A Tri-Modal Vision–Language–Biomechanics Framework for Interpretable Clinical Gait Assessment

Erdong Chen[1,2⋆], Yuyang Ji[1⋆], Jacob K. Greenberg[2], Benjamin Steel[3], Faraz Arkam[2], Abigail Lewis[2], Pranay Singh[2], and Feng Liu[1✉]

[1] Department of Computer Science, Drexel University
[2] Department of Neurological Surgery, Washington University
[3] University of California, Berkeley

Abstract. Video-based Clinical Gait Analysis often suffers from poor generalization as models overfit environmental biases instead of capturing pathological motion. To address this, we propose **BioGait-VLM**, a tri-modal Vision-Language-Biomechanics framework for interpretable clinical gait assessment. Unlike standard video encoders, our architecture incorporates a Temporal Evidence Distillation branch to capture rhythmic dynamics and a Biomechanical Tokenization branch that projects 3D skeleton sequences into language-aligned semantic tokens. This enables the model to explicitly reason about joint mechanics independent of visual shortcuts. To ensure rigorous benchmarking, we augment the public GAVD dataset with a high-fidelity Degenerative Cervical Myelopathy (DCM) cohort to form a unified 8-class taxonomy, establishing a strict subject-disjoint protocol to prevent data leakage. Under this setting, BioGait-VLM achieves state-of-the-art recognition accuracy. Furthermore, a blinded expert study confirms that biomechanical tokens significantly improve clinical plausibility and evidence grounding, offering a path toward transparent, privacy-preserving gait assessment.

Keywords: Clinical Gait Analysis · Vision-Language Models · Biomechanics · Degenerative Cervical Myelopathy · Explainable AI

1 Introduction

Gait is a sensitive functional biomarker of neurological and musculoskeletal health [6,15,14,22], making clinical gait analysis (CGA) a fundamental tool for screening, severity assessment, and longitudinal monitoring of neurodegenerative diseases [24,10,18]. Despite its diagnostic value, gold-standard CGA is restricted to controlled settings using instrumented walkways (*e.g.*, GAITRite [3])

⋆ Equal contribution. Erdong was a visiting research intern at Drexel University during this work.
✉ fl397@drexel.edu

or body-worn inertial sensors. These methods require on-site setup and specialized personnel, limiting their scalability for frequent longitudinal follow-up [3,6].

To enable scalable assessment, researchers have adapted deep learning architectures to clinical gait analysis [17,12,16], such as I3D [7] and SlowFast [4] to extract spatiotemporal features directly from video frames. More recently, Video Transformers like TimeSformer [2] and VideoMAE [19] have been employed to capture long-range temporal dependencies.

While promising, these end-to-end models operate as "black boxes," learning abstract features that correlate with pathology but lack explicit biomechanical grounding. Consequently, they fail to provide the *granular, interpretable kinematic evidence*, such as specific joint angle deviations, that clinicians require for decision-making.

Recently, Large Vision-Language Models (LVLMs) [25,11,9] offer a potential solution to this interpretability gap. However, standard LVLMs primarily rely on static visual cues and lack the domain knowledge to quantify fine-grained temporal abnormalities (*e.g.*, foot drop). Furthermore, when deployed on uncontrolled "in-the-wild" videos, they are prone to *shortcut learning*: attending to environmental confounders (*e.g.*, a hospital bed) rather than the patient's underlying motion mechanics [23,5,13].

