Benchmarking Video Foundation Models for Remote Parkinson's Disease Screening

Md Saiful Islam[1*][0000−0003−3725−3493], Ekram Hossain[1], Abdelrahman Abdelkader[1], Tariq Adnan[1], Fazla Rabbi Mashrur[1], Sooyong Park[1], Praveen Kumar[1], Qasim Sudais[1], Natalia Chunga[2], Nami Shah[3], Christopher Kanan[1], Jan Freyberg[4], Ruth B. Schneider[3], and Ehsan Hoque[1†][0000−0003−4781−4733]

[1] University of Rochester, Rochester, NY, USA
[2] Louisiana State University Health Sciences Center at Shreveport, LA, USA
[3] University of Rochester Medical Center, Rochester, NY, USA
[4] University of Cambridge, Cambridge, UK
*mislam6@ur.rochester.edu, †mehoque@cs.rochester.edu

Abstract. Video-based assessments offer a scalable pathway for remote Parkinson's disease (PD) screening. While traditional approaches rely on handcrafted features mimicking clinical scales, recent advances in video foundation models (VFMs) enable representation learning without task-specific customization. However, the comparative effectiveness of different VFM architectures across diverse clinical tasks remains poorly understood. We present a large-scale systematic study using a novel video dataset from 1,888 participants (727 with PD), comprising 32,847 videos across 16 standardized clinical tasks. We evaluate seven state-of-the-art VFMs – including VideoPrism, V-JEPA, ViViT, and VideoMAE – to determine their robustness in clinical screening. By evaluating frozen embeddings with a linear classification head, we demonstrate that task saliency is highly model-dependent: VideoPrism excels in capturing visual speech kinematics (no audio) and facial expressivity, while V-JEPA proves superior for upper-limb motor tasks. Notably, TimeSformer remains highly competitive for rhythmic tasks like finger tapping. Our experiments yield AUCs of 76.4–85.3% and accuracies of 71.5–80.6%. While high specificity (up to 90.3%) suggests strong potential for ruling out healthy individuals, the lower sensitivity (43.2–57.3%) highlights the need for task-aware calibration and integration of multiple tasks and modalities. Overall, this work establishes a rigorous baseline for VFM-based PD screening and provides a roadmap for selecting suitable tasks and architectures in remote neurological monitoring. Code and anonymized structured data are publicly available:
https://github.com/saiful1105020/park_video_benchmarking

Keywords: Remote screening · Video foundation models · Parkinson's.

1 Introduction

The sharply rising prevalence of Parkinson's disease (PD) [9] is creating an urgent need for scalable and accessible assessment tools. Traditionally, the diagnosis

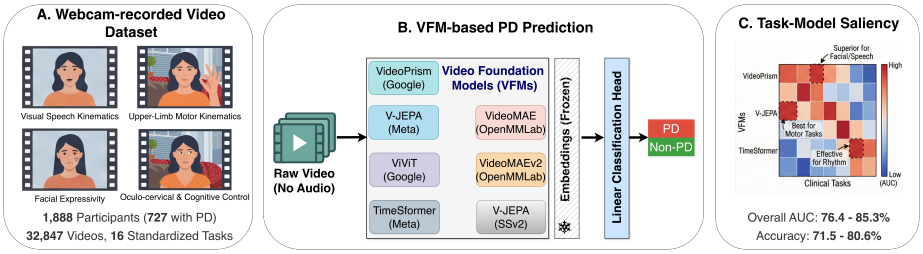


Fig. 1. Overview of the benchmarking framework for PD screening using video foundation models (VFMs). (A) Video Dataset: Our study utilizes a large-scale dataset from 1,888 participants performing 16 standardized clinical tasks. **(B) Evaluation Pipeline**: Raw video data is processed through a suite of state-of-the-art frozen VFMs to extract latent representations. These embeddings are evaluated using a task-specific linear classification head to differentiate between PD and non-PD participants. **(C) Task-Model Saliency**: Systematic evaluation to investigate architecture-specific strengths.

and severity assessment of PD rely on clinical evaluation and assessment of motor signs, including bradykinesia [5], rest tremor, and rigidity. The MDS-UPDRS [12], which serves as the gold standard for assessment of motor signs, requires a trained specialist to rate the performance of standard motor tasks. However, geographic and financial barriers to specialized care prevent timely diagnosis in underserved populations [24].

