GaitLingo: Self-Supervised Gait Representation Learning with Language Priors

Chenye Wang[1] **Zhengxiang Lan**[2] **Saihui Hou**[2] **Zhikang Liu**[3]
Xingqun Qi[4] **Yongzhen Huang**[2,5] **Sirui Han**[1]*
[1] The Hong Kong University of Science and Technology
[2] Beijing Normal University
[3] Afari Intelligent Drive.
[4] North China Electric Power University
[5] WATRIX.AI

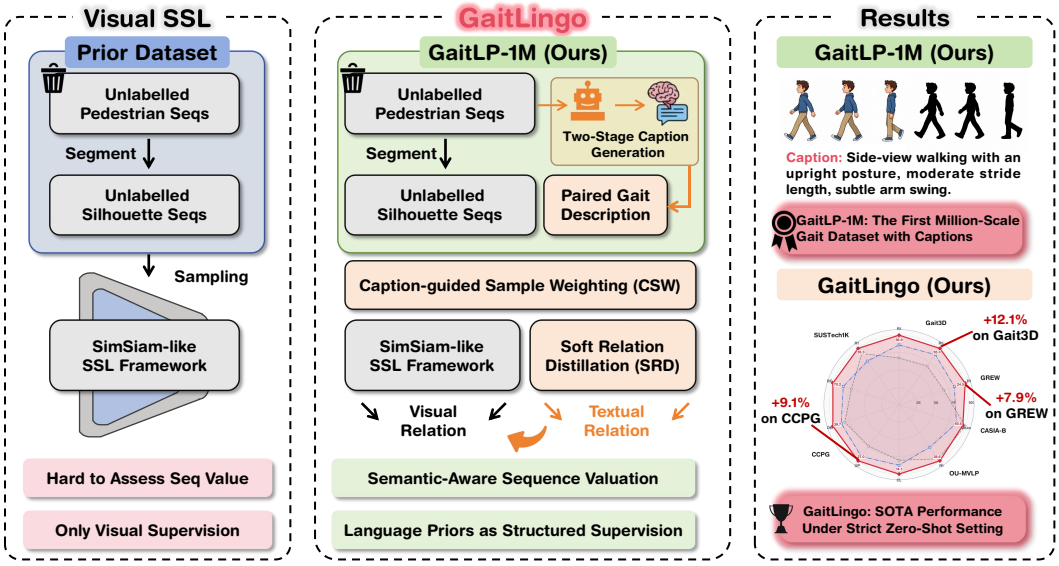


Figure 1: **Overview of GaitLingo.** Conventional visual self-supervised gait representation learning lacks semantic awareness. We introduce GaitLP-1M, **the first million-scale gait dataset paired with gait-centric captions**, and propose GaitLingo to integrate language priors into a self-supervised pretraining framework. By exploiting semantic sequence valuation and structural relations, GaitLingo achieves **superior zero-shot performance** across six gait benchmarks.

Abstract

Self-supervised gait representation learning aims to learn transferable representations that generalize to unseen scenarios. Existing visual-only approaches provide limited guidance for assessing the semantic value of individual sequences and exploiting relations across samples. To address this limitation, we introduce language priors into self-supervised gait pretraining. We construct GaitLP-1M, a million-scale silhouette-based gait pretraining dataset without identity labels. A two-stage caption generation pipeline further produces gait descriptions from the corresponding RGB sequences. To the best of our knowledge, GaitLP-1M is the first million-scale silhouette gait dataset paired with gait-centric captions. Building on this foundation, we propose GaitLingo, a language-guided self-supervised framework for gait representation learning. Specifically, we design a Caption-

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Guided Sample Weighting (CSW) mechanism to prioritize informative samples based on sequence-level semantic richness and attribute rarity. We then introduce a lightweight Temporal Adapter (TA) to better capture motion semantics and improve cross-domain robustness. Finally, we propose Soft Relation Distillation (SRD) to transfer caption-derived relational structures into the visual embedding space. Extensive experiments demonstrate that GaitLingo consistently outperforms prior self-supervised gait pretraining methods in the zero-shot setting across six benchmarks. Compared with GaitSSB, it achieves Rank-1 gains of 12.1 percentage points on Gait3D and 7.9 percentage points on GREW, as well as an average improvement of 9.1 percentage points across four CCPG settings.

1 Introduction

Gait recognition has been widely regarded as a promising biometric technology for open-world scenarios due to its long-range, non-contact, and hard-to-disguise characteristics [1, 48, 51]. Driven by the rapid development of deep learning and the availability of large-scale datasets, gait recognition methods have made remarkable progress and continuously advanced the state of the art on existing benchmarks [74, 33, 11]. However, most gait recognition models still degrade significantly in cross-domain scenarios [12]. The central challenge of gait recognition is no longer merely improving accuracy, but learning robust gait representations that remain stable across domains.

Recently, the pretrain-then-finetune paradigm has become a dominant recipe for visual representation learning [19, 3, 15, 76]. In gait recognition, self-supervised pretraining leverages massive unlabeled data to learn transferable representations without relying on costly identity annotations. A representative work, GaitSSB [12], introduced the first million-scale unlabeled gait pretraining benchmark and demonstrated the effectiveness of instance discrimination for self-supervised gait learning.

However, the supervisory signals in existing self-supervised gait learning methods are still entirely derived from the visual space itself. Positive pairs are typically constructed based on visual augmentation consistency, with no guidance from a higher-level semantic space. Such a visual-only formulation makes it difficult to assess the semantic richness of individual sequences or capture latent semantic relations across samples. Consequently, although current datasets are quantity-rich, they remain structure-poor, which limits pretraining efficiency and prevents models from fully exploiting the supervision hidden in large-scale unlabeled data.

To address this issue, we leverage large language models (LLMs) and vision-language models (VLMs) to generate high-quality textual descriptions for large-scale unlabeled gait sequences. This opens up a new opportunity to introduce language priors into self-supervised gait learning. Compared with purely visual self-supervision, textual descriptions provide higher-level semantic interpretations of gait sequences. Specifically, language priors offer three key advantages. (1) At the intra-sequence level, captions provide rich semantic descriptions for each gait sequence, enabling us to *quantify the semantic richness of unlabeled samples*. (2) At the inter-sequence level, captions provide *continuous semantic relations across samples*. Textual descriptions encode semantic relation priors between sequences, offering structured soft supervision that encourages the model to preserve semantic ranking relations. (3) From a practical perspective, caption-based supervision preserves the privacy-friendly nature of silhouettes while remaining highly scalable to million-scale unlabeled datasets at low cost.

Motivated by the above, we reformulate self-supervised gait pretraining from pure visual contrastive learning into language-guided structure learning. We first construct GaitLP-1M, a million-scale silhouette-based gait pretraining dataset, where each gait sequence is paired with a descriptive caption. These descriptions are generated by a two-stage pipeline that enhances identity discriminability while preserving semantic accuracy. Built upon this dataset, we propose GaitLingo, a language-guided self-supervised framework for gait representation learning. GaitLingo consists of three key components. First, to address the varying semantic richness across samples, we introduce a Caption-Guided Sample Weighting (CSW) mechanism, which assigns larger sampling weights to semantically richer sequences. Second, considering that caption-derived priors also contain dynamic gait semantics, we introduce a lightweight Temporal Adapter (TA) to enhance temporal modeling capacity and improve robustness during dataset transfer. Finally, to exploit semantic relation priors across samples, we introduce Soft Relation Distillation (SRD), which transfers textual similarity structures into visual

