GaVA-CLIP:Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases

Diwei Wang, Kun Yuan, Cedric Bobenrieth, Hyewon Seo

GaVA-CLIP:Refining Multimodal Representations with Clinical Knowledge and Numerical Parameters for Gait Video Analysis in Neurodegenerative Diseases

Diwei Wang, Kun Yuan, Cédric Bobenrieth and Hyewon Seo

***Abstract*— We present GaVA-CLIP, a knowledge augmentation strategy for Gait Video Analysis, designed to assess diagnostic groups and gait impairment. Based on the large-scale pretrained Vision Language Model, CLIP, GaVA-CLIP learns and enhances visual, textual, and numerical representations of patient gait videos through 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 class-specific medical descriptions in guiding text prompt learning. Second, we integrate paired gait parameters as numerical texts to enhance the numeracy of textual representations. Results demonstrate that GaVA-CLIP not only significantly outperforms state-of-the-art (SOTA) 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/`.**

***Index Terms*— Pathological gait classification, MDS-UPDRS gait score, Pretrained VLM prompt tuning, Medical knowledge transfer, Numeracy in language models.**

I. Introduction

WHILE quantitative gait impairment analysis has proven to be an established method for assessing neurodegenerative diseases and gauging their severity [1]–[4], 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 many efforts to develop a single 2D-RGB-camera-based gait analysis system, with the majority leveraging advancements in deep learning. They train classifier models on spatiotemporal features extracted from either image sequences [5] or 3D joint trajectories extracted from videos [6]–[9].

However, 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 [6], or the employment of crafted loss functions [7]. Therefore, the need for data-efficient approaches with superior performance is crucial in video-based pathological gait classification. Meanwhile, the recent emergence of large-scale pretrained vision-language models (VLMs) has demonstrated remarkable performance and transferability to different types of visual recognition tasks [10], [11], thanks to their generalizable visual and textual representations of natural concepts. In the context of medical image analysis, VLMs have been tailored to various medical imaging tasks via finetuning [12], multimodal global and local representation learning [13], knowledge-based prompt learning [14], [15], knowledge-based contrastive learning on decoupled image and text modalities [16], and large-scale noisy video-text pretraining [17].

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. Given the limited size of our clinical data, and to exploit the availability of the gait data accompanying the gait video, we prompt our baseline model using distinct modalities. 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. As a result, we enhance the visual and text representations of the VLM model, thereby improving its understanding of both class-discriminating and numerical features present in gait videos. To our knowledge, our work represents the first attempt to deploy VLM for the analysis of

All authors are affiliated with the ICube laboratory (UMR7357), France.

Diwei Wang (e-mail:d.wang@unistra.fr) is with the University of Strasbourg, France.

Cédric Bobenrieth is with Institut Catholique d'Arts et Métiers (e-mail: cedric.bobenrieth@icam.fr).

Hyewon Seo (seo@unistra.fr) is with the CNRS and the University of Strasbourg.

Kun Yuan is with the University of Strasbourg, France and Universität München, Munich, Germany (e-mail: kun.yuan@ext.ihu-strasbourg.eu).

pathological gait videos. A previous version of this work has been presented in [18].

II. Related Work

A. Gait analysis in neurodegenerative diseases

Quantitative gait impairment analysis is an established method for assessing neurodegenerative diseases such as Dementia with Lewy Bodies (DLB) and Alzheimer's Disease (AD), and gauging their severity, even in the prodromal phase [19]. Such pathological gait studies often rely on data collected from wearable sensors or electronic walkways, which allows for quantitative gait analysis. Inertial measurement unit (IMU) sensors have been widely adopted, due to their lightweight design, high sampling rate, and cost-effectiveness. Mannini et al. [20] trained class-specifc Hidden Markov Models to capture the gait motion and evaluated the log-likelihoods of observing data, which were then combined with twelve time and frequency domain features extracted from IMU data and fed into a Support Vector Machines classifier to distinguish between the gaits of healthy elderly, post-stroke, and Huntington's disease. Hsu et al. [21] used multiple inertial sensors to classify gaits in stroke and neurological disorder patients, highlighting the importance of sensor placement and the effectiveness of combining time-domain features (e.g., kurtosis, variance, mean) with temporal gait parameters (e.g., stance time, stride time, double-limb support). More recently, Mc Ardle et al. [3] studied the impact of walking context, revealing that impairment differences between dementia subtypes are more pronounced in laboratory-based gaits, while in real-world settings, these differences are evident only in short walking bouts. Pressure sensitive walkways are also commonly deployed to obtain insights in the analysis of gait impairment for neurodegenerative diseases. Merory et al. [2] showed that individuals with AD and DLB exhibit comparable spatiotemporal gait characteristics that differ significantly from those of the normal population. Somewhat contrary to their study, Mc Ardle et al. [1] demonstrated that patients with Lewy body disease, which includes both DLB and Parkinson's Disease Dementia, exhibit greater variability in step and stance times compared to those with AD. Muller [4] utilized a decision tree [22] to analyze gait motions of individuals with AD and DLB, revealing that walking speed and asymmetry in left-to-right step lengths are the primary factors for distinguishing between dementia subtypes and estimating disease severity.

While wearable sensors or instrumented walkways can facilitate detailed gait analysis, they can often be costly, intrusive, and cumbersome in terms of wearability and calibration. With the advancement of deep learning in computer vision, recent works have focused on vision-based impairment assessments for pathological gait. This shift is particularly important for monitoring diseases in real-world settings, where the cameras are ubiquitously available. Albuquerque et al. [5] developed a spatiotemporal deep learning approach that combines image features extracted by convolutional neural networks (CNNs) with a temporal encoding based on a recurrent neural network (RNN). Interestingly, most recent works base their estimation on 3D skeleton sequences extracted from video [23]. For example, Lu et al. [7] extracted and tracked 3D body meshes and poses from video frames, and performed classification on the 3D pose sequences using a temporal CNN. Additionally, Sabo et al. [6] demonstrated that Spatiotemporal-Graph Convolution Network models operating on 3D joint trajectories outperform earlier models. Gaitforemer [24] introduces a transformer model that operates on sequences of 3D human body skeletons for the pre-training task of human motion forecasting, which is subsequently adapted to the downstream task of MDS-UPDRS gait score estimation. Wang et al [8] 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.

B. Multimodal contrastive learning in medical image analysis

The use of multimodal data, particularly images paired with expert annotations, is a well-established method to enhance model performance and facilitate downstream tasks, which is also widely adopted in medical imaging [25]. Implementations such as BioViL [26] have introduced benchmarks to evaluate the self-supervised biomedical VLM in chest X-ray application settings. However, the complexity and specificity of categorical notions in medical imaging often hinder the direct application of models pre-trained on natural images. This necessitates advanced techniques such as informative supervisions [16], [27], domain-agnostic descriptive attributes [28], enriched textual descriptions [29], [30], fine-grained alignments [13], [31], and multimodal expert annotations [32] to enhance the alignment and understanding of medical images and texts in diagnostic models.