Video-based assessments offer a scalable route by allowing subjects to perform standardized clinical tasks remotely using everyday devices (e.g., smartphones [6] and webcams [19]) and obtain a preliminary PD risk screening. Computer-vision-based prior works focused on developing handcrafted features designed to mimic clinical observations (e.g., finger-tapping speed) and plugging them into machine learning models for PD screening [22] and severity assessment [16]. In addition, the recent evolution of deep learning has introduced more sophisticated architectures, such as 3D ConvNets (e.g., I3D [7]), which can learn spatiotemporal representations directly from raw videos. However, these task-specific models either cannot generalize across other tasks without custom-engineered features or require large amounts of labeled data.

The emergence of Video Foundation Models (VFMs) may enable a paradigm shift in medical video analysis. VFMs are large-scale vision transformers [10] pre-trained on massive, heterogeneous datasets using self-supervised learning. These models can learn versatile latent representations of visual dynamics, spanning fine-grained motion and high-level semantic interactions, without task-specific customization during pretraining. However, despite their broad capabilities in general action recognition [17], the comparative effectiveness of different VFM architectures across the specific clinical tasks used in PD screening remains poorly understood. These tasks vary in their requirements: some demand high-frequency temporal reasoning (e.g., finger tapping), while others require an understanding of subtle spatial deformations in facial expressivity or complex articulatory kine-

matics during speech. Our systematic benchmarking study addresses this gap by evaluating seven state-of-the-art VFMs – VideoPrism [30], V-JEPA2 [3] (two variants), ViViT [2], VideoMAE [28], VideoMAEv2 [29], and TimeSformer [4].

By curating the largest webcam-recorded video dataset (32,847 videos from 1,888 participants; 727 with PD), we evaluate the ability of frozen embeddings from these models to differentiate between PD and non-PD across 16 standardized tasks (Fig. 1). Our objective is to determine which architectural paradigms – such as the semantic-visual distillation of VideoPrism, the latent-predictive world modeling of V-JEPA, or the divided space-time attention of TimeSformer – are most salient for specific clinical tasks. By establishing a rigorous baseline for VFM-based PD screening, this work provides a roadmap for selecting suitable tasks and architectures for the next generation of remote video-based neurological monitoring systems.

2 Methods

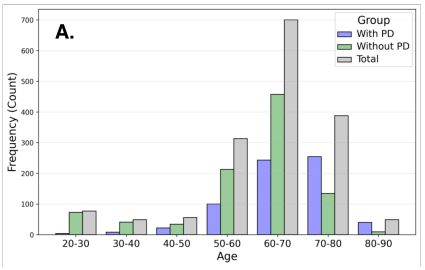
2.1 Dataset and Ethics

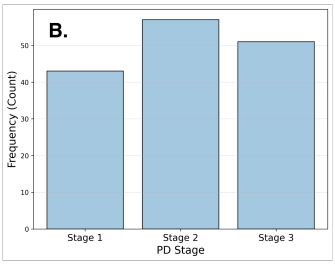
The dataset was compiled from eight independent clinical and non-clinical studies (2017–2025), each approved by an institutional review board. Participants were recruited via global and nationwide channels, including clinician referrals, social media, PD wellness centers, and research cohorts. We utilized a standardized web-based platform for recording tasks mimicking MDS-UPDRS assessments, performed either independently at home or under clinical supervision. PD status was determined through self-reporting (445) or clinical confirmation (282). Overall, the dataset comprises **32,847** total videos from **1,888** (991 female) individuals (see Fig. 2 for age and clinical stage details). The cohort is predominantly White (1,366) with 68 Asian, 77 Black, and 27 individuals from other groups, while 350 participants did not disclose their race.

