Enhancing Gait Video Analysis in Neurodegenerative Diseases by Knowledge Augmentation in Vision Language Model *

Diwei Wang[1], Kun Yuan[1], Candice Muller[2], Frédéric Blanc[1,2], Nicolas Padoy[1], and Hyewon Seo[1]

[1] ICube laboratory, University of Strasbourg, CNRS, France
{d.wang, kyuan, f.blanc, npadoy, seo}@unistra.fr
[2] Hôpital de la Robsertsau, France
{candice.muller, frederic.blanc}@chru-strasbourg.fr

Abstract. We present a knowledge augmentation strategy for assessing the diagnostic groups and gait impairment from monocular gait videos. Based on a large-scale pre-trained Vision Language Model (VLM), our model learns and improves visual, textual, and numerical representations of patient gait videos, through a collective learning across three distinct modalities: gait videos, class-specific descriptions, and numerical gait parameters. Our specific contributions are two-fold: First, we adopt a knowledge-aware prompt tuning strategy to utilize the class-specific medical description in guiding the text prompt learning. Second, we integrate the paired gait parameters in the form of numerical texts to enhance the numeracy of the textual representation. Results demonstrate that our model not only significantly outperforms state-of-the-art methods in video-based classification tasks but also adeptly decodes the learned class-specific text features into natural language descriptions using the vocabulary of quantitative gait parameters. The code and the model will be made available at our project page: https://lisqzqng.github.io/GaitAnalysisVLM/.

Keywords: Pathological gait classification · MDS-UPDRS Gait score · Knowledge-augmented prompt tuning · Numeracy for language model

1 Introduction

While quantitative gait impairment analysis has proven to be an established method for accessing neurodegenerative diseases and gauging their severity [17,19,15,21], current clinical assessments are used in highly restricted contexts, posing significant challenges: Not only do they often require specialized equipment, such as force plates or IMU sensors, but they also struggle to capture moments with prominent symptoms during clinical visits, which are somewhat special occasions for patients. Analysing motor symptoms from video offers new possibilities, enabling cost-effective monitoring, remote surveillance without the need of frequent

in-person clinic visits, thereby facilitating timely and personalized assessment.

Naturally, there have been recent efforts to develop a single 2D-RGB-camera-based gait analysis system, with the majority leveraging advancements in deep learning. Albuquerque et al [2] develop a spatiotemporal deep learning approach by producing a gait representation that combines image features extracted by Convolutional Neural Networks (CNNs), chained with a temporal encoding based on a LSTM (Long Short Term Memory) network. Sabo et al [24] have shown that Spatiotemporal-Graph Convolution Network models operating on 3D joint trajectories outperform earlier models. In the work by Lu et al [14], 3D body mesh and pose are extracted and tracked from video frames, and the sequence of 3D poses is classified based on MDS-UPDRS gait scores [6] using a temporal CNN. Wang et al [25] have developed a dedicated 3D skeleton reconstructor tailored for gait motion, incorporating a gait parameter estimator from videos and a multihead attention Transformer for similar classification tasks. Among methods for non-pathological gait analysis, GaitBase [4] combines improved spatial feature extraction and temporal gait modeling for appearance-based gait recognition, in both indoor and outdoor settings.

Existing works face challenges in handling insufficient pathological gait data and imbalances with normal data, promoting strategies such as a self-supervised pretraining stage prior to the task-specific supervision [24], or the employment of crafted loss functions [14]. Nevertheless, the need for data-efficient approaches with superior performance is crucial in video-based pathological gait classification. Meanwhile, the recent emergence of large-scale pre-trained vision-language models (VLMs) has demonstrated remarkable performance and transferability to different types of visual recognition tasks [23,20], thanks to their generalizable visual and textual representations of natural concepts. In the context of medical image analysis, VLMs tailored to various medical imaging tasks via finetuning [9], multimodal global and local representation learning [8], knowledge-based prompt learning [22,10], knowledge-based contrastive learning on decoupled image and text modality [27], and large-scale noisy video-text pretraining [29].

