

Decoding Children’s Gait Behavior

Yifan Shen^{1,2,*}, Boyi Li^{1,*}, Meihuan Huang^{3,2,4,*}, Yuanzhe Liu^{1,*}, Xu Cao^{1,2,*,§} and Jinyang Jin¹, Zhengyuan Li¹, Anglin Liu⁵, Junho Kim¹, Jingyuan Zhu², Fangzhou Lan², Jianguo Cao^{2,3}, Jintai Chen⁵, Ismini Lourentzou¹, and James M. Rehg^{1,†}

¹ University of Illinois Urbana-Champaign

² PediaMed AI

³ Shenzhen Children’s Hospital

⁴ The Hong Kong Polytechnic University

⁵ The Hong Kong University of Science and Technology (Guangzhou)

* Equal contribution § Project lead † Corresponding author

Abstract. We introduce a new problem domain for human action recognition: the fine-grained analysis of children’s gait behaviors from standard RGB video. We specifically target the ambulatory patterns of children aged 3–17 years. Such behaviors arise naturally in the diagnosis and treatment of several critical developmental and neuromuscular disorders, such as cerebral palsy and hemiplegia. Despite their clinical value, current 3D sensor-based gait analysis systems are expensive, intrusive, and often impractical for young subjects. To address this, we introduce a new dataset comprising over 1,100 high-frame-rate (60 FPS) video sequences from 110 subjects, accompanied by synchronized, anonymized pose sequences. In each session, the child performs a 5-second "walk-around" task, capturing the gait cycle from multiple viewpoints. Crucially, we demonstrate that current state-of-the-art approaches, including gait foundation models and Multimodal Large Language Models (MLLMs), fail to effectively resolve these clinical nuances. We identify the key technical challenges in analyzing these erratic and subtle motor patterns and describe a unified end-to-end framework for decoding fundamental components of pediatric gait. Through comprehensive experimental results, we demonstrate the potential of this dataset to drive novel research questions and establish a rigorous baseline for automated child gait assessment. **Project page:** pediamedai.com/ChildrenGait

Keywords: Children Gait · Visual Gait Score · Foundation Models

1 Introduction

Quantitative gait analysis constitutes a fundamental pillar of clinical diagnostics and physical rehabilitation for movement disorders [5, 6, 9, 84, 97]. In pediatric populations, the challenge of decoding pediatric gait across different developmental milestones is central to the diagnosis and management of a variety of developmental disorders [3, 4, 79]. Moreover, with the advent of Brain-Computer

Interfaces (BCI) and neuroprosthetics, objective gait quantification has evolved into a critical benchmark for evaluating the efficacy of functional restoration in child patients [8, 29, 89]. Currently, the early identification of gait abnormalities in conditions such as cerebral palsy (CP) relies heavily on subjective visual assessment by a pediatrician during a standard office visit [28, 68]. While research utilizing 3D gait analysis and Inertial Measurement Units (IMUs) has successfully identified quantitative warning signs for gait deviations in cerebral palsy [24, 25], deploying these technologies in clinical practice faces significant hurdles. Technologies that require significant physical space (*e.g.*, to obtain multiple unobstructed lines-of-sight in motion capture) are difficult to incorporate and can bias the adoption of technology to large, well-resourced clinical sites, creating potential inequities [55, 92]. Technologies that require an attachment to children’s bodies (*e.g.*, IMU-based measurement) will not be tolerated by all children and may induce reactivity, where the measurement technology alters the natural gait behavior [10, 51]. In this setting, there is immense potential for computer vision-based movement analysis to bridge this gap, as camera hardware is relatively inexpensive and accurate measurements can, in principle, be obtained from only one or two camera views, without encumbering the child. Automated analyses could potentially scale early screening efforts by bringing reliable, rich, and non-intrusive measurement of child gait into everyday clinical settings [23, 105].

Despite this clinical imperative, existing computer vision methods for gait analysis have largely focused on healthy adult subjects [22, 78], relying on the assumption that gait is a mature, stable, and highly periodic process [35]. Such ‘adult-centric’ foundation models fail to capture the high entropy, high intra-class variance, and inconsistent motion patterns inherent to the developing motor system [86, 94, 102]. In this work, we define a new challenge for the computer vision community: *the fine-grained analysis of children’s gait from video, targeting clinically-relevant characterizations of children’s gait quality*.

To address this challenge, we introduce the **Children Gait Video (CGV) dataset**, a repository comprising over 1,100 video sessions from 110 child subjects. Video collection followed a standardized observational protocol: brief (3–5 second) video sequences capturing anterior, posterior, and lateral walking views of children aged 3–17 years. Videos were annotated by expert clinicians for the Edinburgh Visual Gait Score (EVGS) [7, 73]. In addition to clinical assessments, CGV also contains rich, frame-level annotations, including instance segmentation masks, bounding boxes, and anatomical keypoints, all tailored to the nuances of the developing body. We also present a comprehensive evaluation of the effectiveness of modern VLMs in analyzing children’s videos and inferring the EVGS, and we find that state-of-the-art models lack sensitivity to the subtle details of children’s gait. To address this limitation, we introduce a new method, ChildGait-Video, and show that it achieves SoTA performance across all base-lines, reaching a highest accuracy range of 70% to 93% in all 34 scoring items.

In summary, this paper makes the following contributions:

- We introduce the Children Gait Video (CGV) dataset, which is the largest repository of multi-camera gait videos of children with developmental motor conditions, containing both clinically relevant assessments (EVGS) and rich frame-level pose annotations.
- We provide the first comprehensive assessment of the ability of open and closed VLMs to decode the subtle signs of gait abnormalities from video, and find that out-of-the-box models are ineffective for this task.
- We present ChildGait-Video, a novel video analysis method that is the current SoTA for automated inference of EVGS scores from children’s videos.

2 Related Work

Human Gait Datasets. The landscape of gait datasets has progressively evolved from general domain biometric recognition to specialized health and clinical applications [82, 118]. In the general domain, foundational benchmarks such as the CASIA series and the extensive OU-ISIR database families have driven immense progress in appearance-invariant and multi-view identity recognition [1, 48, 69, 96, 107]. Other datasets, including SUSTech1K and CMU MoBo, further expand visual gait analysis into wild and treadmill environments [85, 109]. In addition, recent large-scale benchmarks curated for practical in-the-wild and cross-covariate evaluation have become increasingly central to gait recognition research [34, 60, 61, 120, 127]. Concurrently, the focus has shifted toward clinical and health-oriented gait analysis, evidenced by the emergence of multimodal and sensor-based datasets like WearGait-PD [2], WhuGait [128], and MAREA [53], which target specific pathologies such as Parkinson’s disease and provide quantitative biomarkers for fall risk and symptom monitoring [27]. However, despite the existence of large-scale datasets spanning wide age ranges, for example, OULP-Age [107], there is an absence of comprehensive, multi-view video datasets tailored for clinical gait screening in children. Pediatric gait exhibits unique biomechanical developmental trajectories and specific pathological manifestations [76], rendering the lack of dedicated pediatric multi-view datasets a critical bottleneck. To bridge this gap, we introduce the CGV Dataset, a new resource designed to support fine-grained analysis of children’s gait and enable the development of clinically meaningful assessment models.

Gait Vision Modeling. Visual gait modeling has evolved significantly through deep representations. Open-source benchmarks like OpenGait [33, 35] have unified classic silhouette-based architectures [19, 20, 37, 44], while concurrent studies tackle cross-covariate and in-the-wild challenges [60, 127, 129]. Concurrently, pose-based and skeleton-based models offer robustness against appearance changes by pairing 2D keypoint detectors [16, 38, 49, 93] with 3D pose lifting or triangulation [52, 75, 117, 119, 126]. Recent specialized skeleton maps further improve baselines under viewpoint variations [33, 36, 41, 42]. To model these dynamic sequences, diverse spatiotemporal encoders are utilized, including graph convolutional networks [31, 64, 90, 99, 108] and video-based architectures [39, 87, 100, 115, 116], alongside recent extensions into LiDAR and point-cloud modalities [85].

Most recently, the field has gravitated toward Transformers [18, 114] and massive foundation models [112], leveraging diffusion and large vision pipelines to achieve zero-shot generalization under real-world degradations [50, 111]. However, state-of-the-art architectures are almost exclusively pre-trained on adult datasets [47]. Due to substantial differences in children’s skeletal proportions, applying these pre-trained models to pediatric gait introduces severe domain shifts, necessitating new, domain-specific modeling strategies.

Children’s Gait Screening and Assessment. Gait screening and assessment translate raw motion signals into actionable clinical diagnostics and developmental metrics. Quantitative Gait Analysis typically extracts fundamental spatiotemporal parameters, such as velocity, cadence, step length, and double support time, alongside kinetic ground reaction forces [56, 103, 112]. While generalized functional scales like the Functional Gait Assessment (FGA) [106], and UPDRS [70] are widely used for adult fall risk and neurological evaluation, pediatric gait assessment demands specialized criteria. Observational scales are critical for diagnosing neurodevelopmental disorders such as Cerebral Palsy (CP) [3]. Beyond the widely recognized Edinburgh Visual Gait Score (EVGS) [80], the clinical assessment includes the Visual Gait Assessment Scale (VGAS) [65], the Gait Deviation Index (GDI) [83], and various musculoskeletal rubrics [40]. Modern automated screening endeavors to map video-derived 3D motion directly to these structured scores using ordinal learning techniques [11, 91] and multi-task frameworks [66, 67, 88]. However, most existing recognition-oriented frameworks focus on coarse classification rather than the item-level, phase-specific observational scoring required for reliable pediatric screening, highlighting the necessity for both specialized datasets and dedicated pediatric modeling techniques.

3 Challenges in Pediatric Gait Analysis

From the perspective of action and activity recognition, the automated 2D video analysis of pediatric gait introduces unique technical challenges that are largely absent in existing adult-centric datasets. First, gait modeling of children confronts a significant domain shift in skeletal structure. Significant anthropometric differences exist between pediatric and adult populations. The relative limb proportions and center of mass in children do not scale linearly from adult data. Consequently, foundation models pre-trained on adult kinematics often fail to generalize, a limitation frequently observed in 3D human mesh recovery and child pose estimation tasks [14, 21, 46, 77]. Second, clinical gait assessment is inherently a fine-grained, phase-dependent task. Unlike the coarse disorder classification labels found in healthcare datasets like Scoliosis1K [123, 124], effective gait analysis requires the precise estimation of joint angles at specific instances of the gait cycle (*e.g.*, maximum knee flexion during swing). Third, the capture environment involves severe occlusion and multi-person interaction. Young children frequently require physical guidance or encouragement from parents and clinicians during the walking process, resulting in complex visual clutter and frequent inter-person occlusions that challenge standard tracking and segmentation

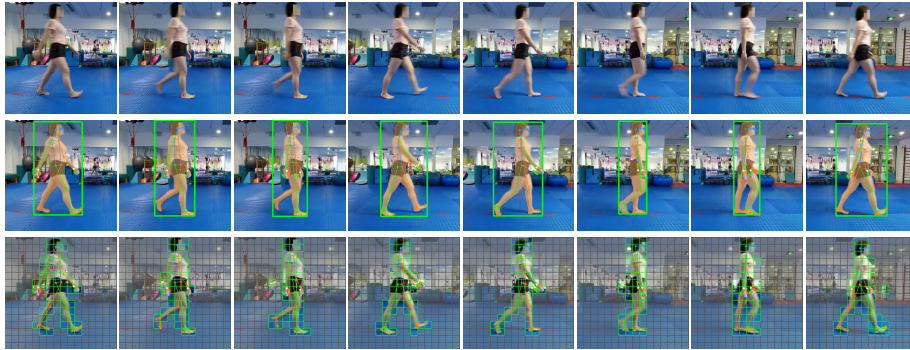


Fig. 1: Video Examples and Annotations. *Top:* An example of a raw video of a single gait cycle. *Middle:* We use SAM 3 to perform instance segmentation and Sapiens-2B to perform pose estimation to obtain bounding boxes, masks, and keypoints. Video then incorporates keypoints as token-level prompts. *Bottom:* We use these annotations to conduct mask-guided pruning to ignore irrelevant background noise.

gait modeling pipelines. Analyzing pediatric gait in therapeutic and diagnostic contexts creates a critical intersection for computer vision researchers and clinical practitioners to investigate fundamental aspects of early motor development. This collaboration opens avenues to explore profound clinical questions, such as whether subtle gait deviations can function as early indicators of autism spectrum disorder, how effectively BCI assistive technologies enhance gait kinematics, or how subsequent walking abnormalities link back to atypical infant general movements. Resolving these child-centered inquiries through a data-driven lens requires the computer vision community to pioneer new, robust methodologies capable of modeling complex, dynamic, developmental gait behaviors from unconstrained video and other sensing modalities.