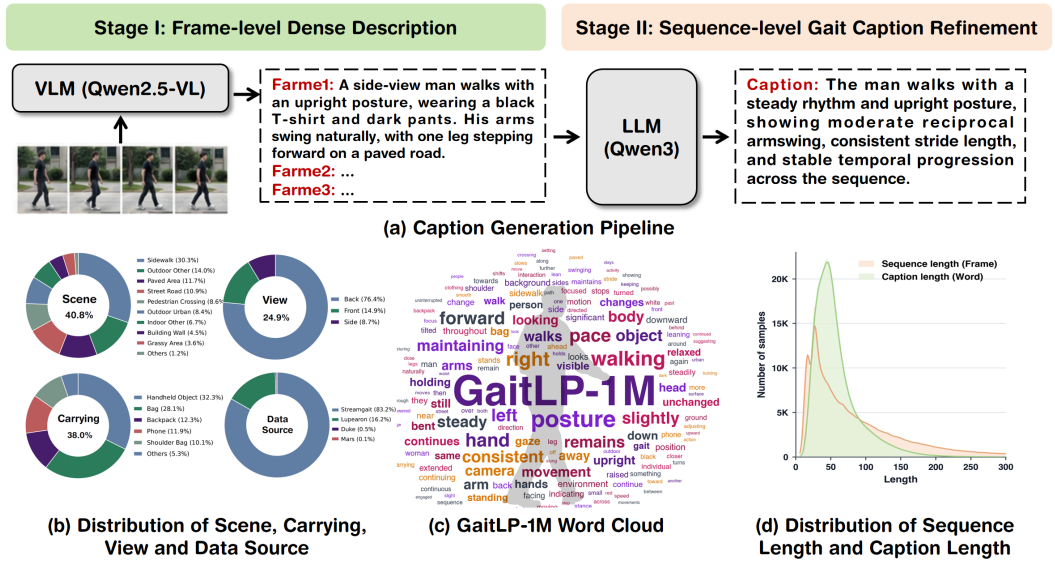


Figure 2: **Overview of GaitLP-1M.** (a) The proposed two-stage caption generation pipeline. (b) Distributions of scene, carrying condition, view, and data source, showing the diversity of the collected gait sequences. (c) A word cloud summarizing frequent semantic concepts in the generated captions. (d) Distributions of sequence and caption lengths, indicating that GaitLP-1M contains temporally diverse sequences and rich language descriptions.

representation learning and encourages the model to preserve continuous semantic relations beyond instance discrimination.

Under a strict zero-shot setting, GaitLingo consistently outperforms the prior self-supervised method GaitSSB across six benchmarks. On Gait3D and GREW, it improves Rank-1 accuracy by 12.1 and 7.9 percentage points, respectively. On CCPG, it achieves an average improvement of 9.1 percentage points. On SUSTech1K, Rank-1 accuracy increases from 50.3% to 55.3%. Our contributions are summarized as follows:

(1) We construct **GaitLP-1M**, the first large-scale unsupervised gait silhouette dataset with sequence-level textual captions. It contains over 1.02 million gait sequences, with an average sequence length of 90 frames and an average caption length of 75 words, providing a data foundation for language-guided unsupervised gait pretraining.

(2) We propose **GaitLingo**, a language-guided self-supervised gait pretraining framework. Specifically, Caption-Guided Sample Weighting improves data utilization by emphasizing semantically informative sequences, Temporal Adapter enhances temporal robustness across datasets, and Soft Relation Distillation transfers caption-derived semantic relations into the visual embedding space.

(3) Extensive experiments demonstrate that GaitLingo achieves state-of-the-art zero-shot performance across six major gait benchmarks. It also remains competitive under standard fine-tuning settings, validating the effectiveness of language priors for learning robust and transferable gait representations.

2 Dataset Construction

To support language-guided self-supervised gait representation learning, we construct GaitLP-1M, a large-scale silhouette gait dataset without identity labels but with sequence-level captions. This dataset serves as the foundation of our language-guided pretraining framework.

2.1 Data Collection

We construct GaitLP-1M by merging several existing large-scale pedestrian video datasets that contain abundant walking sequences captured across diverse environments. Specifically, we collect data from four public datasets: *LUPerson-NL* [16], *MARS-Attribute* [75], *DukeMTMC-VID-Attribute* [64], and *StreamGait* [43]. These datasets cover a wide range of real-world scenarios, including university

Table 1: **Comparison with existing silhouette-based gait datasets.** GaitLP-1M is the first million-scale silhouette gait dataset paired with gait-centric captions.

Dataset	#IDs	#Seqs	Covariates	Unlabeled	Gait Captions
CASIA-B [73] ICPR'06	124	13,607	View, Clothing	✗	✗
OU-MVLP [53] IPSJ T-CVA'18	10,307	267,388	View	✗	✗
GREW [79] ICCV'21	26,000	126,887	Diverse	✗	✗
Gait3D [74] CVPR'22	4,000	25,309	Diverse	✗	✗
CCPG [35] CVPR'23	200	16,365	Clothing	✗	✗
SUSTech1K [50] CVPR'23	1,050	25,239	Diverse	✗	✗
CCGR [80] AAAI'24	970	1,580,617	Diverse	✗	✗
BarbieGait [2] CVPR'26	521	1,203,324	Diverse	✗	✗
MMGait [59] CVPR'26	725	334,060	Diverse	✗	✗
GaitLU-1M [12] TPAMI'23	943,884	943,884	Diverse	✓	✗
GaitLP-1M Ours	**1,027,043**	**1,027,043**	**Diverse**	✓	✓

campuses, street surveillance, and unconstrained in-the-wild environments, providing rich diversity in viewpoints, carrying conditions, occlusions, and motion patterns.

During preprocessing, we use a human segmentation model [38] to extract pedestrian silhouettes, removing irrelevant cues such as background, clothing, and texture. Each silhouette is further aligned and resized to a fixed resolution of 64×44, following the standard setting in silhouette-based gait recognition [13]. To ensure dataset quality, we perform data cleaning to remove noisy samples, including extremely short or long sequences and segmentation failures. After filtering, we retain over one million high-quality gait sequences.

Figure 2 provides an overview of GaitLP-1M. The average silhouette sequence length of GaitLP-1M is approximately 90 frames. The generated captions contain approximately 75 words on average, providing rich sequence-level language supervision. Table 1 compares the statistics of existing public silhouette-based gait datasets. Compared with prior datasets, GaitLP-1M is the first large-scale unlabeled gait dataset paired with reliable textual annotations. More visualizations are provided in Appendix B.

2.2 Caption Generation Pipeline

Recent progress in multimodal foundation models has shown strong potential for extracting high-level semantics from visual signals [47, 65, 52, 7, 4, 46, 37, 72]. However, converting large-scale pedestrian sequences into effective language supervision for gait representation learning remains non-trivial. In our preliminary exploration, directly prompting VLMs to generate gait-centric captions from RGB sequences often produced over-abstracted descriptions. This is because a VLM must simultaneously perceive fine-grained visual details and filter out appearance- or scene-related nuisance factors. As a result, the generated captions tend to collapse into generic motion templates, leading to high similarity across sequences and limited discriminative value for pretraining.

To address this challenge, we decompose caption generation into two stages with clearly separated objectives, as shown in Figure 2. The first stage performs dense visual description, in which the VLM is encouraged to preserve observable information from sampled frames without prematurely filtering potentially useful cues. The second stage performs language-level refinement, in which an LLM removes appearance-related confounders and reorganizes the retained information into sequence-level gait descriptions. This decoupled design better matches the strengths of the two model types: VLMs are used for detailed visual perception, while LLMs are used for semantic abstraction and denoising.

Stage I: Frame-level Dense Description. The goal of the first stage is to convert raw visual observations into a high-recall textual representation before any gait-specific filtering is applied. For each sequence, we sample every third frame throughout the sequence, obtaining temporally ordered observations that cover the complete walking process. The sampled frames are then fed into *Qwen2.5-VL-7B-Instruct* [55] to generate frame-level dense descriptions.

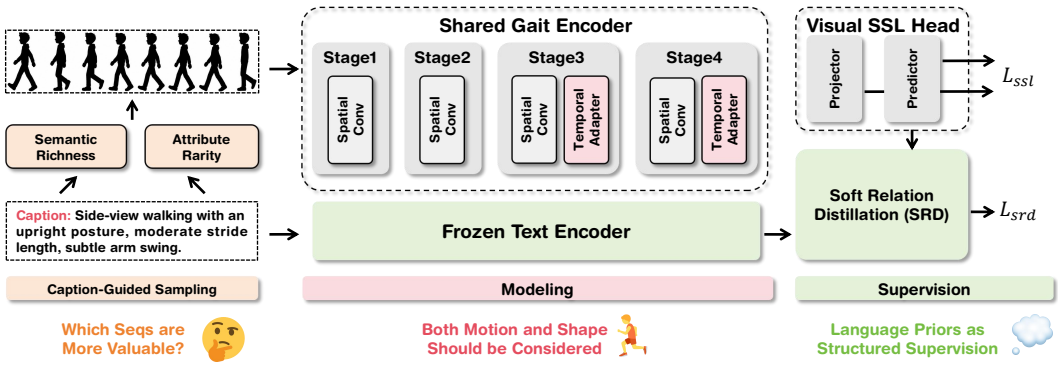


Figure 3: **GaitLingo Framework.** GaitLingo introduces language priors into self-supervised gait pretraining from three perspectives: sampling, modeling, and supervision. Text-related components are removed at inference. Only the shared gait encoder and visual SSL head are used for recognition.