Inspired by these works, we propose a new approach to transfer and improve representations of VLMs for the pathological gait classification task in neurodegenerative diseases. Concretely, we model the prompt's context with learnable vectors, which is initialized with domain-specific knowledge. Additionally, numerical gait parameters paired with videos are encoded and aligned with the text representation with a contrastive learning. During training, the model learns visual and text representations capable of understanding both the class-discriminating and numerical features of gait videos. To our knowledge, our work represents the first attempt to deploy VLM for the analysis of pathological gait videos.

2   Method

An overview of our method is shown in Fig.1. We utilize three distinct modalities to enhance the accuracy and the reliability of the VLM in classifying medical

concepts: gait videos, class-specific medical descriptions and numerical gait parameters. Our knowledge augmentation strategy consists of two parts: First, we adopt a knowledge-aware prompt learning strategy to exploit class-specific description in the text prompts generation, while leveraging the pre-aligned video-text latent space (Sec.2.2). Second, we incorporate the associated numerical gait parameters as numerical texts to enhance the numeracy within the latent space of the text (Sec.2.3). Once the training is complete, we run a classifier solely on the video features extracted from the input video.

2.1 Dataset and preprocessing

Dataset. Our study leverages a dataset comprising 90 gait videos from 40 patients diagnosed with neurodegenerative disorders and 3 healthy controls. Moreover, 28 gait video clips featuring healthy elderly individuals have been added, chosen from the TOAW archive [18] based on specific criteria (Berg Balance Scale ≥ 45, 0-falls during last 6 months, etc.), bringing the total number to 118 clips. All the videos are recorded at 30 fps, each capturing a one-way walking path of an individual. The patients were instructed to walk forth and back on a GAITRite (`https://www.gaitrite.com/`) pressure-sensitive walkway, providing a set of gait parameters as outlined in Table 1 in Supplementary Material.

Preprocessing. We crop the original videos based on bounding boxes, and employ a sliding window scheme (window size: 70 frames) to generate subsequences, with a stride of 25 for training and 0 for validation. This process results in approximately 900 clips of 70 frames for each cross-validation fold. To effectively incorporate the gait parameters into text space, we formulate sentences by combining four gait parameters with "and", connecting names and values with "is", as illustrated in Fig.2. The choice of four parameters per sentence is based on our observation that, in practice, neurologists often label a video by using only a few prominent or representative visual clues rather than exhaustively listing all evidences. Out of the total 29 parameters available, we select 438 combinations, each containing 4 parameters whose Pearson correlation coefficients are within the range of $[-0.4, 0.4]$.