4 Children Gait Video (CGV) Dataset

We introduce a dataset named CGV for the development of children’s gait modeling. This is the first open-sourced children’s gait video dataset. Before the data annotation and model design, the IRB approval is obtained from affiliated hospitals. The dataset contains 339,236 frames (1,185 videos) of size 1920×1080 (2K) with 17 EVGS [80] sub-item annotations per limb and the diagnosis results of multiple gait abnormalities, such as Cerebral Palsy (CP), Traumatic Brain Injury (TBI), Developmental Dysplasia of the Hip (DDH), Toe in, Idiopathic Toe Walking (ITW). The videos were recorded at a children’s hospital in Asia using smartphone cameras and action cameras, positioned simultaneously to capture sagittal and coronal views. Figure 1 shows examples of videos and annotations. In total, 110 patients participated in the data collection. In some cases, multiple recordings per view were available. The dataset provides detailed annotations per frame and per video, including the subject’s body bounding box (detected with SAM 3 [17] and manually selected by human annotators), the 2D human keypoints (detected with Sapiens-2B [54] and manually adjusted by human annotators), and the EVGS sub-items (annotated by an experienced

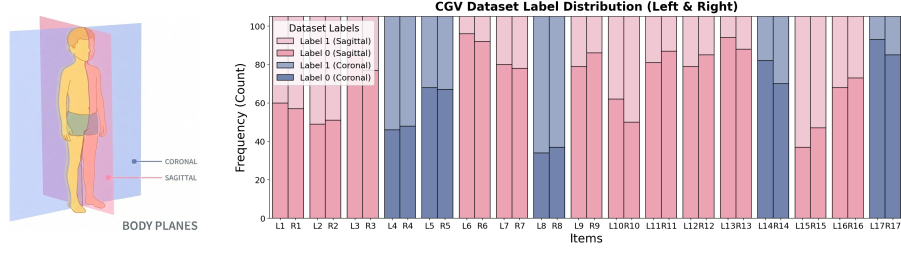


Fig. 2: The Overview of our CGV Dataset. The CGV dataset comprises 110 pediatric patients, with each evaluation covering a total of 2×17 EVGS scoring items (17 items per limb) across two body planes. *Left:* The child’s body planes. *right:* The label distribution of the CGV dataset.

pediatricians in the author team and reviewed by a senior pediatrician author with 40 years of clinical experience), the diagnosis results tracing from the patient’s follow-up record. Figure 2 shows the EVGS scores distribution among all patients. Appendix A shows the demographic statistics of the CGV dataset.

To capture the dataset, a primary camera is positioned at the terminus of an 8-meter walkway to record the coronal (frontal and posterior) view. A secondary camera is oriented orthogonally, facing the center of the walkway, to capture the sagittal (lateral) view. This lateral camera is positioned at a sufficient distance to ensure its field of view encompasses the middle four meters of the trial

Table 1: 17 Scoring Items in CGV Referring to EVGS.

No.	Gait Pattern	Abbr.	Score Scale	View	# Videos
1	Initial Contact in Stance	IC	0, 1, 2	Sagittal	647
2	Heel Lift in Stance	HL	0, 1, 2	Sagittal	647
3	Max Ankle Dorsiflexion in Stance	SAD	0, 1, 2	Sagittal	647
4	Hind-foot Varus/Valgus in Stance	HVV	0, 1, 2	Coronal	261
5	Foot Rotation in Stance	FRT	0, 1, 2	Coronal	277
6	Foot Clearance in Swing	FCL	0, 1	Sagittal	647
7	Max Ankle Dorsiflexion in Swing	WAD	0, 1, 2	Sagittal	647
8	Knee Progression Angle in Mid-Stance	KPA	0, 1, 2	Coronal	277
9	Peak Knee Extension in Stance	KEX	0, 1, 2	Sagittal	647
10	Knee Position in Terminal Swing	KPS	0, 1, 2	Sagittal	647
11	Peak Knee Flexion in Swing	KFX	0, 1, 2	Sagittal	647
12	Peak Hip Extension in Stance	HEX	0, 1, 2	Sagittal	647
13	Peak Hip Flexion during Swing	HFX	0, 1, 2	Sagittal	647
14	Pelvic Obliquity at Mid-Stance	POB	0, 1, 2	Coronal	277
15	Pelvic Rotation at Mid-Stance	PRT	0, 1, 2	Coronal	277
16	Peak Sagittal Trunk Position in Stance	TSG	0, 1, 2	Sagittal	647
17	Maximum Trunk Lateral Shift	TLT	0, 1, 2	Coronal	277

space. This specific distance is calibrated to guarantee the capture of 2-3 complete gait cycles (strides) per subject. Table 1 details the 17 fine-grained gait parameters annotated in the CGV dataset, which strictly adhere to the EVGS reference guide to ensure high clinical validity. While the standard EVGS protocol employs a three-point ordinal scale (0: normal, 1: moderate deviation, 2: severe deviation) for each parameter, the natural distribution of pediatric gait pathologies inherently results in a severe class imbalance, particularly for the most extreme deviations (score 2). To mitigate this imbalance and establish a robust computational benchmark, we binarize the assessment by merging scores 1 and 2. Consequently, the prediction for each fine-grained gait parameter is formulated as a binary classification task (see Figure 2) discriminating between “typical” and “atypical” gait patterns. Detailed EVGS scoring criteria are provided in Appendix B. **Expert Agreement.** To further verify the agreements between experts and the objectiveness of our clinical annotations, we add an additional annotation round from another pediatrician and calculate the intra-class correlation coefficient (ICC) with ours, achieving $ICC = 0.93$. We also invite the pediatrician as a human baseline, achieving an average scoring accuracy of 93.8%.

Ethical Considerations and Data Privacy. Consistent with established ethical frameworks for pediatric facial, pose, and behavioral datasets [14, 45, 71, 72, 81, 95], we have implemented a multi-layered protocol to ensure participant protection. The CGV dataset was approved as a retrospective study by the Institutional Review Board (IRB) of our affiliated clinical institution (approval number: 202106202). Our commitment to participant privacy and data security is reflected in the following measures: **(1) Irreversible De-identification:** To protect the identities of the children, all original video data underwent an irreversible obfuscation process. We employed an automated preprocessing pipeline utilizing RetinaFace [26] for face detection and SAM 3 [17] for precise segmentation, applying a mosaic filter to all identifiable regions. To ensure 100% efficacy, the output was manually verified by two independent human annotators. **(2) Restrictive Licensing and AI Governance:** The dataset is released under the CC BY-NC 4.0 license. The public available data is the children’s pose sequence. Furthermore, we mandate that video data and annotation from this repository may not be used as training material for large-scale foundation models without explicit, separate authorization, safeguarding against unauthorized generative use.

Experimental Setup. To prevent identity leakage and shortcut bias, we implement a strict object-level data split based on the unique **Patient ID** in the experiment. The test set is randomly selected, and we ensure all positive samples and negative sample nearly balanced. All videos and derived gait cycles for any given patient belong exclusively to either the training or the test set. Finally, the ratio of patients in the training set to those in the test set is 6:1. Then, we evaluate all models on the CGV test split on an item-wise basis. We report the per-item percentage accuracy and per-limb average accuracy across all 34 bilateral scoring items to provide a holistic measure of clinical reliability. All experiments are conducted on a cluster equipped with 10 NVIDIA L40S GPUs. For model training in Section 6 and Section 7, we use AdamW as the optimizer with $\eta = 10^{-4}$ and a weight decay of 0.05. We employ a linear warm-up schedule over the first 5% of the total training epochs to stabilize early optimization. The batch size is configured to 8 per GPU.

5 Benchmarking Children’s Gait Analysis

Baselines. We evaluate a wide range of zero-shot multimodal LLMs like Gemini 3 Pro [43], GPT-5.2 [74], Qwen3-VL-235B [58], GLM4.6V [113], Qwen3.5-9B [98], and InternVL3-8B [125], validating if these SoTA methods can resolve the children’s gait analysis problem. All models take as input uniformly sampled $T = 16$ frames from the center window of the downsampled 30 FPS video and the prompt **P**. Appendix C shows the additional details of **P**.

Experimental Results. As demonstrated in Table 2, all listed MLLMs fail to effectively resolve the required clinical nuances, with the average accuracy of each limb range 50% to 60%, which is only marginally above random guessing for binary scoring tasks. This performance gap indicates that, although these models exhibit strong general visual-language reasoning abilities, they struggle

Table 2: Zero-shot Quantitative Evaluation of Mainstream MLLMs. All reported values are **percentages (%)**. We report 17 scoring items for the **Left (L-)** limb (top) and **Right (R-)** limb (bottom). The detailed definition of each item is shown in Table 1. L/R-AVG indicates the average accuracy over all scoring items of the left/right limb.

Method	L-IC	L-HL	L-SAD	L-WAD	L-HVV	L-FRT	L-FCL	L-KPA	L-KEX	L-KPS	L-KFX	L-HEX	L-HFX	L-POB	L-PRT	L-TSG	L-TLT	L-AVG
Gemini 3 Pro [43]	52	48	41	59	39	57	52	46	67	48	52	63	59	46	41	70	54	53
GPT-5.2 [74]	56	63	44	59	42	54	56	54	48	44	59	63	59	54	44	56	54	53
Qwen3-VL-235B [58]	59	44	70	48	58	57	56	46	59	48	56	56	59	50	48	52	57	54
GLM4.6V [113]	56	48	67	41	39	57	63	46	56	48	52	59	63	50	37	52	57	52
Qwen3.5-9B [98]	48	44	63	44	35	57	59	39	52	44	56	56	59	54	41	48	54	50
InternVL3-8B [125]	52	44	63	44	42	61	59	57	52	44	56	56	59	54	41	48	68	53

Method	R-IC	R-HL	R-SAD	R-WAD	R-HVV	R-FRT	R-FCL	R-KPA	R-KEX	R-KPS	R-KFX	R-HEX	R-HFX	R-POB	R-PRT	R-TSG	R-TLT	R-AVG
Gemini 3 Pro [43]	59	59	48	44	42	57	48	46	52	52	59	67	70	54	59	63	64	55
GPT-5.2 [74]	63	59	63	52	46	61	59	46	52	63	52	44	67	50	59	44	64	56
Qwen3-VL-235B [58]	52	59	48	67	58	61	70	43	67	56	56	70	59	46	67	63	57	59
GLM4.6V [113]	44	56	48	59	42	64	56	39	67	48	59	63	63	50	52	59	61	55
Qwen3.5-9B [98]	52	48	48	59	42	57	56	46	59	41	59	56	56	46	59	59	54	53
InternVL3-8B [125]	44	48	48	59	46	57	56	57	59	41	59	56	56	46	59	59	64	54

to capture subtle kinematic deviations and fine-grained temporal dynamics that are critical for clinical gait assessment. More specifically, zero-shot MLLMs tend to correctly identify visually obvious abnormalities like Peak Hip Flexion during Swing (HFX) with an average accuracy of 61%, but frequently misclassify subtle impairments such as Max Ankle Dorsiflexion in Stance (SAD) with an average accuracy of 54%. The results suggest that pre-trained MLLMs are biased toward semantic-level understanding rather than quantitative biomechanical interpretation. Additionally, the lack of task-specific alignment between medical scoring criteria and general-purpose language supervision further limits their discriminative capacity. These findings highlight the necessity of domain-adaptive fine-tuning and structured supervision to bridge the gap between generic multimodal reasoning and clinically grounded gait analysis.