Stage II: Sequence-level Gait Caption Refinement. The dense descriptions produced in Stage I contain useful motion and body-structure information, but they also include redundant frame-level observations and nuisance factors such as color, texture, and background context. The second stage therefore performs language-level refinement using *Qwen3-8B* [66]. Given the frame-level descriptions, the language model is prompted to remove appearance- and scene-related information, merge repetitive content, and reorganize the remaining observations into a temporally coherent sequence-level description.

In Stage I, the VLM extracts observable frame-level information with high recall. In Stage II, explicit instructions remove clothing color, texture, background, and other appearance- or scene-related confounders. Thus, clothing-related appearance cues are explicitly filtered during caption construction rather than directly transferred into gait representation learning.

We further conduct a quality assessment of the generated captions, with detailed results provided in Appendix B. In addition, RGB sequences are used only for caption generation and are removed afterward to protect participant privacy. GaitLP-1M retains only anonymized silhouette sequences and descriptive gait-centric captions. These captions focus on gait-centric semantics while excluding identity-sensitive appearance details, such as facial information, color, and texture.

3 Method

This section presents **GaitLingo**, a language-guided self-supervised framework for gait representation pretraining. As illustrated in Figure 3, GaitLingo retains the visual self-supervised pretraining objective while introducing language priors through three components: Caption-Guided Sample Weighting (CSW), Temporal Adapter (TA), and Soft Relation Distillation (SRD).

3.1 Preliminaries: Self-Supervised Gait Representation Learning

In this section, we review the prevailing framework for self-supervised gait representation learning. Given an unlabeled walking sequence, two non-overlapping clips are sampled as a positive pair, while clips from other sequences are treated as negatives. After augmentation, the sampled clips are processed by a Shared Gait Encoder followed by a Visual SSL Head.

The Shared Gait Encoder consists of a four-stage backbone and temporal-spatial aggregation modules, including Temporal Pooling (TP) and Horizontal Pyramid Pooling (HPP). The Visual SSL Head contains part-wise projectors and predictors. The optimization objective is an instance discrimination task, where two clips from the same sequence are encouraged to be pulled together in the embedding space, while clips from different sequences are pushed apart. This objective is implemented with an InfoNCE loss [21] and a stop-gradient operation to mitigate representation collapse [6].

$$\mathcal{L}_{\text{ssl}} = \frac{1}{2}\left[\text{InfoNCE}(q, \text{sg}(k)) + \text{InfoNCE}(k, \text{sg}(q))\right], \tag{1}$$

where q and k denote the two augmented clip representations, $\text{sg}(\cdot)$ denotes the stop-gradient operation, and $\text{InfoNCE}(\cdot, \cdot)$ contrasts the positive clip pair against clips sampled from other sequences.

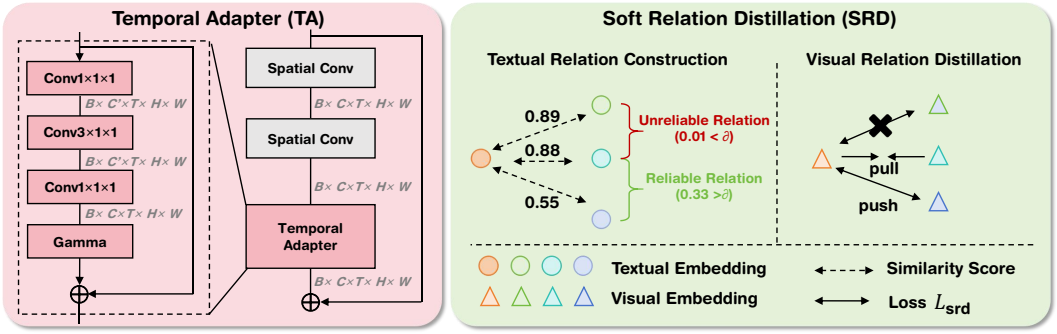


Figure 4: **Details of Temporal Adapter and Soft Relation Distillation.** The Temporal Adapter adopts a bottleneck architecture with γ initialized to zero. Soft Relation Distillation transfers textual priors into the visual space via ranking consistency.

3.2 Caption-Guided Sample Weighting

To convert the generated captions into a lightweight sampling prior, we propose a Caption-Guided Sample Weighting (CSW) mechanism. The key idea is to estimate the training value of each sequence from two complementary perspectives: ***semantic richness*** and ***attribute rarity***. Semantic richness measures how much gait-relevant information a sequence contains, while attribute rarity emphasizes infrequent yet potentially discriminative attributes.

Concretely, given a sequence X_i with frame-level dense captions and a predefined attribute set $\mathcal{A} = \{a_m\}_{m=1}^{M_a}$, we first parse the free-form descriptions into attribute responses. We then quantify the temporal occupancy of each visible attribute as $v_{i,m} \in [0, 1]$, which denotes the proportion of frames in X_i where attribute a_m is present. A larger $v_{i,m}$ indicates a stronger response of attribute a_m in sequence X_i.

Semantic richness. A sequence with more diverse and stable attribute responses usually contains richer semantic information for pretraining. We define the semantic richness score as

$$R_i = \sum_{(a_m, v_{i,m})} v_{i,m}. \tag{2}$$

Attribute rarity. Besides richness, attribute frequency also matters. In large-scale unlabeled data, common patterns may dominate the training distribution, whereas rare covariates may provide stronger cues for robust representation learning. We therefore compute an inverse-frequency weight for each attribute. The rarity score of sequence X_i is then defined as

$$L_i = \sum_{(a_m, v_{i,m})} v_{i,m} q_m, \quad q_m = \log\left(1 + \frac{1}{\sum_{i=1}^{N} v_{i,m} + \epsilon}\right), \tag{3}$$

where N denotes the number of sequences in the dataset, and ϵ is a small constant for numerical stability. We normalize R_i and L_i over the dataset and combine them into a unified sampling weight:

$$w_i = \alpha R_i + (1 - \alpha) L_i, \tag{4}$$

where α controls the trade-off between semantic richness and attribute rarity.

3.3 Temporal Adapter

Caption-derived priors contain not only static cues but also motion semantics, such as walking posture and dynamic patterns. To model both motion and shape information, we introduce a lightweight and plug-and-play Temporal Adapter (TA) into the DeepGaitV2 backbone [11]. The design follows a late-stage adaptation strategy.

As shown on the left of Figure 4, stages 1 and 2 employ standard spatial residual blocks, while stages 3 and 4 insert a Temporal Adapter after the original spatial convolution layers. The Temporal Adapter adopts a bottleneck design for efficiency. It first reduces the channel dimension from C

to $C' = \lfloor C/r \rfloor$ using a $1 \times 1 \times 1$ convolution, followed by BatchNorm and ReLU. A depthwise temporal convolution with kernel size $3 \times 1 \times 1$ is then applied along the temporal dimension. Finally, another $1 \times 1 \times 1$ convolution restores the channel dimension, followed by BatchNorm. The temporal features are injected into the main stream through a gated residual connection [20]: $X + \gamma TA(X)$, where γ is a learnable scalar initialized to zero. This initialization allows the adapter block to start from the original 2D backbone and gradually incorporate temporal information during training.

3.4 Soft Relation Distillation

Beyond measuring the value of individual samples, captions also provide soft relations among different walking sequences. Built upon the instance-discriminative objective in GaitSSB, we introduce a complementary task to model such relational structure.

As shown in Figure 3 and on the right of Figure 4, given a query sequence within a batch, the textual embeddings induce a ranking over other sequences based on their semantic similarity. We transfer this relational structure into the visual domain by encouraging visual embeddings to preserve the same relative ordering. That is, if the captions suggest that one sequence is semantically closer to the query than another, the visual similarity should reflect the same preference.

Textual relation construction. Captions are first encoded by a frozen text encoder to obtain textual embeddings, followed by normalization. We then compute pairwise textual similarities within a batch. Given a query sequence, the textual embeddings induce a ranking over other sequences according to their semantic similarity. We define *reliable relations* as pairs with sufficiently large similarity gaps. Specifically, for a query i, we construct triplets (i, j^+, j^-) such that

$$S^t_{i,j^+} - S^t_{i,j^-} > \delta, \tag{5}$$

where S^t denotes textual similarity and δ is the textual separation threshold. Only these reliable relations are used for supervision. In each triplet, j^+ corresponds to a semantically closer sequence, while j^- is less similar.