2.2 VLM fine-tuning with visual and knowledge-aware prompts

We adopt the prompt learning strategy, keeping the pre-trained VLM frozen to preserve its general representation and leverage the pre-aligned multi-modal latent space. Taking inspiration from KAPT [10], we introduce gait-specific knowledge-based prompts by feeding per-class descriptions $Desc_i$ (See our project page) into the text prompts. These *clinical gait notions* have been generated using ChatGPT-4 [1], then subsequently filtered, modified, and validated by a neurologist. To devise learnable prompts, we use KEPLER [26], similar to [10], on the class descriptions, which are then projected through per-class multi-layer
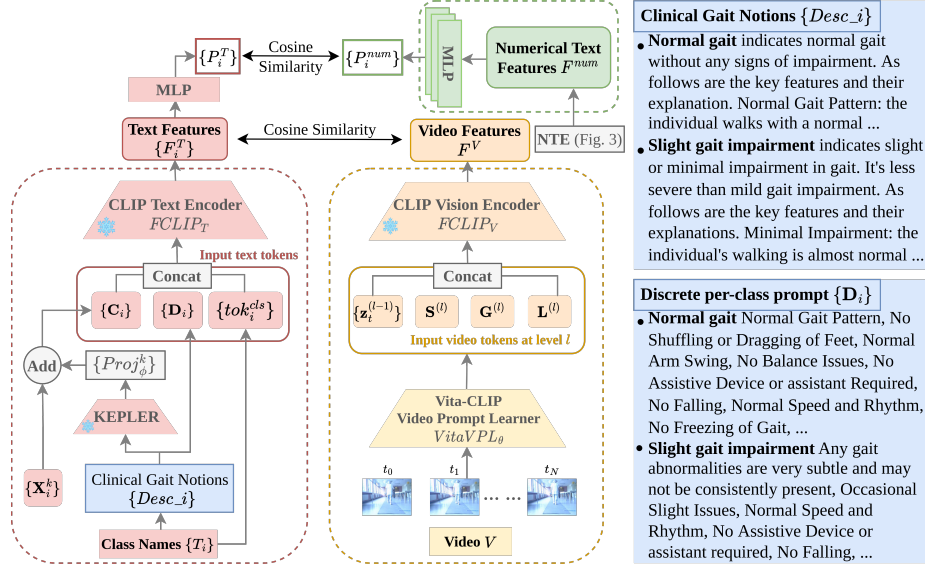
Fig. 1: Overview of our cross-modality model for video-based clinical gait analysis (left), alongside clinical gait notions and per-class descriptions of gait classes utilized for prompt initialization (right). Three colored blocks represent the text- and video encoding pipelines, and the text embedding of numerical gait parameters, respectively.
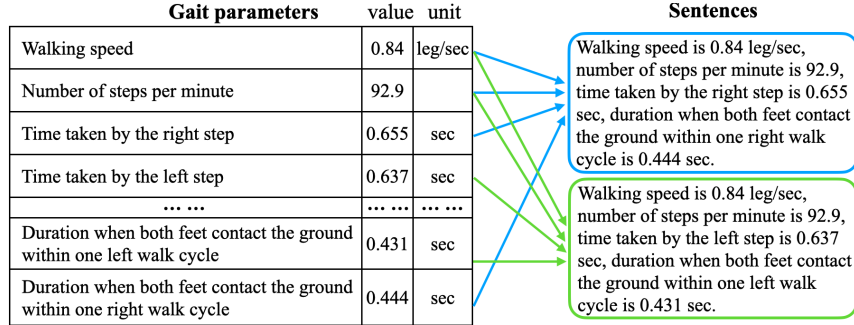

Fig. 2: Translation of gait parameters into text.

perceptrons (MLPs), and added to the learnable parameters $\{X_i^k\}$ to form learnable prompts $\{C_i^k\}$:

$$\{C_i^k\}_{i=1,...,N_{cls}} = Proj_\phi^k(KEPLER(\{Desc_i\})) + \{X_i^k\}, \quad k=1,...,8 \quad (1)$$

where N_{cls} is the number of class, $C_i^k \in \mathbb{R}^{512}$ and $X_i^k \in \mathbb{R}^{512}$ represent the k-th learnable prompts and parameters associated to the i-th class, respectively. For the automatic prompt $\{D_i\}$, we extract keywords from $\{Desc_i\}$, as illustrated in Fig.1. More examples can be found on the project page. These selected texts

then undergo standard tokenization of the frozen CLIP text encoder $FCLIP_T$ to obtain $\{D_i\}$. Similarly, we pass the class names $\{T_i\}$ into the tokenizer of $FCLIP_T$ to generate the class token tok_i^{cls}. As shown in Fig.1, we concatenate $\{C_i\}$, $\{D_i\}$ and $\{tok_i^{cls}\}$ into $FCLIP_T$ to obtain the text features $\{F_i^T\}$:

$$\{F_i^T\} = FCLIP_T([\{C_i\}, \{D_i\}, \{tok_i^{cls}\}]). \tag{2}$$

On the video side, each frame of the input video V goes through the tokenization of the Vision transformer (ViT) [3] and forms a sequence of per-frame representations $z_t^{(0)}$. The visual prompts for the l-th layer of the pretrained CLIP Vision Encoder $FCLIP_V$ are derived by applying Vita-CLIP [28]'s video prompt learner ($VitaVPL$) to the output of the previous layer $\{z_t^{(l-1)}\}$:

$$[S^{(l)}, G^{(l)}, L^{(l)}]_{l=1,\dots,12} = VitaVPL_\theta(\{z_t^{(l-1)}\}), \tag{3}$$

where $S^{(l)}$, $G^{(l)}$, and $L^{(l)}$ respectively denote the learnable summary, global, and local prompt tokens at layer l. As suggested in [28], these prompt tokens are appended to $\{z_t^{(l-1)}\}$ and subsequently fed into $FCLIP_V$ to obtain F^V:

$$F^V = FCLIP_V([\{z_t^{(l-1)}\}, S^{(l)}, G^{(l)}, L^{(l)}]). \tag{4}$$

Moreover, to combat class imbalance, we employ a multi-class focal loss [14] to maximize the cosine similarity of positive pairs:

$$L_k = \sum_{i=1}^{N_{cls}} -\alpha(1-p_i)^\gamma y_i log(p_i), \quad p_i = \frac{exp(<F_i^T|F^V>/\tau)}{\sum_{j=1}^{N_{cls}} exp(<F_j^T|F^V>/\tau)}, \tag{5}$$

where y denotes the one-hot encoded label, $<\cdot|\cdot>$ the cosine similarity, and $\tau = 0.01$ temperature parameter. We set the weighting factor $\alpha = 0.25$ and the focusing parameter $\gamma = 2$.

2.3   Contrastive learning with numerical text embeddings

Text embedding of numerical gait parameters. Starting from the set of sentences each containing four gait parameters, we employ a two-step embedding process as illustrated in Fig.3. Initially, sentences without numerical values are fed into the CLIP text encoder, resulting in a descriptive embedding of the textual content $\{F_{gp}^T\}$. As illustrated in Fig.3, we treat separately the logical conjunction "is" to generate text embedding **[IS]**. Subsequently, number embeddings are generated by multiplying the dedicated embedding base **[NUM]** with the associated numerical values $\{\omega_{gp}\}$. The chosen specialized embedding base is designed to be orthogonal to the position encoding [7], ensuring the efficient transmission of numerical information through the self-attention blocks of the Transformer. The numerical text embedding F^{num} is then obtained by applying the $FCLIP_T$ to the concatenated sentence:

$$F^{num} = FCLIP_T(\{[F_{gp}^T, [\textbf{IS}], \omega_{gp} \cdot [\textbf{NUM}]]\}), \quad gp \in \{1, 2, 3, 4\}. \tag{6}$$
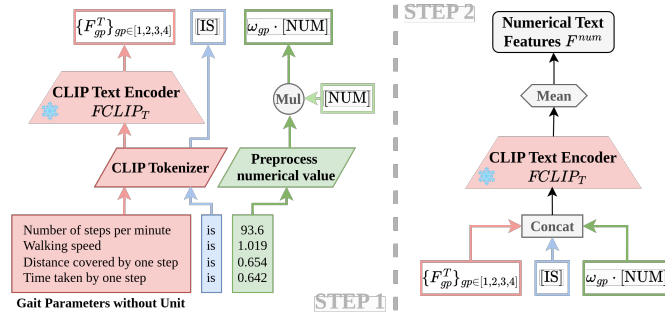
Fig. 3: Numerical text embedding process using the frozen CLIP text encoder.

Fig.1 in Supplementary Material demonstrates the cosine similarities of F^{num} embedded from the text: "the walking speed is [value]", where [value] ranges from 0 to 200, employing different methods to represent the numerical values. Our numerical embedding scheme, in contrast to adopting position encoding or directly encode digit and numerical text with $FCLIP_T$, produces continuous embeddings that best reflect the numerical domain. Given that most gait parameter values are positive, we designate the mean value among healthy controls as the zero reference: $V_{norm} = \alpha \cdot \frac{(V - \overline{V}_{healthy})}{\sigma}$, where σ is the variance of the gait parameter values, and α is the scaling factor to adjust the data range to [-2.5, 2.5], the dynamic range of layer normalization within the self-attention block. The numerical text embeddings for the dementia grouping task are visualized in Fig.4.