Specifically, MedCLIP [16] addresses the challenge of limited paired data in the medical domain by decoupling images and text reports from a same patient and thereby identifying positive pairs based on semantic similarities between all images and reports. Qin et al. [28] propose to leverage *expressive* medical prompts, incorporating visual descriptive attributes that resonate across both natural and medical domains, to effectively bridge domain gap. FLAIR [30] enriches categorical supervision with text descriptions embedding expert knowledge, adapting the idea of [29] to the medical domain.

MedMPG [31] and GLoRIA [13] focus on addressing the fine-grained informational needs of medical diagnostics by enhancing the alignment within local region of image and texts. MedKLIP [27] further develops the method for local alignment by extracting and utilizing complex medical entities (keywords) from diagnostic text reports. This extraction not only allows for a deeper understanding of the medical context but also supports a more nuanced alignment between images and reports. Additionally, Kumar and Marttinen [32] propose integrating eye-gaze heatmaps of radiologist as expert annotations, to diversify text-image alignment supervision with additional positive samples.

These strategies reflect a broader move towards more domain-specialized models that are deeply integrated with domain-specific knowledge, thereby improving the effectiveness and applicability of VLM in medical contexts. Drawing

inspiration from these techniques that bridge the domain gap between natural and medical images, our approach leverages the knowledge transfer capabilities of the VLM, the CLIP [10] model in particular, to analyze gait impairments in neurodegenerative diseases.

III. Method

We utilize three distinct modalities to enhance the accuracy and the reliability of the VLM in classifying gait videos: 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.III-B). Second, we incorporate the associated numerical gait parameters as numerical texts to enhance the numeracy within the latent space of the text (Sec.III-C). The overview of our model is shown in Fig 1.

A. Dataset and preprocessing

1) Dataset: Our study leverages the clinical videos documented in [8], 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 TOAGA archive [33] based on specific criteria (Berg Balance Scale ≥ 45, 0-falls during last 6 months, etc.), totaling 118 clips. All the videos are recorded at 30 fps, each capturing a straight, one-way walking path of an individual. The subjects in the clinical videos were instructed to walk forth and back on a GAITRite (`https://www.gaitrite.com/`) pressure-sensitive walkway [8], which generated a set of 29 gait parameters. From these, we identified 8 basic gait parameters that are also available in the TOAGA dataset, as outlined in Table I.

TABLE I: Eight basic gait parameters used in our work.

ID	Gait parameter description
1	Number of steps per minute
2	Walking speed
3	Distance covered by one step
4	Time taken by one step
5	Difference in time taken between a right step and a left step
6	Difference in distance covered between a right step and a left step
7	Standard variance among step times
8	Standard variance among step lengths

2) Preprocessing: We crop the original videos based on bounding boxes, and employ a sliding window scheme (window size: 70 frames) to generate sub-sequences, 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 gait parameters with "and", connecting names and values with "is", as illustrated in Fig.2. We selected four parameters per sentence, based on our observation that neurologists typically label a video using only a few prominent visual clues rather than exhaustively listing all evidences. Out of the 8 basic parameters, we made 70 combinations, each containing 4 parameters.

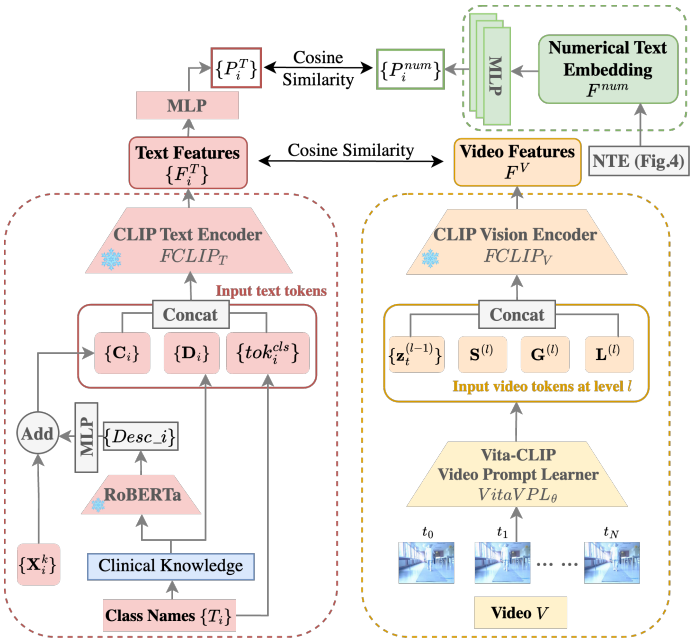


Fig. 1: Overview of our cross-modality model for video-based clinical gait analysis. Three dashed color-blocks (pink, orange, and green) represent the text- and video encoding pipelines, and the text embedding of numerical gait parameters, respectively.

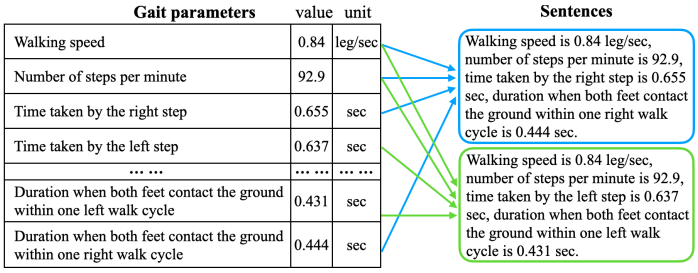


Fig. 2: Gait parameters presented as tabular elements are translated into sentences.

B. Fine-tuning VLM with knowledge-augmented prompts

As demonstrated by ActionCLIP [34], video action recognition can effectively leverage pretrained multimodal models like CLIP [10]. By utilizing appropriate prompts with learnable parameters, ActionCLIP performs end-to-end VLM fine-tuning using downstream training data. In a similar spirit, we adapt the CLIP model for video action recognition but in the context of the pathological gait analysis. Specifically, we employ a prompt learning strategy by creating textual prompts derived from clinical gait notions and using additional prompts on the visual side to manage sequential frame inputs. Note that in our paper, we use "prompt" both as a noun and as a verb, referring to the process of adjusting the pretrained VLM to our gait classification tasks through prompt tuning.

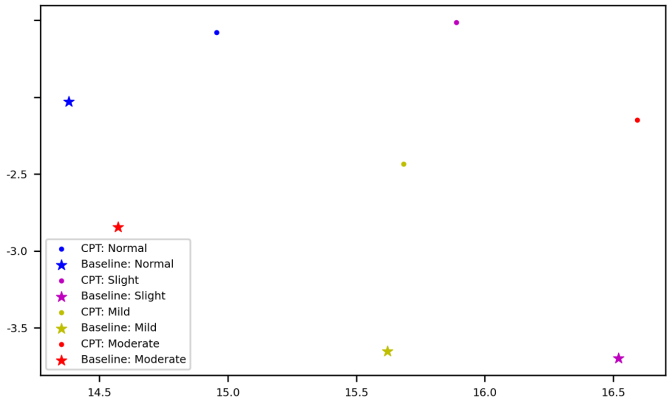
To prompt the frozen CLIP text encoder $FCLIP_T$, we incorporate informative clinical knowledge by constructing per-class learnable prompts $\{C_i\}$. We use ChatGPT-4 [35] to generate categorical descriptions for MDS-UPDRS gait impairment classes and different diagnostic groups. These descriptions are then filtered, modified, and validated by a neurologist. Table V presents an overview of thus obtained clinical knowledge in the form of categorical descriptions. We distill them into category description embeddings $Desc_i$ using a RoBERTa model [36]. Following the unified training strategy introduced in KEPLER [37], the RoBERTa is initialized with the pretrained $\text{RoBERTa}_{\text{BASE}}$ checkpoints, and undergoes an additional training process with joint-objectives of knowledge embedding (KE) and masked language modeling. The KE objective embeds entities and their relation in knowledge graphs, helping to extract inter-class relationships from the category descriptions. To improve contextual understanding and thereby enhance the classification performance, we conditioned text prompts on class names, similar to previous work [38], [39]. Specifically, $\{C_i\}$ are the projections of $\{Desc_i\}$ via a weight-only two-layer multi-layer perceptron MLP^T, added with additional learnable vectors $\{X_i\}$:

$$\{C_i\} = MLP^T(\{Desc_i\}) + \{X_i\}, \tag{1}$$

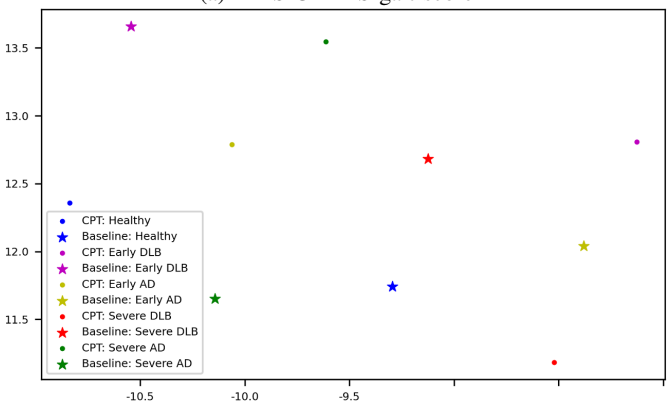
where the index i represents the i-th class, both the MLP and the learnable vectors are defined per-class. We refer to this prompting strategy as Continuous Prompt Tuning (CPT). Fig.3 compares the per-class text features $\{F_i^T\}$ optimized via CPT against those derived through baseline fine-tuning. Utilizing CPT, the relative positioning of text feature distances in a low-dimensional space exhibits a more coherent arrangement. The baseline model trained for gait score classification misrepresents the moderate class as the most relevant to the normal class in Fig. 3(a). Conversely, the CPT model trained for diagnostic group classification in Fig. 3(b) effectively distinguishes gait impairments, showing AD-related impairments as less pronounced than those of DLB. Note that this aligns with the descriptions of clinical knowledge listed in Table V.

Recent work on knowledge-aware prompt tuning (KAPT) [15] indicates that prompts trained on specific data may overfit to seen data. Building on this insight, we tokenize class descriptive texts into discrete prompts $\{D_i\}$ to better leverage clinical knowledge. Considering the 77-word context length limitation of $FCLIP_T$, we have crafted three variants of discrete prompts to condense the text length.

- *KeyPT* (Keyword-wise Prompt Tuning): Unlike the KAPT method [15] where the summaries are extracted by a pretrained language model, our categorical descriptions are a list of criteria characterizing the gait motion, as outlined in Table V. Thus we directly select names of each criterion (such as "Normal gait pattern") to create the summary.
- *SegKPT* (Segmented Knowledge Prompt Tuning): We divide the original descriptions presented in Table V into segments. During contrastive training, $FCLIP_T$ encodes prompts based on the knowledge segments containing one or more sentences. For each class i, we compute the centroid of the text features $\{F_{ij}^T\}$ (j=1,...,N_s) to represent the final $\{F_i^T\}$.



(a) MDS-UPDRS gait score



(b) Diagnostic groups

Fig. 3: The learned per-class text features $\{F_i^T\}$ are visualized using UMAP with 2 components, comparing models prompt-tuned with (circled dots) and without CPT (stars) in two classification tasks.

On the video side, each frame of the input video V goes through the tokenization of the Vision transformer (ViT) [40], collectively forming 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 VitaCLIP [38]'s video prompt learner ($VitaVPL$) to the output of the previous layer $\{z_t^{(l-1)}\}$:

$$[S^{(l)}, G^{(l)}, L^{(l)}]_{l=1,\ldots,12} = \textit{VitaVPL}_\theta(\{z_t^{(l-1)}\}), \tag{2}$$

where $S^{(l)}$, $G^{(l)}$, and $L^{(l)}$ respectively denote the learnable summary, global, and local prompt tokens at layer l. As suggested in [38], these prompt tokens are appended to $\{z_t^{(l-1)}\}$ and subsequently fed into $\textit{FCLIP}_V$ to obtain F^V:

$$F^V = \textit{FCLIP}_V([\{z_t^{(l-1)}\}, S^{(l)}, G^{(l)}, L^{(l)}]). \tag{3}$$

We determine the class label of the visual feature F^V by comparing it to the per-class text features $\{F_i^T\}$ encoded from text prompts. The class with the highest similarity score is chosen as the label. The trainable components of our model are optimized using a contrastive loss, denoted as L_k, to maximize the cosine similarity between class description-video pairs.

Table II shows the class imbalance present in our dataset. To address this imbalance, we implement a multi-class focal loss [7] designed to enhance the cosine similarity of positive pairs. Additionally, we introduce ordinal weights to preserve the intrinsic class ordering in MDS-UPDRS gait scores. The loss L_k for the text-video contrastive learning is formulated as:

$$L_k = \sum_{i=1}^{N_{cls}} [-\alpha(1-p_i)^{\gamma} + \beta \frac{|i - argmax(y)|}{N_{cls}-1}] y_i log(p_i), \quad (4)$$

$$p_i = \frac{exp(< F_i^T | F^V > /\tau)}{\sum_{j=1}^{N_{cls}} exp(< F_j^T | F^V > /\tau)}, \quad (5)$$

where y denotes the one-hot encoded label and p_i the predicted probability of class i. The cosine similarity of feature pair (F_i^T, F^V) is scaled by a learnable temperature parameter τ, initialized to 0.01 to balance the feature distances between intra- and inter-class samples. We set the weighting factors $\alpha = 0.25$ and the focusing parameter $\gamma = 2$, as in Lu et al. [7]. β is set to 0.2 for the gait score classification, while for the classification of diagnostic groups, it is set to 0.

C. Text embedding of numerical gait parameters

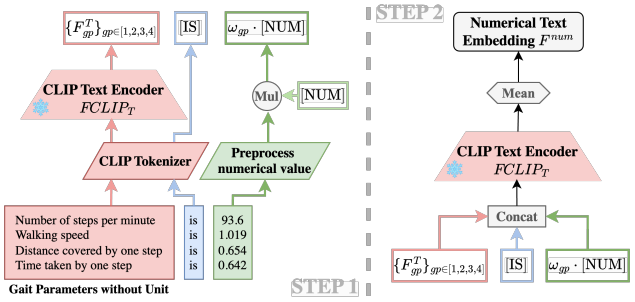


Fig. 4: Our numerical text embedding (NTE) paradigm.