Visual relation distillation. On the visual side, we aggregate part-wise similarities via max pooling to obtain the sequence-level similarity S^v. SRD encourages the visual similarity to preserve the same ordering as the textual relation:

$$\mathcal{L}^{q\to k}_{\mathrm{srd}} = \frac{1}{|\mathcal{T}|} \sum_{(i,j^+,j^-)\in\mathcal{T}} \max\left(0,\ m - (S^v_{i,j^+} - S^v_{i,j^-})\right), \tag{6}$$

where $\mathcal{T}$ denotes the set of reliable triplets constructed from textual similarities and m is the visual ranking margin. The reverse loss $\mathcal{L}^{k\to q}_{\mathrm{srd}}$ is computed symmetrically, and the final relation loss is

$$\mathcal{L}_{\mathrm{srd}} = \frac{1}{2}\left(\mathcal{L}^{q\to k}_{\mathrm{srd}} + \mathcal{L}^{k\to q}_{\mathrm{srd}}\right). \tag{7}$$

SRD avoids explicit cross-modal feature alignment and instead enforces *ordering consistency*, which preserves high-confidence semantic relations while mitigating noise from caption inaccuracies. This design prevents representation collapse and enables effective integration of language-derived relational structure into visual representation learning.

3.5 Training Objective and Inference

The overall pretraining objective combines the original visual self-supervised loss with the caption-induced relation loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ssl}} + \lambda_{\mathrm{srd}}\mathcal{L}_{\mathrm{srd}}, \tag{8}$$

where λ_{srd} controls the strength of relation distillation.

Inference. Textual information is used only during pretraining. At inference time, captions, text embeddings, SRD supervision, and the CSW sampler are removed. The model takes only silhouette sequences as input and introduces no additional dependency on language models.

Downstream evaluation. We follow the same zero-shot and fine-tuning protocols as GaitSSB [12]. In the zero-shot setting, the pretrained model is directly evaluated on target datasets without using identity labels, and the retrieval feature is extracted after the predictor. In the fine-tuning setting, the pretrained encoder and projector initialize the downstream model, while a new recognition head is randomly initialized and trained with standard supervised objectives.

Table 2: **Zero-shot performance comparison.** Yellow cells indicate within-domain evaluations, while all others are cross-domain. **Bold** and underlined numbers denote the best and second-best self-supervised results, respectively.

Source	Method	Target Dataset											
		Gait3D		GREW	CASIA-B			OU-MVLP	CCPG				SUSTech1K
		R1	R5	R1	NM	BG	CL	R1	CL	UP	DN	BG	R1
Supervised Training													
Gait3D [74]	GaitSet [5]	40.5	60.0	19.2	66.6	48.7	12.8	22.9	9.4	17.4	28.8	54.6	27.8
	GaitPart [10]	22.9	41.3	14.2	61.8	47.0	13.7	20.3	9.0	16.6	23.6	42.8	21.4
	DeepGaitV2 [11]	73.4	87.2	35.7	81.7	68.8	15.2	46.4	13.8	27.0	46.2	81.2	50.6
GREW [79]	GaitSet [5]	17.8	31.6	47.9	66.0	45.8	19.8	22.1	7.9	19.0	22.3	41.5	19.8
	GaitPart [10]	14.1	25.8	47.6	69.2	52.1	25.4	25.4	7.6	16.2	17.7	31.2	18.8
	DeepGaitV2 [11]	33.1	49.1	77.8	77.1	56.3	30.2	31.3	13.3	30.6	28.9	61.9	27.7
CASIA-B [73]	GaitSet [5]	6.5	14.4	9.4	95.8	90.0	75.4	9.6	9.1	14.8	15.5	23.8	12.7
	GaitPart [10]	6.1	13.1	9.3	96.1	90.7	78.7	10.8	10.4	15.4	16.1	20.6	13.4
	DeepGaitV2 [11]	12.1	23.4	15.0	93.8	87.1	61.3	15.2	11.9	21.0	21.8	35.8	24.4
OU-MVLP [53]	GaitSet [5]	13.7	23.4	12.3	74.1	55.5	16.4	87.2	10.9	19.0	31.3	47.9	24.4
	GaitPart [10]	11.0	21.9	14.8	73.9	56.9	20.7	88.6	13.3	20.8	37.5	47.9	23.2
	DeepGaitV2 [11]	28.5	44.6	33.5	91.7	78.6	34.7	91.7	19.7	33.0	58.2	76.8	40.1
Self-supervised Pretraining													
GaitLU-1M [12]	GaitSSB [12]	24.8	38.0	16.6	83.8	**75.7**	**28.7**	37.2	11.3	20.0	32.5	50.6	50.3
[GaitLU-1M, GaitLP-1M]	Scaled GaitSSB	31.7	49.2	20.5	75.4	65.0	23.0	29.2	12.3	24.9	34.7	59.2	42.9
	GaitLingo	**36.9**	**55.3**	**24.5**	**84.4**	73.0	24.2	**38.6**	**14.1**	**27.0**	**39.7**	**70.2**	**55.3**

4 Experiments

4.1 Implementation Details

Datasets. We use two large-scale unlabeled datasets for self-supervised pretraining: GaitLU-1M [12] and the proposed GaitLP-1M. None of the target benchmark training or test splits is used during pretraining. For downstream evaluation, we adopt six widely used gait benchmarks covering diverse evaluation scenarios: Gait3D [74] and GREW [79] for in-the-wild recognition, CASIA-B [73] and OU-MVLP [53] for in-the-lab recognition, CCPG [35] for clothing-changing recognition, and SUSTech1K [50] for multi-covariate evaluation.

Baselines. The main self-supervised baseline is GaitSSB [12], which was originally pretrained on GaitLU-1M only. To ensure a fair comparison under the same data and model scale, we construct a scaled version of GaitSSB by merging GaitLU-1M and GaitLP-1M as the pretraining data and replacing the original backbone with DeepGaitV2 [11] using the layer configuration $[1, 4, 8, 4]$. For zero-shot evaluation, all target datasets are strictly unseen during pretraining, including their training splits. Therefore, methods pretrained with access to target benchmark training data, such as FoundationGait [71], are not included in the comparison.

Pretraining and Fine-tuning Settings. Data augmentation follows GaitSSB [12], and all downstream evaluations strictly follow the official protocols of each benchmark. We use DeepGaitV2 [11] with the layer configuration $[1, 4, 8, 4]$ as the shared gait encoder. The model is pretrained for 80K iterations using SGD with a learning rate of 0.05, momentum of 0.9, weight decay of 5×10^{-4}, and a MultiStepLR scheduler with milestones at 40K and 60K iterations. Following GaitSSB, we keep the same batch-size budget but construct each batch with 1,024 sequences and 8 frames per clip to increase sequence-level diversity. For fine-tuning, we adopt the layer-wise optimization strategy of GaitSSB. The model is trained for 9K iterations using SGD and a MultiStepLR scheduler. For SUSTech1K, we additionally adopt part-aware training [71] to mitigate representation collapse. More implementation details and hyperparameter settings are provided in Appendix C.

4.2 Zero-Shot Transfer Performance

Table 2 reports the cross-domain performance of supervised methods [5, 10, 11] and the zero-shot performance of unsupervised methods across six widely used gait benchmarks. Compared with GaitSSB, Scaled GaitSSB achieves substantial improvements on more challenging in-the-wild

Table 3: **Fine-tuning performance on Gait3D.**

Method	R1	R5	mAP
GaitSet [5]	36.7	58.3	30.0
GaitPart [10]	28.2	47.6	21.6
GaitGL [36]	29.7	48.5	22.3
SMPLGait [74]	46.3	64.5	37.2
DANet [41]	48.0	69.7	–
GaitGCI [9]	50.3	68.5	39.5
DyGait [62]	66.3	80.8	56.4
HSTL [61]	61.3	76.3	55.5
GaitMoE [28]	73.7	–	66.2
Free Lunch [60]	70.1	–	61.9
VPNet [42]	75.4	87.1	–
DeepGaitV2 [11]	74.4	88.0	65.8
SwinGait [11]	75.0	–	67.2
Mesh-Gait [63]	75.0	88.0	66.1
GaitLingo	**76.8**	**88.7**	**69.0**

Table 4: **Ablation studies.** Effects of pretraining data and language-guided modules on Gait3D and CCPG.

