, with the teacher updated via EMA.

0.006B vs. GaitSSB-0.03B, purple curves), resulting in limited fusion gains compared with FoundationGait-0.03B (red curves).

We argue that applying part pooling on an already-computed feature map is fundamentally late for large gait models. As networks deepen, part features tend to entangle and become homogeneous, negligible in shallow models but severe at scale. The issue has been ignored for a long time, and cannot be solved by existing part pooling. We therefore propose a part-aware encoding framework throughout the backbone to prevent cross-part interference and preserve the diversity necessary for effective scaling, forming the central motivation of FoundationGait. As further evidenced by Fig. 2 (right), FoundationGait's part-aware design encourages attention to fine-grained details across the entire body, preventing collapse to a few dominant shortcuts regardless of model scale.

3.2 FoundationGait

As shown in Fig. 4, two non-overlap clips (c_i and c_{i+}) are sampled from a sequence as two augmented views of identical walking patterns, where $i \in \{1, ..., N\}$. Each batch contains $2N$ clips forming N positive pairs. All clips are randomly augmented by $\Pi(\cdot)$. Following prior self-supervised works [10, 15], one clip c_{i+} is encoded by the teacher network to produce a projected feature k_{i+} for downstream tasks and a predicted feature q_{i+} for self-supervised test:

$$k_{i+} = H_t(Pool(F_t(\Pi(c_{i+})))), \quad q_{i+} = G_t(k_{i+}), \tag{1}$$

where F_t, H_t, and G_t are the teacher's encoder, projector, and predictor, and $Pool(\cdot)$ is spatiotemporal aggregation [9]. More in **Supplementary Materials**.

To encourage fine-grained part-level description, we introduce a part-aware training strategy in the student. Inspired by GaitPart [13], the other clip c_i is first horizontally divided into p parts (with $p = 4$ in Fig. 4 for simplicity), producing $\{c_{i,j}\}_{j=1}^{p}$. Then, each part is independently fed into the student encoder F_s to obtain part feature maps $\{v_{i,j}\}_{j=1}^{p}$. These detail-enriched tensors are then

Algorithm 1 Pseudocode of FoundationGait Framework.

```
# F_t,H_t,G_t: teacher encoder, projector, predictor
# F_s,H_s,G_s: student encoder, projector, predictor
# P=[1,2,4,8]: part-aware hyperparameter;
for seq in loader:
    c_t, c_s = sample_two_non_overlapping_clips(seq)
    c_t, c_s = aug(c_t), aug(c_s) # two random views
    # ------------ Teacher Branch ------------
    with torch.no_grad():
        v_t = F_t(c_t) # (N,C,S,H,W)
        k_t = H_t(Pool(v_t)) # (N,C1,C2)
        q_t = G_t(k_t) # (N,C1,C2), Self-supervised feature
    # ------------ Student Branch ------------
    feats = []
    subs = split_batch(c_s, len(P))
    # process each sub-batch with p ∈ P
    for clip, p in zip(subs, P):
        parts = Split(clip, p) # (n,C,S,H,W) -> (n*p,C,S,H//p,W)
        f_part = F_s(parts)
        f_clip = Reassemble(f_part) # (n*p,C,S,H//p,W) -> (n,C,S,H,W)
        feats.append(f_clip)
    v_s = Concat(feats) # len(P)*(n,...) -> (N,...)
    q_s = G_s(H_s(Pool(v_s))) # (N,C1,C2)
    # ------------ Optimization ------------
    loss = InfoNCE(q_s, k_t)
    loss.backward(); optimizer.step()
    T = (F_t, H_t, G_t); S = (F_s, H_s, G_s)
    MomentumUpdate(T, S) # update teacher by EMA
```

reassembled back into a unified global representation v_i, that is,

$$\begin{aligned} \{v_{i,j}\}_{j=1}^{p} &= F_s(\text{Split}(\Pi(c_i))), \\ v_i &= \text{Concat}(\{v_{i,j}\}_{j=1}^{p}). \end{aligned} \tag{2}$$

When p=1, no split is applied, while larger p yields finer local granularity. To balance local diversity and global consistency, a hyperparameter P (e.g., $p \in P = \{1, 2, 4, 8\}$) is used during training, where N clips are divided into len(P) subsets, each with its own p. A shift-window operation, similar with SwinViT [40], is also applied to half of the samples in Split($\cdot$) to enhance local variation, only performed along the height (H) dimension following common practice in gait. In brief, the global feature map v_i is constructed from multiple distinct local ones, **naturally inheriting their part-level diversity**.

Finally, we compute the InfoNCE loss [24] between the student's predictor features $q_i = G_s(H_s(\text{Pool}(v_i)))$ and teacher's projected output k_{i+}, formulated:

$$L(k_{i+}, q_i) = -\log \frac{\exp(\text{sim}(k_{i+}, q_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(k_{i+}, q_j)/\tau)}, \tag{3}$$

where sim($\cdot$) denotes cosine similarity, $\tau = 16$ is the temperature, and $j \neq i$ indexes negative sample pairs. Cross-layer loss follows a common self-supervised paradigm, with asymmetric predictors [4,15] preventing representation collapse. The pre-training scheme's details are presented in Alg. 1. For downstream tasks, the teacher model is fine-tuned with its predictor replaced by a new task head; the part-aware training method keeps active for maintaining local diversity. See **Supplementary Materials** for details.

Table 1: Overview of WebGait-2M. It is built from twelve public train sets, **treating each sequence as an individual ID in the self-supervised setting**. NM, BG, CL, UP, and DN denote normal, bag-, clothing-, ups-, and pants-changing.