2.2 Standardized Clinical Tasks and Domain Classification

The study evaluates performance across 16 standardized clinical tasks (see Fig. 2C.) inspired by the MDS-UPDRS [12] and other validated neurological assessments [19]. We have organized these tasks into four broad clinical domains based on the physiological signs they are designed to elicit. To optimize quality and consistency in remote data collection settings, a video instruction demonstrating each specific task was provided to participants (**supplementary materials**). The protocol and instruction set were co-designed with three MDS-certified neurologists.

Upper-Limb Motor Kinematics: This domain includes finger tapping, flip palm (hand pronation-supination), open fist (hand open-close), extending arms, and nose touching. These standard MDS-UPDRS items are utilized to assess bradykinesia [5] (slowness of movement), hypokinesia [27] (reduced amplitude), sequence effect (or decrement), and interruptions.





C.

Task	Number of Videos		
	PD	Non-PD	Total
Extend Arm	872	1,385	2,257
Eye Gaze	448	845	1,293
Facial Expression (Disgust)	581	1,196	1,777
Facial Expression (Smile)	981	1,306	2,287
Facial Expression (Surprise)	577	1,194	1,771
Finger Tapping	1,925	2,049	3,974
Flip Palm	1,009	1,700	2,709
Head Pose	444	844	1,288
Nose Touch	758	1,516	2,274
Open Fist	1,755	1,906	3,661
Pangram Utterance	994	1,322	2,316
Reverse Count	442	845	1,287
Sustained Phonation (A)	1,066	1,123	2,189
Sustained Phonation (E)	446	842	1,288
Sustained Phonation (O)	445	840	1,285
Tongue Twister	413	778	1,191
Grand Total	**13,156**	**19,691**	**32,847**

Fig. 2. Dataset characteristics. (A) The age distribution of participants categorized by PD status; (B) the distribution of clinical PD stage (Hoehn & Yahr scale) when available (n = 158); and (C) the number of videos collected for each task.

Visual Speech Kinematics: Tasks consist of uttering an English pangram (a sentence containing every letter of the alphabet), tongue twisters, and sustained phonation of vowels /a/, /e/, and /o/. While performed with audio, this study focuses on the **video-only** modality to capture articulatory movements of the jaw, lips, and tongue, which are frequently reduced or imprecise in PD.

Facial Expressivity: Tasks that require mimicking different facial expressions (smile, disgust, surprise) are used to evaluate hypomimia, characterized by a reduction in spontaneous and expressive facial movements [23].

Oculomotor, Cervical, and Cognitive Control: This domain assesses eye gaze (following an on-screen target), head pose (horizontal and vertical tilting), and reverse counting. These tasks evaluate a range of symptoms, including cervical dystonia [18], oculomotor control, and cognitive-motor interference – a task requiring mental flexibility that is often impaired in PD [21].

2.3 Video Foundation Model Architectures

We evaluate seven state-of-the-art VFMs spanning diverse architectures and pre-training objectives. Together, they provide a broad benchmark across temporal modeling (e.g., divided attention vs. factorized encoding) and supervision strategies (e.g., pixel reconstruction, latent prediction, and language alignment).

– **VideoPrism [30]:** Employs a two-stage pre-training strategy involving video-text contrastive learning on 36M pairs followed by masked video modeling

with global-local distillation. This hybrid approach allows the model to capture both high-level semantic language cues and fine-grained visual motion.
- **V-JEPA2 [3] and V-JEPA2 (SSv2):** Utilizes a **J**oint **E**mbedding **P**redictive **A**rchitecture that forecasts the latent representations of masked regions from unmasked context rather than reconstructing raw pixels. We include a version fine-tuned on the Something-Something v2 (SSv2 [13]) dataset to assess if motion-centric supervision enhances clinical motor task performance.
- **TimeSformer [4] and ViViT [2]:** These models utilize factorized attention mechanisms. TimeSformer uses "divided space-time attention" to sequentially process spatial and temporal features, while ViViT factorizes the spatial and temporal encoders to handle complexity of video transformers.
- **VideoMAE [28] and VideoMAEv2 [29]:** These models use masked autoencoding to reconstruct the original pixels of highly masked (up to 90%) video frames. VideoMAEv2 further scales this approach using dual masking to efficiently learn detailed visual structures across massive datasets.