To address both robustness and clinical interpretability, we propose **BioGait-VLM**, a tri-modal framework that seamlessly integrates RGB vision, natural language, and explicit biomechanics. Conceptually, our approach mitigates the reliance on superficial visual shortcuts by enforcing a dual-path reasoning process: while the visual stream captures contextual appearance, a dedicated biomechanical stream isolates pure kinematic dynamics. Crucially, this design supports a ***dual-output capability:*** the fused representations drive a specialized classification head for high-precision diagnosis, while the underlying language model retains the generative capacity to output human-readable clinical descriptions, providing transparent reasoning for its predictions. Structurally, BioGait-VLM creates a holistic clinical representation by bridging pixel-level features and medical reasoning. While we leverage a pre-trained Visual Branch for appearance cues, standard VLMs lack temporal sensitivity. To address this, we integrate a Temporal Evidence Distillation Branch that compresses frame features into dynamic gait descriptors, allowing the model to detect subtle pathologies like festination. Uniquely, our Biomechanical Branch employs a novel tokenization strategy that projects 3D kinematics into language-aligned semantic tokens, enabling the LLM to explicitly "read" the patient's motion. To rigorously validate this architecture, we extend the public GAVD benchmark [17] by integrating our collected Degenerative Cervical Myelopathy (DCM) Dataset, a high-fidelity clinical cohort of 30 patients, resulting in a unified 8-class pathological taxonomy comprising a total of $1,181$ video clips. We rectify this combined benchmark with leakage-aware subject-disjoint splits, ensuring that clips from the same patient never overlap between training and testing. By evaluating on this challenging setting and conducting a blinded expert study, we demonstrate that our tri-

modal approach achieves state-of-the-art generalization and generates clinically grounded assessment.

In summary, the contributions of this work include:

⋄ We propose **BioGait-VLM**, a novel tri-modal framework that introduces Biomechanical Tokenization, projecting 3D kinematics into language-aligned tokens to enable robust reasoning and interpretable clinical assessment.

We integrate a specialized *Temporal Evidence Distillation Branch* that compresses frame-level features into compact dynamic descriptors, ensuring the model captures temporal evolution rather than just static appearance.

⋄ We collect and curate the DCM Clinical Dataset, a high-fidelity, expert-validated benchmark for Degenerative Cervical Myelopathy, enabling rigorous evaluation of fine-grained pathological motion in authentic clinical settings.

⋄ We achieve state-of-the-art performance on clinical benchmarks and confirm via a human-expert study that our approach significantly improves clinical correctness and evidence grounding.

2 Methodology

2.1 Problem Formulation and Overview

We formulate video-based clinical gait analysis as a dual-objective task: *classification* and *interpretable reasoning*. Given a gait video $\mathcal{V} = \{\mathbf{x}_t\}_{t=1}^{T}$ and an associated 3D skeleton sequence $\mathcal{S}$ (inferred via HSMR [21]), the model aims to predict a pathology label $y \in \{1, \ldots, K\}$ (where K=8 in our experiments, includes *Abnormal*, *Myopathic*, *Parkinson's*, etc.) while simultaneously generating a textual rationale $\mathcal{T}$ describing the biomechanical tokens.

To achieve this, we construct **BioGait-VLM** upon InternVL [25], a pre-trained LVLM. We select this foundation to leverage its *pre-aligned semantic space*, which naturally accommodates our textualized biomechanical tokens, and its generative capabilities for clinical reporting. Furthermore, by freezing the massive pre-trained parameters and training only lightweight adapters, we reduce the number of task-specific trainable parameters, which is suitable for limited clinical datasets. As illustrated in Fig. 1, the framework processes inputs through three complementary pathways:

1. **Visual Encoding Branch**: A frozen vision encoder extracts spatial features from $\mathcal{V}$ to capture contextual appearance.
2. **Temporal Evidence Distillation (TED) Branch**: A query-based decoder aggregates frame-level features into compact gait descriptors, explicitly modeling *temporal dynamics*.
3. **Biomechanical Tokenization Branch**: A kinematic tokenizer projects the skeleton sequence $\mathcal{S}$ into language-aligned semantic tokens, enabling the model to incorporate joint-mechanics cues in addition to visual appearance.