Using the CLIP text encoder to generate numerical text embeddings do not yield favorable representation for our classification tasks. For instance, the text embeddings form clusters by the textual similarity, i.e. the names of gait parameters. Worse, it fails to effectively capture the numerical information within sentences. When sentences vary only by numerical values, the generated embeddings do not adequately reflect these differences. To illustrate this, we used the pretrained CLIP text encoder to generate embeddings for 200 sentences describing the same gait parameter but with values ranging from 1 to 200. As shown in Fig.5(a) and Fig.5(b), the similarities among the generated embeddings, the values are represented with digits (e.g., "1", "2", etc.) or number words (e.g., "one", "two", etc.), show quasi-repetitive patterns. In fact, these patterns highlight the similarity between representations of individual digits and number words. To improve the numerical accuracy of the representation, we introduce a new embedding paradigm specifically designed for gait parameters. Starting from the set of sentences each containing four gait parameters (Sec.III-A.2), we employ a two-step encoding process as illustrated in Fig.4. We start by feeding sentences without numerical values into the CLIP text encoder, resulting in a descriptive embedding of the textual content $\{F_{gp}^T\}$. We treat the logical conjunction "is" separately 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 positional encoding [41], ensuring the efficient transmission of numerical information through the self-attention blocks of the Transformer. The final numerical text embedding F^{num} is then obtained by applying the $FCLIP_T$ to the concatenated sentence:

$$F^{num} = FCLIP_T(\{[F_{gp}^T, [\text{IS}], \omega_{gp} \cdot [\text{NUM}]]\}), \quad gp \in \{1,2,3,4\}. \quad (6)$$

To highlight the importance of using [NUM] for value representation before feeding the concatenated tokens into $FCLIP_T$, we compare different methods for generating token embeddings for integer values in Fig. 5(c)-(f). Our method (Fig. 5(c)) normalizes the value to some range (Details below), and multiplies it with the dedicated base vector [NUM]. Notice how it produces continuous embeddings that effectively capture numerical polarity. Fig. 5(d) is obtained by using positional encoding, where one can observe the repetitive similarity patterns caused by the periodic nature of the encoding functions. Fig. 5(e) and Fig. 5(f) again illustrate the problem of text-similarity dependency when using the CLIP text encoder for embedding numbers, similar to what was observed in Figs. 5(a) and Fig. 5(b). Overall, our numerical embedding scheme produces continuous embeddings that best reflect the numerical domain.

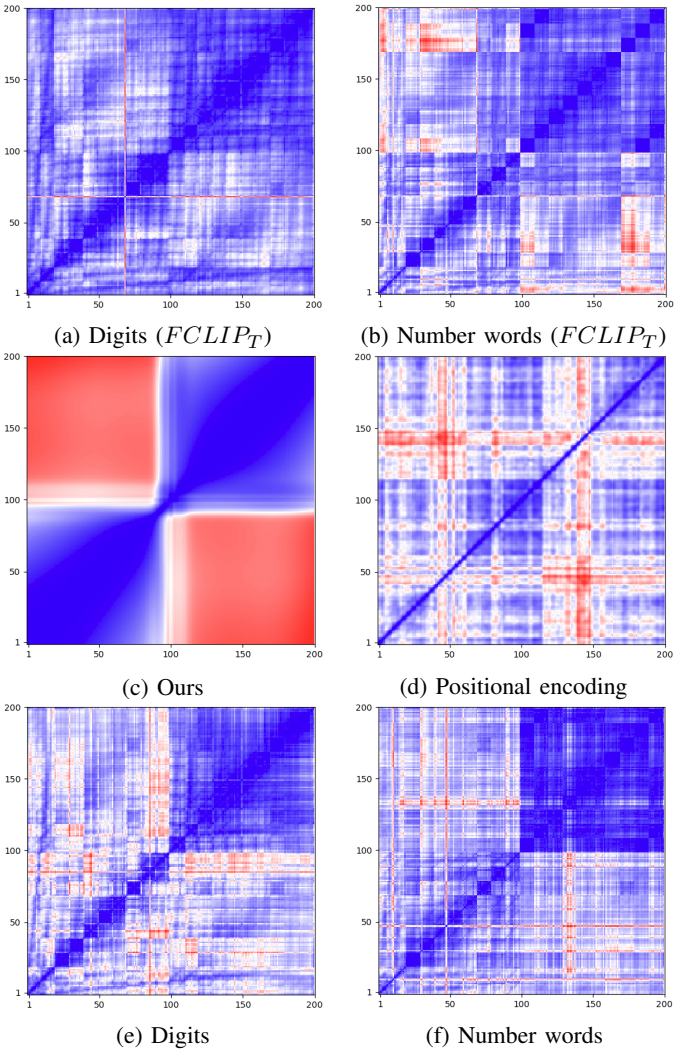
Normalization of parameter values. Given that most gait parameter values are positive, we designate the mean value among healthy controls as the zero reference:

$$V_{norm} = \alpha \cdot \frac{(V - \overline{V}_{healthy})}{\sigma}, \quad (7)$$

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 [41].

D. Multimodal contrastive learning with numerical text embeddings

In our dataset, since each set of gait parameters is assigned a class label, we use these parameters to better align the multimodal representation for our classification tasks. The numerical text embedding as described in SectionIII-C is still insufficient for our classification tasks, as shown in Fig.4(a) and Fig.4(c). Therefore, we train the projections of the generated text features so as to maximize the cosine similarity between the projection of the numerical embedding of gait parameters P^{num} and the projected text feature of their ground-truth class P^T (See Fig.1), using a cross-entropy objective L_{gp}. With this, the global loss function becomes: $L = L_k + \omega \cdot L_{gp}$. We set $\omega = 0.05$ through heuristic analysis. To demonstrate the effect of this learning on alignment of numerical embeddings with the multi-modal space, we visualize the embedding spaces before



(a) Digits ($FCLIP_T$) (b) Number words ($FCLIP_T$)

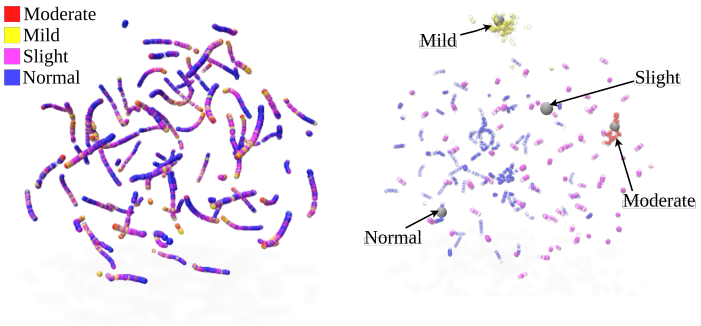
(c) Ours (d) Positional encoding

(e) Digits (f) Number words

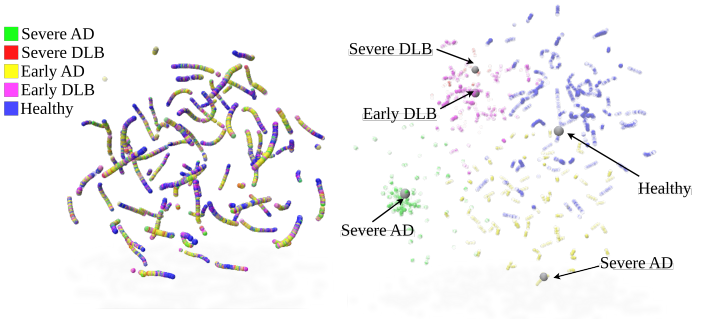
Fig. 5: Different embedding schemes and number representation methods are compared by measuring the cosine similarities among numerical text embeddings (NTE) derived from 200 sentences: "The walking speed is [value]", with [value] ranging from 1 to 200. Subfigures (a) and (b) are obtained by using the pretrained CLIP text encoder $FCLIP_T$, while (c) to (f) adopt our paradigm with different representation methods for numbers: (c) Multiplication with [NUM] (ours), (d) Positional encoding, (e) CLIP token embedding of digits, and (f) CLIP token embedding of number words. Cosine similarity maps are scaled to [-1, 1] which has been color-coded from red (-1) to white (0), and to blue (1).

