BiggerGait: Unlocking Gait Recognition with Layer-wise Representations from Large Vision Models

Dingqiang Ye[1,3*], **Chao Fan**[2*], **Zhanbo Huang**[3],
Chengwen Luo[2], **Jianqiang Li**[2], **Shiqi Yu**[1†], **Xiaoming Liu**[3]
[1] Department of Computer Science and Engineering, Southern University of Science and Technology
[2] School of Artificial Intelligence, Shenzhen University
[3] Department of Computer Science and Engineering, Michigan State University
11810121@mail.sustech.edu.cn, chaofan996@szu.edu.cn, huang247@msu.edu,
{chengwen, lijq}@szu.edu.cn, yusq@sustech.edu.cn, liuxm@cse.msu.edu

Abstract

Large vision models (LVM) based gait recognition has achieved impressive performance. However, existing LVM-based approaches may overemphasize gait priors while neglecting the intrinsic value of LVM itself, particularly the rich, distinct representations across its multi-layers. To adequately unlock LVM's potential, this work investigates the impact of layer-wise representations on downstream recognition tasks. Our analysis reveals that LVM's intermediate layers offer complementary properties across tasks, integrating them yields an impressive improvement even without rich well-designed gait priors. Building on this insight, we propose a simple and universal baseline for LVM-based gait recognition, termed **BiggerGait**. Comprehensive evaluations on CCPG, CAISA-B*, SUSTech1K, and CCGR_MINI validate the superiority of BiggerGait across both within- and cross-domain tasks, establishing it as a simple yet practical baseline for gait representation learning. All the models and code are available at https://github.com/ShiqiYu/OpenGait/.

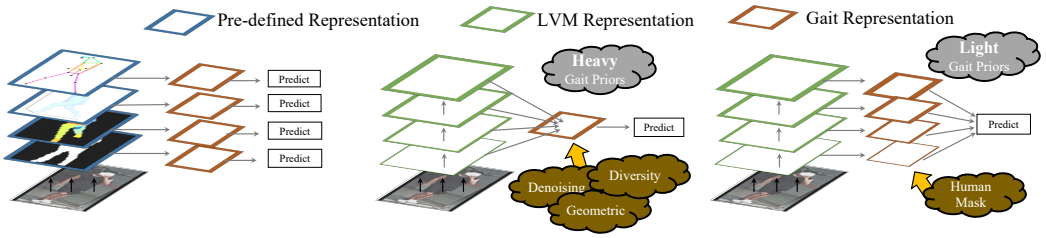
1 Introduction

Gait recognition aims to identify an individual based on the unique patterns in the walk sequence. Unlike other biometric [23] modalities such as face [53, 8, 38, 27, 46, 45, 28, 18], fingerprint [3], or iris [51, 49], gait is unobtrusive and capable of identifying individuals from afar without their active involvement. These unique advantages make gait recognition especially effective for security applications, including suspect tracking and identity verification [47, 58, 39, 40, 29, 79, 22, 21].

To focus on pure gait patterns, early approaches suppress appearance noise by transforming each frame into pre-defined representations, like silhouettes [5, 11, 37, 57, 14], skeleton landmarks [36, 4, 63, 15], body parsing [41, 77, 80], or SMPL meshes [34, 76, 43], before feature extraction, as shown in Figure 1 (a). Although such explicit representations curb distractions from clothing and background, they also discard crucial cues: silhouettes erase body structure, skeletons remove shape information, and SMPL overly smooths personal idiosyncrasies, capping accuracy. Alternatively, recent approaches [74, 72, 26] achieve substantial gains by guiding large vision models (LVM) with human priors such as feature smoothing [74], language guidance [72], and geometry-driven denoising [26] to extract rich and implicit gait features directly from RGB data.

*Equal contribution. Part of the work was conducted when Mr. Ye visited MSU.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Pre-defined Gait (b) Typical LVM Gait (c) Layer-wise LVM Gait (**ours**)

Figure 1: Comparison of gait representation paradigms. (a) Pre-defined gait learning uses explicit gait patterns for prediction, losing essential identity information. (b) Typical LVM gait learning relies heavily on gait priors and treats all LVM layers equally, underutilizing discriminative features. (c) Our layer-wise LVM gait learning fully utilizes intermediate LVM layers with minimal gait priors.

Despite recent advances, LVM-based gait recognition methods [74, 26] still rely heavily on traditional gait priors, whether through supervised or self-supervised training, as shown in Fig 1(b). However, is this dependence truly necessary? Through extensive experiments, we challenge this assumption, showing that LVMs [48, 30, 52] inherently possess rich, layer-wise representations that can support high-performance gait recognition with minimal reliance on human-designed priors. These findings suggest a simpler yet more robust baseline for LVM-based gait recognition, potentially redefining the role of domain knowledge in this field. Our experiments across various LVMs [52, 48, 30] and datasets [75, 56, 80, 25] further validate this remarkable observation, aligning similarly with related studies [7, 16, 59, 60, 62].

To make the above findings comprehensive, this work systematically investigates layer-wise representations in LVMs for gait recognition and uncovers three key insights: (1) Across a range of LVM architectures and scales, intermediate layers consistently yield more discriminative features than the final layer, echoing trends observed in LLMs [7, 16, 60]. (2) Each layer contributes unique, task-dependent information. (3) Intermediate-layer features are highly complementary, and their fusion produces substantial performance gains. Based on these findings, we propose a simple and universal layer-wise baseline, termed **BiggerGait**. However, fully exploiting the advantages of BiggerGait remains a challenge for GPU-limited scenes. The issue lies in the need to add a dedicated gait encoder to every LVM layer, inflating the parameters and computation cost as depth grows.

To mitigate this issue, we propose an optional grouping strategy for BiggerGait to balance performance and efficiency. Previous work [62] suggests that residual connections in LVMs encourage intermediate layers to inhabit a shared feature space. Based on this insight, we contend that a single gait encoder can effectively process features from multiple LVM layers, eliminating the need for layer-specific encoders. Specifically, this grouping strategy to merge neighboring similar LVM layers, replacing per-layer gait encoders with two shared ones: one for shallow layers and one for deep layers. This design delivers the similar performance gains as the standard BiggerGait while saving considerable cost during gait representation extraction. Experiments on CCPG [33], CCGR_MINI [80], CASIA-B* [75] and SUSTech1K [56] datasets validate the effectiveness of BiggerGait and this grouping strategy in both within- and cross-domain evaluations.

This paper makes two important contributions.

- *Comprehensive layer-wise analysis.* We conduct the first systematic examination of LVM layer representations for gait recognition, detailing how different depths affect task-specific performance and discovering that fusing complementary intermediate features unlocks substantial accuracy gains.

- *A simple yet powerful baseline.* We present BiggerGait, a simple and universal framework for LVM-based gait recognition, along with a grouping strategy to balance performance-efficiency trade-offs. Extensive experiments show that BiggerGait establishes state-of-the-art results across multiple RGB-based gait benchmarks, achieving superior performance in both intra-domain and cross-domain evaluation settings.

2 Related Work

Gait Recognition Gait recognition aims to extract subtle gait patterns that remain invariant to background clutter and clothing variations. Recent research has primarily focused on addressing these challenges in RGB video-based gait analysis [33, 56, 19, 61, 69, 66, 20]. Existing approaches can be broadly grouped into two categories: pre-defined [64, 5, 11, 37, 57, 14, 36, 63, 15, 4, 77, 41, 34, 76, 43, 71, 70, 73, 58] and LVMs representations [74, 72, 26]. Pre-defined methods explicitly extract gait-relevant components using segmentation [5, 11, 37, 57, 14], pose estimation [36, 63, 15, 4], and 3D modeling [34, 76, 43], effectively suppressing irrelevant gait information. However, this often comes at the cost of discarding identity-discriminative cues. In contrast, LVMs methods [74, 72, 26] extract implicit gait features from large vision models guided by human priors, preserving richer semantics and achieving stronger performance. Building on the LVMs paradigm, this paper introduces BiggerGait to delve deeper into its potential for gait representation.