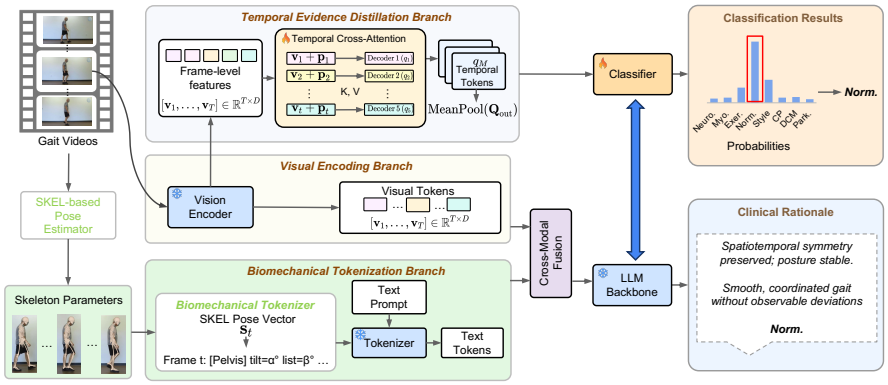

Fig. 1: **Overview of BioGait-VLM.** The framework integrates three streams: (Top) Temporal Evidence Distillation (TED) for rhythmic dynamics; (Middle) Visual Context; and (Bottom) Biomechanical Tokenization, projecting 3D kinematics into semantic tokens. Fused within a frozen LVLM, these modalities enable state-of-the-art classification and evidence-grounded clinical reporting.

2.2 Tri-Modal Vision-Language Architecture

Visual Encoding Branch. Given T sampled frames (T=32 in our experiment), we employ the frozen InternVL vision encoder f_{vis} to process each frame $\mathbf{x}_t$ independently. Unlike standard action recognition backbones that output abstract feature maps, InternVL yields semantically rich visual tokens. We perform spatial mean-pooling on the patch embeddings to obtain a sequence of frame-level visual features $\mathbf{V}$: $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_T] \in \mathbb{R}^{T \times D}$, where $\mathbf{v}_t = \text{MeanPool}(f_{\text{vis}}(\mathbf{x}_t))$ and D denotes the hidden dimension of the language model.

Temporal Evidence Distillation (TED) Branch. Standard VLMs typically process video via naive mean-pooling, which can obscure fine-grained rhythmic cues essential for gait diagnosis, such as tremors or festination. To recover these dynamics, we propose a new Temporal Evidence Distillation module.

Instead of simple pooling, we initialize M=32 learnable motion queries $\mathbf{Q} \in \mathbb{R}^{M \times D}$. These queries function as latent temporal anchors, learning to attend to specific phases of the gait cycle across the visual feature sequence $\mathbf{V}$. To preserve sequential order, we inject learnable temporal positional embeddings $\mathbf{P} \in \mathbb{R}^{T \times D}$:

$$\mathbf{Q}_{\text{out}} = \text{TransformerDecoder}(\mathbf{Q},\ \mathbf{V} + \mathbf{P}). \tag{1}$$

The decoder comprises L=3 layers with H=4 attention heads. Through cross-attention, the queries dynamically aggregate frame-level features based on their relevance to motion patterns rather than visual appearance. The outputs are condensed into a single dynamic gait descriptor via mean-pooling: $\mathbf{f}_{\text{temp}} = \text{MeanPool}(\mathbf{Q}_{\text{out}}) \in \mathbb{R}^D$. This explicitly preserves high-frequency motion anomalies independent of the VLM's semantic context.

Biomechanical Tokenization Branch. To disentangle pathological motion from environmental confounders, we introduce a strategy to project kinematics

into the semantic space of the LLM. First, we extract a 46-dimensional SKEL pose vector $\mathbf{s}_t$ for each frame from the skeleton sequence $\mathcal{S}$ using the off-the-shelf HSMR estimator [8,21]. This covers clinically critical joints including the pelvis, spine, and limbs. We convert these numerical tensors into structured natural-language descriptions using a template-based mapping function. For a given frame t, the joint angles are formatted as:

`Frame` t`: [Pelvis] tilt=`α`° list=`β`° ... [R.Knee] flex=`δ`°`

The descriptions for all T frames are concatenated with a task-specific clinical instruction prompt and tokenized to produce a sequence of biomechanical embedding tokens $\mathbf{E}_{\text{bio}}$: $\mathbf{E}_{\text{bio}} = [\mathbf{e}_1, \ldots, \mathbf{e}_{L_{\text{text}}}] \in \mathbb{R}^{L_{\text{text}} \times D}$. This approach eliminates the need for a separate skeleton encoder and aligns the kinematic data with the pre-trained semantic knowledge of the LLM, effectively allowing the model to "read" the patient's motion.