and after learning in Fig.6. Once the training is complete, we run a classifier on the extracted video feature F^V, obtained by evaluating the video encoding pipeline on the input gait video.

We have attempted to integrate gait parameters into the text modality, by encoding them through a similar pipeline as the text branch in Fig.1. Specifically, we replace the $FCLIP_T$ used in the step 1 of the NTE process (Fig.4) with the RoBERTa model, and add a trainable projection for $\{F_{gp}^T\}$ before the concatenation in step 2. However, we got only suboptimal results in the text-video contrastive learning.



(a) The original NTE (left) and the projected embeddings on the gait scoring (GS) classification



(b) Original NTE (left) and the projected embeddings on the dementia group (DG) classification

Fig. 6: Feature visualization of numerical text embeddings (NTE) derived from gait parameters using UMAP (no. components=3). Compared to the original NTEs on the left, the NTEs projected by the learned MLPs on the right clearly show improved representations for both classification tasks. Yellow points represent the projections of the learned per-class text features. Images rendered with Polyscope.

E. Interpreting the per-class text features

By decoding the text features, we expect to effectively translate the class centers $\{F_i^T\}$ into natural language expressions using the vocabulary of numerical gait parameters. To this end, we trained a text decoder independently from the GaVA-CLIP model for classification. The role of this decoder is to convert numerical text embeddings F^{num} back into their corresponding gait parameters, reverting the encoding scheme shown in Fig.4. A four-layer transformer decoder D_T is employed for text decoding. In line with recent developments in text-only decoder pre-training [42], we train D_T using the prefix language modeling. Starting with the numerical text embedding F^{num}, D_T learns to reconstruct the sequence of token IDs $\{tok_j\}$ which is subsequently mapped into natural language words. During training, to obtain ground-truth token sequences for NTE, we use the $FCLIP_T$ dictionary for words, and expand this dictionary with N_{num} additional tokens to accommodate numerical values. More specifically, we *scale* the normalized numbers to a graduated integer scale of [0,N_{num}]. The token ID tok of a number [num] is defined as: tok = [EOS]+*scale*([num]), where [EOS] = 49407. Moreover, we use an ordinal cross-entropy loss in addition to the vanilla cross-entropy loss, to further penalize the reconstruction error

TABLE II: Average per-fold sequence counts for MDS-UPDRS gait scores and diagnostic groups in our 10-fold cross-validation.

(a) MDS-UPDRS Gait Score

Class name	normal	slight	mild	moderate
No. sequences	202.4	373.5	158.6	167.9

(b) Diagnostic Groups

Class name	healthy	early DLB	early AD	severe DLB	severe AD
No. sequences	98.9	159.7	306.9	156.6	180.3

of the number values:

$$L_{num} = -\frac{|\hat{tok} - tok|}{[\text{EOS}] + N_{num} - 1} \sum_{m=1}^{[\text{EOS}]+N_{num}} y_m log(p_m), \quad (8)$$

where $N_{num} = 200$, $|\hat{tok} - tok|$ represents the absolute distance between the ground-truth token ID tok and the estimated $\hat{tok}$, 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: $\{\hat{Desc}_i\} = \mathbf{D}^T(\{\hat{F}_i^{num}\})$.

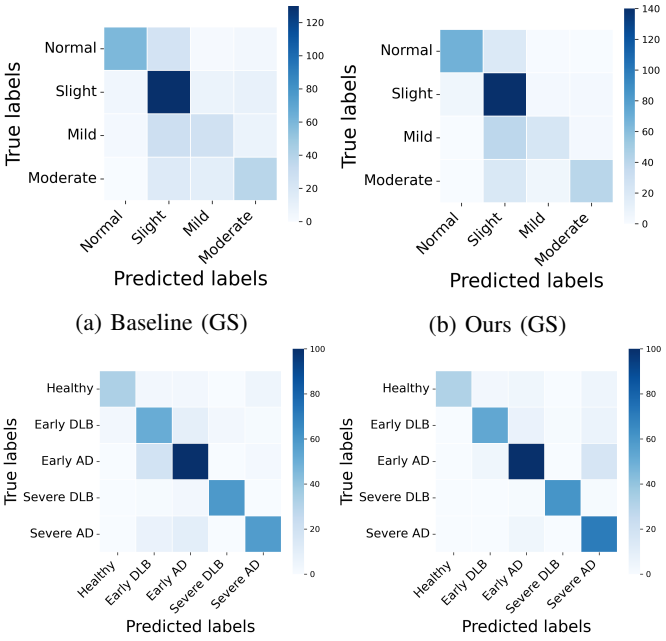
IV. Experiments and Results

Our study includes two classification tests: *Gait scoring (GS)* 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 (part III) [43], and *diagnostic group classification (DG)* to distinguish between different dementia groups and the corresponding phases: normal/ early DLB (Dementia with Lewy Bodies)/ severe DLB/ early AD (Alzheimer's Disease)/ severe AD. 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.

A. Ablation studies

We conduct ablation experiments on different model configurations, with different designs of knowledge-augmented prompts and the integration of NTE. Among the variants of KAPT, CPT model delivers better and more stable classification performance with less cross-fold variance. This performance improvement is especially noticeable in the GS test, where the inter-class relationships are ordinal. However, when incorporating texts as discrete prompts $\{D_i\}$ into the model, using keywords alone does not seem to be sufficient —KeyPT does not improve classification accuracy but helps stabilize model performance by reducing the standard deviation of cross-fold accuracy. This suggests that the keywords related to clinical knowledge provide only limited information, thereby hindering the transfer of knowledge from the pretrained VLM to our downstream tasks. As for the SegKPT, we examined the impact of varying segmentation granularity. Given the number of descriptive criteria ranges from 5 to 7 (Table V), we tried two segmentations: 1) Split the initial description into segments based on each criterion ($N_s > 5$); and 2) Form 5 segments per class ($N_s = 5$) by grouping 2 or 3 criteria together, maintaining an equivalent text length across segments. Note that each criterion can be used in up to three segments. Table III demonstrates that grouping multiple criteria into a segment yielded higher performance compared to per-criterion segmentation. In Fig. 8, we further analyze the precision rates associated with each criterion for the GS classification task. The precision rate of each criterion represents the percentage of true positives for its associated texture feature F^T achieving maximum similarity with the video feature F^V. We observe a significant percentage imbalance across different classes, suggesting that certain criteria do not contribute to the classification tasks. Utilizing these less effective criteria in the discrete prompts affect the calculation of the centroid (Sec.III-B), thereby degrading the estimation of per-class text features $\{F_i^T\}$.