6 Decoding Children’s Gait via Vision-Language Models

We first validate whether Video Vision-Language Models (VLMs) are the best solution for children’s gait visual analysis. We propose a multimodal reasoning framework based on a fine-tuned Qwen3-VL-4B [58] to map pediatric gait patterns to all fine-grained items in CGV, denoted as Qwen3-VL-ChildGait.

Multimodal Representation. To capture the complex kinematics as we described, each video frame at step t is represented as a composite input $\mathcal{F}_t = \{\mathbf{I}_t, \mathbf{K}_t, \mathbf{M}_t\}$, where $\mathbf{I}_t$ denotes the original RGB frame, $\mathbf{K}_t$ denotes the skeletal keypoints, and $\mathbf{M}_t$ denotes the instance segmentation mask. This multi-view visual evidence allows the model’s transformer layers to attend to both fine-grained joint angles and global body morphology, filtering out clinical background noise.

Table 3: Comparison Between Fine-tuned (Or Not) Qwen3-VL-4B. All reported values are **percentages (%)**. We report 17 scoring items for the **Left (L-)** limb (top) and **Right (R-)** limb (bottom). The detailed definition of each item is shown in Table 1. L/R-AVG indicates the average accuracy over all scoring items of the left/right limb.

Method	L-IC	L-HL	L-SAD	L-WAD	L-HVV	L-FRT	L-FCL	L-KPA	L-KEX	L-KPS	L-KFX	L-HEX	L-HFX	L-POB	L-PRT	L-TSG	L-TLT	L-AVG
Qwen3-VL-4B [58]	52	44	63	44	42	54	59	50	52	44	56	56	59	46	41	48	54	51
Qwen3-VL-ChildGait	63	56	59	48	39	57	52	46	52	56	52	56	63	50	41	48	57	53

Method	R-IC	R-HL	R-SAD	R-WAD	R-HVV	R-FRT	R-FCL	R-KPA	R-KEX	R-KPS	R-KFX	R-HEX	R-HFX	R-POB	R-PRT	R-TSG	R-TLT	R-AVG
Qwen3-VL-4B [58]	44	48	48	59	42	46	56	36	59	41	59	56	56	54	59	59	54	52
Qwen3-VL-ChildGait	56	52	56	67	42	61	56	43	59	59	67	56	56	46	59	59	57	53

Clinical Goal-Oriented Fine-tuning. We formulate gait analysis as a sequence-to-label task. Given a video sequence $\mathbf{F} = \{\mathcal{F}_t\}_{t=1}^T$ which is downsampled to 30 FPS and where we randomly crop $M=4$ temporal windows and uniformly sample $T = 16$ per window during training and uniformly sample $T = 16$ in the single central temporal window during testing, the model is fine-tuned to predict a categorical gait score $y \in \mathcal{Y}$. The training objective minimizes the negative log-likelihood of the target tokens:

$$\mathcal{L}_{\text{vlm}}(\boldsymbol{\theta}) = - \sum_j j \log p_{\boldsymbol{\theta}}(w_j \mid w_{<j}, \mathbf{F}, \mathbf{P}), \quad (1)$$

where $\mathbf{P}$ is a task-specific prompt instructing the model to synthesize visual trajectories into gait quality ratings. This approach uses the VLM’s pre-trained spatial reasoning to identify sub-second deviations that traditional adult-centric models often miss. Appendix C shows the additional details of $\mathbf{P}$.

Experimental Results. We compare zero-shot Qwen3-VL-4B [58] with the fine-tuned Qwen3-ChildGait on the CGV test split. As shown in Table 3, fine-tuned Qwen3-VL-4B has no significant improvement in performance compared to the base model, with the average accuracy of left and right limbs increasing 2% and 1%, respectively. While certain scoring items exhibit little gains in accuracy, such as Initial Contact in Stance (IC) with +11.5%, some of the other items show decreased performance after fine-tuning, such as Hind-foot Varus/Valgus in Stance (HVV) with -1.5%, resulting in a mixed overall average. The counterintuitive result demonstrates that VLMs do not perform well on the subtle task of child gait recognition, which aligns with recent findings that VLMs face fundamental limitations in fine-grained action recognition and dynamic spatiotemporal interactions [59, 122]. The core of VLMs’ training lies in modeling high-dimensional discrete tokens. By performing causal language modeling across trillions of text corpora, it fundamentally captures the statistical distribution of human knowledge within logical and semantic spaces. Therefore, knowledge compressed from human language results cannot be generalized to motion recognition [101], much less to the more refined task of gait recognition.

7 Decoding Children’s Gait via Video-based Models

The failure of Video VLMs prompts us to rethink whether existing video-based foundation and gait analysis models can resolve this problem after fully fine-tuning in CGV.

7.1 Fine-tuning Video Foundation and Gait Analysis Models

We establish our video model baseline by directly fine-tuning VideoMAE v2 [104], a state-of-the-art video foundation model. It is pre-trained via masked autoencoding on the UnlabeledHybrid dataset to capture generic spatiotemporal knowledge, and subsequently fine-tuned with supervision on Kinetics-710 [57] to acquire human motion recognition capabilities.

In addition to the video foundation model baseline, we also evaluate a wide range of representative gait analysis baselines. These include the current state-of-the-art (SoTA) model BiggerGait [110], as well as its precursors: GaitSet [19,20], GaitPart [37], GaitGL [63], GaitBase [35], SwinGait [32], and DeepGaitV2 [32]. These models are first pre-trained on various large-scale datasets (*e.g.*, CCPC [60], Gait3D [120], Gait3D-Parsing [121], SUSTech1K [85], GREW [127], OUMVLPP [96]) and subsequently fine-tuned on the CGV dataset.

Video Downsample and Clip. To standardize the temporal input distribution and align it with the model’s expectation, the raw video stream is first downsampled to a fixed frame rate of 30 FPS, ensuring that the physical time stride between adjacent frames remains constant across the dataset. For temporal augmentation during training, we randomly crop $M = 4$ temporal windows and uniformly sample $T = 16$ frames per window. During testing, a single $T = 16$ frame sequence is uniformly extracted from the video’s central temporal window for deterministic evaluation. Thus, the visual input is formally defined as $\mathbf{V} \in \mathbb{R}^{T \times 3 \times H \times W}$, where H and W denote the spatial resolution, which is then flattened to patches. This ensures a sufficient temporal receptive field to capture a complete gait cycle phase without overwhelming the computational budget.

Task-Specific Architecture Adaptation. Let the output from the final transformer block be denoted as $\mathbf{X} \in \mathbb{R}^{B \times N \times D}$, where B is the batch size, N is the total number of spatiotemporal patch tokens, and D is the embedding dimension.

To aggregate the global context, we perform spatiotemporal global average pooling over all N patch tokens, yielding a compact clip-level feature of shape $B \times D$. Finally, a newly initialized linear classification head projects the normalized features into a two-dimensional logit vector:

$$\mathbf{L} = \mathbf{W}_{\text{head}} \left(\frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \right) + \mathbf{b}_{\text{head}}, \quad (2)$$

where $\mathbf{x}_i$ represents the i -th token, and $\mathbf{W}_{\text{head}} \in \mathbb{R}^{2 \times D}$. The final prediction is obtained via the argmax operation over the two logits, and the model is optimized end-to-end using the Soft Target Cross-Entropy loss:

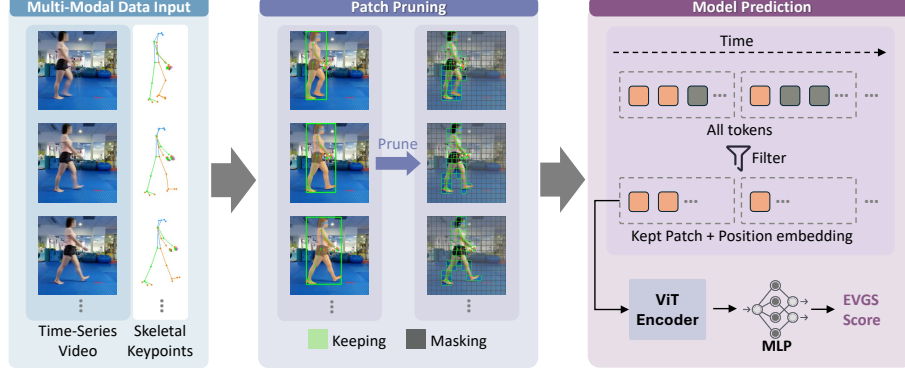


Fig. 3: ChildGait-Video Framework. We render the skeletal keypoints on the original RGB frame as the token-level kinematic prompts and then perform mask-guided patch pruning to force the model to focus on the foreground patient. The processed input is then passed into the model to predict EVGS scoring items.

$$\mathcal{L}_{\text{video}}(\theta) = - \sum_{k \in \{0,1\}} q_k \log (\text{Softmax}(\mathbf{L}(\theta))_k), \quad (3)$$

where $\mathbf{q} = (\lambda, 1 - \lambda)$, $\lambda \in [0, 1]$ is the proportion of positive/negative samples.

7.2 Designing ChildGait-Video Baseline Aligning with Clinical Intuition

Traditional gait analysis methods mostly follow a multi-stage pipeline: pose estimation for keypoint extraction, geometric calculation of joint angles, and rule-based scoring. However, this geometry-centric paradigm suffers from two major limitations. First, compressing high-dimensional video data into sparse skeletal representations inevitably discards rich appearance and texture information. Second, multi-stage pipelines typically evaluate specific joint angles in isolation, neglecting the inherent kinematic synergy of human movement.

As shown in Figure 3, to address these limitations, we propose **ChildGait-Video**, an end-to-end model that directly maps visual features to EVGS items. ChildGait-Video uses the same architecture of vision encoder as VideoMAE v2 [104], followed by an MLP head to predict EVGS items. This end-to-end design inherently mitigates multi-stage error accumulation and uses the strong spatiotemporal representation capabilities of video foundation models to capture the global motion patterns and inter-dependencies among EVGS items.

To unleash the potential of the model for pediatric gait analysis, we propose a two-stage adaption paradigm. We first pre-train ChildGait-Video on Kinetics-710 [57] to acquire human motion recognition capability, and then supervised fine-tuning the model on our proposed CGV dataset with **Token-Level Kinematic Prompting** and **Mask-Guided Patch Pruning** to further make the model learn the subtle gait anomalies as well as focus on the gait transition.

Anatomical Alignment via Token-Level Kinematic Prompting. Although pre-trained on Kinetics-710, the model still lacks an explicit understanding of the underlying anatomical priors of children, which is a critical requirement for identifying subtle gait anomalies. To bridge this semantic gap, we introduce skeletal keypoints as token-level kinematic prompts.