Cross-Modal Fusion We concatenate visual features $\mathbf{V}$ and biomechanical tokens $\mathbf{E}_{\text{bio}}$ along the sequence dimension:

$$\mathbf{Z}_{\text{input}} = [\mathbf{V}; \mathbf{E}_{\text{bio}}] \in \mathbb{R}^{(T+L_{\text{text}}) \times D}. \tag{2}$$

The frozen LLM processes the sequence. We pool visual hidden states into $\mathbf{f}_{\text{vlm}}$ and concatenate this with the explicit temporal descriptor $\mathbf{f}_{\text{temp}}$ to form the holistic representation $\mathbf{f}_{\text{final}} \in \mathbb{R}^{2D}$.

2.3 Clinical Instruction Tuning and Objectives

Training Objective. To mitigate class imbalance, we use weighted cross-entropy loss. A linear head $g : \mathbb{R}^{2D} \to \mathbb{R}^K$ maps $\mathbf{f}_{\text{final}}$ to logits:

$$\mathcal{L} = -\sum_{k=1}^{K} w_k \, y_k \log \frac{\exp(g(\mathbf{f}_{\text{final}})_k)}{\sum_j \exp(g(\mathbf{f}_{\text{final}})_j)}, \quad w_k = \frac{N}{K \cdot n_k}. \tag{3}$$

We update only the Temporal Decoder and linear head, freezing InternVL.

Evidence-Grounded Prompting. We employ a *Clinical Instruction Prompt* that defines each gait class using kinematic descriptors (*e.g.*, *"Parkinsonian: shuffling steps, reduced arm swing"*). Injecting these definitions alongside $\mathbf{E}_{\text{bio}}$ primes the model to attend to specific joint anomalies during the forward pass.

2.4 The DCM Clinical Gait Dataset

To bridge the gap between uncurated web videos and clinical reality, we introduce the Degenerative Cervical Myelopathy (DCM) Dataset, a high-fidelity cohort comprising 239 video recordings from 30 patients collected during routine neurosurgical assessments. Unlike variable internet clips, these videos are captured using a standardized sagittal-view protocol in a hospital environment, yet retain realistic background clutter to rigorously test model robustness against environmental shortcuts. Ground-truth DCM labels are established based on

clinical diagnosis by spine specialists. These samples form the DCM class within the unified 8-class taxonomy and provide clinically collected examples of spinal-cord-related gait impairment. All data collection is conducted under Institutional Review Board (IRB) approval with strict de-identification protocols, establishing a trustworthy benchmark for fine-grained pathological gait analysis.

2.5 Implementation Details

We implement the framework using InternVL3.5-1B [25] as the backbone, with input frames resized to 448×448. The Temporal Evidence Distillation decoder is configured with L=3 layers and hidden dimension D=2048. Biomechanical text sequences are truncated to a maximum length of 1024 tokens. Model training is conducted on a single NVIDIA A100 GPU using the AdamW optimizer (lr=5×10^{-4}, batch size 2) for 40 epochs.

3 Experiments

3.1 Experimental Settings

We evaluate on a unified clinical gait benchmark spanning an 8-class taxonomy: *Abnormal* (Abnorm.), *Normal* (Norm.), *Style*, *Exercise*, *Parkinson's* (Park.), *Cerebral Palsy (CP)*, *Myopathic* (Myo.), and our newly introduced *DCM*.

GAVD [17] (Public Component). We utilize the Gait Abnormality Video Dataset to source established gait patterns. This includes *Normal* controls and broad *Abnormal* gait cases, non-pathological variations (*Style*, *Exercise*), and standard neuropathologies (*Parkinson's*, *CP*, *Myopathic*).