The combination of NTE and SegKPT optimizes performance better in the DG test compared to the GS test, which can be attributed to the frequent occurrence of comparative adjectives in the descriptions of DG class categories (Table V). Overall, the models tend to perform better in DG test, supposedly due to the more distinctive per-class descriptions and more objective ground truth labeling in that classification. Confusion matrices are provided in Fig.7.



(a) Baseline (GS) (b) Ours (GS)

(c) Baseline (DG) (d) Ours (DG)

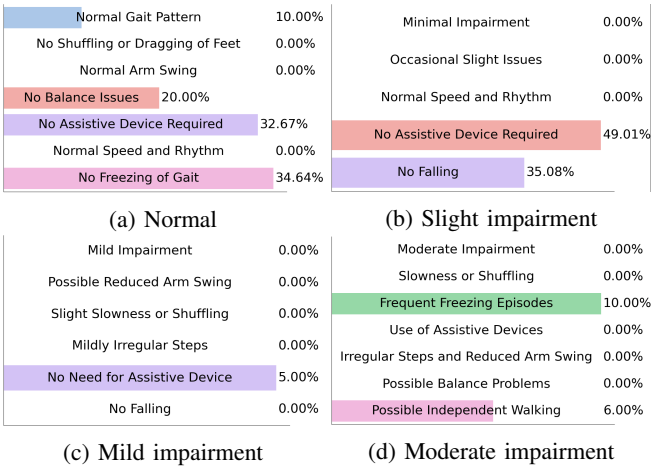
Fig. 7: Confusion matrices for the classification tasks.

B. Comparison with state-of-the-art

We compare our model with several related state-of-the-art (SOTA) models. Four of these models ([6]–[8], [24])

TABLE III: Comparative analysis on model configurations. Model performance is evaluated using top-1 accuracy (%), F1-score, weighted F1-score ("w.F1"), and cross-fold accuracy standard deviation ("cf.std"). Best performances are highlighted in **bold**.

Method	Gait scoring			
	Accuracy	F1-score	w.F1	cf.std
Baseline	67.59	0.636	0.668	12.39
+CPT	70.59	0.673	0.692	11.90
+KeyPT	69.77	0.668	0.697	**10.51**
+SegKPT ($N_S = 5$)	73.05	0.654	0.691	12.92
+SegKPT ($N_S > 5$)	66.47	0.602	0.639	13.03
+NTE	71.55	**0.705**	0.716	14.57
+SegKPT+NTE ($N_S = 5$)	**73.27**	0.693	**0.716**	13.08
	Diagnostic groups classification			
	Accuracy	F1-score	w.F1	cf.std
Baseline	80.94	0.820	0.819	11.81
+CPT	81.62	0.816	0.812	9.91
+KeyPT	80.40	0.809	0.794	**8.79**
+SegKPT ($N_S = 5$)	83.00	0.843	0.835	10.33
+SegKPT ($N_S > 5$)	81.41	0.825	0.824	11.20
+NTE	84.33	0.850	0.847	10.68
+SegKPT+NTE ($N_S = 5$)	**85.72**	**0.863**	**0.860**	11.56



(a) Normal (b) Slight impairment

(c) Mild impairment (d) Moderate impairment

Fig. 8: We visualize the classification precision for each class of SegKPT ($N_s > 5$) model, based on the categorical criteria (see Table V), for the gait scoring test.

are specifically designed for the classification of Parkinsonism severity on 3D skeletons. As shown in Table IV, our method achieves the best overall results in both tasks with comparable computational complexity (30.2M trainable parameters compared to $48.3textM + 0.6$M of the VIBE and OF-DDNet combination). This is somewhat expected, as other models are not designed for, and do not adapt well to, the constraints of limited data size. Note that the chosen SOTA methods operate on 3D reconstructed poses, for which we used VIBE [44], MAX-GRNet [8] or PoseFormerV2 [45]. Since PoseFormerV2 is designed to estimate 3D joint positions from sequential 2D joint positions, we used the 2D joint positions projected from the MAX-GRNet 3D outputs. As shown in Table IV, hanging among these algorithms did not result in meaningful differences in performance.

TABLE IV: Comparison with state-of-the-art methods. Model performance is evaluated using top-1 accuracy ("Acc.", %) and F1-score. The size and the complexity of each model is measured by the total number of trainable parameters ("#Params"). For classifiers, #Params are those specifically used in the gait score classification task. Best performances are in **bold**.

Method		GS		DG	
Reconstructor *(#Params)*	Classifier *(#Params)*	Acc.	F1-score	Acc.	F1-score
VIBE [44] *(48.3M)*		54.73	0.486	56.35	0.517
MAX-GRNet [8] *(32.9M)*	OF-DDNet [7] *(567.7K)*	52.48	0.439	56.78	0.582
PoseFormerV2 [45] *(14.4M)*		54.53	0.496	48.93	0.439
VIBE [44]		46.18	0.386	60.79	0.334
MAX-GRNet [8]	ST-GCN [6]	49.08	0.439	59.69	0.342
PoseFormerV2 [45]	*(162.8K)*	49.00	0.433	43.70	0.390
VIBE [44]		53.69	0.449	50.41	0.462
MAX-GRNet [8]	KShapeNet [46]	50.52	0.395	45.18	0.381
PoseFormerV2 [45]	*(835.9K)*	57.99	0.445	42.51	0.365
VIBE [44]		42.82	0.426	34.54	0.360
MAX-GRNet [8]	GaitForeMer [24]	40.65	0.418	40.67	0.398
PoseFormerV2 [45]	*(5.967M)*	37.76	0.374	35.20	0.349
GaVA-CLIP *(30.2M)*		**73.27**	**0.693**	**85.72**	**0.863**

Normal -- 0
Number of steps per minute is 107.1, difference in distance covered between a right step and a left step is -0.033 leg, standard variance among step times is 0.029, **time taken by one step is 0.805 s**.

Slight -- 1
Number of steps per minute is 95.268, difference in distance covered between a right step and a left step is -0.033 leg, standard variance among step times is 0.029, time taken by one step is 1.279 s.

Mild -- 2
Number of steps per minute is 59.076, difference in distance covered between a right step and a left step is -0.033 leg, standard variance among step times is 0.029, time taken by one step is 1.279 s.

Moderate -- 3
Number of steps per minute is 59.076, difference in distance covered between a right step and a left step is -0.033 leg, standard variance among step times is 0.029, time taken by one step is 1.279 s.

Fig. 9: MDS-UPDRS gait score descriptions generated from per-class text features through the pretrained text decoder. Key criteria are highlighted in the respective class color.