Dataset	Framework	Gait3D		CCPG			
		R1	R5	CL	UP	DN	BG
GaitLU-1M	GaitSSB	24.8	38.0	11.3	20.0	32.5	50.6
GaitLP-1M	GaitSSB	25.2	39.6	12.6	25.5	36.4	53.0
[GaitLU-1M, GaitLP-1M]	GaitSSB	25.9	38.9	11.9	23.4	33.0	48.1
	Scaled GaitSSB	**31.7**	**49.2**	**12.3**	**24.9**	**34.7**	**59.2**

(a) Effect of pretraining datasets and visual scaling.

Dataset	Framework	Gait3D		CCPG			
		R1	R5	CL	UP	DN	BG
GaitLP-1M	Scaled GaitSSB	29.4	47.2	11.5	23.9	36.2	60.1
GaitLP-1M	+TA	30.7	47.8	12.1	24.6	33.3	58.8
GaitLP-1M	+CSW	33.7	50.4	12.3	26.5	36.0	64.4
GaitLP-1M	+SRD	**35.2**	**53.4**	**13.6**	**27.2**	**41.4**	**68.5**

(b) Effect of the proposed language-guided modules.

datasets. Specifically, the Rank-1 accuracy on Gait3D increases from 24.8% to 31.7%, and on GREW from 16.6% to 20.5%. It also consistently outperforms GaitSSB on the clothing-changing dataset CCPG. However, performance drops on controlled datasets, suggesting that simply scaling in-the-wild pretraining data may not uniformly improve transfer to controlled scenarios. Building on Scaled GaitSSB, GaitLingo further improves model performance across benchmarks. The Rank-1 accuracy reaches 36.9% on Gait3D and 24.5% on GREW. On CCPG, consistent gains are observed under all four evaluation conditions. GaitLingo also surpasses GaitSSB on OU-MVLP and SUSTech1K. On CASIA-B, although GaitLingo does not surpass GaitSSB under the BG and CL conditions, it remains the second-best self-supervised method.

4.3 Fine-Tuning Performance

We further evaluate whether the learned representations remain effective after supervised target-domain adaptation. In this setting, the pretrained model is fine-tuned using labeled training data from the target domain and evaluated following the official protocols.

Tables 3 and 5 report the fine-tuning results on Gait3D and SUSTech1K, respectively. After fine-tuning, GaitLingo achieves a Rank-1 accuracy of 76.8% on Gait3D, outperforming existing supervised methods. On SUSTech1K, it reaches 84.3% overall Rank-1 accuracy and achieves the best performance across all probe conditions. These results demonstrate that GaitLingo is fully compatible with supervised adaptation, suggesting that the structures distilled from textual descriptions improve transferability without sacrificing downstream discriminative power.

4.4 Ablation Study

To better understand the effectiveness of the proposed pretraining data and modules, we conduct ablation studies under the zero-shot transfer setting. We select two challenging benchmarks, Gait3D and CCPG, to evaluate in-the-wild and clothing-changing transfer performance.

Table 4 presents ablation results from two complementary aspects: pretraining data/model scaling and language-guided modules. Subtable (a) compares the impact of different pretraining datasets and visual scaling within the same framework. The first two rows show that GaitLP-1M yields better performance than GaitLU-1M under pure visual self-supervised learning. Simply combining GaitLU-1M and GaitLP-1M brings only marginal improvement in Gait3D Rank-1 accuracy and even degrades several CCPG metrics, indicating that data scaling alone is insufficient under the original model capacity. We therefore introduce Scaled GaitSSB as a stronger visual-only baseline to disentangle the effect of language priors from data and model scaling.

Table 5: **Fine-tuning performance on SUSTech1K.**

Method	Probe Sequence (R1)								Overall	
	NM	BG	CL	CA	UM	UN	OC	NG	R1	R5
GaitSet [5] AAAI'19	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
GaitPart [10] CVPR'20	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
GaitGL [36] ICCV'21	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
GaitBase [13] CVPR'23	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
DeepGaitV2 [11] TPAMI'25	87.4	84.1	53.4	81.3	86.1	84.8	88.5	28.8	82.3	92.5
GaitLingo Ours	**88.3**	**86.3**	**60.4**	**83.6**	**87.4**	**87.5**	**88.6**	**31.3**	**84.3**	**93.7**

For module ablations, experiments are conducted only on GaitLP-1M, as it is the only dataset equipped with textual annotations. Subtable (b) reports the incremental effect of each proposed component. Starting from Scaled GaitSSB, adding TA improves temporal modeling and enhances Gait3D transfer performance. Introducing CSW further improves data utilization by prioritizing semantically informative samples. Finally, adding SRD further improves all evaluated metrics by transferring caption-derived ranking relations into the visual embedding space, highlighting the importance of inter-sequence semantic structure.

5 Discussion

Limitations. Recent studies have reported reliability limitations in multimodal reasoning, particularly under incomplete or misleading visual evidence [78, 23, 22]. In particular, these models may generate fluent descriptions that are not fully grounded in visual evidence, introducing plausible but unsupported details [77, 14, 52]. They may also struggle to capture subtle temporal variations, potentially overlooking gait-specific motion patterns or relying excessively on static visual cues. Although the two-stage pipeline reduces generic descriptions and filters appearance-related cues, caption noise and missing fine-grained motion information may still affect sample weighting and relation distillation. Caption quality also depends on the models used in the two-stage pipeline. Evaluating stronger models and combining descriptions from multiple models are directions for future work, with the aim of improving caption quality and assessing how these changes affect downstream gait representation learning.

Broader Impacts. This work may have positive impacts on long-range person retrieval, public safety, and human motion analysis by improving transferable gait representation learning under limited annotation. At the same time, gait recognition is a biometric technology and may introduce negative societal impacts. Incorrect predictions may lead to false matches, which can be harmful in security-sensitive scenarios.

Code of Ethics. This work uses existing public research datasets and processes pedestrian videos into silhouettes to reduce appearance, texture, background, and facial information. The generated captions focus on gait-related cues and are refined to remove appearance- and scene-related nuisance factors. We recognize the privacy risks of gait recognition. To protect privacy, we will not release the original RGB videos or raw data. Any released data will be limited to processed silhouettes and derived annotations, subject to the licenses and usage agreements of the original datasets.

6 Conclusion

We presented GaitLingo, a language-guided self-supervised framework for large-scale gait pretraining. To provide language supervision for unlabeled gait sequences, we constructed GaitLP-1M, a million-scale silhouette gait dataset with paired sequence-level captions generated by a two-stage captioning pipeline. Built on this data foundation, GaitLingo introduces Caption-Guided Sample Weighting, Temporal Adapter, and Soft Relation Distillation to improve data utilization, enhance temporal robustness, and transfer semantic relations. Experiments under zero-shot and fine-tuning settings show that language priors provide complementary supervision beyond visual scaling, leading to more transferable and structured gait representations.

Acknowledgments and Disclosure of Funding

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. T2600237, the Early Career Scheme (ECS) of the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China, under Project No. 26601126, and the research funding under the HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01).

References

[1] Imed Bouchrika, Michaela Goffredo, John Carter, and Mark Nixon. On using gait in forensic biometrics. *Journal of forensic sciences*, 56(4):882–889, 2011.

[2] Qingyuan Cai, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Barbiegait: An identity-consistent synthetic human dataset with versatile cloth-changing for gait recognition. *arXiv preprint arXiv:2604.12221*, 2026.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Chi-Min Chan, Han Zhu, Chunyang Jiang, Jiaming Ji, Juntao Dai, Wei Xue, Sirui Han, and Yike Guo. Benchmarking fine-grained error detection in multimodal reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44672–44702. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2068.

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[7] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow: Towards a world omniscient world model through embodied interaction, 2025.

[8] Carlo Dindorf, Jonas Dully, Rebecca Keilhauer, Michael Lorenz, and Michael Fröhlich. Evaluating large language models for gait classification using text-encoded kinematic waveforms. *arXiv preprint arXiv:2603.13317*, 2026.

[9] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, and Xi Li. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5578–5588, 2023.

[10] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[11] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023.

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

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

[14] Sitong Fang, Wenjing Cao, Jiahao Li, Xuyao Wang, Chi-Min Chan, Sirui Han, Juntao Dai, Yike Guo, Yaodong Yang, and Jiaming Ji. When slower isn't truer: Inverse scaling law of truthfulness in multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 1236–1263, 2026.

[15] Chuyao Fu, Xiaowei Chi, Yuhan Rui, Yu kai Wang, Zezhong Qian, Xiaojie Zhang, Yunfan Lou, Kevin Zhang, Kuangzhi Ge, Chak Wing Mak, Zhiyang Chen, Athena Zhuoming Zhong, Hongyang Chen, Haoran Li, Yike Guo, Sirui Han, and Shanghang Zhang. Token-World: World modeling in vision-language model token space for robot manipulation, 2026.