Large Vision Models Motivated by the success of LLMs [2, 9, 24, 31, 54], the vision community has increasingly turned its attention to building large-scale foundation models for visual understanding [48, 30]. These models aim to learn transferable and general-purpose visual representations from massive web-scale datasets. Representative LVMs include CLIP [52], which leverages language supervision to guide visual representation learning; SAM [30], a promptable segmentation model trained on a large-scale annotated dataset for strong generalization; and DINOv2 [48], a self-supervised model that learns highly transferable features from vast, diverse image collections. The features extracted from these LVMs are termed all-purpose, as they effectively transfer to a range of downstream tasks, including image classification, semantic segmentation, and depth prediction. Our aim is to explore how these all-purpose representations can be adapted to the task of gait recognition, harnessing the broad advantages offered by LVMs. In this paper, we conduct a comprehensive investigation of how different types and scales of three representative LVMs (CLIP, SAM, and DINOv2) can be leveraged for downstream gait recognition. Further, we propose a unified baseline, BiggerGait, that consistently excels across multiple LVM architectures and model sizes.

Layer-wise Analysis in Large Models Recent works [60, 16, 62, 42, 17, 10, 1] have increasingly focused on layer-wise representation from large models, as intermediate features often show surprising robustness, challenging the traditional final layer representations. In NLP, researchers [42] have found that lower layers tend to encode more syntactic information, while higher layers specialize in semantic features. Others suggest that residual connections encourage layers to share a common feature space while still specializing in distinct sub-tasks [62], or that attention sink effects may weaken final-layer performance [17]. Similar trends emerge in vision domains: Head2Toe [10] selects the most useful representations from intermediate layers in transfer learning, outperforming the final layer. A fresh work [1] on LVMs again suggests that the final layer may not contain the most robust visual features, and addresses this by distilling optimal intermediate features back into the final layer. Unlike prior works [10, 1] that focus on coarse-grained vision tasks (classification, detection, and tracking), our study goes a step further by validating and advancing this insight in a significantly more demanding setting, *i.e.*, a highly fine-grained recognition task. **Beyond broadening this insight, we further reveal more interesting and unexplored findings unique to gait tasks in Sec. 3.4 & 3.5.**

3 Layer-wise Representation Analysis

First of all, we hypothesize that the unique layer-wise heterogeneity observed in large language models [60, 62, 16, 7, 59] (LLM) may also exist in large vision models [48, 30, 52] (LVM), potentially influencing downstream gait recognition. To verify our conjecture, we introduce a simple layer-wise gait baseline, a comprehensive experiment setting, and three key questions to systematically explore the impact of different LVM layers on gait recognition. Eventually, we further analyse the computational overhead inherent in layer-wise methods and propose a mitigation option.

3.1 BiggerGait: A Layer-wise LVM-based Gait Baseline

We construct a simple layer-wise gait baseline, called **BiggerGait**, illustrated in Fig. 2. Given an image $x \sim p(x)$ from a walking video, a LVM projects it into multiple intermediate feature maps $\{f_i \mid i \in \{1, 2, ..., N\}\}$ with the corresponding semantic hierarchy spanning from low to high levels.

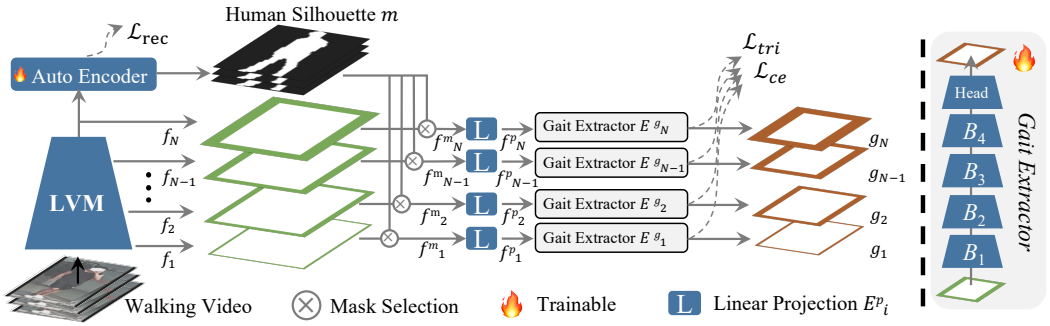


Figure 2: Overview of the proposed BiggerGait. An LVM extracts multi-level features from RGB videos. Human silhouettes are generated using an unsupervised auto-encoder [74], serving as the only human-designed gait prior. Each level's features, with background noise removed, are processed by separate linear projection layers and gait extractors to obtain the final gait representations.

In BiggerGait, we set $N = 12$, uniformly sampling 12 layers from the LVMs. Followed the design of BigGait [74], the last feature map f_N which contains the highest-level semantic information, is fed into an auto-encoder to generate the human silhouette m:

$$m = \text{softmax}(E(f_N)),\ \bar{f}_N = D(m),\ \mathcal{L}_{rec} = \left\| f_N - \bar{f}_N \right\|_2^2, \tag{1}$$

where E and D are 1×1 convolution layers, E outputs 2 channels, and D restores the original channel dimension. **Notably, unlike traditional LVM gait methods [74, 26] reliant on heavy gait priors (*e.g.*, geometric, denoising, and diversity constraints), this human mask is the only gait prior we employed.** The softmax function is conducted along the channel dimension, and $\mathcal{L}_{rec}$ presents the reconstruction loss. After masking the background noise in $\{f_i\}$, we obtain $\{f_i^m\}$:

$$f_i^m = m \cdot f_i, \tag{2}$$

where $\cdot$ denote the multiplication. To reduce the GPU memory consumption, each f_i^m is fed into a lightwight linear projection layer:

$$f_i^p = \text{sigmoid}(E_i^{\text{p}}(f_i^m)), \tag{3}$$

where E_i^{p} consists of two 1×1 convolutions, two batch-normalization layers, a GELU activation. Its output channel is set to C. Here the hyper-parameter C is set to 16, following [74]. Each f_i^p is upsampled by bilinear interpolation to improve the resolution, *i.e.,* exhibited as a 3-D tensor with a size of $16 \times 64 \times 32$ while the first dimension denotes the output channel of the linear projection. Finally, we feed the f_i^p into gait extractors E_i^g (GaitBase [14]) to obtain gait representation g_i. Overall, the gait representation R of the BiggerGait can be formulated as:

$$R = \{g_i = E_i^{\text{g}}(\text{sigmoid}(E_i^{\text{p}}(f_i^m))) \mid i \in \{1, 2, ..., N\}\}. \tag{4}$$

Consistent with recent works [76, 44, 67, 68, 81], triplet losses $\mathcal{L}_{tri}$ and cross-entropy losses $\mathcal{L}_{ce}$ are used for gait training. The overall loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{tri} + \mathcal{L}_{ce} + \mathcal{L}_{rec}. \tag{5}$$

In summary, the central innovation of BiggerGait lies in treating intermediate layers independently to fully unlock the power of large vision models.

3.2 Experimental Setting

Separate Testing. To assess each layer's discriminative power, we adopt a separate testing strategy at inference. Given a probe sample $x \sim \mathcal{X}$ and a gallery sample $y \sim \mathcal{Y}$, layer i produces gait features g_i^x and g_i^y. Their Euclidean distance serves as the similarity score for layer i:

$$d_i\,(x, y) = \|g_i^x - g_i^y\|_2. \tag{6}$$

Since embeddings vary across depths, each layer yields a distinct score for this pair (x, y).