Datasets	Type	Scene	Condition	Train Set		Test Set		Frames
				#IDs	#Seq	#IDs	#Seq	
CASIA-B [87]	Recognition	in-the-lab	NM, BG, CL	74	8,107	50	5,500	1,117,083
CCPG [35]	Recognition	in-the-lab	CL, UP, DN, BG	100	8,187	100	8,178	1,753,431
CASIA-E [57]	Recognition	in-the-lab	NM, BG, CL	200	152,697	814	626,055	95,581,249
OUMVLP [62]	Recognition	in-the-lab	NM	5,153	133,531	5,154	133,857	18,814,160
SUSTech1K [55]	Recognition	in-the-lab	diverse	250	6011	800	19,228	763,416
CCGR [100]	Recognition	in-the-lab	diverse	571	908,322	399	672,295	150,632,991
Gait3D [94]	Recognition	in-the-wild	diverse	3,000	18,940	1,000	6,369	3,279,239
GREW [99]	Recognition	in-the-wild	diverse	20,000	102,887	6,000	24,000	13,946,946
D-Gait [39]	Healthcare	in-the-lab	Depression	194	18,081	98	9,039	2,857,343
RA-GAR [67]	Healthcare	in-the-lab	Gait Attributes	250	57,155	283	65,912	21,304,199
Scoliosis1K [97]	Healthcare	in-the-lab	Scoliosis	745	745	748	748	447,900
GaitLU-1M [10]	Self-supervised	in-the-wild	diverse	943,884	943,884	-	-	87,235,933
WebGait-2M	Self-supervised	mixed	diverse	2,358,547	2,358,547	-	-	231,675,844

4 Experiment

4.1 Datasets

Pretraining Datasets. As shown in Tab. 1, we construct WebGait-2M, a large-scale gait database for pretraining, by integrating 12 public datasets spanning both human identification [10, 35, 55, 57, 62, 87, 94, 99, 100] and healthcare domains [39, 67, 97]. Consisting of 2.35M seq. and 0.23B frames, it spans an abundant range of scenes, subjects, motion patterns, clothing styles, viewpoints, and gait attributes, playing a key role in training gait foundation models. Our FoundationGait targets a universal gait representation generalizable to both recognition and healthcare. Thus, we deliberately avoid identity/disease labels, liberating it to learn task-agnostic, bias-free gait patterns.

Recognition Datasets. After WebGait-2M pretraining, we evaluate the self-supervised capability on six popular gait datasets, including four in-the-lab datasets (CASIA-B [87], OUMVLP [62], CCPG [35] and SUSTech1K [55]) and two in-the-wild datasets (Gait3D [94] and GREW [99]). In fine-tuning tasks, experiments are conducted on three representative and challenging datasets: CCPG [35], offering the richest clothing diversity (coats, trousers, backpacks, and full-body outfits); Gait3D [94], targeting real-world pedestrian recognition in complex supermarket scenes with frequent occlusions and dynamic motions; and CCGR-MINI [100], a compact subset of CCGR that preserves its testing difficulty and variations in viewpoints, speeds, and terrains.

Healthcare Datasets. In healthcare, three gait-based medical datasets are used: Scoliosis1K [97], a large-scale adolescent gait dataset for non-invasive scoliosis screening in real-world settings; D-Gait [39], designed for depression risk recognition, capturing diverse views and appearances under varying walking conditions; and RA-GAR [67], a richly annotated dataset with 15 gait attributes, *e.g.*, gender, age, forward head, humpback, and limping.

Table 2: (a) **FoundationGait Configuration.** Mapping denotes the projector (H) and predictor (G); FLOPs are computed on a 64×44 silhouette. (b) **Fine-tuning for Recognition.** Batch (p, q, j): p subjects, q sequences each, j frames per sequence. (c) **Fine-tuning for Healthcare.** (d) **Inference Time.** Averaged over 20 runs after 30 warm-ups on an RTX5090 (FP32) with $30 \times 64 \times 44$ input.

(a) FoundationGait Configuration.

Model	Layer Depth	Parameter Size			Flops
		Backbone	Mapping	Total	
0.006B	(1, 1, 1, 1)	5.9M	32.6M	38.5M	1.42G
0.01B	(1, 4, 4, 1)	11.1M	32.6M	43.7M	2.88G
0.03B	(1, 4, 8, 4)	33.1M	32.6M	65.7M	6.76G
0.13B	(1, 4, 32, 16)	132.3M	32.6M	164.9M	24.21G

(b) Fine-tuning Configuration for Recognition.

Dataset	Model	Batch Size	Milestones	Total	P
CCPG [35]	0.03B	(16, 32, 30)	(2K)	3K	[1, 2, 4, 8]
	0.13B	(8, 32, 30)	(2K)	3K	[1, 2, 4, 8]
Gait3D [94]	0.03B	(128, 4, 30)	(3K)	4K	[1, 2]
	0.13B	(128, 4, 30)	(3K, 4K)	5K	[1, 2]
CCGR-MINI [100]	0.03B	(32, 16, 30)	(3K, 5K)	6K	[1, 2, 4, 8]
	0.13B	(32, 16, 30)	(3K, 5K)	6K	[1, 2, 4, 8]

(c) Fine-tuning Configuration for Healthcare.

Mode	Dataset	Model	Milestones	Total	P
Linear Probe	Scoliosis1K [97]	0.03B	(1K, 2K)	3K	[1]
		0.13B	(1K, 2K)	3K	[1]
	D-Gait [39]	0.03B	(1K, 2K)	3K	[1]
		0.13B	(3K, 4K)	5K	[1]
	RA-GAR [67]	0.03B	(1K)	2K	[1]
		0.13B	(1K)	2K	[1]
Fine-tuning	Scoliosis1K [97]	0.03B	(1K)	2K	[1, 2, 4, 8]
		0.13B	(1K, 2K)	3K	[1, 2, 4, 8]
	D-Gait [39]	0.03B	(1K)	2K	[1, 2, 4, 8]
		0.13B	(1K, 2K)	3K	[1, 2, 4, 8]
	RA-GAR [67]	0.03B	(1K)	2K	[1, 2, 4, 8]
		0.13B	(1K)	2K	[1, 2, 4, 8]

(d) Inference Time Comparison (ms).

Models	0.006B	0.01B	0.03B	0.13B
GaitSSB	33.36	52.81	76.65	268.27
FoundationGait	29.83	49.78	82.92	273.72

Table 3: Self-supervised Performance Comparison. FoundationGait is evaluated against self-supervised [10] method on six public benchmarks.