[16] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2476–2486, 2022.

[17] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19595–19604, 2023.

[18] Yang Fu, Saihui Hou, Shibei Meng, Xuecai Hu, Chunshui Cao, Xu Liu, and Yongzhen Huang. Cut out the middleman: Revisiting pose-based gait recognition. In *European Conference on Computer Vision*, pages 112–128. Springer, 2024.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[20] Ayush Gupta, Siyuan Huang, and Rama Chellappa. Mind the gap: Bridging occlusion in gait recognition via residual gap correction. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2025.

[21] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[22] Sirui Han, Yidan Huang, Guoying Lu, Shuchao Wu, Zefeng Chen, Yujin Zhou, Chuxue Cao, Yuyao Zhang, Mingxuan Zheng, Bubu Hou, Ruoxi Li, Xiangyu Long, Boqin Yin, Kit Yeung Fung, Xinyi Wang, and Yike Guo. Trustworthy legal reasoning in 2026: A mid-year review of research, products, events, and governance. Preprints.org, 2026. Preprint, version 1.

[23] Sirui Han, Zhizhuo Kou, Ruoxi Li, Yuyao Zhang, Yujin Zhou, Chuxue Cao, Han Zhu, Kunhao Pan, Haoran Li, Conghui He, Haitian Lu, and Yike Guo. Trustworthy legal reasoning: A comprehensive survey. Preprints.org, 2026. Preprint, version 1.

[24] Xiao Han, Yiming Ren, Peishan Cong, Yujing Sun, Jingya Wang, Lan Xu, and Yuexin Ma. Gait recognition in large-scale free environment via single lidar. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 380–389, 2024.

[25] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024.

[26] Saihui Hou, Chenye Wang, Wenpeng Lang, Zhengxiang Lan, and Yongzhen Huang. Gaitsnippet: Gait recognition beyond unordered sets and ordered sequences. *arXiv preprint arXiv:2508.07782*, 2025.

[27] Saihui Hou, Chenye Wang, Aoqi Li, Jilong Wang, Liang Wang, and Yongzhen Huang. Gaitasset: In defense of regarding gait as a set. *IEEE Transactions on Information Forensics and Security*, 20:12301–12316, 2025. doi: 10.1109/TIFS.2025.3630892.

[28] Panjian Huang, Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Zhiqiang He, and Yongzhen Huang. Occluded gait recognition with mixture of experts: an action detection perspective. In *European Conference on Computer Vision*, pages 380–397. Springer, 2024.

[29] Panjian Huang, Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Vocabulary-guided gait recognition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] Dongyang Jin, Chao Fan, Jingzhe Ma, Jingkai Zhou, Weihua Chen, and Shiqi Yu. On denoising walking videos for gait recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12347–12357, 2025.

[31] Wenpeng Lang, Saihui Hou, and Yongzhen Huang. Beyond sparse keypoints: Dense pose modeling for robust gait recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 669–678, 2025.

[32] Aoqi Li, Saihui Hou, Chenye Wang, Qingyuan Cai, and Yongzhen Huang. Aerialgait: Bridging aerial and ground views for gait recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1139–1147, 2024.

[33] Huakang Li, Wing WY Ng, Yidan Qiu, Ying Gao, Jianjun Zhang, and Sam Kwong. Gaitss: Gait slices as a set for gait recognition in interfered sequences. *IEEE Transactions on Artificial Intelligence*, 2026.

[34] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413, 2023.

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

[36] Beibei Lin, Shunli Zhang, Ming Wang, Lincheng Li, and Xin Yu. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv preprint arXiv:2208.01380*, 2022.

[37] Hengyu Liu, Chenxin Li, Zhengxin Li, Yipeng Wu, Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yunlong Lin, Sirui Han, and Brandon Feng. IR3D-Bench: Evaluating vision-language model scene understanding as agentic inverse rendering. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0249.

[38] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. Paddleseg: A high-efficient development toolkit for image segmentation. *arXiv preprint arXiv:2101.06175*, 2021.

[39] Yiqun Liu, Yi Zeng, Jian Pu, Hongming Shan, Peiyang He, and Junping Zhang. Selfgait: A spatiotemporal representation learning method for self-supervised gait recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2570–2574. IEEE, 2021.

[40] Zhiyang Lu, Wankang Zeng, Ming Cheng, and Cheng Wang. Language as a bridge: Semantic-guided cross-modal gait recognition via text prototype and feature decoupling. *IEEE Transactions on Information Forensics and Security*, 2026.

[41] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22076–22085, 2023.

[42] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 593–603, 2024.

[43] Shibei Meng, Saihui Hou, Yang Fu, Xuecai Hu, Junzhou Huang, and Yongzhen Huang. Seeing from magic mirror: Contrastive learning from reconstruction for pose-based gait recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7719–7728, 2025.

[44] Wei Ming Neo, Koichi Shinoda, and Tat-Jen Cham. Diffusion pretraining for gait recognition in the wild. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 1295–1300. IEEE, 2025.

[45] Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24245–24254, 2025.

[46] Runqi Ouyang, Haoyun Li, Zhenyuan Zhang, Xiaofeng Wang, Zeyu Zhang, Zheng Zhu, Guan Huang, Sirui Han, and Xingang Wang. Motion-R1: Enhancing motion generation with decomposed chain-of-thought and RL binding. In *International Conference on Learning Representations*, 2026.

[47] Xingqun Qi, Hengyuan Zhang, Yatian Wang, Jiahao Pan, Chen Liu, Muyi Sun, Wei Xue, Shanghang Zhang, Sirui Han, Qifeng Liu, et al. Cocogesture: Towards coherent co-speech 3d gesture generation in the wild. *Information Fusion*, page 103613, 2025.

[48] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):264–284, 2022.

[49] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11174–11184, 2023.

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

[51] Chuanfu Shen, Shiqi Yu, Jilong Wang, George Q Huang, and Liang Wang. A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[52] Qianpu Sun, Xiaowei Chi, Yuhan Rui, Ying Li, Kuangzhi Ge, Jiajun Li, Sirui Han, and Shanghang Zhang. LABSHIELD: A multimodal benchmark for safety-critical reasoning and planning in scientific laboratories, 2026.

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

[54] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17127–17137, 2024.

[55] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[56] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[57] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1569–1577, 2022.

[58] Chenye Wang, Saihui Hou, Aoqi Li, Qingyuan Cai, and Yongzhen Huang. Ra-gar: A richly annotated benchmark for gait attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7591–7599, 2025.

[59] Chenye Wang, Qingyuan Cai, Saihui Hou, Aoqi Li, and Yongzhen Huang. Mmgait: Towards multi-modal gait recognition. *arXiv preprint arXiv:2604.15979*, 2026.

[60] Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Free lunch for gait recognition: A novel relation descriptor. In *European Conference on Computer Vision*, pages 39–56. Springer, 2024.

[61] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19582–19592. IEEE, 2023.

[62] Ming Wang, Xianda Guo, Beibei Lin, Tian Yang, Zheng Zhu, Lincheng Li, Shunli Zhang, and Xin Yu. Dygait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13424–13433, 2023.

[63] Zhao-Yang Wang, Jieneng Chen, Jiang Liu, Yuxiang Guo, and Rama Chellappa. Mesh-gait: A unified framework for gait recognition through multi-modal representation learning from 2d silhouettes. *arXiv preprint arXiv:2510.10406*, 2025.

[64] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5177–5186, 2018.

[65] Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. EgoTwin: Dreaming body and view in first person. In *International Conference on Learning Representations*, 2026.

[66] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[67] Shaopeng Yang, Jilong Wang, Saihui Hou, Xu Liu, Chunshui Cao, Liang Wang, and Yongzhen Huang. Bridging gait recognition and large language models sequence modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3460–3469, 2025.

[68] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17343–17353, 2024.

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

[70] Dingqiang Ye, Chao Fan, Zhanbo Huang, Chengwen Luo, Jianqiang Li, Shiqi Yu, and Xiaoming Liu. Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132*, 2025.

[71] Dingqiang Ye, Chao Fan, Kartik Narayan, Bingzhe Wu, Chengwen Luo, Jianqiang Li, and Vishal M Patel. Silhouette-based gait foundation model. *arXiv preprint arXiv:2512.00691*, 2025.