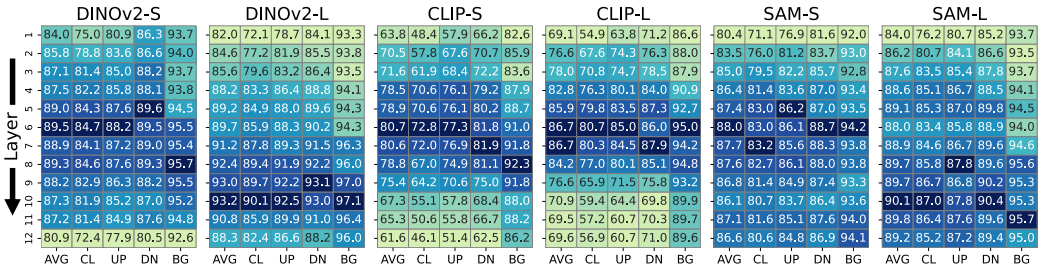


Figure 3: Layer-wise Performance Across LVMs. This figure presents the gait recognition accuracy of six large vision models, *i.e.*, DINOv2-Small/Large [48], CLIP-Small/Large [52], and SAM-Small/Large [30], across 12 intermediate layers, evaluated on the CCPG [33] dataset. Each cell lists the Rank-1 accuracy for full clothing (CL), top (UP), pants (DN), and bag (BG) changes, along with the average (AVG). A downward arrow on the left indicates increasing network depth. In each column, the best score is shown in black color.

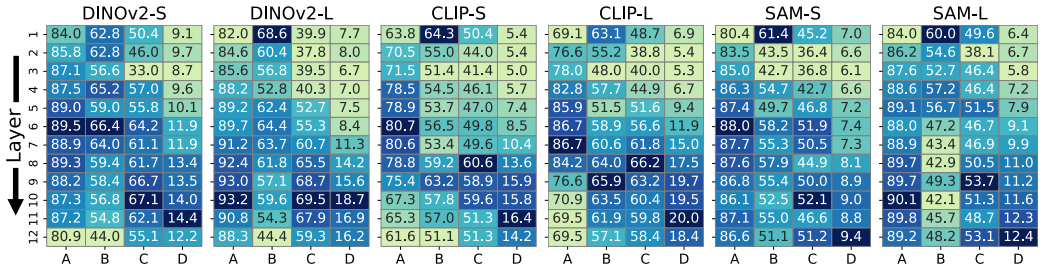


Figure 4: Layer-wise Performance Across Datasets. The columns (A, B, C, D) represent the four test sets, CCPG [33], SUSTech1K [56], CASIA-B* [75], and CCGR_MINI [80]. All models are trained on CCPG dataset. Column A reports within-domain performance, whereas columns B–D present cross-domain results.

Target LVMs. Three representative LVMs with distinct pretraining strategies are evaluated in our experiment: SAM [30] is trained under supervised segmentation objectives, CLIP [52] follows an image-text contrastive learning approach, and DINOv2 [48] adopts self-supervised knowledge distillation. In this paper, LVM-S and LVM-L refer to the small and large versions of LVM. We test every one of the 12 layers in LVM-S, but uniformly sample 12 layers from LVM-L.

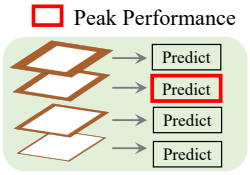
Dataset. Subsequent experiments are mainly conducted on four widely used clothing-variation and multi-view gait datasets: CCPG [33], CASIA-B* [75], SUSTech1K [56], and CCGR_MINI [80]. CCPG serves as the cornerstone benchmark, offering diverse full body clothing variations while masking faces and shoes to simulate real-world cloth-changing scenarios.

3.3 Do Middle Layers Outperform the Final Layer?

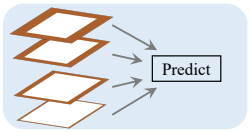
As shown in Fig. 3, to investigate whether middle layers offer stronger discriminative power than the final layer, we evaluate gait recognition accuracy across 12 layers of three popular LVMs [30, 52, 48] in different model sizes. Clearly, all LVMs achieve peak performance at middle layers rather than at the deepest one, echoing similar findings in LLMs [60, 16, 7]. For instance, in DINOv2-S [48], the highest average accuracy of 89.5% occurs at Layer 6, while the final layer drop to 80.9%.

This similar trend is also observed in both CLIP [52], which uses image-text pretraining, and SAM [30], trained with supervision, despite their different paradigms from the self-supervised DINOv2, highlighting the generality of this interesting phenomenon. Notably, even when model size increases, this middle-layer advantage remains, suggesting that deeper depth does not eliminate the representational superiority of intermediate layers. This effect is particularly evident in CLIP, where middle layers outperform both shallow and final layers by a significant margin.

Our answer: Yes, middle layers consistently outperform the final layer for gait recognition, regardless of LVM type and size.



(a) Separate Testing



(b) Ensemble Testing

LVM	Ensemble Testing	CCPG				SUSTech1k		CASIA-B			CCGR*		Avg
		CL	UP	DN	BG	CL	R-1	NM	BG	CL	R-1	R-5	
DINOv2-S	×	84.7	88.2	89.5	95.5	48.6	66.4	84.7	76.7	39.8	14.4	25.4	59.3
	✓	**89.8**	**92.1**	**93.7**	**97.5**	**70.9**	**79.6**	**93.0**	**90.8**	**55.6**	**15.5**	**26.9**	**67.0**
DINOv2-L	×	90.1	92.5	93.0	97.1	48.8	68.6	85.7	79.5	43.5	**18.7**	**30.5**	62.5
	✓	**92.8**	**94.6**	**95.7**	**98.2**	**76.4**	**85.3**	**95.4**	**93.7**	**62.2**	15.7	26.9	**70.0**
CLIP-S	×	72.8	77.3	81.8	91.0	39.5	64.3	77.4	72.1	32.4	**16.4**	**29.2**	55.5
	✓	**81.0**	**86.2**	**89.7**	**96.9**	**67.0**	**84.2**	**91.5**	**87.8**	**47.3**	15.2	26.1	**65.8**
CLIP-L	×	80.3	84.5	87.9	94.2	35.2	65.9	82.6	77.2	38.8	**20.0**	**33.2**	59.7
	✓	**87.5**	**90.7**	**92.9**	**97.4**	**72.6**	**88.3**	**95.0**	**92.6**	**51.7**	18.3	30.4	**69.6**
SAM-S	×	83.0	86.1	88.7	94.2	51.0	61.4	69.8	62.8	23.8	9.4	17.4	52.7
	✓	**86.8**	**89.5**	**91.8**	**95.9**	**64.6**	**75.5**	**80.1**	**76.2**	**31.8**	**9.5**	**17.8**	**59.7**
SAM-L	×	87.0	87.8	90.4	95.3	45.7	60.1	68.2	61.9	30.9	**12.4**	**21.6**	54.1
	✓	**90.0**	**91.2**	**93.8**	**97.0**	**64.1**	**73.3**	**81.3**	**77.7**	**42.2**	11.9	21.3	**61.3**

(c) Comparison of Separate vs. Ensemble Testing

Figure 5: Separate vs. Ensemble Layer Testing (a) The red box highlights the peak performance achieved by these layers. This peak performance represents the final result of separate testing. (b) Ensemble testing works likes a fair voting manner. (c) All models are trained on CCPG [33], and tested on four datasets [33, 80, 75, 56]. CCGR* indicates CCGR_MINI dataset.

3.4 Do Middle Layers Contribute Similarly?

To further evaluate the contribution of different LVM layers to gait recognition, we adopt additional testing datasets, each reflecting distinct characteristics. Specifically, SUSTech1K [56] emphasizes LiDAR-accessible scenes, CASIA-B* [75] represents controlled indoor environments, CCPG [33] provides extensive clothing variations, and CCGR_MINI [80] integrates diverse covariates.