2.4 Experimental Protocol

To evaluate foundation models for PD risk screening, we use a "frozen-backbone" protocol: VFM weights remain fixed after pretraining and are used only as feature extractors. We resize input videos and sample frames at 15 FPS, following model-specific requirements, to obtain spatiotemporal embeddings. **Pretrained model weights and implementation are available in our public repository.**

A neural network featuring one hidden layer (ReLU activation, random dropout) and a linear classification head is trained on these embeddings to minimize binary cross-entropy loss against ground-truth PD/non-PD labels. The hidden layer dimension is treated as a tunable hyperparameter. Data is randomly partitioned into training (60%), validation (20%), and test (20%) folds **based on unique participants** to ensure **evaluation on unseen subjects**. We optimize hyperparameters, including learning rate, hidden nodes, and dropout, using Weights & Biases, allocating equal compute time across all tasks and VFMs (script for hyperparameter search is provided in our repository). The final model for each task is selected based on peak validation set area under the receiver operating characteristic curve (AUC). When feasible, results are reported as the mean and 95% confidence interval (CI) obtained across 30 different random seeds.

Additionally, we explore multi-view training, where a secondary linear head aggregates likelihoods from individual views, and assess oversampling strategies (e.g., RandomOverSampler [20], weighted loss) to address class imbalance. We avoided undersampling to maintain the diversity of the non-PD class, aiming to provide the linear classifier with the broadest possible distribution of healthy movement patterns. All experiments are conducted locally on a 32-core AMD Ryzen Threadripper PRO 5975WX workstation equipped with dual NVIDIA RTX A6000 (each with 48 GB vRAM) and 256 GB RAM. Local deployment guarantees that **sensitive patient videos are processed on-premises** without exposure to cloud or third-party servers.

3 Results and Discussions

3.1 Task-wise Saliency Patterns and Clinical Implications

Performance results across the 16 tasks (see Table 1) indicate that upper-limb motor maneuvers exhibit the highest saliency. Specifically, `Flip Palm` and `Open Fist` yielded the highest AUCs (95% CI: $85.3 \pm 0.2\%$ and $84.3 \pm 0.1\%$), with `Flip Palm` achieving a peak specificity of $90.3 \pm 0.5\%$. These findings align with clinical literature identifying limb bradykinesia as assessed with rapid alternating movements as a high-yield diagnostic markers during clinical assessments [26].

Table 1. Performance of Best Models across standardized tasks. Results are reported for the model best performing on the validation set (AUC) on each task. Bold denotes the overall best result, and ± indicates 95% confidence intervals.