[72] Yifan Ye, Yankai Fu, Yaoxu Lv, Bohan Hou, Jun Cen, Lingdong Kong, Duo Zheng, Tianxing Chen, Jiaming Liu, Ziang Cao, Yunfan Lou, Wei Chow, Xian Sun, Yingshuo Wang, Kuangzhi Ge, Xiaowei Chi, Xidong Zhang, Zhibo Pang, Yiwu Zhong, Sirui Han, Zhihe Lu, Weihao Yuan, Qifeng Chen, Michael Yu Wang, Yao Mu, Ziwei Liu, Jianfei Yang, Ping Luo, and Shanghang Zhang. Data pyramid for embodied manipulation: A survey, 2026.

[73] Shiqi Yu, Daoliang Tan, and Tieniu 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.

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

[75] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer, 2016.

[76] Rui Zhou, Yibo Yuan, Junkai Zhao, Fangyuan Zhao, Xiaoguang Zhao, Shanghang Zhang, and Sirui Han. DexRoam: Learning mobile bimanual dexterous manipulation from egocentric whole-body human demonstrations, 2026.

[77] Yujin Zhou, Pengcheng Wen, Jiale Chen, Boqin Yin, Han Zhu, Jiaming Ji, Juntao Dai, Chi-Min Chan, and Sirui Han. What, whether and how? unveiling process reward models for thinking with images reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29071–29079, 2026.

[78] Yujin Zhou, Mingxuan Zheng, Chuxue Cao, Yidan Huang, Jiale Chen, Yike Guo, and Sirui Han. LexAgentHallu: A hierarchical benchmark for profiling hallucinations in legal agents, 2026.

[79] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021.

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

[81] Jialong Zuo, Jiahao Hong, Feng Zhang, Changqian Yu, Hanyu Zhou, Changxin Gao, Nong Sang, and Jingdong Wang. Plip: Language-image pre-training for person representation learning. *Advances in Neural Information Processing Systems*, 37:45666–45702, 2024.

A Related Work

A.1 Gait Recognition

Supervised Gait Recognition. Supervised gait recognition can be categorized according to the input modality, including silhouette-based methods [5, 10, 36], pose-based methods [56, 57, 17, 18, 31], and RGB-based methods [69, 70, 30]. Among them, silhouette-based settings dominate the literature because binary silhouettes preserve identity-related body shape and motion patterns while largely removing distracting appearance cues caused by clothing texture, illumination, and background variations.

Representative examples include GaitBase [13], which provides a unified and strong baseline for gait modeling, and DeepGaitV2 [11], which systematically explores stronger backbone designs and training strategies for large-scale gait recognition. Based on these mature frameworks, recent studies focus more on extracting finer-grained local-global representations [42, 41], achieving more robust temporal aggregation [27, 26], and introducing additional semantic priors to enhance discriminative learning [29, 67]. These efforts continuously improve the robustness of gait representations under challenging factors such as viewpoint changes, clothing variation, carrying conditions, and occlusions.

Progress in supervised gait recognition is supported by several benchmarks [32, 58, 24], including large-scale in-the-wild datasets such as Gait3D [74] and GREW [79], clothing-changing benchmarks such as CCPG [35], and multimodal datasets such as SUSTech1K [50]. These benchmarks have significantly advanced model development and fair evaluation. However, constructing supervised benchmarks is highly labor-intensive and relies heavily on fixed acquisition environments and expensive identity annotations. As a result, models trained under supervised settings often suffer from limited generalization and struggle to maintain robustness when transferred to unseen domains or real-world cross-scene scenarios.

Self-Supervised Gait Recognition. Unlike supervised approaches that rely on identity labels from a target benchmark, self-supervised gait recognition aims to learn transferable gait priors from large collections of unlabeled walking videos. Early attempts mainly adapted generic self-distillation objectives to gait data. SelfGait [39] introduced a BYOL-style pretraining strategy for spatiotemporal silhouette backbones. A more influential line formulates gait-specific contrastive learning at scale. GaitSSB [12] constructed the large-scale GaitLU-1M corpus and showed that contrastive pretraining with gait-aware augmentations can learn transferable walking priors that substantially improve performance on both constrained and in-the-wild benchmarks after fine-tuning. More recent studies have further expanded the design space beyond standard contrastive objectives. Diffusion-based generative pretraining has been explored as an alternative way to initialize gait backbones for unconstrained recognition [44]. FoundationGait [71] is an important foundation-scale silhouette pretraining effort with promising transferability. However, its unlabeled pretraining dataset overlaps with the training splits of target benchmarks. Since our zero-shot protocol requires all target data to be strictly unseen during pretraining, we exclude FoundationGait from the comparison.

Despite these advances, existing self-supervised gait methods remain visual-only, limiting their ability to exploit semantic cues in unlabeled sequences. This motivates us to introduce language priors into self-supervised gait representation learning.

A.2 Text-Guided Pedestrian Representation Learning

Recent person re-identification studies have explored language as a semantic prior for pedestrian representation learning. CLIP-ReID [34] adapts CLIP to supervised ReID by learning identity-specific textual prompts, while PromptSG [68] introduces image-conditioned pseudo tokens to provide more adaptive semantic guidance. Other works scale person-domain image-text pretraining with pseudo or synthetic descriptions. UniPT [49] constructs pseudo-text annotations for pedestrian images, PLIP [81] introduces large-scale person language-image pretraining with fine-grained pretext tasks, and MLLM4Text-ReID [54] leverages multimodal LLMs to improve description diversity and handle caption noise. More recent methods, such as Instruct-ReID [25] and ChatReID [45], further use instructions or conversational interfaces to unify multiple retrieval settings.

Despite their progress, these methods are mainly designed for appearance-based pedestrian retrieval. Their textual supervision typically describes clothing, colors, accessories, and other visual appearance

cues, and the learning objectives often rely on identity labels, paired image-text retrieval data, or large vision-language models at inference. In contrast, our work introduces language priors into self-supervised gait representation learning, where the input is a silhouette sequence and the target representation should capture body structure and motion rather than appearance.

Language models have also been explored in gait analysis. GaitLLM [67] exploits the sequence-modeling capability of pretrained LLMs to process gait features, without using explicit textual descriptions as supervision. Dindorf et al. [8] encode kinematic waveforms as text to evaluate whether general-purpose LLMs can directly classify gait patterns. Lu et al. [40] use textual prototypes to bridge camera- and LiDAR-based gait representations, aiming to reduce the modality gap in cross-modal gait recognition. GaitLingo investigates how automatically generated gait descriptions can guide large-scale silhouette representation pretraining without identity annotations. Caption-Guided Sample Weighting uses the semantic richness and corpus-relative attribute rarity of each sequence's description to emphasize informative and underrepresented training samples, while Soft Relation Distillation transfers confidence-filtered relative semantic relationships across sequences to the visual representation. The captions thus provide auxiliary supervision for sample weighting and relation learning, rather than serving as direct classifier inputs, identity labels, or textual prototypes.

B More Details of GaitLP-1M

Visualization. We provide qualitative examples from GaitLP-1M in Figure 5. Each example consists of a silhouette sequence and its corresponding sequence-level gait caption generated by our two-stage captioning pipeline. The captions capture gait-related semantics that are difficult to obtain from identity labels alone, including walking pace, body posture, limb movement, view direction, stopping or turning behaviors, and interactions with carried objects. These examples show that the generated descriptions preserve fine-grained sequence dynamics while suppressing irrelevant RGB appearance cues, making them suitable as language supervision for self-supervised gait pretraining. The prompts used in the two-stage caption generation pipeline are listed in Table 6.



Figure 5: **Examples from GaitLP-1M.** Each sample contains a silhouette sequence and a sequence-level gait caption generated by our two-stage captioning pipeline. The captions describe gait-relevant semantics such as walking pace, limb motion, carrying objects, stopping behavior, and view changes.