Source Dataset	Method	Target Dataset											
		CASIA-B [87]			OU-MVLP [62]	Gait3D [94]		GREW [99]	CCPG [35]				SUSTech1K [55]
		NM	BG	CL	Rank-1	Rank-1	Rank-5	Rank-1	CL	UP	DN	BG	Rank-1
GaitLU-1M	GaitSSB [10]	83.8	**75.7**	**28.7**	**37.2**	24.8	38.0	16.6	11.3	20.0	**32.5**	50.6	50.3
	GaitSSB-0.13B	76.1	64.9	23.4	26.3	25.1	39.9	**18.7**	11.1	20.0	31.0	49.4	41.3
	FoundationGait-0.13B	**85.0**	71.2	25.6	35.5	**26.6**	**40.1**	18.3	**12.3**	**23.0**	32.1	**53.4**	**52.5**
WebGait-2M	GaitSSB [10]	88.9	74.8	33.5	44.4	27.4	43.3	21.5	16.3	25.2	48.1	57.6	38.5
	GaitSSB-0.03B	84.2	64.8	26.0	39.2	33.2	51.2	24.7	16.0	24.6	44.1	64.9	41.1
	GaitSSB-0.13B	82.8	64.6	25.9	38.9	34.6	53.2	23.0	16.3	24.2	45.3	65.9	36.9
	FoundationGait-0.03B	92.0	72.9	39.4	57.0	41.1	59.0	29.1	26.8	37.9	61.8	77.1	48.1
	FoundationGait-0.13B	**94.0**	**75.4**	**40.2**	**64.5**	**48.0**	**66.5**	**29.4**	**27.6**	**39.2**	**64.9**	**80.8**	**49.0**

4.2 Implementation Details

Protocols. Silhouettes are size-normalized following [11], resized to 64×44, and augmented following [10]. All experiments strictly follow the official protocols.
Network. FoundationGait uses the popular DeepGaitV2's backbone [9] as its encoder, scaled up by stacking layers. See Tab. 2 (a) and (d) for exact model size, computation costs and inference time.
Pretraining. SGD is used with an initial learning rate of 0.05, momentum 0.9, and weight decay 5×10^{-4}. The 0.13B model is trained for 80K iterations with a cosine annealing scheduler ($T_{max} = 80$K, $\eta_{min} = 1 \times 10^{-5}$), while the others are trained for 40K iterations using MultiStepLR (20K, 30K; $\gamma = 0.1$). The EMA momentum rises from 0.99 to 1.0 via a cosine schedule. Batch size is 512 with 16 frames per clip. In the part-aware student, $P = [1, 2, 4, 8]$.
Sampling. To balance data diversity, each batch samples from the 12 datasets with softmax-normalized probabilities proportional to $\log(\text{size})/3.0$, while *GaitLU-1M* is downweighted by $10\times$ due to its limited viewpoint diversity [42]. The 0.13B

Table 4: Gait Recognition Performance. Fine-tuning FoundationGait on three challenging public gait datasets: CCPG [35], Gait3D [94] and CCGR-MINI [100]. FoundationGait and GaitSSB are pretrained on WebGait-2M.

(a) Fine-tuning on CCPG [35].

Input	Method	Venue	CL	UP	DN	BG	Mean
RGB	AP3D [16]	ECCV'20	53.4	57.3	69.7	91.4	67.8
	PSTA [72]	ICCV'21	42.2	52.2	60.3	84.5	59.8
	PiT [89]	TII'22	41.0	47.6	64.3	91.0	61.0
	BigGait [85]	CVPR'24	82.6	85.9	87.1	93.1	87.2
	BiggerGait* [84]	NeurIPS'25	89.0	91.9	94.0	97.2	93.0
Ske.	GaitGraph2 [64]	CVPRW'22	5.0	5.3	5.8	6.2	5.6
	Gait-TR [90]	ES'23	15.7	18.3	18.5	17.5	17.5
	MSGG [52]	MTA'23	29.0	34.5	37.1	33.3	33.5
	SkeletonGait [12]	AAAI'24	40.4	48.5	53.0	61.7	50.9
Sil.	GaitSet [2]	AAAI'19	60.2	65.2	65.1	68.5	64.8
	GaitPart [13]	CVPR'20	64.3	67.8	68.6	71.7	68.1
	OGBase [35]	CVPR'23	52.1	57.3	60.1	63.3	58.2
	GaitBase [11]	CVPR'23	71.6	75.0	76.8	78.6	75.5
	DeepGaitV2 [9]	TPAMI'25	78.6	84.8	80.7	89.2	83.3
	GaitSSB-0.13B [10]	TPAMI'23	50.8	56.3	55.2	61.5	56.0
	FoundationGait-0.03B	ours	80.0	86.0	83.7	90.8	85.1
	FoundationGait-0.13B	ours	**84.0**	**87.5**	**85.5**	**91.7**	**87.2**

(b) Fine-tuning on CCGR-MINI [100].

Method	Venue	Rank-1	mAP	mINP
GaitSet [2]	AAAI'19	13.77	15.39	5.75
GaitPart [13]	CVPR'20	8.02	10.12	3.52
GaitGL [37]	ICCV'21	17.51	18.12	6.85
GaitBase [11]	CVPR'23	26.99	24.89	9.72
DeepGaitV2 [9]	TPAMI'25	39.37	36.01	16.77
GaitSSB-0.13B [10]	TPAMI'23	26.06	26.69	15.66
FoundationGait-0.03B	ours	45.66	40.97	26.36
FoundationGait-0.13B	ours	**53.25**	**48.00**	**33.29**

(c) Fine-tuning on Gait3D [94].

Method	Venue	Rank-1	Rank-5	mAP
GaitSet [2]	AAAI'19	36.7	58.3	30.0
GaitPart [13]	CVPR'20	28.2	47.6	47.6
GaitGL [37]	ICCV'21	29.7	48.5	22.3
GaitContour [20]	WACV'25	25.3	41.3	-
SMPLGait [94]	CVPR'22	46.3	64.5	37.2
MTSGait [93]	MM'22	48.7	67.1	37.6
QAGait [74]	AAAI'24	67.0	81.5	56.5
GaitBase [11]	CVPR'23	64.6	-	-
DANet [42]	CVPR'23	48.0	69.7	-
GaitGCI [7]	CVPR'23	50.3	68.5	39.5
DyGait [71]	ICCV'23	66.3	80.8	56.4
HSTL [70]	ICCV'23	61.3	76.3	55.5
VPNet [41]	CVPR'24	75.4	87.1	-
DeepGaitV2 [9]	TPAMI'25	74.4	88.0	65.8
CLTD [80]	ECCV'24	69.7	85.2	-
GaitMoE [26]	ECCV'24	73.7	-	66.2
Free Lunch [69]	ECCV'24	70.1	-	61.9
VM-Gait [76]	WACV'25	75.4	87.5	66.4
Mesh-Gait [75]	ArXiv'25	75.0	88.0	66.1
GaitSSB-0.13B [10]	TPAMI'23	53.7	75.5	48.7
FoundationGait-0.03B	ours	77.7	**89.6**	71.1
FoundationGait-0.13B	ours	**79.3**	89.5	**74.0**

model takes roughly 178 hours on 16× 32GB V100 GPUs. Fine-tuning 0.13B on CCPG takes 9.5 h and on Scoliosis1K 2.2 h, on 8×24GB RTX A5000.