We see two interesting finding in Fig. 4: (1) The best layers for cross-domain performance often differ from those for within-domain performance, occurring with a probability of 83.3%. This means that, for domain-specific tasks, the contribution of layers is inconsistent, where the most effective layer may change. (2) Notably, SUSTech1K achieves peak performance in shallow layers, most frequently at Layer 1 (66.7%), while CCGR_MINI performs better in deeper ones. We consider that deeper LVM layers capture stronger semantic features, while shallower ones preserve appearance details. To verify this, we carefully check its layer-wise performance on SUSTech1K, where the optimal layer shifts from the 1st (Same-clothing condition, easy appearance task) to the 7th (Changing-clothing condition, harder semantic task). **These results suggest that all LVM layers, from the shallowest to the deepest, may potentially benefit different domain-specific recognition tasks.**

Our answer: No, layer contributions vary by task, and should be treated independently.

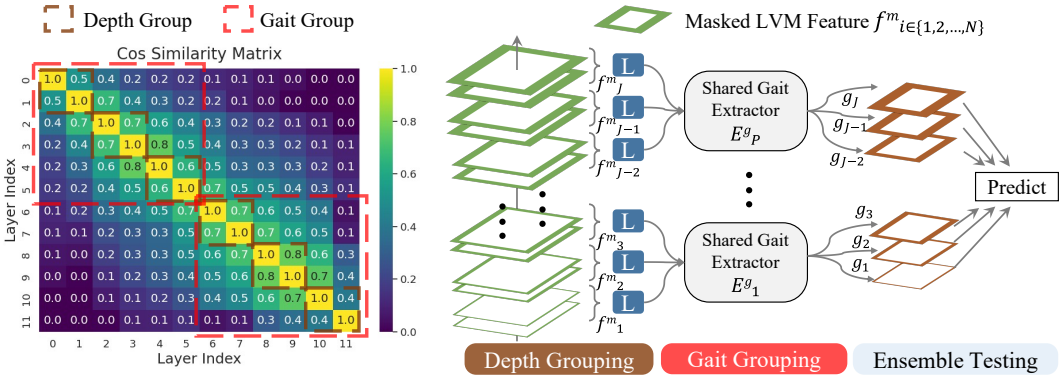
3.5 Are Middle Layers Complementary?

To evaluate the complementarity of these layers, we introduce ensemble testing during inference and compare it with separate testing, as illustrated in Fig. 5.

Ensemble Testing. Unlike separate testing in Sec. 3.2, which scores each layer independently, ensemble testing pools all layer distances and yields a single unified result. For a probe-gallery pair (x, y) with per-layer scores $\{d_i(x, y)\}_{i=1}^{N}$, the final similarity of ensemble testing is their mean:

$$D(x, y) = \frac{1}{N} \sum_{i=1}^{N} d_i(x, y). \tag{7}$$

Fig. 5 (c) shows that ensemble testing often offers an impressive performance gain in both within- and cross-domain settings, raising accuracy by $+7.7\%$ on DINOv2-S, $+10.3\%$ on CLIP-S, and $+7.0\%$ on SAM-S. Meanwhile, an abnormal observation arises in the CCGR_MINI, where the fusion results sometimes slightly drop compared to the separated ones. Sec. 3.4 shows that the challenging CCGR_MINI prefers deeper semantic-based layers due to its complex data variations. We further experimented with fusing only the deeper half of the layers, which alleviate this problem on CCGR but harm generalization on other datasets such as SUSTech1K. Thus, our BiggerGait still adopts all layers for simplicity, consistency, and stronger generalization.



(a) Layer Similarity & Grouping Regions (b) The architecture of BiggerGait*

Figure 6: (a) Average pairwise cosine similarity across the 12 layers of DINOv2-S [48], with depth and gait groups highlighted. (b) BiggerGait*: A grouped based BiggerGait that clusters similar layers ($J = 6$), uses two shared gait encoders ($P = 2$) to replace original per-layer ones, and applies ensemble testing during inference.

Our answer: Yes, middle layers are highly complementary, and aggregating them often yields significant gains in both within- and cross-domain tasks.

3.6 Efficiency Discussion and Mitigation

Although multi-layer features are highly complementary, assigning a separate gait head to every LVM layer inflates computation and memory. Recent work [62] in LLMs shows that ViT residual connections encourage middle layers in a shared feature space. Fig. 6 (a) shows that the same pattern may also exist in LVM: for DINOv2-S [48], similarity remains strong between adjacent middle layers but decays quickly. To suit the need of GPU-limited scene, we equip BiggerGait with a grouping strategy inspired by this shared space hypothesis. To distinguish it from the standard version, we denote this grouping-based variant as **BiggerGait***. As shown in Fig. 6 (b), the strategy has two components: (1) depth grouping, which lowers the computational cost of gait heads; and (2) gait grouping, which reduces parameter overhead.

Depth Grouping. Adjacent, similar layer features $\{f_i^m\}$ are divided into J continuous depth groups, and the features within each group are concatenated:

$$f_j^m = \text{concat}(f_{i_1}^m, f_{i_2}^m, ..., f_{i_k}^m) \mid j \in \{1, 2, ..., J\}, \tag{8}$$

where f_j^m represents the concatenated feature for group j. The subscript , $i_1, i_2, ..., i_k$ refer to the specific layers. Depth grouping cuts the number of gait heads' forward passes from N to J.

Gait Grouping. The multiple gait encoders are also merged from $\{E_i^g\}$ into P groups $\{E_p^g \mid p \in \{1, 2, ..., P\}\}$. Gait grouping reduces parameters of the gait encoders from Ns to Ps, where s is the size of one gait encoder. The gait results R' of BiggerGait* is reformulated as:

$$R' = \{g_j = E_p^g(\text{sigmoid}(E_j^{\text{p}}(f_j^m))) \mid j \in \mathcal{J}, p \in \mathcal{P}\}, \tag{9}$$

where $\mathcal{J} = \{1, 2, ..., J\}$ and $\mathcal{P} = \{1, 2, ..., P\}$, denoting the number of depth groups and gait groups.

The objective of the grouping paradigm is to minimize the number of groups J and P while preserving the fidelity of the output representations, *i.e.*, ensuring that new gait result R' closely approximates its original counterpart R in Eq. (4). This objective can be expressed as a bi-level optimization problem:

$$\min_{J,\,P} \text{ s.t. } \min_{\{E_p^g\},\,\{E_j^p\}} \|R - R'\|_2^2 \leq \epsilon. \tag{10}$$

In Sec. 4.3, we experimentally conclude that $J = 6$ and $P = 2$ represent an effective configuration. Our current grouping design prioritizes feasibility, reproducibility, and efficiency. More sophisticated grouping strategies, like explicitly maximizing representational dissimilarity or learnable grouping, are also promising future directions worth specific exploration.

Table 1: Performance comparison across methods and datasets. Yellow regions indicate within-domain evaluations, others are cross-domain. The last column reports the overall average accuracy. CCGR* indicates CCGR_MINI dataset. BiggerGait* refers the grouping-based BiggerGait.