We directly render the extracted joint coordinates $\mathbf{K}_t$ and their connectivity graph onto the original RGB frame $\mathbf{I}_t$. Let $\mathcal{R}(\cdot)$ denote this rendering operation. The prompted input frame is formulated as $\hat{\mathbf{I}}_t = \mathcal{R}(\mathbf{I}_t, \mathbf{K}_t)$. During the flattening process in Vision Transformer (ViT) encoder [30], the image is divided into patches that encompass the rendered skeletal keypoints and connections, inherently encapsulating both local anatomical topology and the original appearance features. When projected into the embedding space, these specific patches act as token-level kinematic prompts mixed with standard visual tokens.

Background Suppression via Mask-Guided Patch Pruning. Gait videos are often cluttered with irrelevant background noises, which easily introduce biases during fine-tuning. Inspired by the inherent sparsity of the MAE architecture, we utilize instance segmentation to extract binary subject masks $\mathbf{M} \in \{0, 1\}^{H \times W}$. Rather than employing random masking, we perform deterministic spatial pruning guided by these masks.

Specifically, a processed RGB frame $\hat{\mathbf{I}}_t$ is divided into a grid of non-overlapping patches. A patch is retained only if its corresponding region in $\mathbf{M}$ contains a sufficient proportion of foreground pixels. Let $\mathbf{E}_v \in \mathbb{R}^{N_{\text{foreground}} \times D}$ denote the embedded representations of the retained foreground patches, where $N_{\text{foreground}} \ll N_{\text{total}}$. This mask-guided token dropping strategy not only explicitly eliminates background noise, forcing the transformer’s self-attention to strictly allocate its representational capacity to the subject’s gait, but also substantially reduces the computational overhead during fine-tuning.

Experimental Results. As shown in Table 4, gait analysis models also struggle to perform precise classification in a clinical setting, yielding low average accuracies that range from just 45% to 56%. Notably, SwinGait pre-trained on Gait3D and GaitBase pre-trained on GREW achieve marginally better results, peaking at average accuracies of 54% and 56% for the left and right limbs, respectively. Although being the current SoTA gait analysis model, BiggerGait still fails to do the classification precisely with low average accuracies of 54% and 53%, respectively, for left and right limbs. We attribute their low accuracy to the inherent design objectives of traditional gait recognition networks, which focus on extracting global spatiotemporal representations to distinguish identities, thereby neglecting the fine-grained, localized kinematic details essential for clinical scoring. Furthermore, there exists a domain gap between the healthy adult subjects in the pre-training datasets and the distinct pathological patterns of children’s gait in the CGV dataset, which limits their feature representation and generalization capabilities.

On the other hand, fine-tuned VideoMAE v2 has surpassed all previous baselines in almost every scoring item with average accuracies of 69% and 72%, increased by 15% and 19% compared to BiggerGait.

Table 4: Quantitative Evaluation of Video Analysis Models. All reported values are **percentages (%)**. We report 17 scoring items for the **Left (L-)** limb (top) and **Right (R-)** limb (bottom). The detailed definition of each item is shown in Table 1. L/R-AVG indicates the average accuracy over all scoring items of the left/right limb. L/R-F1 indicates the average F1-score across all items of the left/right limb.

Method	Pretrain	L-IC	L-HL	L-SAD	L-WAD	L-HVY	L-FRT	L-FCL	L-KPA	L-KEX	L-KPS	L-KFX	L-HEX	L-HFX	L-POB	L-PRT	L-TSG	L-TLT	L-AVG	L-F1
GaitSet [19, 20]	Gait3D-Parsing [121]	48	44	48	56	54	43	56	54	52	63	48	59	44	50	41	52	57	51	-
GaitPart [37]	Gait3D-Parsing [121]	37	44	37	26	30	61	59	50	52	33	63	70	52	50	56	48	43	48	-
GaitGL [63]	Gait3D-Parsing [121]	52	56	37	44	39	54	56	54	48	44	37	56	41	50	59	74	32	49	-
GaitBase [35]	Gait3D [120]	56	63	48	44	39	68	22	46	44	48	52	37	41	54	70	44	64	50	-
GaitBase [35]	GREW [127]	41	70	37	44	69	61	52	57	52	59	44	67	59	50	59	41	43	53	-
GaitBase [35]	OUMVLP [96]	56	70	37	37	54	50	44	54	52	59	44	41	41	46	26	52	57	48	-
SwinGait [32]	CCPG [60]	52	56	30	56	62	57	67	50	48	44	41	44	41	57	59	40	39	50	-
SwinGait [32]	Gait3D [120]	63	44	52	52	39	43	59	50	52	41	44	85	60	57	59	59	57	54	-
SwinGait [32]	SUSTech1K [85]	63	44	48	44	19	43	59	50	59	56	59	59	52	46	67	48	57	52	-
DeepGaitV2 [32]	CCPG [60]	30	52	63	56	62	54	59	43	74	37	44	44	41	61	30	44	54	50	-
DeepGaitV2 [32]	Gait3D [120]	48	67	37	44	62	57	70	32	22	56	44	48	41	50	59	48	61	50	-
DeepGaitV2 [32]	GREW [127]	52	41	41	56	62	57	59	54	52	41	44	56	41	50	33	48	57	50	-
DeepGaitV2 [32]	OUMVLP [96]	48	41	48	52	62	64	48	54	48	44	37	44	59	54	37	52	43	49	-
DeepGaitV2 [32]	SUSTech1K [85]	52	44	37	48	35	43	59	50	48	56	56	59	50	59	59	57	51	-	-
BiggerGait [110]	CCPG [60]	48	56	63	44	62	43	59	54	52	56	56	59	50	59	52	57	54	0.53	-
VideoMAE v2 [104]	Kinetics-710 [57]	72	70	70	63	67	68	81	80	60	60	63	67	74	68	80	67	68	69	0.70
ChildGait-Video (Ours)	Kinetics-710 [57]	85	92	81	70	88	92	89	84	74	82	78	85	93	89	89	85	76	84	0.83

Method	Pretrain	R-IC	R-HL	R-SAD	R-WAD	R-HVY	R-FRT	R-FCL	R-KPA	R-KEX	R-KPS	R-KFX	R-HEX	R-HFX	R-POB	R-PRT	R-TSG	R-TLT	R-AVG	R-F1
GaitSet [19, 20]	Gait3D-Parsing [121]	44	48	63	41	58	39	56	57	59	67	41	59	52	46	30	37	57	50	-
GaitPart [37]	Gait3D-Parsing [121]	52	56	52	44	42	43	56	54	41	37	56	67	48	46	41	59	43	49	-
GaitGL [63]	Gait3D-Parsing [121]	44	48	52	59	42	57	52	57	41	41	48	59	44	46	37	52	32	48	-
GaitBase [35]	Gait3D [120]	44	41	37	37	42	50	26	43	59	48	59	41	37	50	63	33	50	45	-
GaitBase [35]	GREW [127]	56	74	52	59	73	57	52	61	59	59	41	63	56	54	41	59	36	56	-
GaitBase [35]	OUMVLP [96]	67	70	63	56	42	54	52	57	56	44	41	37	59	50	52	41	57	53	-
SwinGait [32]	CCPG [60]	44	52	56	41	58	68	63	46	48	41	37	41	44	61	41	30	39	48	-
SwinGait [32]	Gait3D [120]	67	44	67	44	42	39	67	54	44	52	59	85	56	61	56	52	57	56	-
SwinGait [32]	SUSTech1K [85]	78	59	67	63	23	40	56	46	59	59	37	59	52	50	52	56	57	54	-
DeepGaitV2 [32]	CCPG [60]	22	70	48	48	65	50	67	46	56	30	44	44	44	57	37	52	54	49	-
DeepGaitV2 [32]	Gait3D [120]	41	74	52	59	58	61	70	29	48	33	41	44	44	46	41	59	46	50	-
DeepGaitV2 [32]	GREW [127]	44	48	52	41	58	61	56	57	48	48	41	56	44	54	48	59	57	51	-
DeepGaitV2 [32]	OUMVLP [96]	56	52	41	41	58	54	52	57	41	41	52	44	56	57	59	41	43	50	-
DeepGaitV2 [32]	SUSTech1K [85]	44	56	52	59	46	39	56	46	59	59	63	56	56	54	41	56	57	53	-
BiggerGait [110]	CCPG [60]	56	52	48	59	58	39	56	57	59	59	59	56	56	46	41	41	57	53	0.51
VideoMAE v2 [104]	Kinetics-710 [57]	67	71	61	81	60	68	85	72	78	81	63	67	74	72	70	85	68	72	0.71
ChildGait-Video (Ours)	Kinetics-710 [57]	89	85	74	85	83	88	93	80	92	86	81	81	89	80	74	92	76	84	0.83

Finally, ChildGait-Video achieves SoTA performance across all baselines, reaching a highest accuracy range of 70% $\sim$ 93% in all 34 scoring items. It gains average accuracy improvements of 15% and 12% for left and right limbs compared to the fine-tuned VideoMAE v2. The average F1-scores also achieve a high value of 0.83. This SoTA performance can be attributed to the interaction within the self-attention mechanism. We also evaluate skeleton baselines in Appendix D.

Statistical Analysis. We employ McNemar’s test on ChildGait-Video and fine-tuned VideoMAE v2. Let b denote the number of data ours gets correct while the baseline gets wrong, and c the opposite. We compute the p -value using a binomial test under the null hypothesis that both models have equal accuracy, and get $(b, c) = (28, 6), p = 2.3 \times 10^{-4} < 0.001$, demonstrating that the performance improvement is highly statistically significant.

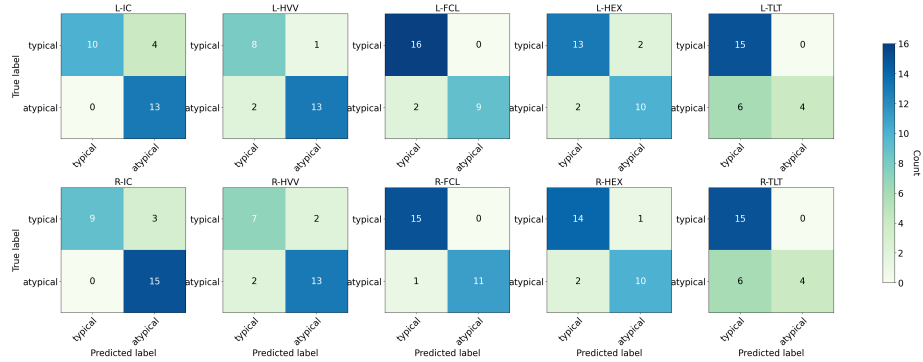


Fig. 4: Visualization of Confusion Matrices. We evaluate the performance of ChildGait-Video on EVGS scoring classification in L/R-IC, L/R-HVV, L/R-FCL, L/R-HEX, and L/R-TLT.

Confusion Matrices. We also visualize confusion matrices across L/R-IC, L/R-HVV, L/R-FCL, L/R-HEX, and L/R-TLT, which cover both sagittal and coronal views of scoring items in EVGS. As depicted in Figure 4, our framework maintains high sensitivity for both sagittal view items and coronal view items. The low error rate of off-diagonal for these items suggests that our method effectively regularizes the self-attention mechanism to focus on gait kinematics. More confusion matrices across all items can be found in Appendix E.

7.3 Ablation Study on Frames

The temporal resolution of the input sequence determines the model’s capacity to capture the dynamic evolution of gait pathologies. We investigate the impact of the number of sampled frames $T \in \{8, 16, 32\}$ while maintaining a constant sampling rate of 30 FPS through training and evaluating on item **L/R-IC**. All variants leverage the proposed mask-guided pruning and visual prompting with ChildGait-Video. We also conduct a module ablation study in Appendix F.