Task	Best Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Extend Arm	V-JEPA2-SSv2	83.2 ± 0.1	78.2 ± 0.3	55.8 ± 1.4	87.8 ± 0.6	66.4 ± 0.8	82.3 ± 0.4
Eye Gaze	VideoPrism	81.0 ± 0.1	75.2 ± 0.3	$\mathbf{57.3 \pm 1.1}$	83.9 ± 0.8	63.6 ± 0.9	80.1 ± 0.3
Facial Expression Disgust	V-JEPA2	79.2 ± 0.1	78.7 ± 0.3	$\mathbf{57.3 \pm 0.8}$	87.8 ± 0.4	66.7 ± 0.6	82.9 ± 0.2
Facial Expression Smile	VideoPrism	80.3 ± 0.3	74.3 ± 0.3	51.4 ± 2.2	84.1 ± 1.1	58.3 ± 0.9	80.3 ± 0.5
Facial Expression Surprise	V-JEPA2-SSv2	80.2 ± 0.1	77.4 ± 0.4	55.3 ± 1.4	86.6 ± 0.8	63.5 ± 1.0	82.3 ± 0.4
Finger Tapping	TimeSformer	76.4 ± 0.3	73.0 ± 0.4	43.2 ± 1.2	85.8 ± 0.9	57.0 ± 1.1	77.8 ± 0.2
Flip Palm	V-JEPA2-SSv2	$\mathbf{85.3 \pm 0.2}$	$\mathbf{80.6 \pm 0.3}$	56.2 ± 0.9	$\mathbf{90.3 \pm 0.5}$	$\mathbf{69.7 \pm 0.9}$	$\mathbf{83.9 \pm 0.2}$
Head Pose	VideoPrism	79.3 ± 0.1	71.7 ± 0.4	52.5 ± 1.1	81.0 ± 1.0	57.3 ± 0.9	77.9 ± 0.2
Nose Touch	V-JEPA2-SSv2	83.0 ± 0.1	79.3 ± 0.2	54.2 ± 1.0	89.3 ± 0.3	66.8 ± 0.5	83.1 ± 0.3
Open Fist	VideoPrism	84.3 ± 0.1	76.8 ± 0.3	54.6 ± 1.9	85.7 ± 0.5	60.5 ± 0.4	82.5 ± 0.5
Pangram Utterance	VideoPrism	81.3 ± 0.2	75.5 ± 0.4	50.1 ± 1.0	86.8 ± 0.9	63.2 ± 1.4	79.6 ± 0.2
Reverse Count	VideoPrism	80.0 ± 0.1	73.2 ± 0.3	52.8 ± 0.9	83.0 ± 0.5	60.0 ± 0.6	78.5 ± 0.3
Sustained Phonation A	VideoPrism	79.2 ± 0.2	71.5 ± 0.5	54.8 ± 2.9	79.7 ± 1.5	57.2 ± 1.2	78.5 ± 0.9
Sustained Phonation E	VideoPrism	80.7 ± 0.1	73.1 ± 0.3	51.9 ± 1.0	83.4 ± 0.6	60.5 ± 0.6	78.0 ± 0.3
Sustained Phonation O	VideoPrism	81.0 ± 0.1	73.4 ± 0.3	56.3 ± 1.2	81.7 ± 0.5	60.1 ± 0.5	79.3 ± 0.4
Tongue Twister	VideoPrism	77.1 ± 0.2	73.4 ± 0.3	54.0 ± 1.2	82.4 ± 0.7	58.9 ± 0.6	79.4 ± 0.4

Conversely, fine-motor tasks like `Finger Tapping` and vocal tasks like `Tongue Twister` showed lower saliency (AUCs < 78). This likely reflects the difficulty of capturing subtle movements or dysarthric shifts [8] via general video-based embeddings (we did not use the audio). While models showed relatively higher sensitivity in facial tasks (i.e., `Eye Gaze` and `Disgust`), overall performance suggests that self-supervised models require enhanced temporal resolution or localized attention to match expert clinical observation. In addition, domain-specific feature extraction [1] or advanced architecture design may be necessary to outperform the VFM baselines established here. Finally, the discrepancy between high AUC and low sensitivity suggests non-optimal decision thresholds. Clinical utility in screening will require improved model calibration (i.e., post-hoc scaling of predicted probabilities [25]). Prior literature also suggests that integrating multiple tasks and modalities can significantly improve performance [15].

3.2 Architecture-Specific Strengths Across Clinical Domains

The comparative evaluation of VFM architectures across diverse domains highlights specific structural advantages for PD screening. **VideoPrism** demon-

strated strong generalization, ranking first across 10 of 16 tasks, while providing competitive performance in other tasks. Its superiority is most pronounced in `Visual Speech Kinematics` and `Facial Expressivity` domains, where the model needed to capture the subtle, low-amplitude spatio-temporal features associated with hypomimia and dysarthria. In addition, given its robust cross-domain performance, we recommend VideoPrism as the primary baseline for general-purpose PD screening or research focused on orofacial manifestations [11].