DCM (Ours). To address the lack of spinal cord pathologies in public data, we integrate our newly collected Degenerative Cervical Myelopathy dataset. In our 8-class benchmark, these samples constitute the distinct *DCM* class, representing a subtle pathological motion often confused with generic abnormalities.

Leakage-Aware Protocol. Unlike [17], to prevent the model from memorizing patient identities, we re-partition the benchmark to strictly enforce subject-disjoint splits. For GAVD, we allocate 158 subjects (728 sequences) to training and hold out 40 subjects (214 sequences) for testing. For DCM, we adhere to an 8 : 2 patient-level split, assigning 24 patients (187 sequences) to training and 6 patients (52 sequences) to testing. The final combined benchmark comprises 915 training and 266 testing sequences, ensuring zero subject overlap.

Baselines. We compare BioGait-VLM against two categories of methods: (1) *End-to-End Video Encoders*: We re-train standard 3D-CNN and Transformer architectures, specifically SlowFast [4] and TSN [20], on our leakage-aware splits. These represent "black-box" approaches that learn abstract spatiotemporal features from pixels without explicit biomechanics. (2) *General-Purpose LVLMs*: We evaluate two SoTA foundational large vision-language models: Qwen3-VL [1] and InternVL3.5 [25], in a zero-shot setting to assess whether general pre-training suffices for fine-grained clinical diagnosis.

Table 1: **Results & Ablation.** Performance on the 8-class clinical gait benchmark. Top: Comparison against state-of-the-art video and VLM baselines. Bottom: Ablation study isolating the impact of the *Temporal Evidence Distillation* (TED) and *Biomechanical Tokenization* branches.

Method	Acc. (%)	Macro F1	Per-Class F1 Score (%)							
			DCM	Myo.	Abnorm.	CP	Park.	Norm.	Style	Exer.
End-to-End Video Encoders (Fine-tuned)										
SlowFast [4]	32.2	31.1	88.2	54.1	21.5	0.0	57.1	0.0	3.9	24.3
TSN [20]	37.1	35.9	100.0	47.6	23.4	12.1	75.0	0.0	0.0	29.2
General-Purpose LVLMs (Zero-Shot)										
Qwen3-VL-2B [1]	21.9	7.8	0.0	0.0	46.7	0.0	0.0	15.7	0.0	0.0
InternVL3.5-1B [25]	32.5	9.6	0.0	0.0	52.3	0.0	0.0	7.6	8.0	9.2
BioGait-VLM Ablation Study (Component Analysis))										
- *w/o* TED Branch	63.8	42.4	93.3	35.5	71.9	10.0	64.1	0.0	4.2	60.2
- *w/o* Biomechanical Branch	64.3	43.8	98.1	33.3	73.4	16.0	65.7	0.0	0.0	64.4
- *w/o* both	60.0	44.2	**100.0**	30.0	72.9	0.0	66.7	21.1	0.0	63.3
BioGait-VLM (Full)	**68.1**	**52.9**	98.1	**36.4**	**80.4**	**35.3**	**66.7**	**31.8**	**4.3**	**70.0**

3.2 Diagnostic Classification Performance

Tab 1 (top) presents results on the unified 8-class benchmark. Standard video encoders (SlowFast, TSN) and zero-shot VLMs struggle to generalize under the rigorous subject-disjoint protocol, achieving accuracies below 40%. This highlights their reliance on environmental shortcuts and inability to capture fine-grained kinematic deviations. In contrast, **BioGait-VLM** achieves a state-of-the-art accuracy of ***68.1%***, outperforming the strongest baseline by $+$***31.0***%. Crucially, our framework delivers substantial gains on subtle pathologies, achieving F1 scores of ***98.1%*** for *DCM* and ***66.7%*** for *Parkinson's*. This confirms that explicitly modeling biomechanics and temporal dynamics enables the system to disentangle pathological motion from complex environmental noise.