As shown in Table 5, increasing the frame count from 8 to 16 yields a significant performance gain with an increase of 12.9%. The performance is further increased to 90.7% if the number of frames is increased to 32. We attribute this to the fact that 8 frames are insufficient to encompass a meaningful kinematic transition in pediatric gait. In contrast, 16 frames effectively capture critical sub-phases of the gait cycle. Interestingly, further increasing the temporal window to 32 frames results in only a slight improvement in performance of 3.7%, as computational complexity grows quadratically, although the field of view has doubled.

Table 5: Ablation Study on the Number of Input Frames. We train ChildGait-Video with different numbers of frames and evaluate the performance on the item L/R-IC with the metric Accuracy and F1-Score.

Frames (T)	Temporal Span (s)	Accuracy (%)	F1-Score
8	0.26	74.1	0.73
16	0.53	87.0	0.89
32	1.06	90.7	0.92

8 Discussion

The CGV dataset and its accompanying framework are fundamentally motivated by the potential of artificial intelligence to democratize and enhance pediatric healthcare [13,62]. A primary objective is to automate clinical gait evaluation, ultimately scaling these capabilities to unconstrained, “in-the-wild” environments. Currently, traditional 2D observational tools, such as the Edinburgh Visual Gait Score (EVGS), rely heavily on expert visual inspection. This reliance introduces inherent subjectivity and inter-rater variability, limiting reproducibility across different evaluators and institutions. Furthermore, human-administered video assessments are also highly sensitive to environmental confounders.

To overcome these limitations, we establish the first comprehensive visual benchmark for children’s gait analysis. We evaluate the efficacy of contemporary foundation models and introduce a specialized, multi-stage pipeline, ChildGait-Video, integrating instance segmentation, anatomical pose estimation, and spatiotemporal modeling. Empirical results demonstrate that our framework achieves an agreement with expert annotations ranging from 70% to 93% in all items for typical and atypical distinction, with an average 0.83 F1 score. Crucially, unlike marker-based 3D motion capture systems that demand expensive, specialized infrastructure, our approach operates on standard RGB camera setups. This drastically lowers both technical and economic barriers, facilitating scalable deployment across heterogeneous and resource-constrained clinical settings.

While the current study establishes robust model-level validation, future efforts will extend this framework to open-scene data collected directly from home, rehabilitation centers, and community environments. This expansion will improve algorithmic generalization across a broader spectrum of complex pathological gait patterns and unconstrained acquisition scenarios. Deploying such technology in the wild enables reproducible longitudinal tracking and cross-site standardization, which are essential for large-scale outcome evaluation and data-driven rehabilitation research. Ultimately, the integration of robust computer vision methodologies into routine pediatric care heralds a paradigm shift: moving from subjective, observer-dependent evaluations toward highly scalable, objective, and computationally reproducible clinical workflows [12, 15].

9 Conclusion

In this work, we introduce fine-grained pediatric gait understanding from standard RGB videos as a new computer vision problem and present CGV, a large-scale multi-view dataset with synchronized anonymized pose, segmentation, and clinically grounded EVGS-derived annotations. Through systematic benchmarking, we show that contemporary zero-shot MLLMs and fine-tuned VLMs remain unreliable for phase-sensitive clinical gait scoring. To address these challenges, we propose ChildGait-Video, an end-to-end adaptation paradigm. This design substantially improves per-item EVGS classification. We expect CGV and the proposed framework to serve as a rigorous benchmark for pediatric gait research and to facilitate scalable.

References

1. Aman, N., Islam, M.R., Ahamed, M.F., Ahsan, M.: Performance evaluation of various deep learning models in gait recognition using the casia-b dataset. *Technologies* **12**(12), 264 (2024)
2. Anderson, A.J., Eguren, D., Gonzalez, M.A., Caiola, M., Khan, N., Watkinson, S., Zuccaroli, I., Hirczy, S.S., Zabetian, C.P., Mills, K., et al.: Weargait-pd: An open-access wearables dataset for gait in parkinson’s disease and age-matched controls. *Scientific Data* (2026)
3. Armand, S., Decoulon, G., Bonnefoy-Mazure, A.: Gait analysis in children with cerebral palsy. *EFORT open reviews* **1**(12), 448–460 (2016)
4. Baker, R.: Gait analysis methods in rehabilitation. *Journal of neuroengineering and rehabilitation* **3**(1), 4 (2006)
5. Baker, R., Esquenazi, A., Benedetti, M.G., Desloovere, K., et al.: Gait analysis: clinical facts. *Eur. J. Phys. Rehabil. Med* **52**(4), 560–574 (2016)
6. Ben Chaabane, N., Conze, P.H., Lempereur, M., Quellec, G., Rémy-Néris, O., Brochard, S., Cochener, B., Lamard, M.: Quantitative gait analysis and prediction using artificial intelligence for patients with gait disorders. *Scientific Reports* **13**(1), 23099 (2023)
7. Beynon, S., McGinley, J.L., Dobson, F., Baker, R.: Correlations of the gait profile score and the movement analysis profile relative to clinical judgments. *Gait & posture* **32**(1), 129–132 (2010)
8. Blanco-Diaz, C.F., da Silva Serafini, E.R., Bastos-Filho, T., de Azevedo Dantas, A.F.O., do Espirito Santo, C.C., Delisle-Rodriguez, D.: A gait imagery-based brain–computer interface with visual feedback for spinal cord injury rehabilitation on lokomat. *IEEE Transactions on Biomedical Engineering* **72**(1), 102–111 (2024)
9. Bonanno, M., De Nunzio, A.M., Quartarone, A., Militi, A., Petralito, F., Calabrò, R.S.: Gait analysis in neurorehabilitation: from research to clinical practice. *Bioengineering* **10**(7), 785 (2023)
10. Bourgeois, A.B., Mariani, B., Aminian, K., Zambelli, P., Newman, C.: Spatio-temporal gait analysis in children with cerebral palsy using, foot-worn inertial sensors. *Gait & posture* **39**(1), 436–442 (2014)
11. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters* **140**, 325–331 (2020)
12. Cao, X., Cao, J.: Commentary: Machine learning for autism spectrum disorder diagnosis—challenges and opportunities—a commentary on schulte-rüther et al.(2022). *Journal of Child Psychology and Psychiatry* **64**(6), 966–967 (2023)
13. Cao, X., Chen, J., Ye, W., Jojic, A., Owusu, S.A., Li, S., Coffee, M., Zhao, S., Rehg, J.M.: Workshop on ai for children: Healthcare, psychology, education. In: *ICLR 2025 Workshop Proposals* (2025)
14. Cao, X., Li, X., Ma, L., Huang, Y., Feng, X., Chen, Z., Zeng, H., Cao, J.: Aggpose: Deep aggregation vision transformer for infant pose estimation. *arXiv preprint arXiv:2205.05277* (2022)
15. Cao, X., Ye, W., Sizikova, E., Bai, X., Coffee, M., Zeng, H., Cao, J.: Vitasd: Robust vision transformer baselines for autism spectrum disorder facial diagnosis. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2023)
16. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* **43**(1), 172–186 (2019)

17. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
18. Catruna, A., Cosma, A., Radoi, E.: Gaitpt: Skeletons are all you need for gait recognition. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024)
19. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019)
20. Chao, H., Wang, K., He, Y., Zhang, J., Feng, J.: Gaitset: Cross-view gait recognition through utilizing gait as a deep set. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3467–3478 (2021)
21. Chatzichristodoulou, G., Efthymiou, N., Filntisis, P., Pavlakos, G., Maragos, P.: Age-inclusive 3d human mesh recovery for action-preserving data anonymization. arXiv preprint arXiv:2512.05259 (2025)
22. Chen, B., Chen, C., Hu, J., Sayeed, Z., Qi, J., Darwiche, H.F., Little, B.E., Lou, S., Darwish, M., Foote, C., et al.: Computer vision and machine learning-based gait pattern recognition for flat fall prediction. Sensors **22**(20), 7960 (2022)
23. Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.I.: A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. Sports medicine-open **4**(1), 24 (2018)
24. Cook, R.E., Schneider, I., Hazlewood, M.E., Hillman, S.J., Robb, J.E.: Gait analysis alters decision-making in cerebral palsy. Journal of pediatric orthopaedics **23**(3), 292–295 (2003)
25. DeLuca, P.A., Davis, R.B., Öunpuu, S., Rose, S., Sirkin, R.: Alterations in surgical decision making in patients with cerebral palsy based on three-dimensional gait analysis. Journal of Pediatric Orthopaedics **17**(5), 608–614 (1997)
26. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
27. Di Biase, L., Di Santo, A., Caminiti, M.L., De Liso, A., Shah, S.A., Ricci, L., Di Lazzaro, V.: Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. Sensors **20**(12), 3529 (2020)
28. Dickens, W.E., Smith, M.F.: Validation of a visual gait assessment scale for children with hemiplegic cerebral palsy. Gait & posture **23**(1), 78–82 (2006)
29. Do, A.H., Wang, P.T., King, C.E., Chun, S.N., Nenadic, Z.: Brain-computer interface controlled robotic gait orthosis. Journal of neuroengineering and rehabilitation **10**(1), 111 (2013)
30. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
31. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2969–2978 (2022)
32. Fan, C., Hou, S., Huang, Y., Yu, S.: Exploring deep models for practical gait recognition. arXiv preprint arXiv:2303.03301 (2023)

33. Fan, C., Hou, S., Liang, J., Shen, C., Ma, J., Jin, D., Huang, Y., Yu, S.: Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
34. Fan, C., Hou, S., Wang, J., Huang, Y., Yu, S.: Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(12), 14920–14937 (2023)
35. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9707–9716 (2023)
36. Fan, C., Ma, J., Jin, D., Shen, C., Yu, S.: Skeletongait: Gait recognition using skeleton maps. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 38, pp. 1662–1669 (2024)
37. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14225–14233 (2020)
38. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alpha-pose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence* **45**(6), 7157–7173 (2022)
39. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6202–6211 (2019)
40. Field-Fote, E.C., Fluett, G.G., Schafer, S.D., Schneider, E.M., Smith, R., Downey, P.A., Ruhl, C.D.: The spinal cord injury functional ambulation inventory (sci-fai). *Journal of rehabilitation medicine* **33**(4), 177–181 (2001)
41. Fu, Y., Hou, S., Meng, S., Hu, X., Cao, C., Liu, X., Huang, Y.: Cut out the middleman: Revisiting pose-based gait recognition. In: *European Conference on Computer Vision*. pp. 112–128. Springer (2024)
42. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19595–19604 (2023)
43. Google: Gemini 3 pro: the frontier of vision ai. Online (2025), <https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-vision/>
44. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: *European conference on computer vision*. pp. 382–398. Springer (2020)
45. Huang, X., Luan, L., Hatamimajoumerd, E., Wan, M., Kakhaki, P.D., Obeid, R., Ostadabbas, S.: Posture-based infant action recognition in the wild with very limited data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4912–4921 (2023)
46. Hui, Z., Wang, J., Dong, X., Zhang, G., Wang, X., Peng, J., Liu, Q., Yuan, X., Zhang, Y., Yin, J.O., et al.: Infantnet: A large scale dataset for infant body pose and shape estimation (2025)
47. Hulzinga, F., Nieuwboer, A., Dijkstra, B.W., Mancini, M., Strouwen, C., Bloem, B.R., Ginis, P.: The new freezing of gait questionnaire: unsuitable as an outcome in clinical trials? *Movement disorders clinical practice* **7**(2), 199–205 (2020)
48. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recogni-