Training Dataset	Input	Method	LVM	Venue	Testing Dataset											Avg
					CCPG [33]				CCGR* [80]		SUSTech1k [56]		CASIA-B* [75]			
					CL	UP	DN	BG	R-1	R-5	CL	R-1	NM	BG	CL	
CCPG	Silh.	GaitSet [5]	-	PAMI'22	60.2	65.2	65.1	68.5	2.4	6.9	8.2	12.8	47.4	40.9	25.8	29.5
	Silh.	GaitPart [11]	-	CVPR'20	64.3	67.8	68.6	71.7	2.4	6.9	8.1	13.5	51.2	41.9	26.0	30.9
	Silh.	GaitGL [37]	-	ICCV'21	68.3	76.2	67.0	76.7	3.3	8.4	25.4	33.6	63.1	58.5	46.3	41.2
	Silh.	GaitBase [13]	-	CVPR'23	71.6	75.0	76.8	78.6	2.8	7.3	9.5	16.8	59.1	52.7	30.4	35.6
	Silh.	DeepGaitV2 [12]	-	Arxiv	78.6	84.8	80.7	89.2	3.7	9.1	27.0	38.4	74.6	67.2	50.2	47.4
	Silh.+Skel.	BiFusion [50]	-	MTA'23	62.6	67.6	66.3	66.0	-	-	-	-	-	-	-	-
	Silh.+Skel.	SkeletonGait++ [15]	-	AAAI'24	79.1	83.9	81.7	89.9	-	-	-	-	-	-	-	-
	Silh.+Pars.	XGait [78]	-	MM'24	72.8	77.0	79.1	80.5	-	-	-	-	-	-	-	-
	Silh.+Pars.+Flow	MultiGait++ [25]	-	AAAI'25	83.9	89.0	86.0	91.5	-	-	-	-	-	-	-	-
	RGB+Silh.	GaitEdge [35]	-	ECCV'22	66.9	74.0	70.6	77.1	-	-	8.9	19.6	66.5	58.7	44.8	-
	RGB+Silh.	DenoisingGait [26]	SD [55]	CVPR'25	84.0	88.0	90.1	95.9	-	-	37.3	59.1	83.9	76.1	34.8	-
	RGB	BigGait [74]	DINOv2-S	CVPR'24	82.6	85.9	87.1	93.1	7.4	16.3	43.7	56.4	77.4	71.5	33.6	53.0
	RGB	BiggerGait	SAM-S [30]	Ours	86.8	89.5	91.8	95.9	9.5	17.8	64.6	75.5	80.1	76.2	31.8	59.7
	RGB	BiggerGait	CLIP-S [52]	Ours	81.0	86.2	89.7	96.9	15.2	26.1	67.0	**84.2**	91.5	87.8	47.3	65.8
	RGB	BiggerGait	DINOv2-S [48]	Ours	**89.8**	**92.1**	93.7	**97.5**	**15.5**	**26.9**	**70.9**	79.6	**93.0**	**90.8**	**55.6**	**67.0**
	RGB	BiggerGait*	SAM-S	Ours	86.9	89.4	92.3	95.8	9.1	17.2	60.8	74.4	79.7	74.9	28.9	59.0
	RGB	BiggerGait*	CLIP-S	Ours	78.9	83.8	87.9	96.1	13.9	24.2	63.1	81.5	92.3	87.1	42.9	64.0
	RGB	BiggerGait*	DINOv2-S	Ours	89.0	91.9	**94.0**	97.2	14.5	25.3	69.5	80.4	91.6	87.7	54.7	66.5
CCGR*	Silh.	GaitBase	-	CVPR'23	20.8	31.3	**38.3**	70.2	21.1	37.8	28.7	48.0	66.3	57.7	33.9	40.5
	Silh.	DeepGaitV2	-	Arxiv	23.0	37.1	36.5	69.9	26.4	45.2	32.1	51.7	72.0	62.0	36.6	44.1
	RGB	BigGait	DINOv2-S	CVPR'24	22.9	42.2	25.0	80.5	**88.0**	**95.9**	71.9	85.6	90.1	87.9	58.4	73.7
	RGB	BiggerGait	SAM-S	Ours	15.8	32.8	23.1	69.6	85.1	94.0	72.7	89.2	87.6	85.3	48.4	70.8
	RGB	BiggerGait	CLIP-S	Ours	21.7	**46.5**	28.6	**88.4**	80.7	92.0	81.2	93.3	94.2	92.8	60.9	75.7
	RGB	BiggerGait	DINOv2-S	Ours	**23.2**	44.5	29.5	86.7	85.7	94.1	82.1	**93.8**	**97.2**	**96.4**	**66.7**	78.1
	RGB	BiggerGait*	SAM-S	Ours	15.2	33.3	24.8	71.1	86.3	95.0	73.7	88.2	84.2	82.0	46.9	70.4
	RGB	BiggerGait*	CLIP-S	Ours	19.6	42.3	29.3	85.9	82.2	93.2	80.6	92.4	92.9	91.5	60.4	75.1
	RGB	BiggerGait*	DINOv2-S	Ours	22.6	44.9	31.5	85.8	87.8	95.8	**83.7**	**93.8**	96.9	96.3	65.6	**78.5**

4 Experiments

We conduct our experiments on four widely used clothing-variation and multi-view gait datasets: CCPG [33], CASIA-B* [75], SUSTech1K [56], and CCGR_MINI [80]. CCPG acts as our cornerstone benchmark, since it offers the richest wardrobe diversity, covering an array of coats, trousers, and bags in assorted color and styles, and faces and shoes are masked to emulate real-world cloth-changing scenarios. Although outfit changes are few in CCGR, it excels in covariate variety: abundant viewpoints, complex ground conditions, different walking speeds, and composite covariate scene. The full CCGR set is too large for quick testing, so we use the official mini version, which keeps the challenge but significantly drops the bulk. Therefore, the challenge CCPG and CCGR_MINI sets supply the training data, whereas evaluation is performed on all four datasets. Every experiment strictly follows the official protocols released by the owner. Gait evaluation protocols is reported for multi-view settings, and rank-1 accuracy serves as the principal metric.

4.1 Implementation Details

All input frames are resized to 448×224 for DINOv2 [48], 224×224 for CLIP [52] and 512×256 for SAM [30]. The training runs for 30k iterations using SGD (momentum $= 0.9$, weight-decay $= 5 \times 10^{-4}$) with an initial learning rate of 0.1, which is dropped by 10× at 15k and 25k steps. Each mini-batch adopts the tuple $(p, k, l) = (8, 4, 30)$, which is 8 identities, 4 sequences per identity, and 30 frames per sequence. Frame sampling follows the protocol of GaitBase [13], and the sole augmentation is a random horizontal flip applied consistently to every frame within a sequence. Using eight 24GB RTX 6000 GPUs, the DINOv2-S-based BiggerGait requires approximately 8.8 hours to train on CCPG. Ensemble testing method presented in Sec. 3.5 is used during inference.

4.2 Main Results

To show our superiority, BiggerGait and its grouping-based variant are compared with diverse SoTA methods, including the silhouette-based [5, 11, 37, 13, 12], multimodal-based [50, 15, 35], and RGB-based [74, 35] gait methods. Due to the lack of multimodal data, cross-domain results for multimodal-based methods are unavailable.

Table 2: All models are evaluated on CCPG [33]. (a) & (b) Hyperparameter search for BiggerGait*. (c) Ablation study on larger LVMs. (d) Efficiency comparison across gait methods. The yellow cells in (a) & (b) mark the final setting for BiggerGait*. FLOPs are computed for an input resolution of 448×224.