3.3 Clinical Validation via Blinded Expert Assessment

Quantitative metrics (accuracy/F1) do not fully capture a model's utility in a clinical workflow. To assess the interpretability and trustworthiness of the generated rationales, we conduct a *blinded qualitative study* with four evaluators possessing specific domain expertise in DCM.

Study Protocol. We utilize the complete set of 52 video sequences derived from all 6 patients in the held-out DCM test cohort. To ensure exhaustive coverage without fatigue, the dataset is partitioned into non-overlapping subsets, with each evaluator reviewing 13 unique cases. For every sequence, experts review anonymized diagnostic rationales from three ***anonymized*** models: (1) *InternVL-3.5 (Zero-Shot):* The generic backbone without adaptation. (2) *BioGait-VLM (w/o biomechanical branch):* fine-tuned model without the biomechanical branch. (3) *BioGait-VLM (Full):* The proposed framework. Evaluators rate the responses on a 5-point Likert scale across four clinical dimensions:

Table 2: **Human Expert Evaluation.** Average likert scores (Range: 1–5, higher is better). Our *Biomechanical Tokenization Branch* significantly boosts Evidence Grounding and Usefulness compared to vision-only approaches, reducing hallucinations.

Model Variant	Grounding	Explainability	Usefulness	Consistency
InternVL3.5-1B [25] (Zero-Shot)	2.31	2.98	2.46	3.85
BioGait-VLM (*w/o Biomechanical Branch*)	1.98	2.12	1.83	2.13
BioGait-VLM (Full)	**4.23**	**3.98**	**3.69**	**4.15**

1. **Evidence Grounding:** Does the rationale cite observable biomechanical metrics (*e.g.*, "reduced knee flexion") rather than generic statements?
2. **Explainability:** Is the explanation coherent, logical, and easy for a clinician to follow?
3. **Clinical Usefulness:** Would this output aid real-world screening or documentation?
4. **Consistency:** Is the predicted diagnosis internally consistent with the supporting text?

Finally, experts select the single "Best Model" for each case and provide open-ended comments to justify their selection.

Results and Analysis. As detailed in Tab. 2, our BioGait-VLM dominates across all metrics, particularly in Evidence Grounding (***4.23*** vs. 2.31 for the strongest baseline). In the comparative ranking, experts select BioGait-VLM as the superior model in ***69.2%*** of cases (36/52), compared to just 26.9% for the zero-shot baseline. Qualitative feedback highlights the critical role of the Biomechanical Branch. Evaluators note that while baseline models often provide "generic" or "vague" explanations (*e.g.*, "walking difficulty"), the Full Model produced rationales anchored in "concrete metrics." One expert comment: *"Even though [the baseline] is generic but makes sense, [BioGait-VLM] wins with concrete metrics."* Another notes that even when visual cues are ambiguous, *"the biomechanical metrics present a better rationale... the data suggests otherwise."* This confirms that textualized kinematics transform the system from a black-box classifier into a transparent, evidence-based clinical assistant.

3.4 Ablation Analysis

As detailed in the bottom section of Tab. 1, we isolate the contribution of each modality. Impact of Temporal Modeling: Removing the *Temporal Evidence Distillation (TED)* branch causes a sharp performance drop, particularly in the *Myopathic* class (F1 decreases significantly). This confirms that static VLM features struggle to capture rhythmic gait anomalies (*e.g.*, waddling) without explicit temporal modeling. Impact of Biomechanical Tokenization: Excluding the *Biomechanical Branch* degrades performance on structurally defined pathologies like *Parkinson's* and *DCM*. The full tri-modal framework achieves the highest overall accuracy (***68.1%***), validating that textualized joint-angle tokens provide critical geometric grounding that complements visual appearance.