- tion. *IEEE Transactions on Information Forensics and Security* **7**(5), 1511–1521 (2012)
49. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtm-pose: Real-time multi-person pose estimation based on mmPose. *arXiv preprint arXiv:2303.07399* (2023)
 50. Jin, D., Fan, C., Ma, J., Zhou, J., Chen, W., Yu, S.: On denoising walking videos for gait recognition. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 12347–12357 (2025)
 51. Kanko, R.M., Laende, E.K., Davis, E.M., Selbie, W.S., Deluzio, K.J.: Concurrent assessment of gait kinematics using marker-based and markerless motion capture. *Journal of biomechanics* **127**, 110665 (2021)
 52. Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., Tuthill, J.C.: Anipose: A toolkit for robust markerless 3d pose estimation. *Cell reports* **36**(13) (2021)
 53. Khandelwal, S., Wickström, N.: Evaluation of the performance of accelerometer-based gait event detection algorithms in different real-world scenarios using the marea gait database. *Gait & posture* **51**, 84–90 (2017)
 54. Khiradkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: *European Conference on Computer Vision*. pp. 206–228. Springer (2024)
 55. Lam, W.W., Tang, Y.M., Fong, K.N.: A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation. *Journal of NeuroEngineering and Rehabilitation* **20**(1), 57 (2023)
 56. Li, B., Shen, Y., Yang, H., Cao, X., Yun, G., Gao, L., Chen, T., Xu, L., Cao, J., Huang, M.: The 1st ai children challenge. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5564–5570 (2026)
 57. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552* (2022)
 58. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720* (2026)
 59. Li, V., Kamalakannan, N., Parnandi, A., Schambra, H., Fernandez-Granda, C.: The potential and limitations of vision-language models for human motion understanding: A case study in data-driven stroke rehabilitation. *arXiv preprint arXiv:2511.17727* (2025)
 60. Li, W., Hou, S., Zhang, C., Cao, C., Liu, X., Huang, Y., Zhao, Y.: An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13824–13833 (2023)
 61. Li, X., Makiyara, Y., Xu, C., Yagi, Y.: Multi-view large population gait database with human meshes and its performance evaluation. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **4**(2), 234–248 (2022)
 62. Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., et al.: Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine* **25**(3), 433–438 (2019)
 63. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 14648–14656 (2021)

64. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
65. Lord, S., Halligan, P., Wade, D.: Visual gait analysis: the development of a clinical assessment and scale. *Clinical rehabilitation* **12**(2), 107–119 (1998)
66. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5137–5146 (2018)
67. Luvizon, D.C., Picard, D., Tabia, H.: Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence* **43**(8), 2752–2764 (2020)
68. Maathuis, K.G., Van Der Schans, C.P., Van Iperen, A., Rietman, H.S., Geertzen, J.H.: Gait in children with cerebral palsy: observer reliability of physician rating scale and edinburgh visual gait analysis interval testing scale. *Journal of Pediatric Orthopaedics* **25**(3), 268–272 (2005)
69. Makihara, Y., Mannami, H., Tsuji, A., Hossain, M.A., Sugiura, K., Mori, A., Yagi, Y.: The ou-isir gait database comprising the treadmill dataset. *IPSI Transactions on Computer Vision and Applications* **4**, 53–62 (2012)
70. Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez-Sanchez, M., Arakaki, T., Bergareche-Yarza, A., Chade, A., Garretto, N., Gershanik, O., Kurtis, M.M., Martinez-Castrillo, J.C., et al.: Expanded and independent validation of the movement disorder society–unified parkinson’s disease rating scale (mds-updrs). *Journal of neurology* **260**(1), 228–236 (2013)
71. Medvedev, I., Shadmand, F., Gonçalves, N.: Young labeled faces in the wild (ylfw): a dataset for children faces recognition. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024)
72. Nojavanasghari, B., Baltrušaitis, T., Hughes, C.E., Morency, L.P.: Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proceedings of the 18th acm international conference on multimodal interaction. pp. 137–144 (2016)
73. Ong, A., Hillman, S., Robb, J.: Reliability and validity of the edinburgh visual gait score for cerebral palsy when used by inexperienced observers. *Gait & posture* **28**(2), 323–326 (2008)
74. OpenAI: Introducing gpt-5.2. Online (2025), <https://openai.com/index/introducing-gpt-5-2/>
75. Pavlo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7753–7762 (2019)
76. Pistacchi, M., Gioulis, M., Sanson, F., De Giovannini, E., Filippi, G., Rossetto, F., Marsala, S.Z.: Gait analysis and clinical correlations in early parkinson’s disease. *Functional neurology* **32**(1), 28 (2017)
77. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14184–14194 (2025)
78. Ranjan, R., Ahmedt-Aristizabal, D., Armin, M.A., Kim, J.: Computer vision for clinical gait analysis: A gait abnormality video dataset. *IEEE Access* **13**, 45321–45339 (2025)

79. Rathinam, C., Bateman, A., Peirson, J., Skinner, J.: Observational gait assessment tools in paediatrics—a systematic review. *Gait & posture* **40**(2), 279–285 (2014)
80. Read, H.S., Hazlewood, M.E., Hillman, S.J., Prescott, R.J., Robb, J.E.: Edinburgh visual gait score for use in cerebral palsy. *Journal of pediatric orthopaedics* **23**(3), 296–301 (2003)
81. Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al.: Decoding children's social behavior. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3414–3421 (2013)
82. Sankhla, E.K., Kumawat, B., Mahalakshmi, S.: Human gait datasets: A review (2022)
83. Schwartz, M.H., Rozumalski, A.: The gait deviation index: a new comprehensive index of gait pathology. *Gait & posture* **28**(3), 351–357 (2008)
84. Sharma, Y., Cheung, L., Patterson, K.K., Iaboni, A.: Factors influencing the clinical adoption of quantitative gait analysis technology with a focus on clinical efficacy and clinician perspectives: A scoping review. *Gait & Posture* **108**, 228–242 (2024)
85. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3d gait recognition with point clouds. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1054–1063 (2023)
86. Shen, Y., Dong, C., Zhong, Z., Tian, P., Yu, T., Liu, J., Fang, B., Li, X., Liu, Y., Li, Z., et al.: Position: Multimodal llms should learn from children. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5619–5630 (2026)
87. Shen, Y., Liu, J., Li, X., Liu, Y., Li, B., Yang, H., Jia, W., Li, Y., Yu, T., Rehg, J.M., et al.: Egoforge: Goal-directed egocentric world simulator. *arXiv preprint arXiv:2603.20169* (2026)
88. Shen, Y., Zhang, J., Xu, J., Kim, J., Lourentzou, I., Cao, X., Huang, M.: Evaluating cognitive age alignment in interactive ai agents. *arXiv preprint arXiv:2605.17894* (2026)
89. Shen, Y., Zhang, Q., Zhang, Y., Zhou, J., Yang, H., Wang, Y., Zhu, J., Susladkar, O.K., Tian, Y., Huang, M., et al.: A survey on embodied ai for pediatrics. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5605–5618 (2026)
90. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12026–12035 (2019)
91. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications* **26**(3), 941–955 (2023)
92. States, R.A., Krzak, J.J., Salem, Y., Godwin, E.M., Bodkin, A.W., McMulkin, M.L.: Instrumented gait analysis for management of gait disorders in children with cerebral palsy: A scoping review. *Gait & Posture* **90**, 1–8 (2021)
93. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5693–5703 (2019)
94. Sutherland, D.: The development of mature gait. *Gait & posture* **6**(2), 163–170 (1997)

95. Tafasca, S., Gupta, A., Odobez, J.M.: Childplay: A new benchmark for understanding children’s gaze behaviour. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20935–20946 (2023)
96. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications* **10**(1), 4 (2018)
97. Tan, T., Van Wouwe, T., Werling, K.F., Liu, C.K., Delp, S.L., Hicks, J.L., Chaudhari, A.S.: Gaitdynamics: A generative foundation model for analyzing human walking and running. *Nature Biomedical Engineering* pp. 1–13 (2026)
98. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), <https://qwen.ai/blog?id=qwen3.5>
99. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE international conference on image processing (ICIP). pp. 2314–2318. IEEE (2021)
100. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* **35**, 10078–10093 (2022)
101. Upadhyay, U., Ranjan, M., Shen, Z., Elhoseiny, M.: Time blindness: Why video-language models can’t see what humans can? *arXiv preprint arXiv:2505.24867* (2025)
102. Vilemeyer, J., Tronicke, L., Schreff, L., Abel, R., Lechler, K., Müller, R.: A full-body motion capture gait dataset of healthy young adults walking ramps up and down. *Scientific Data* (2026)
103. Wang, L., Chen, J., Chen, Z., Liu, Y., Yang, H.: Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition. *Connection Science* **34**(1), 652–669 (2022)
104. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14549–14560 (2023)
105. Wishaupt, K., Schallig, W., van Dorst, M.H., Buizer, A.I., van der Krogt, M.M.: The applicability of markerless motion capture for clinical gait analysis in children with cerebral palsy. *Scientific reports* **14**(1), 11910 (2024)
106. Wrisley, D.M., Marchetti, G.F., Kuharsky, D.K., Whitney, S.L.: Reliability, internal consistency, and validity of data obtained with the functional gait assessment. *Physical therapy* **84**(10), 906–918 (2004)
107. Xu, C., Makihara, Y., Ogi, G., Li, X., Yagi, Y., Lu, J.: The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation. *IPSJ Transactions on Computer Vision and Applications* **9**(1), 24 (2017)
108. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
109. Yang, M., Zhu, P., Van Gool, L., Zhang, L.: Face recognition based on regularized nearest points between image sets. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp. 1–7. IEEE (2013)
110. Ye, D., Fan, C., Huang, Z., Luo, C., Li, J., Yu, S., Liu, X.: Biggergait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132* (2025)

111. Ye, D., Fan, C., Ma, J., Liu, X., Yu, S.: Biggait: Learning gait representation you want by large vision models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 200–210 (2024)
112. Ye, D., Fan, C., Narayan, K., Wu, B., Luo, C., Li, J., Patel, V.M.: Silhouette-based gait foundation model. *arXiv preprint arXiv:2512.00691* (2025)
113. Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471* (2025)
114. Zhang, C., Chen, X.P., Han, G.Q., Liu, X.J.: Spatial transformer network on skeleton-based gait recognition. *Expert Systems* **40**(6), e13244 (2023)
115. Zhang, H., Li, Y., He, S., Nagarajan, T., Chen, M., Lu, J., Li, A., Fu, Y.: Thinkjapa: Empowering latent world models with large vision-language reasoning model. *arXiv preprint arXiv:2603.22281* (2026)
116. Zhang, H., Lu, Y., Wang, L., Li, Y., Chen, D., Xu, Y., Fu, Y.: Linkedout: Linking world knowledge representation out of video llm for next-generation video recommendation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7111–7121 (2026)
117. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 13232–13242 (2022)
118. Zhao, Y., Liu, R., Xue, W., Yang, M., Shiraishi, M., Awai, S., Maruyama, Y., Yoshioka, T., Konno, T.: Effective fusion method on silhouette and pose for gait recognition. *IEEE Access* **11**, 102623–102634 (2023)
119. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 11656–11665 (2021)
120. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 20228–20237 (2022)
121. Zheng, J., Liu, X., Wang, S., Wang, L., Yan, C., Liu, W.: Parsing is all you need for accurate gait recognition in the wild. In: *Proceedings of the 31st ACM International Conference on Multimedia*. pp. 116–124 (2023)
122. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 8600–8612 (2025)
123. Zhou, Z., Liang, J., Peng, Z., Fan, C., An, F., Yu, S.: Gait patterns as biomarkers: A video-based approach for classifying scoliosis. In: *International conference on medical image computing and computer-assisted intervention*. pp. 284–294. Springer (2024)
124. Zhou, Z., Peng, Z., Jin, D., Fan, C., An, F., Yu, S.: Pose as clinical prior: Learning dual representations for scoliosis screening. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 464–474. Springer (2025)
125. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025)