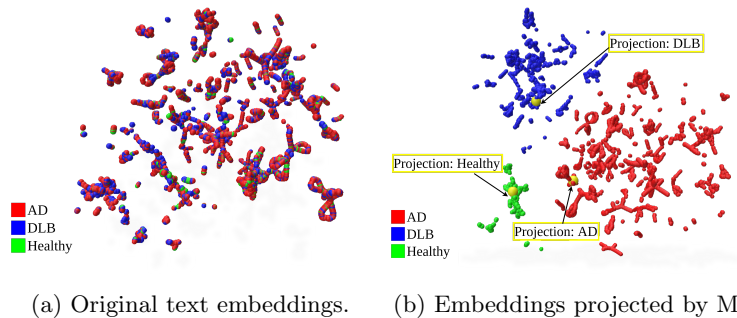(a) Original text embeddings.   (b) Embeddings projected by MLPs.

Fig. 4: Feature visualization using UMAP (no. components=3) for numerical text embeddings derived from gait parameters. Yellow points in (b) represent the projections of the learned per-class text features. Images rendered with Polyscope.

Cross-modal contrastive learning. Since each set of gait parameters is assigned a class label in our training data, we encode these gait parameters into

numerical text embeddings using the embedding method described earlier, and introduce classification tasks thereon.

As illustrated in Fig.1, the projection of the generated text features F_i^T and that of numerical embedding F^{num} are trained in a way that the cosine similarity between P^{num} and the projected text feature of its ground-truth class P^T is maximized, by using a cross-entropy objective L_{gp}. The global loss function becomes: $L = L_k + \omega \cdot L_{gp}$. We set $\omega = 0.05$ through heuristic analysis. To demonstrate alignment of numerical embeddings with the multi-modal space, we visualize the embedding spaces before and after learning in Fig.4.

Interpreting the per-class text embedding. We aim at the effective translation of per-class text features $\{F_i^T\}$ into natural language expressions using the vocabulary of numerical gait parameters. To this end, we trained a text decoder from scratch to transform numerical text embeddings F^{num} back into their corresponding gait parameters, reverting the embedding scheme shown in Fig.3. An 4-layer transformer decoder D_T is employed for text decoding. In line with recent developments in text-only decoder pre-training [13], we train D_T using the prefix language modeling. Specifically, given a sentence composed of gait parameters $\mathbf{s} = \{word_1, word_2, ..., word_L\}$, we generate a sequence of token IDs using the dictionary of $FCLIP_T$. For the numbers, we *scale* the numbers, which had been previously normalized to [-2.5, 2.5], to a graduated integer scale of $[0, N_{num}]$. The token ID tok of a number [**num**] is defined as: $tok = [\mathbf{EOS}] + scale([\mathbf{num}])$, where $[\mathbf{EOS}] = 49407$. D_T learns to reconstruct the sequence of token IDs $\{tok_j\}$ starting from the numerical text embedding F^{num}. In addition to the vanilla cross-entropy loss [13], we leverage an ordinal cross-entropy loss to further penalize the reconstruction error of the number values:

$$L_{num} = -\frac{|\hat{tok} - tok|}{[\mathbf{EOS}] + N_{num} - 1} \sum_{m=1}^{[\mathbf{EOS}]+N_{num}} y_m log(p_m), \tag{7}$$

where $N_{num} = 200$, $|\hat{tok} - tok|$ represents the absolute distance between the ground-truth token ID tok and the estimation $\hat{tok}$, y denotes the one-hot encoded ground-truth label, and p the estimated probability.

Benefiting from the proposed cross-modal contrastive learning scheme, $\{F_i^T\}$ can be represented as a linear combination of the numerical text embeddings F^{num}, with weights computed by measuring the cosine similarity between $\{P_i\}$ and P^{num}. Subsequently, we apply D_T on $\{\hat{F}_i^{num}\}$ to generate natural language descriptions: $\{\hat{Desc}_i\} = \mathbf{D}^T(\{\hat{F}_i^{num}\})$.