4.3 Self-supervised Performance

For brevity, only the two largest FoundationGait versions (0.03B and 0.13B) are compared against representative self-supervised [10] methods across six widely used benchmarks in Tab. 3. Note that self-supervised gait pretraining remains in its infancy, with GaitSSB [10] being the only reproducible baseline.

First, when pre-training on GaitLU-1M [10], scaling GaitSSB [10] from 0.006B to 0.13B yields significant performance drops across most datasets — −7.7% on CASIA-B (NM), −10.8% (BG), −10.9% on OU-MVLP, and −9.0% on SUSTech1K. By contrast, FoundationGait-0.13B remains competitive with the small GaitSSB, owing to its stronger retention of fine-grained details. However, a key observation is that neither scaled-up model benefits meaningfully from GaitLU-1M beyond the small baseline. We attribute this to the dataset's limited diversity: sourced from handheld internet videos, it offers constrained viewpoint variation and is prone to causing overfitting in larger models [42]. Therefore, we turn to the larger and more diverse WebGait-2M for scaling experiments.

Second, when pre-training on WebGait-2M, the scaling trend begins to pay off. Compared with the SoTA self-supervised method GaitSSB [10], our 0.13B model achieves notable gains: +20.6% on the challenging in-the-wild Gait3D [94], +20.1% on the largest indoor OU-MVLP [62], and +11.3% on the CCPG

(CL) [35]. However, even with richer pretraining data, scaling up GaitSSB to 0.13B yields limited gains and even regressions (*e.g.*, –5.5% on OU-MVLP). This demonstrates that simply enlarging model size and pretraining data is insufficient — **method-level innovations, as in FoundationGait, are essential**.

4.4 Fine-tuning on Gait Recognition

Implementation. Following GaitSSB [10], we adopt layer-wise fine-tuning. Only the last two backbone blocks are fine-tuned (lr = 0.001, 0.005), while the projector and head use 0.01 and 0.1. SGD (lr = 0.1, momentum = 0.9) with a MultiStepLR scheduler is used. More details are in Tab. 2 (b).
CCPG. In Tab. 4 (a), our 0.13B model delivers notable improvements in this challenging clothing-changing scene: +5.4% in full-changing (CL), +2.7% in ups-changing (UP), +4.8% in pants-changing (DN), and +3.9% on average. Despite relying solely on informative-sparse silhouettes as input, our FoundationGait-0.13B achieves performance comparable to BigGait@CVPR'24—which leverages RGB inputs and large vision model features (84.0% *vs.* 82.6% on CL; 87.2% *vs.* 87.2% on Mean). Though a gap remains with the SoTA RGB foundation model, BiggerGait [84], we consider that advances in Silhouette-based methods often inspire RGB ones (*e.g.*, BiggerGait builds upon GaitBase).
CCGR-MINI. In Tab. 4 (b), even in this complex covariate setting, existing gait methods have long struggled (Rank-1 $< 40\%$), but FoundationGait offers a significant breakthrough, *i.e.*, 45.66% / 53.25% on Rank-1 for our 0.03B / 0.13B.
Gait3D. In Tab. 4 (c), FoundationGait establishes a new SoTA on this challenging real world scene. Our 0.13B model achieves 79.3% Rank-1 and 74.0% mAP, significantly outperforming the previous best methods, VPNet [41] (+3.9% in Rank-1) and GaitMoE [26] (+7.8% in mAP).

4.5 Fine-tuning on Healthcare Tasks

Implementation. Two settings (linear probing and fine-tuning) are evaluated. For linear probing, only the classification head is trained (lr = 0.1). For fine-tuning, the model is initialized from the linear probe stage and trained with layer-wise learning rates. The 0.03B model uses learning rates of 1×10^{-3}, 1×10^{-2}, and 1×10^{-1} for the backbone, projector, and head, respectively. Other settings follow Sec. 4.4. Batch size is (8, 8, 30), following the setting in Tab. 2 (b). More details are in Tab. 2 (c).
Linear Probe. As demonstrated in Tables 5 (a-c), the features extracted by FoundationGait exhibit strong linear separability across a variety of gait healthcare tasks. Specifically, for gait attribute estimation [67] (+2.2% in F1 compared to CLIP-GAR [67]), Foundation-0.13B performs on par with, or even surpasses, previous SoTA methods that require full training. These results underscore the model's superior ability to capture abundant and informative gait characteristics that are **already highly linearly separable in its feature space**, contributing to its strong generalizability across vision-based gait tasks.

Table 5: Healthcare Performance. Linear-probe and fine-tuning results of FoundationGait on scoliosis screening [97], depression prediction [39], and gait attribute [67]. FoundationGait and GaitSSB are pretrained on WebGait-2M.

(a) Evaluation on Scoliosis1K [97].

Mode	Method	Acc.	Precision	Recall	F1
Full-training	ScoNet [97]	93.3	83.4	92.5	86.4
	ScoNet-MT [97]	95.2	80.2	96.6	84.9
Linear Probe	GaitSSB-0.13B	72.5	56.7	69.8	58.9
	FoundationGait-0.03B	71.1	59.7	75.3	56.7
	FoundationGait-0.13B	73.5	61.0	68.5	56.9
Fine-tuning	GaitSSB-0.13B	93.7	91.0	74.6	79.5
	FoundationGait-0.03B	**97.3**	**95.4**	88.9	**91.7**
	FoundationGait-0.13B	96.0	80.1	**97.5**	85.1

(b) Evaluation on D-Gait [39].

Mode	Method	Acc.	Precision	Recall	F1
Full-training	GaitBase [11]	-	51.7	60.4	55.7
	GaitSet [2]	-	53.8	65.7	59.2
	GaitGL [37]	-	57.5	36.8	44.9
Linear Probe	GaitSSB-0.13B	65.0	56.9	57.3	57.1
	FoundationGait-0.03B	66.3	57.7	57.8	57.8
	FoundationGait-0.13B	66.0	57.7	58.0	57.8
Fine-tuning	GaitSSB-0.13B	72.9	64.9	62.9	63.6
	FoundationGait-0.03B	**75.1**	**67.9**	65.1	66.1
	FoundationGait-0.13B	72.2	66.0	**67.4**	**66.5**

(c) Evaluation on RA-GAR [67].

