Vision-Language Models for Gait-Based Screening via Zero-Shot and Multimodal In-Context Learning

MCMI Workshop @ ICPR 2026

Moreno La Quatra, **Vito Cammarata**, Gabriele Trovato,
Vincenzo Conti, Valerio Mario Salerno, Salvatore Sorce, Nicole Dalia Cilia

Kore University of Enna, Italy

VERA | Interreg Italia - Malta | Cofinanziato dall'Unione Europea Co-funded by the European Union | COESIONE ITALIA 21-27 SICILIA | FONDI.eu

Outline

Motivation: Camera-Based Gait Analysis

- **Gait as a Clinical Indicator:** Walking patterns carry rich diagnostic signals for musculoskeletal (e.g., Knee Osteoarthritis) and neurological (e.g., Parkinson's) disorders.
- **Scalable Screening:** Video gait analysis allows contact-free, low-cost early screening without expensive motion-capture hardware.
- **Supervised Bottleneck:** Current classifiers require large labeled datasets for every single target condition.

Key Research Question

Can general-purpose **Vision-Language Models (VLMs)** perform clinical gait screening via zero-shot prompts or in-context learning, without task-specific fine-tuning?

Background & Challenges

VLMs in Medical Domains

- VLMs excel in general video understanding and QA.
- *Challenge:* Zero-shot VLMs often fail on fine-grained, out-of-domain clinical tasks.

Modality Bias

- Multimodal LLMs tend to prioritize textual priors over visual context.
- Scaling model parameters alone does not resolve visual perception bottlenecks.

Objective

Systematically evaluate open/closed VLMs across zero-shot prompting levels vs. **Similarity-Guided Multimodal In-Context Learning (ICL)**.

Dataset & Evaluation Protocol

- **Benchmark Dataset:** KOA-PD-NM Gait Video Dataset (96 subjects, 190 usable clips, fixed frontal camera).
- **Classes:** Healthy Controls (**Normal**), Knee Osteoarthritis (**KOA**), Parkinson's Disease (**PD**).

Subject-Disjoint Partition

Strict separation to prevent identity leakage:

- **ICL Support Set:** 14 subjects (8 KOA, 4 Normal, 2 PD).
- **Test Set:** 82 subjects (42 KOA, 26 Normal, 14 PD).

Evaluation Metrics

- **Macro-F1** (Primary metric due to class imbalance).
- Balanced Accuracy & Accuracy.
- Per-class F1 metrics.

Evaluated Models

Family	Model	Key Characteristics
Open VLMs	**Gemma 4** (E2B, E4B, 31B)	Dense models, native video support, runtime thinking mode toggle.
	Qwen3-VL (8B, 32B)	Native video, interleaved multimodal contexts, thinking checkpoints.
Closed VLM	**Gemini 2.5 Flash**	Proprietary API, ingests native video, internal reasoning mode.
Reference	**V-JEPA 2 + kNN**	Self-supervised video encoder (1.2B), **language-free reference**.

Prompt Design Spectrum

We evaluate four prompt configurations of increasing structure and context:

1. **L0 (Direct Classification):** Asks for label directly in JSON (no explicit reasoning).
2. **L1 (Describe then Classify):** Prompts free-text description of walking pattern before classification.
3. **L2 (Structured Gait Analysis):** Analyzes 6 clinical features (stride, arm swing, posture, foot clearance, cadence, base of support) prior to output.
4. **L3 (Multimodal ICL):** Prepends $k = 2$ labeled support clips before query clip.

Similarity-Guided Multimodal ICL (L3)

Retrieval Strategy

- Support clips retrieved using **SigLIP 2** frame-level embeddings.
- Nearest demonstrations selected via average cosine similarity ($k = 2$).

Prompt Context Structure

- `Support 1:` [Clip 1] $\rightarrow$ {`"label": "KOA"`}
- `Support 2:` [Clip 2] $\rightarrow$ {`"label": "Normal"`}
- `Query:` [Test Clip] $\rightarrow$ Classify gait

Core Advantage

Demonstrations anchor the VLM to real domain visual examples **without gradient updates** or parameter fine-tuning.

Main Results: Three-Way Gait Screening

Model	Macro-F1	Bal. Acc.	Acc.	KOA F1	Normal F1	PD F1
Zero-Shot VLMs (L2 Canonical Prompt)						
Qwen3-VL 8B	0.229	0.370	0.333	0.000	0.486	0.200
Gemma 4 31B	0.314	0.484	0.364	0.000	0.613	0.330
Gemini 2.5 Flash†	0.243	0.375	0.321	0.000	0.479	0.250
Multimodal ICL VLMs (L3 2-Shot Similarity)						
Qwen3-VL 8B	0.771	0.768	0.852	0.951	0.826	0.537
Qwen3-VL 32B	0.501	0.536	0.611	0.704	0.622	0.176
Gemma 4 31B	0.600	0.617	0.685	0.794	0.662	0.343
Gemini 2.5 Flash†	0.454	0.470	0.494	0.515	0.513	0.333
Language-Free Reference						
V-JEPA 2 + kNN	**0.791**	0.782	0.883	0.976	0.857	0.540

† Closed proprietary model.