C. Decoding the per-class description

We apply the pretrained text decoder D_T (Sec.III-E) on the per-class text features $\{F_i^T\}$ obtained through the cross-modal contrastive learning in Sec.III-C. Examples of the decoded texts for $\{F_i^T\}$ leanred through the gait scoring task are shown in Fig.9. In general, the distinctive criteria in the decoded texts follow the ordinal relationship within the classes. Certain criteria in the clinical knowledge have been mapped to specific quantitative gait parameters, such as the slowness in *mild* and *moderate* impairments. However, the decoded texts for *moderate* impairment are identical to those for *mild*. This issue can be attributed to the limited availability of moderate samples during training, as shown in Fig. 6. Additionally, using only 8 basic gait parameters, as outlined in Table I, can hinder the effective decoding of nuanced differences.

V. Conclusion and Discussion

We presented a knowledge augmentation strategy to enhance the adaptability of a large-scale pre-trained Vision-

TABLE V: Clinical knowledge descriptions for gait impairment based on MDS-UPDRS [43], and for different diagnostic groups. The keywords (criteria) in the descriptions are colored in *blue*.

Category	Clinical knowledge descriptions
Normal	*Normal Gait Pattern:* The individual walks with a normal gait pattern, which includes regular, rhythmic steps with a typical step length and height. *No Shuffling or Dragging of Feet:* There is no shuffling or dragging of feet while walking. *Normal Arm Swing:* ... *No Balance Issues:* ... *No Assistive Device Required:* ... *Normal Speed and Rhythm:* ... *No Freezing of Gait:* ...
Slight impairment	*Minimal Impairment:* It's less severe than mild gait impairment. The individual's walking is almost normal. Any gait abnormalities are very subtle and may not be consistently present. *Occasional Slight Issues:* There might be occasional problems with gait, such as a slight drag of one foot or a minimal reduction in arm swing, but these are not consistently observable. *Normal Speed and Rhythm:* ... *No Assistive Device Required:* ... *No Falling:* ...
Mild impairment	*Mild Impairment:* The impairment in walking is noticeable but not severe. It's less severe than moderate gait impairment but more severe than slight gait impairment. The person can walk without assistance, but gait abnormalities are apparent. *Possible Reduced Arm Swing:* One or both arms may not swing normally while walking. There might be a reduced arm swing on one side or both sides. *Slight Slowness or Shuffling:* ... *Mildly Irregular Steps:* ... *No Need for Assistive Device:* ... *No Falling:* ...
Moderate impairment	*Moderate Impairment:* It's more severe than mild gait impairment. The individual's gait is noticeably impaired, and these impairments are consistent and evident. *Marked Slowness or Shuffling:* The person may walk with a marked slowness. The shuffling quality of the gait can be more pronounced, with reduced step height and length. *Frequent Freezing Episodes:*... *Use of Assistive Devices:* ... *Irregular Steps and Reduced Arm Swing:* ... *Possible Balance Problems:* ... *Independent Walking May Still Be Possible:* ...
Healthy	*General Stable Gait Patterns:* Generally stable, with only minor changes compared to younger adults, ensuring consistent stride lengths and minimal sway. *Longer Stride Length than in DLB and AD:* While there may be a slight decrease compared to younger adults, healthy elderly tend to have longer stride lengths than those with DLB and AD. *Regular and Consistent Cadence:* Healthy people maintain a regular and consistent cadence. *Faster and More Consistent Speed than in AD and DLB:* ... *Even Weight Distribution and Movement:* ... *Consistent Rhythm:* ... *Arm Swing Naturally Synchronized with Leg Movements:* ...
Early DLB	*More Noticeable Gait Changes than in Early AD:* Gait changes in early DLB are subtle and may be easily overlooked, but the alterations are more noticeable than those in early AD. *Slight Speed Reduction:* There may be a slight reduction in walking speed, which impacts the overall fluidity and pace of gait. *Minor Balance Issues:* Minor issues with balance are present, which can affect stability and confidence in movement. *Less Fluidity than Normal:* ... *Occasional Hesitations in Initiating Movement:* ... *Slightly Reduced Arm Swing:* ... *Less Severe Mobility Impairment than Severe DLB:* ...
Early AD	*Less Pronounced Gait Changes than in Early DLB:* Gait changes in early AD are generally subtle and may not be readily apparent, especially when compared to early DLB. *Slight Speed Reduction:* There is a slight reduction in walking speed, which subtly influences the overall pace. *Minor Decrease in Fluidity:* A minor decrease in the fluidity of movement contributes to a less smooth walking experience. *Mild Balance Problems in Complex Conditions:* ... *Less Pronounced Changes Compared with early DLB:* ... *Less Severe Gait Impairment than Severe AD:* ...
Severe DLB	*More Severe than Early DLB:* Gait impairment in severe DLB is considerably more severe compared to early DLB with greater problems in mobility and stability. *Shuffling Gait:* Individuals might exhibit a shuffling gait, characterized by taking small steps and having difficulty lifting their feet off the ground. *En Bloc Turning:* Turning might involve a series of small steps, sometimes referred to as "en bloc" turning, instead of a fluid motion. *Significant Balance Issues:* ... *Freezing Movement or Frequent Hesitation:* ...
Severe AD	*More Severe than Early AD:* Gait impairment in severe AD is considerably more severe compared to early AD with greater problems in mobility and stability. *Greatly Reduced Speed and Irregular steps:* Walking speed is greatly reduced, with steps becoming irregular and uncoordinated, contributing to difficulty in maintaining a steady pace. *Significant Balance Issue:* The Balance is heavily compromised, elevating the risk of falls and requiring constant vigilance and support. *Loss the Ability of Independent Walking:* ... *Fewer freezing episodes Compared with DLB:* ... *Profound Mobility Impairment and caregiver dependence:* ...

Language Model for video-based gait analysis in neurodegenerative diseases. 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. Second, we incorporate the associated numerical gait parameters as numerical texts to enhance the numeracy within the latent space of the text, addressing the challenge of data scarcity in the medical domain. Notably, on two video-based gait classifications tasks, our model significantly outperformed other strong SOTA methods, given only slightly more than 100 videos, and led to representations with higher quality. Our work demonstrates how to efficiently enhance multimodal representation learning and introduces a novel approach for incorporating metadata, especially in tabular form, which is common in the medical domain.

Our proposed knowledge augmentation method introduces additional prompts and a separate modality to the CLIP model without altering its original architecture, demonstrating generalizability and compatibility with different pretrained VLMs. However, current experiments reveal the sensitivity of GaVA-CLIP to integrated domain knowledge, suggesting a need for further improvements in distillation and aggregation of domain knowledge.

The promising results in this paper, combined with the advantages of video-based analysis, such as being sensor-free and widely accessible, pave the way for future developments in pathological gait analysis. Moreover, our method currently relies on limited public datasets of elderly gait videos, highlighting the need for further community contributions to expand data resources. Such efforts could significantly benefit practical applications, including monitoring disease progression, early detection of neurodegenerative diseases, and improving care for elderly subjects in residential settings.

ACKNOWLEDGEMENT

This work has been funded by the French national project "ArtIC" (Artificial Intelligence for Care, ANR-20-THIA-0006) and the binational project "Synthetic Data Generation and Sim-to-Real Adaptive Learning for Real-World Human Daily Activity Recognition of Human-Care Robots (21YS2900)"

granted by the ETRI, South Korea. Kun Yuan has been supported by the EU-funded ERC projet CompSURG (European Union, 101088553).