Mode	Method	Instance-based				Attr.-based
		Acc.	Prec.	Recall	F1	mA
Full-training	GaitSet [2]	84.4	75.7	66.1	70.2	64.3
	GaitGL [37]	85.0	78.0	64.9	70.4	62.6
	GaitTR [90]	83.6	73.6	65.0	68.7	62.3
	GPGait [14]	84.3	75.0	66.3	70.0	63.9
	GAR-Net [58]	84.3	77.4	74.7	73.2	62.2
	DeepGaitV2 [9]	85.0	76.6	66.7	70.9	61.7
	CLIP-GAR [67]	84.3	77.4	74.7	76.0	65.6
Linear Probe	GaitSSB-0.13B	84.0	77.1	74.2	75.5	61.8
	FoundationGait-0.03B	85.2	78.6	76.4	77.4	64.3
	FoundationGait-0.13B	**85.7**	**79.3**	**77.3**	**78.2**	65.6
Fine-tuning	GaitSSB-0.13B	84.7	78.7	74.4	76.4	62.9
	FoundationGait-0.03B	84.8	78.0	75.9	76.9	65.1
	FoundationGait-0.13B	84.7	78.1	75.5	76.7	**65.7**

Fine-tuning. After fine-tuning, FoundationGait demonstrates significant performance gains, with +6.8% in F1 on the Scoliosis1K [97] and +7.3% in F1 on the D-Gait [39] datasets. However, FoundationGait-0.13B tends to be overly optimistic in predicting positive cases, leading to high recall but lower accuracy compared to the 0.03B version on these class-imbalanced benchmarks. This suggests that the model size may influence the tendency to classify more samples as positive in long-tailed distributions. For the gait attribute estimation task (RA-GAR [67]), where labels are already well separable through linear probing of the pretrained FoundationGait, further fine-tuning yields limited gains.

In summary, FoundationGait demonstrates remarkable transferability across multiple gait healthcare tasks, achieving strong performance with simple linear probing or fine-tuning without task-specific designs. We believe that more refined strategies, such as LoRA [6] and Adapter [3], could further unlock its potential for healthcare gait analysis.

4.6 Ablation Study

We systematically evaluate: (a) the effectiveness of part-aware student and its hyperparameter P, (b) part-aware extensions in supervised models, (c) cross-domain generalization, and (d) cross-modality generalization of FoundationGait. See **Supplementary Materials** for experiment details.

Part-aware Effectiveness. Compared to cropping [91] or masking silhouettes (Tab. 6 (a)), our part-aware student cleverly preserves all details, achieving stronger self-supervised results. Expanding P from [1] to $[1, 2, 4, 8]$ provides clear gains on indoor CASIA-B [87] and OU-MVLP [62].

Extension in Supervised Models. Beyond the self-supervised setting, can part-aware training be applied to large supervised models? Table 6 (b) demonstrates that our part-aware training significantly benefits DeepGaitV2 with large

Table 6: Ablation Study. (a) Exploring the effectiveness of part-aware training and its P , using FoundationGait-0.13B for pretraining. (b) Extending part-aware training to DeepGaitV2, where w/ denotes the use of part-aware training. (c) Evaluating FoundationGait on the unseen domain datasets [34]. (d) Fine-tuning FoundationGait on Gait3D with taking the unseen human parsing as input.

(a) Ablation of part-aware training and P.

Method	P	CASIA-B [87]			Gait3D [94]	
		NM	BG	CL	Rank-1	mAP
Crop	-	70.8	59.3	27.9	17.2	12.3
Mask	-	87.6	68.1	36.2	38.1	29.4
Part-aware	[1]	86.6	67.0	28.7	39.6	30.2
	[1, 4]	92.2	73.8	37.7	45.3	60.3
	[1, 2, 4, 8]	**94.0**	**75.4**	**40.2**	**48.0**	**66.5**

(b) Extending to supervised methods.

Model	w/	CASIA-B [87]			Gait3D [94]	
		NM	BG	CL	Rank-1	mAP
DeepGaitV2 -0.03B	×	84.5	72.3	46.5	73.3	65.6
	✓	**92.4**	**87.0**	**66.2**	**74.3**	**67.3**
DeepGaitV2 -0.13B	×	79.2	68.0	42.2	70.8	64.0
	✓	**90.0**	**81.4**	**62.4**	**76.3**	**68.7**

(c) Evaluate on the unseen dataset [34].

DroneGait [34]	Mid-pitch views			High-pitch views		
	NM	BG	CL	NM	BG	CL
GaitSSB-0.03B	91.1	85.5	51.4	34.8	25.9	17.6
GaitSSB-0.13B	89.0	82.5	50.6	36.0	27.1	17.8
FoundationGait-0.03B	95.3	90.5	57.8	36.8	26.9	17.7
FoundationGait-0.13B	**96.9**	**93.2**	**63.4**	**50.4**	**37.5**	**24.8**

(d) Extending to the unseen modality.

Modality	Method	Venue	Rank-1	Rank-5	mAP
Sil.	VPNet [41]	CVPR'24	75.4	87.1	-
	DeepGaitV2 [11]	TPAMI'25	74.4	88.0	65.8
	GaitMoE [26]	ECCV'24	73.7	-	66.2
Parsing	FoundationGait-0.03B	ours	82.0	91.6	76.1
	FoundationGait-0.13B	ours	83.8	92.7	79.3
Sil. +Parsing	XGait [96]	MM'24	80.5	91.9	73.3
	MultiGait++ [28]	AAAI'25	85.4	**94.9**	80.5
	FoundationGait-0.03B	ours	84.2	92.8	78.8
	FoundationGait-0.13B	ours	**86.5**	93.8	**82.6**

model sizes. For example, the 0.13B version shows improvements of +5.5% on Gait3D and up to +20.2% on CASIA-B. This underscores the importance of preserving part diversity in large gait models, both in supervised and self-supervised settings, and our part-aware training offers a flexible and promising solution.

Cross-domain Generalization. Results on DroneGait further validate FoundationGait's strong generalization to unseen UAV-view domains (Table 6 (c)), consistently outperforming GaitSSB across both mid- and high-pitch views.