3   Experiments and Results

Our study includes two classification tests: *Gait scoring* to estimate the severity of a patient's condition based on a 4-class gait scoring (normal–0, slight–1, mild–2, and moderate–3) following MDS-UPDRS III [6], and *dementia subtyping* to distinguish between different dementia groups: normal/DLB(Dementia with

Lewy Bodies)/AD(Alzheimer's Disease). See the project page for detailed clinical gait descriptions on each class. Due to its limited size (a total of 120 videos), we divide our video dataset into training and validation sets and conduct 10-fold cross-validation for each classification task.

Table 1: Comparative analysis on two classification tasks: Gait score ('Gait scoring') and dementia subtyping ('Dem. group'). Model performance is evaluated using top-1 accuracy ('acc',%) and F1-score ('Fscore',%).

(a) Different model configurations

Model configurations	Gait scoring		Dem. group	
	Acc.	Fscore	Acc.	Fscore
Baseline	64.78	60.75	86.27	79.24
Baseline+KAPT	65.98	61.97	87.29	78.48
Baseline+NTE	64.44	57.64	88.26	81.34
Ours	**67.76**	**62.59**	**90.08**	**83.86**

(b) SOTA methods

State-of-the-art models	Gait scoring		Dem. group	
	Acc.	Fscore	Acc.	Fscore
OF-DDNet[14]	54.73	48.59	68.92	65.38
ST-GCN [24]	49.08	43.87	61.46	56.99
KShapeNet[5]	53.69	44.85	65.27	54.86
GaitBase[4]	43.48	30.25	53.42	41.76

3.1   Ablation studies

We conduct ablation experiments on four model configurations, with different combinations of knowledge augmented prompt tuning (KAPT) and numerical text embedding (NTE). As shown in Table 1(a), the combination of both KAPT and NTE yield the best performance, whereas NTE alone or KAPT alone can sometimes slightly worsen the performance. In general, the models tend to perform better for dementia group task, which can be attributed to the more distinctive per-class descriptions and more objective ground truth labeling in that classification.

3.2   Comparison with state-of-the-art

We compare our model with several related state-of-the-art (SOTA) models. Three of these models ([14,24,25]) are specifically designed for the classification of Parkinsonism severity on 3D skeletons, whereas GaitBase [4] is for the gait recognition and silhouette-based. As shown in Table 1, our method achieves the best overall results in both tasks. This is somewhat expected, as other models are not designed for, and do not adapt well to, the constraints of limited data size. Differing from other models where the performance is averaged across 10 folds, result of GaitBase[4] represents the best across 5 folds. Note that the chosen SOTA methods operate on 3D reconstructed poses or extracted 2D silhouettes, for which we used VIBE[11], MAX-GRNet[25] or PARE[12]. Changing among these algorithms did not result in meaningful differences in performance. The results in Table 1(b) represent the best among these combinations.

3.3   Decoding the per-class description

We apply the pretrained text decoder D_T on the per-class text features $\{F_i\}$ obtained through the cross-modal contrastive learning in Sec.2.3. Examples of the decoded texts are shown in Fig.2 of Supplementary Material and in the accompanying video. For the gait scoring task, descriptions of *slight* impairment bear a closer similarity to *normal* than to *mild* or *moderate* impairments, conforming to the severity levels. Most notably, descriptions on *moderate* manifest significant abnormalities, including marked deviation of the foot angle from the progression line. Additionally, we observe that certain criteria in the clinical gait notions have been mapped to some quantitative gait parameters, such as the slowness in *mild* impairment. For the dementia subtype task, the decoded sentences for the DLB clearly show that the model has learned the distinctive gait characteristics. This aligns with findings in [16], where the motor symptom in the DLB group is greater compared to the AD group. The decoded texts for the AD group, on the other hand, are less distinctive and are rather close to the healthy group. This may be attributed to the limited availability, with only 6 healthy elderly videos having corresponding gait parameters during training.