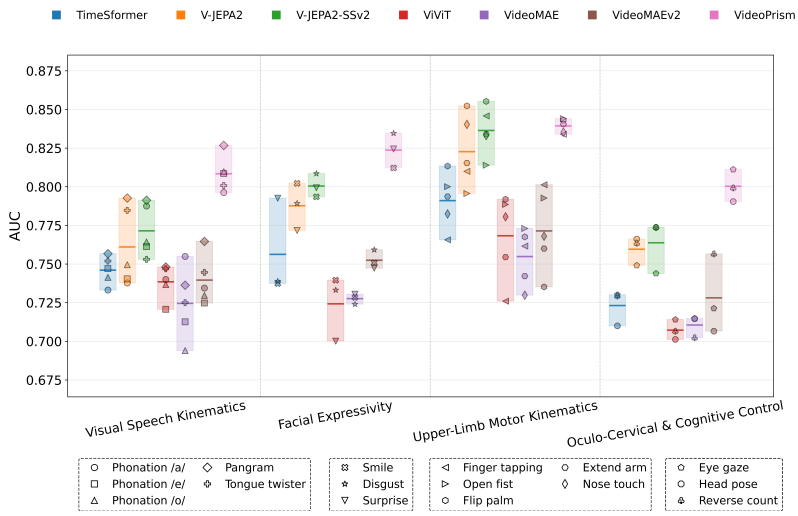


Fig. 3. Comparative performance of VFMs across broad clinical domains. This plot organizes results by broad clinical category along the x-axis, with color-coded boxes showing the performance distribution of each VFM within that domain (horizontal lines showing mean). Individual data points, distinguished by unique marker shapes (defined in the bottom legend), represent the specific AUC achieved by a model on a distinct clinical task. High-level analysis reveals that the "Upper-Limb Motor Kinematics" domain generally yields the highest predictive performance. At the task level, distinct model specializations emerge: VideoPrism shows superior consistency across "Facial Expressivity," "Visual Speech Kinematics," and "Oculo-Cervical & Cognitive Control". In contrast, V-JEPA models dominate in "Upper-Limb Motor Kinematics," achieving the highest overall scores on tasks like flip-palm. V-JEPA-SSv2 generally outperforms V-JEPA, suggesting additional benefits of training on the SSv2 dataset.

For `Upper-Limb Motor Kinematics`, specifically tasks involving high-amplitude movements like `Flip Palm`, the **V-JEPA2** family proved most effective. Notably, V-JEPA2-SSv2 consistently outperformed the base V-JEPA2 model in tasks requiring precise limb coordination and trajectory tracking, such as `Extend Arm` and `Nose Touch`. This performance gap suggests that the additional fine-tuning on the Something-Something v2 (SSv2) dataset – which is heavily oriented toward human-object interactions and directional motion – likely enhanced the

model's ability to represent complex temporal dynamics and motion causality. Consequently, V-JEPA2-SSv2 may serve as the optimal baseline for research targeting bradykinesia and motor coordination.

Ultimately, foundation model selection should depend on the clinical domain. Although VideoPrism provides the strongest baseline, VJEPA2-SSv2's motion-focused pretraining better captures rhythmic irregularities in gross motor symptoms. For fine-motor tasks such as `Finger Tapping`, where **TimeSformer** performed best, current VFMs may benefit from architectures that emphasize high-resolution temporal attention rather than broad semantic embeddings.

3.3 Multi-View and Oversampling Ablations

We evaluated multi-view aggregation and oversampling to handle videos longer than a single view supports and to mitigate class imbalance. The single-view configuration without oversampling generally provided robust results, achieving a mean accuracy of 74.97% (std: 3.12%) and a mean AUC of 80.54% (std: 2.42%) across all the 16 tasks. While multi-view training was expected to improve performance by capturing more temporal context, it did not significantly outperform single-view metrics (p-value$>$0.70 for both accuracy and AUC), suggesting that salient diagnostic features maybe often captured within short segments. An alternative explanation is we need more complex architecture to integrate multiple views, as they substantially increase feature dimension (which remains out of scope for this study). Furthermore, oversampling strategies were found to slightly (non-significant) degrade mean accuracy (74.51%; std = 2.99%) and AUC (79.83%; std = 2.82%), indicating that the frozen VFM embeddings are already robust enough to handle existing class imbalances, and artificial data expansion may introduce noise into the latent feature space.