(a) Ablation of Gait Group ($J = 12$)

LVM	P	#Params	CL	UP	DN	BG
SAM-S [30]	12	209.7M	**86.8**	**89.5**	**91.8**	95.9
	2	112.2M	85.8	**89.5**	**91.8**	**96.3**
	1	101.7M	85.2	88.6	91.3	95.8
CLIP-S [52]	12	208.6M	**81.0**	**86.2**	**89.7**	**96.9**
	2	110.8M	80.1	85.5	89.4	96.3
	1	100.6M	78.4	82.9	87.7	95.0
DINOv2-S [48]	12	142.2M	**89.8**	92.1	93.7	97.5
	2	43.6M	89.5	**92.5**	**94.0**	**97.6**
	1	33.4M	86.3	90.3	92.5	97.1

(b) Ablation of Depth Group ($P = 2, 2, 3$)

LVM	J	FLOPS	CL	UP	DN	BG
SAM-S	12	95.1G	85.8	**89.5**	91.8	**96.3**
	6	79.2G	**86.9**	89.4	**92.3**	95.8
	3	71.2G	84.6	87.5	90.5	94.4
CLIP-S	12	49.3G	**80.1**	**85.5**	**89.4**	**96.3**
	6	33.4G	78.9	83.8	87.9	96.1
	3	25.5G	75.4	80.6	86.8	95.8
DINOv2-S	12	45.7G	**89.5**	**92.5**	**94.0**	**97.6**
	6	29.8G	89.0	91.9	**94.0**	97.2
	3	21.9G	88.8	91.7	93.1	97.5

(c) Ablation of Scaling LVM's Size

LVM	Method	#Params	FLOPs	CL	UP	DN	BG
SAM-L	BiggerGait	436.4M	258.7G	**90.0**	**91.2**	**93.8**	**97.0**
	BiggerGait*	337.7M	246.1G	89.0	**91.2**	93.5	96.6
CLIP-L	BiggerGait	430.9M	111.3G	**87.5**	**90.7**	**92.9**	**97.4**
	BiggerGait*	332.2M	97.0G	85.6	89.7	91.5	97.2
DINOv2-L	BiggerGait	429.9M	203.4G	**92.8**	**94.6**	95.7	**98.2**
	BiggerGait*	339.6M	190.7G	92.7	93.9	**96.2**	97.8

(d) Parameter and GFLOPs Comparison

Method	Upstream	Downstream	#Params	FLOPs
Silh.-based	DeepLabV3+ [6]	GaitBase	34.2M	45.4G
Parsing-based	SCHP [32]	GaitBase	74.0M	35.2G
Skeleton-based	HRNet-W32 [65]	Gait-TR	29.0M	31.2G
BigGait [74]	DINOv2-S	GaitBase	30.8M	12.7G
BiggerGait	SAM-S	12 x GaitBase	209.7M	95.1G
BiggerGait	CLIP-S	12 x GaitBase	208.6M	49.3G
BiggerGait	DINOv2-S	12 x GaitBase	142.2M	45.7G
BiggerGait*	SAM-S	2 x GaitBase	112.2M	79.2G
BiggerGait*	CLIP-S	2 x GaitBase	110.8M	33.4G
BiggerGait*	DINOv2-S	2 x GaitBase	43.6M	29.8G

Cross-domain Evaluation. The final column of Tab. 1 highlights BiggerGait's impressive results. Trained on CCPG, the SAM-S-, CLIP-S-, and DINOv2-S-based BiggerGait impressively boost rank-1 by $+6.7\%$, $+12.8\%$, and $+14.0\%$ over prior work. With CCGR_MINI as the train set, CLIP-S- and DINOv2-S-based BiggerGait push SoTA up by $+2.0\%$ and $+4.4\%$, respectively. Such huge jumps confirm that BiggerGait learns a robust gait embedding that travels well across datasets.

Like BigGait [74], BiggerGait also shows a data-bias limitation, *i.e.*, the distribution of training data influences outcomes. Trained on CCGR_MINI and tested on CCPG, RGB-based methods exhibit less impressive results in some cases, *e.g.*, performing well in the ups-changing (UP) and bag-changing (BG), but poorly in the full-changing (CL) and pants-changing (DN). The limited clothing diversity in CCGR_MINI (9.4% of pairs, exclusively UP changes) probably accounts for this result.

Within-domain Evaluation. As highlighted in the yellow block of Tab. 1, BiggerGait shines on the challenge CCPG dataset. SAM-S- and DINOv2-S-based BiggerGait outperforms other methods on every metric. Notably, DINOv2-S-based BiggerGait shows significant improvements of $+5.9\%$ on CL, $+3.1\%$ on UP, $+6.6\%$ on DN, and $+4.4\%$ on BG. These results highlight BiggerGait's effectiveness in learning subtle clothing-irrelevant gait representations.

On the CCGR_MINI, BiggerGait slightly struggles. As discussed on Sec. 3.4 and 3.5, CCGR_MINI with diverse covariant only prefers deeper semantic-based layers, and other datasets prefers shallow appearance-based layers. For simplicity, consistency, and stronger generalization, BiggerGait adopts all layers, resulting slightly drop on CCGR_MINI. We consider that this performance gap is acceptable: with the same DINOv2-S backbone, BiggerGait and BiggerGait* achieves 2.3% and 0.2% lower rank-1 accuracy than BigGait, respectively, comparing their larger overall performance gains of $+4.4\%$ and $+4.8\%$.

Comparing Different LVMs. A clear domain pattern emerges: (1) the text-aligned CLIP [52] excels in cross-domain tests but lags within-domain; (2) the segmentation-supervised SAM [30] exhibits the reverse trend; (3) the self-supervised DINOv2 [48] balances both. This implies that LVM supervision strategies probably shape their domain adaptation properties thereby affecting gait recognition.

4.3 Ablation Study

All experiments in this section are performed on the CCPG benchmark [33]. We systematically evaluate: (1) an effective configuration for BiggerGait*; (2) the efficiency issue of BiggerGait.

Gait & Depth Group. Tab. 2(a) shows, using just two gait encoders delivers an accuracy similar to that of using twelve. Remarkably, even with all layer-wise features share one single gait encoder, the DINOv2-S- and SAM-S-based BiggerGait still delivers SOTA results, outperforming the methods listed in Tab. 1. Tab. 2(b) reveals that six depth groups achieve an accuracy comparable to the

Table 3: Model size and computation cost comparison across methods. All methods trained on CCPG, and tested on four datasets. This is a supplement for Tab. 1. This report follows the settings in Table 2 (d), with FLOPs computed at an input resolution of 448 × 224.

Input	Method	Upstream	#Param	FLOPs	Testing Dataset											Avg
					CCPG				CCGR*		SUSTech1k	CASIA-B*				
					CL	UP	DN	BG	R-1	R-5	CL	R-1	NM	BG	CL	
Silh.	GaitSet	DeepLabV3+	29.4M	50.3G	60.2	65.2	65.1	68.5	2.4	6.9	8.2	12.8	47.4	40.9	25.8	29.5
Silh.	GaitPart	DeepLabV3+	31.6M	53.0G	64.3	67.8	68.6	71.7	2.4	6.9	8.1	13.5	51.2	41.9	26.0	30.9
Silh.	GaitGL	DeepLabV3+	30.1M	88.7G	68.3	76.2	67.0	76.7	3.3	8.4	25.4	33.6	63.1	58.5	46.3	41.2
Silh.	GaitBase	DeepLabV3+	34.2M	45.4G	71.6	75.0	76.8	78.6	2.8	7.3	9.5	16.8	59.1	52.7	30.4	35.6
Silh.	DeepGaitV2	DeepLabV3+	35.2M	93.2G	78.6	84.8	80.7	89.2	3.7	9.1	27.0	38.4	74.6	67.2	50.2	47.4
Silh.+Skel.	BiFusion	-	-	-	62.6	67.6	66.3	66.0	-	-	-	-	-	-	-	-
Silh.+Skel.	SkeletonGait++	-	-	-	79.1	83.9	81.7	89.9	-	-	-	-	-	-	-	-
Silh.+Pars.	XGait	-	-	-	72.8	77.0	79.1	80.5	-	-	-	-	-	-	-	-
Silh.+Pars.+Flow	MultiGait++	-	-	-	83.9	89.0	86.0	91.5	-	-	-	-	-	-	-	-
RGB+Silh.	GaitEdge	UNet	-	-	66.9	74.0	70.6	77.1	-	-	8.9	19.6	66.5	58.7	44.8	-
RGB+Silh.	DenoisingGait	SD & DeepLabV3+	-	-	84.0	88.0	90.1	95.9	-	-	37.3	59.1	83.9	76.1	34.8	-
RGB	BigGait	DINOv2-S	30.8M	12.7G	82.6	85.9	87.1	93.1	7.4	16.3	43.7	56.4	77.4	71.5	33.6	53.0
RGB	BiggerGait	SAM-S	209.7M	95.1G	86.8	89.5	91.8	95.9	9.5	17.8	64.6	75.5	80.1	76.2	31.8	59.7
RGB	BiggerGait	CLIP-S	208.6M	49.3G	81.0	86.2	89.7	96.9	15.2	26.1	67.0	**84.2**	91.5	87.8	47.3	65.8
RGB	BiggerGait	DINOv2-S	142.2M	45.7G	**89.8**	**92.1**	93.7	**97.5**	**15.5**	**26.9**	**70.9**	79.6	**93.0**	**90.8**	**55.6**	**67.0**
RGB	BiggerGait*	SAM-S	112.2M	79.2G	86.9	89.4	92.3	95.8	9.1	17.2	60.8	74.4	79.7	74.9	28.9	59.0
RGB	BiggerGait*	CLIP-S	110.8M	33.4G	78.9	83.8	87.9	96.1	13.9	24.2	63.1	81.5	92.3	87.1	42.9	64.0
RGB	BiggerGait*	DINOv2-S	43.6M	29.8G	89.0	91.9	**94.0**	97.2	14.5	25.3	69.5	80.4	91.6	87.7	54.7	66.5