We would like to thank Dr. Candice Muller and Prof. Dr. Frédéric Blanc at the Robsertsau hospital for sharing patient data and their valuable expertise. We extend our gratitude to Dr. Duc-Hoan Nguyen from the Johann Radon Institute for Computational and Applied Mathematics for our discussions on strategizing the use of segmented knowledge.

REFERENCES

[1] R. Mc Ardle, B. Galna, P. Donaghy, A. Thomas, and L. Rochester, "Do alzheimer's and lewy body disease have discrete pathological signatures of gait?" *Alzheimer's & Dementia*, vol. 15, no. 10, pp. 1367–1377, 2019.

[2] J. Merory, J. Wittwer, C. Rowe, and K. Webster, "Quantitative gait analysis in patients with dementia with lewy bodies and alzheimer's disease," *Gait & posture*, vol. 26, pp. 414–9, 10 2007.

[3] R. Mc Ardle, S. Del Din, P. Donaghy, B. Galna, A. J. Thomas, and L. Rochester, "The impact of environment on gait assessment: considerations from real-world gait analysis in dementia subtypes," *Sensors*, vol. 21, no. 3, p. 813, 2021.

[4] C. Muller, J. Perisse, F. Blanc, M. Kiesmann, C. Astier, and T. Vogel, "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*, vol. 174, pp. S2–S3, 2018.

[5] P. Albuquerque, T. T. Verlekar, P. L. Correia, and L. D. Soares, "A spatiotemporal deep learning approach for automatic pathological gait classification," *Sensors*, vol. 21, no. 18, p. 6202, 2021.

[6] A. Sabo, S. Mehdizadeh, A. Iaboni, and B. Taati, "Estimating parkinsonism severity in natural gait videos of older adults with dementia," *IEEE journal of biomedical and health informatics*, vol. 26, no. 5, pp. 2288–2298, 2022.

[7] M. Lu, K. Poston, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, J. C. Niebles, and E. Adeli, "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*. Springer, 2020, pp. 637–647.

[8] D. Wang, C. Zouaoui, J. Jang, H. Drira, and H. Seo, "Video-based gait analysis for assessing alzheimer's disease and dementia with lewy bodies," in *Applications of Medical Artificial Intelligence*, S. Wu, B. Shabestari, and L. Xing, Eds. Cham: Springer Nature Switzerland, 2024, pp. 72–82.

[9] V. Adeli, S. Mehraban, I. Ballester, Y. Zarghami, A. Sabo, A. Iaboni, and B. Taati, "Benchmarking skeleton-based motion encoder models for clinical applications: Estimating parkinson's disease severity in walking sequences." [Online]. Available: http://arxiv.org/abs/2405.17817

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[11] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9879–9889.

[12] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[13] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "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*, 2021, pp. 3942–3951.

[14] Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical image understanding with pretrained vision language models: A comprehensive study," *arXiv preprint arXiv:2209.15517*, 2022.

[15] B. Kan, T. Wang, W. Lu, X. Zhen, W. Guan, and F. Zheng, "Knowledge-aware prompt tuning for generalizable vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 670–15 680.

[16] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887.

[17] K. Yuan, V. Srivastav, T. Yu, J. Lavanchy, P. Mascagni, N. Navab, and N. Padoy, "Learning multi-modal representations by watching hundreds of surgical video lectures," *arXiv preprint arXiv:2307.15220*, 2023.

[18] D. Wang, K. Yuan, C. Muller, F. Blanc, N. Padoy, and H. Seo, "Enhancing gait video analysis in neurodegenerative diseases by knowledge augmentation in vision language model," *arXiv preprint arXiv:2403.13756*, 2024.

[19] I. G. McKeith, T. J. Ferman, A. J. Thomas, F. Blanc, B. F. Boeve, H. Fujishiro, K. Kantarci, C. Muscio, J. T. O'Brien, R. B. Postuma *et al.*, "Research criteria for the diagnosis of prodromal dementia with lewy bodies," *Neurology*, vol. 94, no. 17, pp. 743–755, 2020.

[20] A. Mannini, D. Trojaniello, A. Cereatti, and A. M. Sabatini, "A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington's disease patients," *Sensors*, vol. 16, no. 1, p. 134, 2016.

[21] W.-C. Hsu, T. Sugiarto, Y.-J. Lin, F.-C. Yang, Z.-Y. Lin, C.-T. Sun, C.-L. Hsu, and K.-N. Chou, "Multiple-wearable-sensor-based gait classification and analysis in patients with neurological disorders," *Sensors*, vol. 18, no. 10, p. 3397, 2018.

[22] D. Von Winterfeldt and W. Edwards, *Decision Analysis and Behavioral Research*. Cambridge University Press, 1986.

[23] V. Adeli, S. Mehraban, Y. Zarghami, I. Ballester, A. Sabo, A. Iaboni, and B. Taati, "Benchmarking skeleton-based motion encoder models for clinical applications: Estimating parkinson's disease severity in walking sequences," *arXiv preprint arXiv:2405.17817*, 2024.

[24] M. Endo, K. L. Poston, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, and E. Adeli, "Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 130–139.

[25] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[26] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle *et al.*, "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[27] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training in radiology," *arXiv preprint arXiv:2301.02228*, 2023.

[28] Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical image understanding with pretrained vision language models: A comprehensive study," *arXiv preprint arXiv:2209.15517*, 2022.

[29] S. Menon and C. Vondrick, "Visual classification via description from large language models," *arXiv preprint arXiv:2210.07183*, 2022.

[30] J. Silva-Rodriguez, H. Chakor, R. Kobbi, J. Dolz, and I. B. Ayed, "A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision," *arXiv preprint arXiv:2308.07898*, 2023.

[31] Z. Chen, Y. Zhou, A. Tran, J. Zhao, L. Wan, G. S. K. Ooi, L. T.-E. Cheng, C. H. Thng, X. Xu, Y. Liu *et al.*, "Medical phrase grounding with region-phrase context contrastive alignment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 371–381.

[32] Y. Kumar and P. Marttinen, "Improving medical multi-modal contrastive learning with expert annotations," *arXiv preprint arXiv:2403.10153*, 2024.

[33] S. Mehdizadeh, H. Nabavi, A. Sabo, T. Arora, A. Iaboni, and B. Taati, "The toronto older adults gait archive: video and 3d inertial motion capture data of older adults' walking," *Scientific data*, vol. 9, no. 1, p. 398, 2022.

[34] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[35] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[37] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "Kepler: A unified model for knowledge embedding and pre-trained language representation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.

[38] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23 034–23 044.

[39] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

[42] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding clip latents for zero-shot captioning via text-only training," *arXiv preprint arXiv:2303.03032*, 2023.

[43] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel *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*, vol. 23, no. 15, pp. 2129–2170, 2008.

[44] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.

[45] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8877–8886.

[46] R. Friji, H. Drira, F. Chaieb, H. Kchok, and S. Kurtek, "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*, 2021, pp. 12 611–12 620.