4 Conclusion

We present **BioGait-VLM**, a tri-modal framework that addresses the generalization failure of standard video models by integrating explicit biomechanical reasoning into a Large Vision-Language Model (LVLM). By projecting 3D kinematics into language-aligned semantic tokens, our approach successfully decouples pathological motion from environmental shortcuts. Extensive experiments on our rigorous 8-class benchmark, which includes the newly collected DCM dataset, demonstrate state-of-the-art performance on unseen patients. Furthermore, our blinded expert study confirms that this architecture transforms the model into a transparent clinical assistant capable of generating evidence-grounded assessments. These results suggest a practical path toward interpretable, privacy-preserving digital biomarkers for neurology.

Acknowledgments. This work was supported in part by the Google Cloud Research Credits program and the NVIDIA Academic Grant Program Award.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML (2021)
3. Bilney, B., Morris, M., Webster, K.: Concurrent related validity of the GAITRite® walkway system for quantification of the spatial and temporal parameters of gait. Gait & posture **17**(1), 68–74 (2003)
4. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
6. Horak, F.B., Mancini, M.: Objective biomarkers of balance and gait for parkinson's disease using body-worn sensors. Movement Disorders **28**(11), 1544–1551 (2013)
7. Joao Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
8. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. ACM Transactions on Graphics (TOG) **42**(6), 1–12 (2023)
9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2023)
10. Li, H., Guo, Y., Jiang, F., Dang, T.M., Ma, H., Zhou, Q., Gao, J., Huang, J.: Text-guided multi-instance learning for scoliosis screening via gait video analysis. In: MICCAI (2025)
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
12. Mahfouf, Z., Jarraya, I., Dhieb, T., Neji, M., Farhat, N., Smaoui, E., M. Hamdani, T., Damak, M., Mhiri, C., M. Alimi, A.: PDGV dataset: Investigating deep learning cnn for gait-based parkinson's disease recognition. Arabian Journal for Science and Engineering pp. 1–21 (2025)
13. Mc Ardle, R., Del Din, S., Donaghy, P., Galna, B., Thomas, A.J., Rochester, L.: The impact of environment on gait assessment: considerations from real-world gait analysis in dementia subtypes. Sensors **21**(3), 813 (2021)
14. Mc Ardle, R., Galna, B., Donaghy, P., Thomas, A., Rochester, L.: Do alzheimer's and lewy body disease have discrete pathological signatures of gait? Alzheimer's & Dementia **15**(10), 1367–1377 (2019)
15. Montero-Odasso, M.M., Sarquis-Adamson, Y., Speechley, M., Borrie, M.J., Hachinski, V.C., Wells, J., Riccio, P.M., Schapira, M., Sejdic, E., Camicioli, R.M., et al.: Association of dual-task gait with incident dementia in mild cognitive impairment: results from the gait and brain study. JAMA neurology **74**(7), 857–865 (2017)
16. Ranjan, R., Ahmedt-Aristizabal, D., Armin, M.A., Kim, J.: Developing normative gait cycle parameters for clinical analysis using human pose estimation. In: DICTA (2024)
17. 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)

18. Studenski, S., Perera, S., Patel, K., Rosano, C., Faulkner, K., Inzitari, M., Brach, J., Chandler, J., Cawthon, P., Connor, E.B., et al.: Gait speed and survival in older adults. Jama **305**(1), 50–58 (2011)
19. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS (2022)
20. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2740–2755 (2018)
21. Xia, Y., Zhou, X., Vouga, E., Huang, Q., Pavlakos, G.: Reconstructing humans with a biomechanically accurate skeleton. In: CVPR (2025)
22. Yin, X., Yang, B., Liu, W., Xue, Q., Alamri, A., Fiedler, G., Gao, W.: ProGait: A multi-purpose video dataset and benchmark for transfemoral prosthesis users. In: ICCV (2025)
23. Zhang, W., Cheng, Z., Lei, F., Liu, B., Gu, D., Wang, Z.: Enhancing logical reasoning in large language models via multi-stage ensemble architecture with adaptive attention and decision voting. In: BDEIM (2024)
24. Zhou, Z., Liang, J., Peng, Z., Fan, C., An, F., Yu, S.: Gait patterns as biomarkers: A video-based approach for classifying scoliosis. In: MICCAI (2024)
25. 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)