ungrouped setup, except the CLIP-S-based one. Therefore, we set $P = 2$ and $J = 6$ for BiggerGait*, cutting roughly 108.8M and 23.8G FLOPs for DINOv2-S-based one. In this setting, DINOv2-S-based BiggerGait* achieves a 44% speedup (29.89 ms / image), approaching BigGait (21.64 ms).

Scaling the LVM Size Tab. 2(c) shows BiggerGait* offers marginal benefits, saving FLOPs limited while hurting performance. We consider that the upstream LVM dominates computation ($\approx 87.5\%$ for DINOv2-L), basicly making BiggerGait's overhead negligible compared to the expensive LVM's cost. Therefore, for larger LVM cases, the standard BiggerGait is recommended.

Efficiency Comparison. Tab. 2(d) shows that BiggerGait*, especially for DINOv2-S-based one, has similar FLOPs as the popular gait methods. This result indicates that the BiggerGait's superiority stems not from increased parameters or FLOPs, but from diverse layer-wise LVM features.

Parameter & FLOPs Comparison. Tab. 3 further confirms that despite its significant performance gains, the DINOv2-S-based BiggerGait* maintains FLOPs and size comparable to popular gait methods. We include the cost of DeepLabV3+ (26.8M parameters, 43.7 GFLOPs) for mask extraction in silhouette-based methods. Statistics for multimodal approaches are unavailable due to reproduction difficulty, yet they clearly incur higher computation costs from additional preprocessing models.

5 Conclusions

This work shifts the attention of LVM-based gait research from well-designed gait priors to the fundamental properties of LVMs itself. Our comprehensive study shows that layer-wise representations in LVM contain rich, distinct gait semantics. Without relying on elaborate gait priors, integrating these diverse layer-wise features delivers substantial gains. Building on these insights, we propose BiggerGait, a simple yet universal layer-wise LVM framework for gait recognition. We systematically analyze the inherent efficiency challenges of layer-wise methods and introduce an optional mitigation strategy. Comprehensive evaluations on CCPG, CASIA-B*, SUSTech1K and CCGR_MINI reveal BiggerGait's advance in most within- and cross-domain tasks. The work may also provide inspiration for employing the layer-wise knowledge produced by LVMs for other vision tasks.

Limitation. While BiggerGait sets impressive results in gait tasks, its feature extraction is mainly at the image level. The temporal feature remains underexplored in this work and deserves further study. Meanwhile, predicting the most effective layers for new, unseen datasets remains an open challenge, as the optimal layer often shifts with training data and task types.

Acknowledgement. This work was supported partially by National Natural Science Foundation of China (Grant 62476120, 62325307, 62422312, and 62506236) and partially by National Key R&D Program of China (Grant 2020YFA0908700 and 2023YFB4704900).

References

[1] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 3

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. 3

[3] K. Cao and A. K. Jain. Automated latent fingerprint recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(4):788–800, 2018. 1

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017. 1, 3

[5] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(7):3467–3478, 2021. 1, 3, 8, 18

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 9

[7] N. Chen, N. Wu, S. Liang, M. Gong, L. Shou, D. Zhang, and J. Li. Is bigger and deeper always better? probing llama across scales and layers. *arXiv preprint arXiv:2312.04333*, 2023. 2, 3, 5

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] U. Evci, V. Dumoulin, H. Larochelle, and M. C. Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning (ICML)*, pages 6009–6033. PMLR, 2022. 3

[11] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14233, 2020. 1, 3, 8, 18

[12] C. Fan, S. Hou, Y. Huang, and S. Yu. Exploring Deep Models for Practical Gait Recognition. *arXiv preprint arXiv:2303.03301*, 2023. 8

[13] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 8, 18

[14] C. Fan, S. Hou, J. Liang, C. Shen, J. Ma, D. Jin, Y. Huang, and S. Yu. Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *arXiv preprint arXiv:2405.09138*, 2024. 1, 3, 4

[15] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 38, pages 1662–1669, 2024. 1, 3, 8

[16] S. Fan, X. Jiang, X. Li, X. Meng, P. Han, S. Shang, A. Sun, Y. Wang, and Z. Wang. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*, 2024. 2, 3, 5

[17] X. Gu, T. Pang, C. Du, Q. Liu, F. Zhang, C. Du, Y. Wang, and M. Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024. 3

[18] X. Guo, X. Song, Y. Zhang, X. Liu, and X. Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116, 2025. 1

[19] Y. Guo, S. Huang, R. Prabhakar, C. P. Lau, R. Chellappa, and C. Peng. Distillation-guided representation learning for unconstrained gait recognition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11, 2024. 3

[20] Y. Guo, A. Shah, J. Liu, A. Gupta, R. Chellappa, and C. Peng. Gaitcontour: Efficient gait recognition based on a contour-pose representation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1051–1061, 2025. 3

[21] Z. Huang and X. Liu. Generalizable object re-identification via visual in-context prompting. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, page [in press], 2025. 1

[22] Z. Huang, X. Liu, and Y. Kong. H-more: Learning human-centric motion representation for action analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22702–22713, 2025. 1

[23] A. K. Jain, A. A. Ross, K. Nandakumar, and T. Swearingen. Additional biometric traits. In *Introduction to Biometrics*, pages 245–287. Springer, 2024. 1

[24] C. Jin, J. Xu, B. Liu, L. Tao, O. Golovneva, T. Shu, W. Zhao, X. Li, and J. Weston. The era of real-world human interaction: Rl from user conversations. *arXiv preprint arXiv:2509.25137*, 2025. 3

[25] D. Jin, C. Fan, W. Chen, and S. Yu. Exploring more from multiple gait modalities for human identification. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025. 2, 8

[26] D. Jin, C. Fan, J. Ma, J. Zhou, W. Chen, and S. Yu. On denoising walking videos for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12347–12357, 2025. 1, 2, 3, 4, 8

[27] M. Kim, Y. Su, F. Liu, A. Jain, and X. Liu. Keypoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–255, 2024. 1

[28] M. Kim, A. Jain, and X. Liu. 50 years of automated face recognition. *arXiv preprint arXiv:2505.24247*, 2025. 1