Cross-modality Generalization. Although FoundationGait has never been pretrained on the gait modality of human parsing, we found that it performs well in both parsing-only and multi-modal settings after fine-tuning (Tab. 6 (d)). This highlights FoundationGait's ability to transform raw pixel-based shapes, across diverse data types, into fine-grained semantic features.

Diversity vs. Scale. To separate the effects of diversity and scale, we subsample WebGait-2M to 1M (1/2.5 per dataset), matching GaitLU-1M. At equal scale (Tab. 7 (a)), both methods favor WebGait-1M: GaitSSB-0.13B by 6.6% and FoundationGait-0.13B by 19.5% on average. This confirms that diversity, not scale alone, is critical for large-scale gait learning.

Head Initialization. By default we initialize the head via linear probing before fine-tuning, akin to two-stage tuning in VLMs. As shown in Tab. 7 (b), this slightly outperforms fine-tuning from scratch (+0.3% F1), though both remain competitive, indicating robustness to head initialization.

5 Challenges and Limitations

We introduce FoundationGait, a scalable self-supervised gait foundation model that addresses key challenges in size scalability and cross-task generalization.

Table 7: Ablation Study. (a) Effect of pretraining data diversity at matched 1M scale. (b) Effect of head initialization before fine-tuning.

(a) Effect of pretraining data diversity at matched 1M scale.

Source Dataset	Method	OU-MVLP	Gait3D		CCPG			
		Rank-1	Rank-1	Rank-5	CL	UP	DN	BG
GaitLU-1M	GaitSSB-0.13B	26.3	25.1	39.9	11.1	20.0	31.0	49.4
	FoundationGait-0.13B	35.5	26.6	40.1	12.3	23.0	32.1	53.4
WebGait-1M	GaitSSB-0.13B	34.5	30.2	48.4	14.7	22.5	41.1	57.9
	FoundationGait-0.13B	**58.6**	**42.5**	**60.1**	**24.3**	**35.3**	**60.8**	**77.9**

(b) Effect of head initialization before fine-tuning.

Mode	Head Initialization	Method	Acc.	Precision	Recall	F1
Fine-tuning	Scratch	FoundationGait-0.13B	94.4	**83.0**	94.5	84.8
	Linear-probing	FoundationGait-0.13B	**96.0**	80.1	**97.5**	**85.1**

FoundationGait shows that a unified model can effectively tackle both recognition and healthcare tasks, paving the way for more practical gait applications.

In the end, three issues warrant further investigation:

ViT-based Gait Foundation Model. This issue is crucial for bridging gait-specific foundation models with those from other domains, particularly LLMs [66], yet it remains largely unexplored within the gait community.

FoundationGait-1B. Due to computational constraints, we only scaled the model up by 20× to 0.13B and verified that its performance still follows the scaling trends in Fig. 3. Can we scale it beyond 1B? This is an ongoing goal.

Toward RGB. While silhouette-based gait methods remain dominant in top venues (62.8% *vs.* 5.1% for RGB [9]), the performance and generalization advantages of RGB-based approaches are becoming increasingly evident. We hope this work serves as a principled foundation and source of inspiration for the design of RGB-based gait foundation models in the future.

References

1. AI, L.: Domain-specific large vision models (lvms). https://landing.ai/domain-specific-lvms (2024), accessed: 2025-11-07
2. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI. pp. 8126–8133 (2019)
3. Chen, H., Tao, R., Zhang, H., Wang, Y., Li, X., Ye, W., Wang, J., Hu, G., Savvides, M.: Conv-adapter: Exploring parameter efficient transfer learning for convnets. In: CVPR. pp. 1551–1561 (2024)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
5. Dey, S., Paez-Granados, D.: Pathofm: Toward a foundation model for pathological gait. In: NeurIPS 2025 Workshop on Learning from Time Series for Health (2025)
6. Ding, C., Cao, X., Xie, J., Fan, L., Wang, S., Lu, Z.: Lora-c: Parameter-efficient fine-tuning of robust cnn for iot devices. arXiv preprint arXiv:2410.16954 (2024)
7. Dou, H., Zhang, P., Su, W., Yu, Y., Lin, Y., Li, X.: Gaitgci: Generative counterfactual intervention for gait recognition. In: CVPR. pp. 5578–5588 (2023)

8. Fan, C., Hou, S., Huang, Y., Yu, S.: Exploring deep models for practical gait recognition. arXiv preprint arXiv:2303.03301 (2023)
9. Fan, C., Hou, S., Liang, J., Shen, C., Ma, J., Jin, D., Huang, Y., Yu, S.: Opengait: A comprehensive benchmark study for gait recognition towards better practicality. IEEE TPAMI (2025)
10. Fan, C., Hou, S., Wang, J., Huang, Y., Yu, S.: Learning gait representation from massive unlabelled walking videos: A benchmark. IEEE TPAMI **45**(12), 14920–14937 (2023)
11. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: CVPR. pp. 9707–9716 (2023)
12. Fan, C., Ma, J., Jin, D., Shen, C., Yu, S.: Skeletongait: Gait recognition using skeleton maps. In: AAAI. pp. 1662–1669 (2024)
13. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: CVPR. pp. 14225–14233 (2020)
14. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. In: ICCV. pp. 19595–19604 (2023)
15. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inform. Process. Syst. **33**, 21271–21284 (2020)
16. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In: ECCV. pp. 228–243 (2020)
17. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: detecting industrial anomalies using large vision-language models. In: AAAI (2024)
18. Guo, H., Ji, Q.: Physics-augmented autoencoder for 3d skeleton-based gait recognition. In: ICCV. pp. 19627–19638 (2023)
19. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: CVPR. pp. 27672–27683 (2024)
20. Guo, Y., Shah, A., Liu, J., Gupta, A., Chellappa, R., Peng, C.: Gaitcontour: Efficient gait recognition based on a contour-pose representation. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 1051–1061. IEEE (2025)
21. Gupta, A., Chellappa, R.: You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 5893–5902 (2024)
22. Gupta, A., Chellappa, R.: Mimicgait: A model agnostic approach for occluded gait recognition using correlational knowledge distillation. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 4757–4766. IEEE (2025)
23. Gupta, A., Huang, S., Chellappa, R.: Mind the gap: Bridging occlusion in gait recognition via residual gap correction. In: IEEE Int. Joint Conf. Biom. (2025)
24. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304 (2010)
25. Huang, P., Hou, S., Huang, J., Huang, Y.: Learning a unified template for gait recognition. In: ICCV. pp. 12459–12469 (2025)
26. Huang, P., Peng, Y., Hou, S., Cao, C., Liu, X., He, Z., Huang, Y.: Occluded gait recognition with mixture of experts: an action detection perspective. In: ECCV. pp. 380–397. Springer (2024)