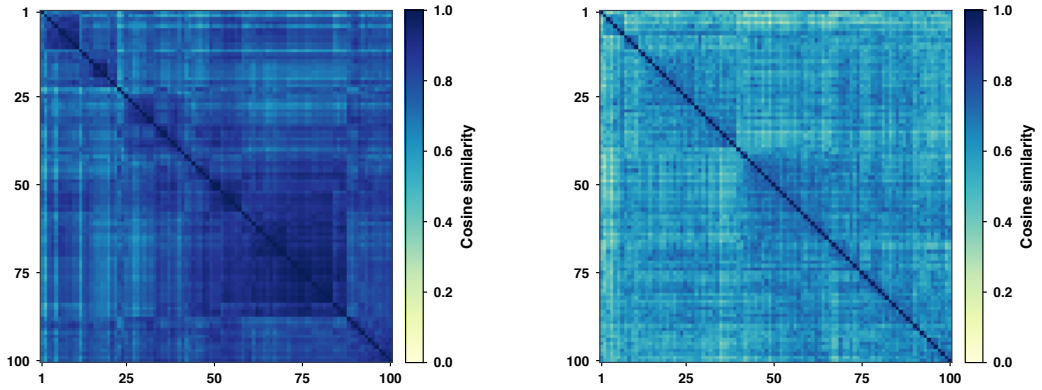
Comparison with Direct VLM Captioning. Direct single-stage captioning with Qwen2.5-VL produces descriptions such as *"The person walks upright with moderate strides and limited arm swing"*, *"The person walks naturally with a slight backward lean"*, and *"The person walks with an upright posture and steady steps"*.

Table 6: Prompts used in the two-stage caption generation pipeline.

Stage	Prompt
Stage I	You will see a sequence of video frames, each centered on a person. Your task is to describe what is clearly visible in each frame, using full, natural, and detailed sentences. For every frame, focus on the person's posture, movement, appearance, and any fine-grained visual details. Pay attention to the continuity between frames and describe how the person's pose or position evolves over time. Do not make assumptions or guesses about things that are unclear or not shown. Only describe what you can directly observe in the image. Label each description with "Frame 1", "Frame 2", etc.
Stage II	System prompt: You are a precise assistant responsible for extracting detailed and accurate action descriptions from frame-by-frame person-centered video captions for gait recognition and video attribute pretraining. Remove all color, texture, and background environment information. Retain identity-relevant features, such as body shape, body type, posture changes, and gait characteristics. Focus on describing the actual actions and postures, such as walking, turning, stopping, bending, raising arms, looking down, and real interactions with objects. The description can be rich and detailed, capturing as much information about the person's actions as is described, but do not imagine or infer information not explicitly stated. Merge repetitive frames and produce a temporally coherent action narrative. Output only the action description without "Frame X" or any explanations. User prompt: Below is a frame-by-frame caption of a person-centered video: {raw_caption}.

Although generally plausible, these descriptions are brief and similar across sequences, largely summarizing common walking patterns. Our two-stage pipeline consolidates detailed frame-level observations while filtering out appearance-related cues. As illustrated in Figure 5, the resulting sequence-level captions describe more specific gait characteristics and their changes over time.

Caption Quality Analysis. To assess the quality of generated captions, we employ the CLIP text encoder as a caption feature extractor and compute pairwise cosine similarities between caption embeddings. Figure 6 compares the resulting similarity heatmaps for captions directly generated by VLMs and those produced by our two-stage pipeline. Direct VLM captioning leads to highly concentrated similarities across sequences, suggesting that many captions collapse into generic walking descriptions. By contrast, our two-stage pipeline produces more diverse and gait-centered captions, yielding a more informative textual embedding space for relation transfer.



(a) Direct VLM captioning (b) Our two-stage pipeline

Figure 6: **Caption quality analysis.** Pairwise cosine similarity heatmaps of CLIP text embeddings. Direct VLM captions exhibit concentrated similarities, whereas two-stage captions are more diverse.

C Implementation Details of GaitLingo

Pretraining Setting. Silhouettes are size-normalized following OpenGait [13] and resized to 64×44. Data augmentation follows GaitSSB [12], and all downstream evaluations strictly follow the official protocols of each benchmark. We use DeepGaitV2 [11] as the shared gait encoder with the layer configuration $[1, 4, 8, 4]$, while keeping other architectural settings unchanged. The model is pretrained for 80K iterations using SGD with a learning rate of 0.05, momentum of 0.9, and weight decay of 5×10^{-4}. The learning rate is scheduled by MultiStepLR with milestones at 40K and 60K iterations. Following GaitSSB, we keep the same batch-size budget, but adjust the batch construction to 1,024 sequences with 8 frames per clip to increase sequence-level diversity.

Table 7: Semantic attributes used for caption parsing in CSW.

Group	Attribute
Viewpoint	front, back, side
Walking Direction	walking toward, walking away, leftward, rightward, turning, crossing
Motion Speed	slow, normal speed, fast, brisk, hurried
Clothing Type	coat, jacket, suit, dress, skirt, t-shirt, hoodie, uniform
Clothing Length	long-sleeve, short-sleeve, sleeveless, long coat, short coat
Lower-body Clothing	pants, jeans, shorts, skirt, trousers
Footwear	sneakers, boots, leather shoes, sandals, high heels
Carrying Condition	backpack, phone, handbag, shoulder bag, suitcase, luggage, briefcase
Occlusion	partially occluded, heavily occluded, occluded by person, occluded by object
Body Shape	slim, thin, average build, heavy, muscular, stocky
Pose / Posture	upright, leaning forward, hunched, swinging arms, restricted arms
Gait Style	limping, asymmetric gait, steady gait, exaggerated motion, dragging feet
Motion Dynamics	large stride, small stride, high leg lift, low leg lift, rhythmic
Scene Context	indoor, outdoor, street, paved area, crossing, sidewalk, building wall, grassy area
Lighting Condition	bright, dim, shadowed, backlit, low-light
Crowd Density	alone, sparse crowd, crowded, group walking

Hyperparameters of GaitLingo. For Caption-Guided Sample Weighting (CSW), we set the trade-off parameter between semantic richness and attribute rarity to $\alpha = 0.5$, giving equal importance to the two sampling cues. Table 7 summarizes the semantic attribute vocabulary used for caption parsing in CSW. For the Temporal Adapter, the channel reduction ratio is set to $r = 4$. The residual scaling parameter γ is initialized to zero and learned during pretraining. For Soft Relation Distillation (SRD), caption embeddings are extracted using a frozen Sentence-BERT encoder. We set the visual ranking margin to $m = 0.05$, the textual separation threshold to $\delta = 0.2$, and the loss weight to $\lambda_{\text{srd}} = 0.1$. Unless otherwise specified, all hyperparameters are kept fixed across datasets and evaluation protocols.

Fine-tuning Setting. We adopt the same layer-wise fine-tuning strategy as GaitSSB [12]. Specifically, the four backbone stages are fine-tuned with learning rates of $[0, 0.001, 0.001, 0.001]$, while the projector and classification head are optimized with learning rates of 0.01 and 0.1, respectively. We use SGD with an initial learning rate of 0.1 and momentum of 0.9, together with a MultiStepLR scheduler. The model is trained for 9K iterations, with learning-rate milestones set at 3K and 6K. For SUSTech1K, we additionally adopt part-aware training [71] to mitigate representation collapse.

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 and introduction clearly state the two main contributions, GaitLP-1M and GaitLingo, and the claims are supported by the experimental results and ablation studies.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper discusses its limitations in Section 5.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] 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 paper provides the dataset construction pipeline, model architecture, training objective, implementation details and evaluation protocols needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: We provide code, data access instructions, and configuration files to support faithful reproduction of the main experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, preprocessing procedure, training schedules, optimizer, learning rates, batch construction, evaluation protocols, and hyperparameters used in the experiments.

   Guidelines:

   - The answer [N/A] 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: We do not report error bars or statistical significance tests because the main experiments involve large-scale pretraining and evaluation across multiple benchmarks, making repeated runs computationally expensive.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: We do not provide detailed per-experiment compute cost, memory usage, or runtime estimates in the current version, although the implementation details describe the training setup.

   Guidelines:

   - The answer [N/A] 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: The research follows the NeurIPS Code of Ethics. The dataset construction and evaluation are conducted for academic research, and privacy-related concerns are considered in data processing and release.

   Guidelines:

   - The answer [N/A] 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: The paper discusses positive impacts such as improving transferable gait representation learning, as well as potential negative impacts related to privacy and misuse in surveillance scenarios.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk generative models or dual-use foundation models that require special safeguards beyond standard responsible data and model release practices.

Guidelines:

- The answer [N/A] 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: Existing datasets, codebases, and pretrained models used in this work are properly cited, and their licenses or usage terms are respected where available.

Guidelines:

- The answer [N/A] 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: [N/A]

    Justification: The paper does not release new assets in the current submission version; therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or direct interaction with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: The paper does not involve crowdsourcing or direct human-subject studies requiring IRB approval or equivalent review.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the use of Qwen2.5-VL and an LLM-based summarization stage in the caption generation pipeline, which is an important component of GaitLP-1M construction.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.