Key Findings: Zero-Shot vs. ICL

Zero-Shot Models Fail

- All zero-shot VLMs score below 0.314 Macro-F1.
- Assign **zero F1 score to KOA** (defaulting to majority class).

ICL Closes the Gap

- Qwen3-VL 8B reaches **0.771 Macro-F1**, rivaling the V-JEPA 2 reference (**0.791**).
- Visual grounding is the single most critical factor.

Inverse Model Scaling Pattern:

- Qwen3-VL 8B (0.771) outperforms Qwen3-VL 32B (0.501) under ICL.
- Larger language backbones impose stronger textual priors that compete with fine visual signals.

Binary Task (Normal vs. Abnormal) & Clinical Bias

Merging KOA and PD into a single **Abnormal** class isolates normal vs. pathological decision-making:

Model Setting	Macro-F1	Bal. Acc.	Abnormal F1	Normal F1
Zero-Shot (Gemma 4 31B)	0.598	0.675	0.615	0.581
ICL (Qwen3-VL 8B)	<u>0.767</u>	0.833	0.800	0.734
V-JEPA 2 + kNN (Reference)	**0.887**	0.923	0.917	0.857

Systematic Clinical Bias

VLMs exhibit high Normal recall but lower precision $\rightarrow$ **tendency to under-flag pathological gait**. In clinical screening, false negatives are a primary failure mode to prevent.

Ablation 1: Prompt Complexity

Does structured clinical prompting improve zero-shot performance?

Model	Thinking Mode	L0 (Direct)	L1 (Describe)	L2 (Structured)
Qwen3-VL 8B	Disabled	0.208	0.226	0.229
Gemma 4 E2B	Disabled	0.153	**0.358**	0.226
Gemma 4 31B	Enabled	0.249	0.265	**0.314**
Gemini 2.5 Flash	Enabled	0.266	**0.360**	0.243

- **L1 (Describe then classify)** achieves higher scores than L2 for most models.
- **Takeaway:** Rigid feature templates (L2) can force confident but incorrect feature assessments, whereas free-form description (L1) imposes fewer constraints.

Ablation 2: Temporal Resolution

Evaluating zero-shot Macro-F1 across frame counts (4, 8, 16, 32 frames):

	Qwen3-VL 8B		Gemma 4 31B	
Frames	*No Think*	*Think*	*No Think*	*Think*
4	0.160	0.160	0.249	0.268
8	0.207	0.160	0.239	0.295
16	**0.229**	0.160	0.301	**0.314**
32	**0.229**	0.160	0.286	0.304

Observations

- Performance plateaus or peaks around **16 frames**.
- Increasing temporal resolution alone does not bridge the clinical domain gap.
- Thinking mode in Qwen3-VL 8B outputs a constant 0.160 across frame counts.

Ablation 3: Impact of Reasoning Mode

Setting	Qwen3 8B	Qwen3 32B	Gemma E2B	Gemma E4B	Gemma 31B
Zero-Shot (No Think)	0.229	0.229	0.226	0.162	0.301
Zero-Shot (Think)	0.160	0.227	0.205	0.196	0.314
Δ Zero-Shot	-0.069	-0.001	-0.021	+0.034	+0.014
ICL (No Think)	**0.771**	0.501	0.238	0.259	0.339
ICL (Think)	0.604	0.327	0.564	0.329	0.600
Δ ICL	-0.167	-0.174	+0.326	+0.070	+0.261

Divergent Family Dynamics:

- Reasoning mode **benefits Gemma 4** under ICL (up to +0.326).
- Reasoning mode **hurts Qwen3-VL** under ICL (down to −0.174), overriding support demonstrations.

Discussion

- **Domain Gap Primacy:** Text prompt engineering (L0–L2) cannot bridge the domain gap between general web video pretraining and clinical gait analysis.
- **Visual Grounding vs. Scale:** Grounded visual support clips (ICL) provide much larger gains than scaling model parameters or adding internal reasoning loops.
- **Textual Overrule:** In larger backbones, language priors tend to dominate, overriding fine visual evidence provided by support demonstrations.

Conclusion & Future Work

Summary

1. Zero-shot VLMs are unreliable for clinical gait screening (Macro-F1 ≤ 0.314).
2. **Similarity-guided 2-shot ICL** brings the best VLM to **0.771 Macro-F1**, close to the language-free reference (0.791).
3. Grounded visual examples are the single dominant factor for transfer.

Future Directions:

- Clinical prompt tuning incorporating biomechanical knowledge.
- Enhanced temporal representations for subtle conditions (Parkinson's disease).
- Lightweight parameter-efficient fine-tuning (PEFT).

Thank You!

Questions & Discussion

Presenter: Vito Cammarata (vito.cammarata@unikore.it)
Kore University of Enna, Italy