- 126. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15085–15099 (2023)
- 127. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14789–14799 (2021)
- 128. Zou, Q., Wang, Y., Wang, Q., Zhao, Y., Li, Q.: Deep learning-based gait recognition using smartphones in the wild. *IEEE Transactions on Information Forensics and Security* **15**, 3197–3212 (2020)
- 129. Zou, S., Fan, C., Xiong, J., Shen, C., Yu, S., Tang, J.: Cross-covariate gait recognition: A benchmark. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 7855–7863 (2024)

Appendix

A CGV Details

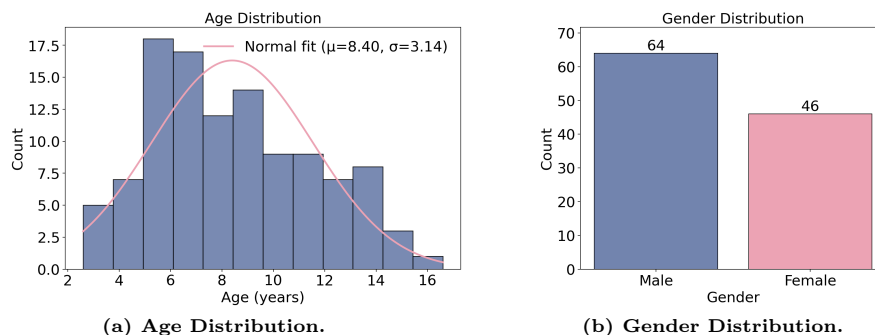


Fig. A1: Demographic Statistics of the ChildGait Video Dataset. *left:* The age distribution of the subjects with a fitted normal distribution curve, highlighting a primary focus on the pediatric population. *right:* The gender distribution across the entire dataset. Overall, the dataset maintains a nearly balanced demographic profile, providing the necessary diversity for training or evaluation in visual gait analysis.

In addition to the overall dataset scale, annotation procedures, and label distributions discussed in Section 4, we provide further details regarding the demographic composition of the participants in the Children Gait Video (CGV) dataset. As illustrated in Figure A1, the dataset consists of 110 pediatric patients, aiming to represent the target clinical population as fully as possible.

Age Distribution. Figure A1a visualizes the age profile of the patients. The histogram is overlaid with a fitted normal distribution curve to illustrate the central tendency of the patient demographics. The participants span a critical developmental window from 2.6 to 16.6 years old, with a mean age of $\mu = 8.40$, $\sigma = 3.14$ years, which is highly clinically relevant and highlights a primary focus on the pediatric population. To explore the relationship between age and performance, we split the test set into younger children (≤ 8 years) and older children (≥ 8 years) subsets, achieving average accuracies of 82.4% and 85.2%, respectively, proving that gait analysis is more difficult for younger children.

Gender Distribution. Figure A1b details the gender composition of the dataset. The patients consist of 58.2% male and 41.8% female subjects. This nearly balanced gender distribution helps to prevent the model from learning shortcuts for specific genders, ensuring sufficient diversity and fairness for training and evaluating models for visual gait analysis.

Table B1: EVGS Three-point Scale. Each scoring item is assessed and categorized into one of three scales: Normal, Moderate deviation, or Marked deviation.

Ordinal Scale	Measurement
0	Normal (within +/- 1.5 standard deviations (SD) of normal mean)
1	Moderate deviation (between 1.5 and 4.5 SD of normal mean)
2	Marked deviation (greater than 4.5 SD of normal mean)

B EVGS Scoring Criteria

The Edinburgh Visual Gait Score (EVGS) [80] is a clinical tool developed to visually assess gait deviations in ambulatory children using coronal and sagittal video recordings. It evaluates 17 observational scoring items for each limb that are graded on a three-point ordinal scale. Table B1 details this specific ordinal scale used for the evaluation process. Table B2 provides the comprehensive clinical criteria, explanations, and score mappings for the gait scoring items.

Table B2: EVGS Items and Scores. We list all EVGS scoring items, including explanations and scoring details.

No.	Scoring Items	Explanation	Score
1	Initial Contact in Stance	The heel normally contacts first. The toe describes that portion of the foot distal to the metatarsophalangeal joints. Simultaneous contact with the heel and toe comprises flat-foot contact.	– Heel contact: 0 – Flatfoot contact: 1 – Toe contact: 2
2	Heel Lift in Stance	If there is no heel contact during stance, there can be no heel lift (<i>i.e.</i> , ‘No heel contact’). Heel lift normally occurs between the opposite foot level and the opposite foot contact (‘Normal’). ‘Early’ heel lift indicates that heel lift precedes the opposite foot being level with the stance foot. ‘Delayed’ heel lift is present if heel lift occurs with or after opposite foot contact. ‘No forefoot contact’ describes the rare occasion of a calcaneus foot when the forefoot does not contact during stance.	– No forefoot contact: 2 – Delayed: 1 – Normal: 0 – Early: 1 – No heel contact: 2

No.	Scoring Items	Explanation	Score
3	Max Ankle Dorsiflexion in Stance	There is normal forward progression of the tibia over the planted hind-foot from slight plantar flexion at initial contact to dorsiflexion at terminal stance. Describe the maximum angle of dorsiflexion between the hind foot and the shaft of the tibia during stance. In pathological gait, lack of heel contact may be caused by either excessive plantar flexion of the foot or excessive knee flexion. The tibial hind-foot angle is therefore analyzed irrespective of the position of the foot on the floor.	– Excessive dorsiflexion ($>40^\circ$ df): 2 – Increased dorsiflexion (26°- 40° df): 1 – Normal dorsiflexion (5°- 25° df): 0 – Reduced dorsiflexion (10° pl - 4° df) – Marked plantar flexion ($>10^\circ$ pl): 2
4	Hind-foot Varus/Valgus in Stance	In the coronal plane, the normal hind-foot is in neutral or very slight valgus.	– Severe valgus ($>15^\circ$ valgus): 2 – Mod valgus (6°- 15° valgus): 1 – Neutral/slight valgus (0°- 5° valgus): 0 – Mild varus (1°- 10° varus): 1 – Severe varus ($>10^\circ$ varus): 2
5	Foot Rotation in Stance	The normal foot is slightly externally rotated relative to the Knee Progression Angle (KPA, <i>i.e.</i> , the direction in which the knee points during gait).	– Marked ext. $>$KPA (by $>40^\circ$): 2 – Mod ext. $>$KPA (by 21°- 40°): 1, – Slightly more ext. than KPA (by 0°- 20° extension): 0 – Mod int. $>$KPA (by 1°- 25°): 1 – Marked int. $>$KPA (by $>25^\circ$): 2

No.	Scoring Items	Explanation	Score
6	Foot Clearance in Swing	The whole foot, including the toe, should clear the foot and not make contact during the swing phase. 'None' should be recorded if there is continuous contact between some part of the foot and the floor throughout the swing phase. 'Reduced' indicates that there is a shortened but definite period of clearance during some part of the swing phase between the whole foot and the floor. 'Full' or normal clearance is when the foot does not touch at all in swing; however, normal clearance is a very small amount. 'High steps' describes excessive lifting of the foot from the floor. When there is reduced clearance followed by high stepping, circle both, giving a score of 2 for this combination of features.	– High Steps: 1 – Full: 0 – Reduced: 1
7	Max Ankle Dorsiflexion in Swing	The ankle is normally approximately neutral in swing, but very slight plantar flexion (5°) is acceptable.	– Excessive dorsiflexion ($>30^\circ$ df): 2 – Increased dorsiflexion (16°- 30° df): 1 – Normal dorsiflexion (15° df - 5° pl): 0 – Mod plantar flexion (6°- 20° pl): 1 – Marked plantar flexion ($>20^\circ$ pl): 2
8	Knee Progression Angle in Mid-Stance	The knee normally points forward during gait. Record the position in which the knee appears to point during most of the stance phase. When either internal or external rotation is present, but the whole knee cap is visible, score 1. When rotation is present to such an extent that the knee cap is partially out of view (external or internal, part of the cap visible), score 2.	– External, part of the knee cap visible: 2 – External, all of the knee cap visible: 1 – Neutral, knee cap midline: 0 – Internal, all of the knee cap visible: 1 – Internal, part of the knee cap visible: 2

No.	Scoring Items	Explanation	Score
9	Peak Knee Extension in Stance	The knee approaches full extension in terminal stance. In pathological gait, the knee may remain more flexed throughout stance. Alternatively, hypertension can occur as femoral progression proceeds over an arrested tibia.	– Severe flexion ($>25^\circ$): 2 – Mod flexion (16°- 25°): 1 – Normal (0°- 15° flexion): 0 – Mod hyperextension (1°- 10°): 1 – Severe hyperextension ($<10^\circ$): 2
10	Knee Position in Terminal Swing	The knee is normally in slight flexion immediately before heel strike.	– Severe flexion ($>30^\circ$): 2 – Mod flexion (16°- 30°): 1 – Normal (5°- 15° flexion): 0 – Mod overextension (4° flexion - 10° extension): 1 – Severe hyperextension ($>10^\circ$ extension): 2
11	Peak Knee Flexion in Swing	The normal range is 50° to 70° .	– Severely increased ($>85^\circ$ flexion): 2 – Mod increased (71°- 85° flexion): 1 – Normal (50°- 70° flexion): 0 – Mod reduced (35°- 49° flexion): 1 – Severely reduced ($<35^\circ$ flexion): 2
12	Peak Hip Extension in Stance	The hip normally extends in stance to between neutral and 20° of extension.	– Severe flexion ($>15^\circ$ flexion): 2 – Mod flexion (1°- 15° flexion): 1 – Normal (0°- 20° extension): 0 – Mod hyperextension (21°- 35° extension): 1 – Marked hyperextension ($>35^\circ$ extension): 2

No.	Scoring Items	Explanation	Score
13	Peak Hip Flexion during Swing	Normal flexion is between 25° and 45°.	– Marked increased flexion (>60° flexion): 2 – Increased flexion (46°-60° flexion): 1 – Normal flexion (25°- 45° flexion): 0 – Reduced flexion (10°- 24° flexion): 1 – Severely reduced (<10° flexion): 2
14	Pelvic Obliquity at Mid-Stance	The pelvis normally drops slightly on the opposite side during loading, becoming level by terminal stance. Estimate the position in mid stance. ‘Up’ and ‘down’ refer to the position of the ASIS on the stance side, relative to the opposite side ASIS.	– Marked down (>10°): 2 – Mod down (1°- 10°): 1 – Normal obliquity (0°- 5° up): 0 – Mod up (6°- 15°): 1 – Marked up (>15°): 2
15	Pelvic Rotation at Mid-Stance	In mid stance, the pelvis should be at approximately neutral rotation, between 5° backward rotation (retraction) of the stance leg, and 10° forward rotation (protraction).	– Marked retraction (>15°): 2 – Mod retraction (6°- 15°): 1 – Normal (5° retraction - 10° protraction): 0 – Mod protraction (11°- 20°): 1 – Marked protraction (>20°): 2
16	Peak Sagittal Trunk Position in Stance	The trunk is erect during the stance and swing phases.	– Marked forward lean (>15° forward): 2 – Mod forward lean (between 6° and 15° forward): 1 – Normal upright (vertical to 5° forward or backward): 0 – Mod backward lean (>5° backward): 1

No.	Scoring Items	Explanation	Score
17	Maximum Trunk Lateral Shift	Normally, the trunk displaces laterally approximately 25 mm during stance, towards the stance leg.	
		‘Excessive’ thoracic shift laterally or lateral flexion should be considered when recording observations.	– Marked: 2 – Mod: 1
		‘Reduced’ describes those cases in which the trunk remains leaning over the swinging leg.	– Normal: 0 – Reduced: 1

C Prompt Design

As stated in Section 5 and Section 6, we design a structured prompt to instruct the MLLMs to perform visual gait analysis. The prompt architecture is composed of three core components: System Persona, Task Instructions, and Output Formatting.