27. Huang, Z., Liu, X., Kong, Y.: H-more: Learning human-centric motion representation for action analysis. In: CVPR. pp. 22702–22713 (2025)
28. Jin, D., Fan, C., Chen, W., Yu, S.: Exploring more from multiple gait modalities for human identification. In: AAAI. pp. 4120–4128 (2025)
29. Jin, D., Fan, C., Ma, J., Zhou, J., Chen, W., Yu, S.: On denoising walking videos for gait recognition. In: CVPR. pp. 12347–12357 (2025)
30. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision. pp. 206–228. Springer (2024)
31. Kim, M., Ye, D., Su, Y., Liu, F., Liu, X.: Sapiensid: Foundation for human recognition. In: CVPR. pp. 13937–13947 (2025)
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
33. Konieczny, M.R., Senyurt, H., Krauspe, R.: Epidemiology of adolescent idiopathic scoliosis. Journal of Children's Orthopaedics **7**(1), 3–9 (2013). https://doi.org/10.1007/s11832-012-0457-4, https://doi.org/10.1007/s11832-012-0457-4, pMID: 24432052
34. Li, A., Hou, S., Cai, Q., Fu, Y., Huang, Y.: Gait recognition with drones: A benchmark. IEEE Transactions on Multimedia **26**, 3530–3540 (2023)
35. Li, W., Hou, S., Zhang, C., Cao, C., Liu, X., Huang, Y., Zhao, Y.: An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: CVPR. pp. 13824–13833 (2023)
36. Liang, J., Fan, C., Hou, S., Shen, C., Huang, Y., Yu, S.: Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In: ECCV. pp. 375–390. Springer (2022)
37. Lin, B., Zhang, S., Wang, M., Li, L., Yu, X.: Gaitgl: Learning discriminative global-local feature representations for gait recognition. arXiv preprint arXiv:2208.01380 (2022)
38. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: ICCV. pp. 14648–14656 (2021)
39. Liu, X., Li, Q., Hou, S., Ren, M., Hu, X., Huang, Y.: Depression risk recognition based on gait: A benchmark. Neurocomputing **596**, 128045 (2024)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
41. Ma, K., Fu, Y., Cao, C., Hou, S., Huang, Y., Zheng, D.: Learning visual prompt for gait recognition. In: CVPR. pp. 593–603 (2024)
42. Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., Huang, Y.: Dynamic aggregated network for gait recognition. In: CVPR. pp. 22076–22085 (2023)
43. Mei, K., Zhou, M., Patel, V.M.: Field-dit: Diffusion transformer on unified video, 3d, and game field generation. In: ICLR (2025)
44. Meng, S., Hou, S., Fu, Y., Hu, X., Huang, J., Huang, Y.: Seeing from magic mirror: Contrastive learning from reconstruction for pose-based gait recognition. In: ACM MM (2025)
45. of Mental Health, N.I.: Major depression. https://www.nimh.nih.gov/health/statistics/major-depression (2022), accessed: 2023-11-08
46. Narayan, K., VS, V., Chellappa, R., Patel, V.M.: Facexformer: A unified transformer for facial analysis. In: ICCV. pp. 11369–11382 (2025)
47. Nixon, M.S., Tan, T., Chellappa, R.: Human identification based on gait, vol. 4. Springer Science & Business Media (2010)

48. Niyogi, Adelson: Analyzing and recognizing walking figures in xyt. In: CVPR. pp. 469–474. IEEE (1994)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
50. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4172–4182 (2023). https://doi.org/10.1109/ICCV51070.2023.00387
51. Peng, G., Wang, Y., Zhao, Y., Zhang, S., Li, A.: Glgait: a global-local temporal receptive field network for gait recognition in the wild. In: ACM MM. pp. 826–835 (2024)
52. Peng, Y., Ma, K., Zhang, Y., He, Z.: Learning rich features for gait recognition by integrating skeletons and silhouettes. Multimedia Tools and Applications **83**(3), 7273–7294 (2024)
53. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
55. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3d gait recognition with point clouds. In: CVPR. pp. 1054–1063 (2023)
56. Shen, C., Wang, R., Duan, L., Yu, S.: Lidargait++: Learning local features and size awareness from lidar point clouds for 3d gait recognition. In: CVPR. pp. 6627–6636 (2025)
57. Song, C., Huang, Y., Wang, W., Wang, L.: Casia-e: A large comprehensive dataset for gait recognition. IEEE TPAMI **45**(3), 2801–2815 (2022)
58. Song, X., Hou, S., Huang, Y., Cao, C., Liu, X., Huang, Y., Shan, C.: Gait attribute recognition: A new benchmark for learning richer attributes from human gait patterns. IEEE Transactions on Information Forensics and Security **19**, 1–14 (2023)
59. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al.: Bioclip: A vision foundation model for the tree of life. In: CVPR. pp. 19412–19424 (2024)
60. Su, Y., Kim, M., Liu, F., Jain, A., Liu, X.: Open-set biometrics: Beyond good closed-set models. In: ECCV. pp. 243–261. Springer (2024)
61. Su, Y., Shi, Y., Liu, F., Liu, X.: Hamobe: Hierarchical and adaptive mixture of biometric experts for video-based person reid. In: ICCV. pp. 11525–11536 (2025)
62. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ transactions on Computer Vision and Applications **10**(1), 4 (2018)
63. Tan, T., Van Wouwe, T., Werling, K.F., Liu, C.K., Delp, S.L., Hicks, J.L., Chaudhari, A.S.: Gaitdynamics: A generative foundation model for analyzing human walking and running. Nature Biomedical Engineering pp. 1–13 (2026)
64. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: ICIP. pp. 2314–2318. IEEE (2021)
65. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1473–1488 (2008)