[29] M. Kim, D. Ye, Y. Su, F. Liu, and X. Liu. Sapiensid: Foundation for human recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13937–13947, 2025. 1

[30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything, 2023. 2, 3, 5, 8, 9

[31] Z. Kong, Y. Li, F. Zeng, L. Xin, S. Messica, X. Lin, P. Zhao, M. Kellis, H. Tang, and M. Zitnik. Token reduction should go beyond efficiency in generative models – from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025. 3

[32] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 9

[33] W. Li, S. Hou, C. Zhang, C. Cao, X. Liu, Y. Huang, and Y. Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13824–13833, 2023. 2, 3, 5, 6, 8, 9

[34] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren. End-to-end Model-based Gait Recognition. In *Asian Conference on Computer Vision (ACCV)*, page [no pagination], 2020. 1, 3

[35] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 8

[36] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition*, pages 474–483. Springer, 2017. 1, 3

[37] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021. 1, 3, 8, 18

[38] F. Liu, M. Kim, A. Jain, and X. Liu. Controllable and guided face synthesis for unconstrained face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 701–719. Springer, 2022. 1

[39] F. Liu, R. Ashbaugh, N. Chimitt, N. Hassan, A. Hassani, A. Jaiswal, M. Kim, Z. Mao, C. Perry, Z. Ren, et al. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *IEEE Winter Conference on Applications of Computer Vision*, pages 6227–6236, 2024. 1

[40] F. Liu, N. Chimitt, L. Guo, J. Jain, A. Kane, M. Kim, W. Robbins, Y. Su, D. Ye, X. Zhang, et al. Person recognition at altitude and range: Fusion of face, body shape and gait. *arXiv preprint arXiv:2505.04616*, 2025. 1

[41] K. Liu, O. Choi, J. Wang, and W. Hwang. CDGNet: Class Distribution Guided Network for Human Parsing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4473–4482, 2022. 1, 3

[42] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019. 3

[43] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6):[no pagination], 2015. 1, 3

[44] J. Ma, D. Ye, C. Fan, and S. Yu. Pedestrian attribute editing for gait recognition and anonymization. *arXiv preprint arXiv:2303.05076*, 2023. 4

[45] K. Narayan, V. VS, and V. M. Patel. Facexbench: Evaluating multimodal llms on face understanding. *arXiv preprint arXiv:2501.10360*, 2025. 1

[46] K. Narayan, V. Vs, and V. M. Patel. Segface: Face segmentation of long-tail classes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 6182–6190, 2025. 1

[47] M. S. Nixon and J. N. Carter. Automatic Recognition by Gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006. 1

[48] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 5, 7, 8, 9

[49] L. Parzianello and A. Czajka. Saliency-guided textured contact lens-aware iris recognition. In *IEEE Winter Conference on Applications of Computer Vision*, pages 330–337, 2022. 1

[50] Y. Peng, K. Ma, Y. Zhang, and Z. He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, 83(3):7273–7294, 2024. 8

[51] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha. Secure and robust iris recognition using random projections and sparse representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(9):1877–1893, 2011. 1

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PmLR, 2021. 2, 3, 5, 8, 9

[53] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(1):121–135, 2017. 1

[54] Z. Ren, Y. Su, and X. Liu. Chatgpt-powered hierarchical comparisons for image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:69706–69718, 2023. 3

[55] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 8

[56] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1063, 2023. 2, 3, 5, 6, 8

[57] C. Shen, B. Lin, S. Zhang, X. Yu, G. Q. Huang, and S. Yu. Gait recognition with mask-based regularization. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 1, 3

[58] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1, 3

[59] O. Skean, M. R. Arefin, Y. LeCun, and R. Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024. 2, 3

[60] O. Skean, M. R. Arefin, D. Zhao, N. Patel, J. Naghiyev, Y. LeCun, and R. Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025. 2, 3, 5

[61] Y. Su, M. Kim, F. Liu, A. Jain, and X. Liu. Open-set biometrics: Beyond good closed-set models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 243–261, 2024. 3

[62] Q. Sun, M. Pickett, A. K. Nain, and L. Jones. Transformer layers as painters. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 25219–25227, 2025. 2, 3, 7

[63] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Gaitgraph: Graph Convolutional Network for Skeleton-Based Gait Recognition. In *International Conference on Image Processing (ICIP)*, pages 2314–2318, 2021. 1, 3

[64] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473–1488, 2008. 3

[65] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10):3349–3364, 2020. 9

[66] J. Wang, S. Hou, X. Guo, Y. Huang, Y. Huang, T. Zhang, and L. Wang. Gaitc 3 i: Robust cross-covariate gait recognition via causal intervention. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2025. 3

[67] R. Wang, C. Shen, C. Fan, G. Q. Huang, and S. Yu. Pointgait: Boosting end-to-end 3d gait recognition with point clouds via spatiotemporal modeling. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 4

[68] R. Wang, C. Shen, M. J. Marin-Jimenez, G. Q. Huang, and S. Yu. Cross-modality gait recognition: Bridging lidar and camera modalities for human identification. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2024. 4

[69] Z.-Y. Wang, J. Liu, R. P. Kathirvel, C. P. Lau, and R. Chellappa. Hypergait: A video-based multitask network for gait recognition and human attribute estimation at range and altitude. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2024. 3

[70] Z.-Y. Wang, J. Liu, Y. Guo, J. Chen, and R. Chellappa. Unigait: A unified transformer-based multitask framework for gait analysis in the wild. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2025. 3

[71] Z.-Y. Wang, Z. Shao, J. Chen, and R. Chellappa. Combo-gait: Unified transformer framework for multi-modal gait recognition and attribute analysis. *arXiv preprint arXiv:2510.10417*, 2025. 3

[72] S. Yang, J. Wang, S. Hou, X. Liu, C. Cao, L. Wang, and Y. Huang. Bridging gait recognition and large language models sequence modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, 2025. 1, 3

[73] D. Ye, J. Ma, C. Fan, and S. Yu. Gaitediter: Attribute editing for gait representation learning. *arXiv e-prints*, pages arXiv–2303, 2023. 3

[74] D. Ye, C. Fan, J. Ma, X. Liu, and S. Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210, 2024. 1, 2, 3, 4, 8, 9

[75] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006. 2, 5, 6, 8

[76] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 4

[77] J. Zheng, X. Liu, S. Wang, L. Wang, C. Yan, and W. Liu. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 116–124, 2023. 1, 3

[78] J. Zheng, X. Liu, B. Zhang, C. Yan, J. Zhang, W. Liu, and Y. Zhang. It takes two: Accurate gait recognition in the wild via cross-granularity alignment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 8786–8794, 2024. 8

[79] J. Zhu, Y. Su, M. Kim, A. Jain, and X. Liu. A quality-guided mixture of score-fusion experts framework for human recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, page [in press], 2025. 1

[80] S. Zou, C. Fan, J. Xiong, C. Shen, S. Yu, and J. Tang. Cross-Covariate Gait Recognition: A Benchmark. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 38, pages 7855–7863, 2024. 1, 2, 5, 6, 8

[81] S. Zou, J. Xiong, C. Fan, C. Shen, S. Yu, and J. Tang. A multi-stage adaptive feature fusion neural network for multimodal gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 4

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly claims our main contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Sec. 5, we discuss the some potential challenges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For all theoretical components in Sec. 3, we provide relative experimental results in Sec. 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The reproduce details are shown in Sec. 4.1, and all the source code will be released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: While all the source code will be released, it is not included in the paper now.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting can be found in Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper use Rank-1 as main results, which is following prior works[5, 11, 37, 13].

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are reported in Sec. 4.1, and Tab. 2 (d).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Gait recognition helps security applications, which is mentioned in Sec. 1.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used in this paper are public and cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: No new human subject assets are collected in this study. All human subject assets are exclusively from publicity available datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.