4   Conclusion

We presented a knowledge augmentation strategy to enhance the adaptability of a large-scale pre-trained Vision-Language Model for video-based gait analysis in neurodegenerative diseases. Our method makes use of class-specific descriptive text and numerical gait parameters associated with patient videos, via prompt learning and numeracy-enhanced text representation, respectively. On two video-based gait classifications tasks, our model significantly outperformed other strong SOTA methods, given only slightly more than 100 videos. We believe that our work demonstrates how to efficiently enhance representation learning and offers a novel alternative to incorporating patient metadata, particularly in tabular form.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Albuquerque, P., Verlekar, T.T., Correia, P.L., Soares, L.D.: A spatiotemporal deep learning approach for automatic pathological gait classification. Sensors **21**(18), 6202 (2021)
3. 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. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
4. 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)

5. Friji, R., Drira, H., Chaieb, F., Kchok, H., Kurtek, S.: Geometric deep neural network using rigid and non-rigid transformations for human action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12611–12620 (2021)

6. Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., et al.: Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. Movement disorders: official journal of the Movement Disorder Society **23**(15), 2129–2170 (2008)

7. Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L., et al.: xval: A continuous number encoding for large language models. arXiv preprint arXiv:2310.02989 (2023)

8. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)

9. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

10. Kan, B., Wang, T., Lu, W., Zhen, X., Guan, W., Zheng, F.: Knowledge-aware prompt tuning for generalizable vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15670–15680 (2023)

11. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020)

12. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021)

13. Li, W., Zhu, L., Wen, L., Yang, Y.: Decap: Decoding clip latents for zero-shot captioning via text-only training. arXiv preprint arXiv:2303.03032 (2023)

14. Lu, M., Poston, K., Pfefferbaum, A., Sullivan, E.V., Fei-Fei, L., Pohl, K.M., Niebles, J.C., Adeli, E.: Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 637–647. Springer (2020)

15. 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)

16. Mc Ardle, R., Del Din, S., Galna, B., Thomas, A., Rochester, L.: Differentiating dementia disease subtypes with gait analysis: feasibility of wearable sensors? Gait & posture **76**, 372–376 (2020)

17. 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)

18. Mehdizadeh, S., Nabavi, H., Sabo, A., Arora, T., Iaboni, A., Taati, B.: The toronto older adults gait archive: video and 3d inertial motion capture data of older adults' walking. Scientific data **9**(1), 398 (2022)

19. Merory, J., Wittwer, J., Rowe, C., Webster, K.: Quantitative gait analysis in patients with dementia with lewy bodies and alzheimer's disease. Gait & posture **26**, 414–9 (10 2007). https://doi.org/10.1016/j.gaitpost.2006.10.006

20. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
21. Muller, C., Perisse, J., Blanc, F., Kiesmann, M., Astier, C., Vogel, T.: Corrélation des troubles de la marche au profil neuropsychologique chez les patients atteints de maladie d'alzheimer et maladie à corps de lewy. Revue Neurologique **174**, S2–S3 (2018)
22. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. arXiv preprint arXiv:2209.15517 (2022)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
24. Sabo, A., Mehdizadeh, S., Iaboni, A., Taati, B.: Estimating parkinsonism severity in natural gait videos of older adults with dementia. IEEE journal of biomedical and health informatics **26**(5), 2288–2298 (2022)
25. Wang, D., Zouaoui, C., Jang, J., Drira, H., Seo, H.: Video-based gait analysis for assessing alzheimer's disease and dementia with lewy bodies. In: Wu, S., Shabestari, B., Xing, L. (eds.) Applications of Medical Artificial Intelligence. pp. 72–82. Springer Nature Switzerland, Cham (2024)
26. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., Tang, J.: Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics **9**, 176–194 (2021)
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
28. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23034–23044 (2023)
29. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. arXiv preprint arXiv:2307.15220 (2023)