66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
67. Wang, C., Hou, S., Li, A., Cai, Q., Huang, Y.: Ra-gar: A richly annotated benchmark for gait attribute recognition. In: AAAI. pp. 7591–7599 (2025)
68. Wang, J., Hou, S., Huang, Y., Cao, C., Liu, X., Huang, Y., Wang, L.: Causal intervention for sparse-view gait recognition. In: ACM MM. pp. 77–85 (2023)
69. Wang, J., Hou, S., Huang, Y., Cao, C., Liu, X., Huang, Y., Zhang, T., Wang, L.: Free lunch for gait recognition: A novel relation descriptor. In: ECCV. pp. 39–56. Springer (2024)
70. Wang, L., Liu, B., Liang, F., Wang, B.: Hierarchical spatio-temporal representation learning for gait recognition. In: ICCV. pp. 19582–19592. IEEE (2023)
71. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: Dygait: Exploiting dynamic representations for high-performance gait recognition. In: ICCV. pp. 13424–13433 (2023)
72. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In: ICCV. pp. 12026–12035 (2021)
73. Wang, Z., Hou, S., Zhang, M., Liu, X., Cao, C., Huang, Y.: Gaitparsing: Human semantic parsing for gait recognition. IEEE TMM **26**, 4736–4748 (2023)
74. Wang, Z., Hou, S., Zhang, M., Liu, X., Cao, C., Huang, Y., Li, P., Xu, S.: Qagait: Revisit gait recognition from a quality perspective. In: AAAI. pp. 5785–5793 (2024)
75. Wang, Z.Y., Chen, J., Liu, J., Guo, Y., Chellappa, R.: Mesh-gait: A unified framework for gait recognition through multi-modal representation learning from 2d silhouettes. arXiv preprint arXiv:2510.10406 (2025)
76. Wang, Z.Y., Liu, J., Chen, J., Chellappa, R.: Vm-gait: Multi-modal 3d representation based on virtual marker for gait recognition. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 5326–5335. IEEE (2025)
77. Wang, Z.Y., Liu, J., Guo, Y., Chen, J., Chellappa, R.: Unigait: A unified transformer-based multitask framework for gait analysis in the wild. In: Int. Conf. Autom. Face Gesture Recognit. pp. 1–9. IEEE (2025)
78. Wang, Z.Y., Liu, J., Kathirvel, R.P., Lau, C.P., Chellappa, R.: Hypergait: A video-based multitask network for gait recognition and human attribute estimation at range and altitude. In: IEEE Int. Joint Conf. Biom. pp. 1–9. IEEE (2024)
79. Wang, Z.Y., Shao, Z., Chen, J., Chellappa, R.: Combo-gait: Unified transformer framework for multi-modal gait recognition and attribute analysis. arXiv preprint arXiv:2510.10417 (2025)
80. Xiong, H., Feng, B., Wang, X., Liu, W.: Causality-inspired discriminative feature learning in triple domains for gait recognition. In: ECCV. pp. 251–270. Springer (2024)
81. Xu, C., Makihara, Y., Li, X., Yagi, Y.: Occlusion-aware human mesh model-based gait recognition. IEEE transactions on information forensics and security **18**, 1309–1321 (2023)
82. Xu, J., Lo, S.Y., Safaei, B., Patel, V.M., Dwivedi, I.: Towards zero-shot anomaly detection and reasoning with multimodal large language models. In: CVPR. pp. 20370–20382 (2025)
83. Yamada, Y., Kobayashi, M., Nemoto, M., Ota, M., Nemoto, K., Arai, T.: Toward a foundation model for healthcare applications on single-camera gait analysis. In: 2025 IEEE International Conference on Digital Health (ICDH). pp. 1–10. IEEE (2025)

84. Ye, D., Fan, C., Huang, Z., Luo, C., Li, J., Yu, S., Liu, X.: Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. Adv. Neural Inform. Process. Syst. (2025)
85. Ye, D., Fan, C., Ma, J., Liu, X., Yu, S.: Biggait: Learning gait representation you want by large vision models. In: CVPR. pp. 200–210 (2024)
86. Ye, D., Ma, J., Fan, C., Yu, S.: Gaitediter: Attribute editing for gait representation learning. arXiv e-prints pp. arXiv–2303 (2023)
87. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: ICPR. vol. 4, pp. 441–444. IEEE (2006)
88. Yu, W., Yu, H., Huang, Y., Wang, L.: Generalized inter-class loss for gait recognition. In: ACM MM. pp. 141–150 (2022)
89. Zang, X., Li, G., Gao, W.: Multi-direction and Multi-scale Pyramid in Transformer for Video-based Pedestrian Retrieval. IEEE Transactions on Industrial Informatics **18**(12), 8776–8785 (2022)
90. Zhang, C., Chen, X.P., Han, G.Q., Liu, X.J.: Spatial transformer network on skeleton-based gait recognition. Expert Systems **40**(6), e13244 (2023)
91. Zhang, K., Ding, T., Jiang, J., Chen, T., Zharkov, I., Patel, V.M., Liang, L.: Procrop: Learning aesthetic image cropping from professional compositions. AAAI (2026)
92. Zhang, S., Wang, Y., Li, A.: Cross-view gait recognition with deep universal linear embeddings. In: CVPR. pp. 9095–9104 (2021)
93. Zheng, J., Liu, X., Gu, X., Sun, Y., Gan, C., Zhang, J., Liu, W., Yan, C.: Gait recognition in the wild with multi-hop temporal switch. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6136–6145 (2022)
94. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: CVPR. pp. 20228–20237 (2022)
95. Zheng, J., Liu, X., Wang, S., Wang, L., Yan, C., Liu, W.: Parsing is all you need for accurate gait recognition in the wild. In: ACM MM. pp. 116–124 (2023)
96. Zheng, J., Liu, X., Zhang, B., Yan, C., Zhang, J., Liu, W., Zhang, Y.: It takes two: Accurate gait recognition in the wild via cross-granularity alignment. In: ACM MM. pp. 8786–8794 (2024)
97. Zhou, Z., Liang, J., Peng, Z., Fan, C., An, F., Yu, S.: Gait patterns as biomarkers: A video-based approach for classifying scoliosis. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. pp. 284–294. Springer (2024)
98. Zhu, J., Su, Y., Kim, M., Jain, A., Liu, X.: A quality-guided mixture of score-fusion experts framework for human recognition. In: ICCV. pp. 13076–13086 (2025)
99. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: ICCV. pp. 14789–14799 (2021)
100. Zou, S., Fan, C., Xiong, J., Shen, C., Yu, S., Tang, J.: Cross-covariate gait recognition: A benchmark. In: AAAI. pp. 7855–7863 (2024)