System Persona. We set the model’s role as an expert pediatrician specializing in observational gait analysis. This initialization ensures that the model leverages its pre-trained domain knowledge of anatomical priors, gait cycle phases, and kinematic deviations.

Task Instructions. We explicitly define the evaluation workflow, asking the model to systematically process both coronal and sagittal video recordings to assess the 17 scoring items for each limb. To ground the model’s reasoning and ensure standardized evaluation, we embed the exact EVGS scoring criteria shown in Appendix B into the context. The model evaluates each item on a three-point ordinal scale, where 0 indicates a normal condition, 1 represents a moderate deviation, and 2 signifies a marked deviation, as stated in Appendix B. We enforce that the model uniformly outputs 1 for both 1 and 2 to align with our requirements mentioned in Section 4.

Output Formatting. The model is asked to return its assessment in a strict JSON format for easy evaluation and inspection.

Overall, the designed prompts are shown in Figure C1, where Scoring Item, Explanation, and Scoring Criteria all originate from Table B2.

D Evaluation of Skeleton Baselines

To evaluate the performance of skeleton baselines in the field of children’s gait analysis, we evaluate strong baselines SkeletonGait++ [36] and ScoNet [123,124].

Experimental Results. As shown in Table D1, although general skeleton-based baselines have demonstrated strong capabilities in standard pose and gait recognition tasks, they still struggle to perform precise clinical scoring consistently across all joints, yielding sub-optimal average accuracies that range from 58% to

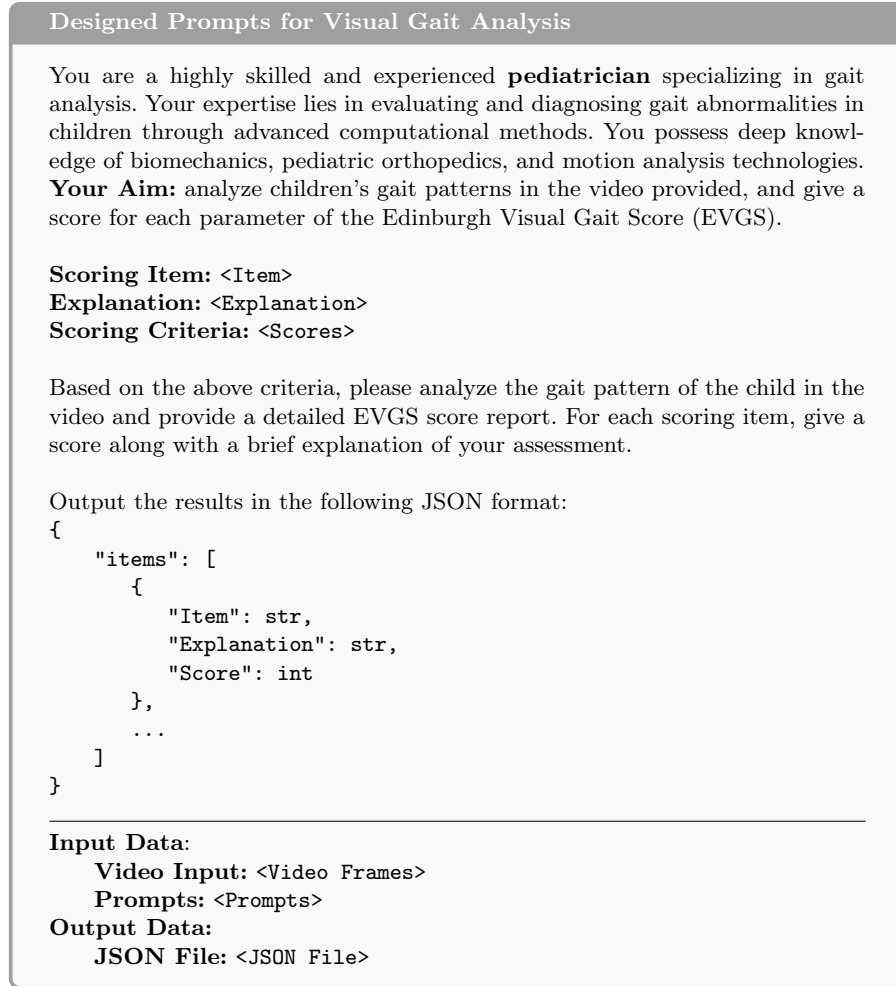


Fig. C1: Designed Prompts for Visual Gait Analysis. The model takes both the video frames and designed prompts as input to generate an explanation and a score for each scoring item.

65% for the left limb and 60% to 64% for the right limb. Notably, ScoNet pre-trained on Scoliosis1K achieves the best results, peaking at average accuracies of 68% and 70% for the left and right limbs, respectively. We attribute the limited performance of the baselines to the inherent design objectives of traditional skeleton-based networks, which focus on extracting global structural representations for macro-level classification, thereby neglecting the fine-grained, localized kinematic anomalies essential for clinical assessment, especially at distal joints (*e.g.*, FRT, POB), where accuracies drop sharply.

Table D1: Quantitative Evaluation of Skeleton Baselines. All reported values are **percentages (%)**. We report 17 scoring items for the **Left (L-)** limb (top) and **Right (R-)** limb (bottom). The detailed definition of each item is shown in Table 1. L/R-AVG indicates the average accuracy over all scoring items of the left/right limb. L/R-F1 indicates the average F1-score across all items of the left/right limb.

Method	Pretrain	L-IC	L-HL	L-SAD	L-WAD	L-HVV	L-FRT	L-FCL	L-KPA	L-KEX	L-KPS	L-KFX	L-HEX	L-HEX	L-POB	L-PRT	L-TSG	L-TLT	L-AVG	L-F1
SkeletonGait++ [36]	Gait3D [120]	68	70	63	67	53	48	50	54	60	56	62	64	61	43	46	54	67	58	0.44
SkeletonGait++ [36]	SUSTech1K [85]	75	77	70	73	59	54	57	61	68	63	70	71	68	52	54	61	72	65	0.61
SkeletonGait++ [36]	GREW [127]	70	73	66	69	56	51	53	58	64	59	65	67	64	48	50	56	68	61	0.54
SkeletonGait++ [36]	CCPG [60]	68	70	63	67	53	48	49	55	60	56	62	64	61	44	45	54	67	58	0.47
ScoNett [123, 124]	Scoliosis1Kt [123, 124]	77	80	73	76	63	58	60	64	71	66	73	74	71	54	57	64	75	68	0.67

Method	Pretrain	R-IC	R-HL	R-SAD	R-WAD	R-HVV	R-FRT	R-FCL	R-KPA	R-KEX	R-KPS	R-KFX	R-HEX	R-HEX	R-POB	R-PRT	R-TSG	R-TLT	R-AVG	R-F1
SkeletonGait++ [36]	Gait3D [120]	71	75	67	69	57	52	55	58	65	60	66	67	64	49	50	58	71	62	0.60
SkeletonGait++ [36]	SUSTech1K [85]	73	77	69	71	59	54	57	59	67	62	68	69	66	51	53	60	73	64	0.53
SkeletonGait++ [36]	GREW [127]	73	77	68	72	58	54	57	60	66	63	68	69	66	51	52	61	73	64	0.55
SkeletonGait++ [36]	CCPG [60]	69	72	65	67	55	51	53	56	62	58	63	65	63	47	48	56	70	60	0.46
ScoNett [123, 124]	Scoliosis1Kt [123, 124]	79	82	75	77	65	60	63	67	75	70	76	77	74	59	61	69	81	70	0.69

E More Confusion Matrices

As shown in Figure E1, we visualize the confusion matrices for ChildGait-Video evaluation across all remaining items in addition to Figure 4. Overall, the model demonstrates robust classification capabilities across these additional items, as evidenced by the high concentration of predictions along the main diagonal.

F Module Ablation Study

To comprehensively evaluate the individual contributions of our proposed modules and justify our design choices, we conduct a detailed module ablation study. As shown in Table F1, Token-Level Kinematic Prompting (TKP) helps to gain a performance improvement, raising the average accuracy (L-AVG/R-AVG) from 69%/72% to **72%/74%**. This demonstrates that TKP effectively provides necessary anatomical priors derived from expert annotations, guiding the network to focus on clinically relevant joint dynamics rather than generic spatial features. We also validate the Mask-Guided Patch Pruning (MPP) module designed for noise mitigation. We compare our MPP with two alternative strategies: a standard *bounding box* mask and a *random* mask. While the *bounding box* mask brings a moderate gain by removing some background (achieving 75%/77% for L/R-AVG), our MPP outperforms it by **3%** and **2%** on the left and right limbs, respectively, proving its superior capability in adaptively filtering background noise while preserving structural integrity. Conversely, employing the *random* mask strategy degrades the overall performance, causing a drop of **1%** to **2%** compared to the baseline (down to 68%/70%). This decline is expected, as random masking inevitably obscures crucial kinematic joints or fails to adequately

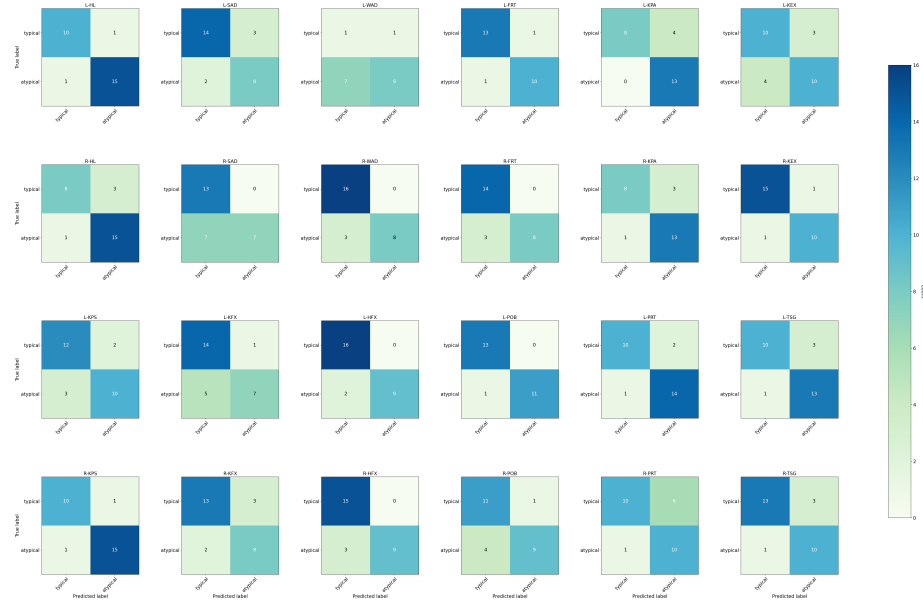


Fig.E1: More Visualization of Confusion Matrices. We further evaluate the performance of ChildGait-Video on EVGS scoring classification in all remaining items.

Table F1: Module Ablation Study Results. We evaluate the individual contributions of each proposed module, including Token-Level Kinematic Prompting (TKP) and Mask-Guided Patch Pruning (MPP), against various masking baselines.

Variant	L-AVG (%)	R-AVG (%)	L-F1	R-F1
VideoMAE v2 (Base)	69	72	0.70	0.71
VideoMAE v2 + TKP	72	74	0.72	0.73
VideoMAE v2 + MPP	78	79	0.77	0.78
VideoMAE v2 + Random Mask	68	70	0.69	0.69
VideoMAE v2 + Bounding Box Mask	75	77	0.74	0.76
ChildGait-Video (Ours)	84	84	0.83	0.83

eliminate background noise, thereby destroying the details essential for accurate clinical assessment.