3.4 Limitations

While this work establishes a rigorous baseline, several limitations remain. First, we instructed participants to record videos in the OFF state (i.e., when Parkinson's symptoms return between doses) to ensure medications did not mask symptoms. However, we did not enforce compliance due to safety considerations (i.e., medication withdrawal). Therefore, occasional ON state recordings are possible and may constitute label noise. Next, the "frozen evaluation" protocol, although ideal for benchmarking the inherent representational power of VFMs, does not explore the performance ceilings achievable through task-specific fine-tuning or parameter-efficient methods such as LoRA [14]. Moreover, to safeguard patient privacy and avoid sending sensitive videos to third-party services, we restricted experiments to open-weight models that can be deployed locally. As the VFM ecosystem evolves, future work should evaluate newer architectures as they become available. In addition, while self-reported PD diagnosis labels helped scale up the study, they may introduce some label noise. Finally, despite a diverse, open recruitment process, the cohort remains predominantly white, which may limit the generalizability of our findings to more diverse populations.

4 Conclusion

We introduce a large-scale benchmark of video foundation models for remote Parkinson's disease screening. Results suggest that frozen embeddings from VFMs contain information relevant to Parkinson's disease classification, even without task-specific fine-tuning. Across seven state-of-the-art VFMs and 16 standardized tasks, performance depended on both the physiological domain and model architecture. Motion-predictive models such as V-JEPA performed best on upper-limb motor tasks, whereas semantic-visual models such as VideoPrism showed advantages for facial and speech-related assessments. While the moderate sensitivity observed across models indicates that further advances are needed before deployment as standalone screening tools, this benchmark provides practical guidance for future research on video-based assessment of movement disorders.

Acknowledgments. The project was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award P50NS108676, Gordon and Betty Moore Foundation, and Google Faculty Research Award. The lead author is supported by a Google PhD fellowship.

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

References

1. Adnan, T., Islam, M.S., Lee, S., Wasifur Rahman Chowdhury, E., Tithi, S.D., Noshin, K., Islam, M.R., Sarker, I., Rahman, M.S., Schneider, R.B., et al.: Ai-enabled parkinson's disease screening using smile videos. NEJM AI **2**(7) (2025)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: IEEE ICCV. pp. 6836–6846 (2021)
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. Preprint arXiv:2506.09985 (2025)
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
5. Bologna, M., Espay, A.J., Fasano, A., Paparella, G., Hallett, M., Berardelli, A.: Redefining bradykinesia. Movement disorders: official journal of the Movement Disorder Society **38**(4), 551 (2023)
6. Bot, B.M., Suver, C., Neto, E.C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E., et al.: The mpower study, parkinson disease mobile data collected using researchkit. Scientific data **3**(1), 1–9 (2016)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE CVPR. pp. 6299–6308 (2017)
8. Di Pietro, D.A., Olivares, A., Comini, L., Vezzadini, G., Luisa, A., Petrolati, A., Boccola, S., Boccali, E., Pasotti, M., Danna, L., et al.: Voice alterations, dysarthria, and respiratory derangements in patients with parkinson's disease. Journal of Speech, Language, and Hearing Research **65**(10), 3749–3757 (2022)
9. Dorsey, E., Sherer, T., Okun, M.S., Bloem, B.R.: The emerging evidence of the parkinson pandemic. Journal of Parkinson's disease **8**(s1), S3–S8 (2018)

10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Friedlander, A.H., Mahler, M., Norman, K.M., Ettinger, R.L.: Parkinson disease: systemic and orofacial manifestations, medical and dental management. The Journal of the American Dental Association **140**(6), 658–669 (2009)
12. 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)
13. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: IEEE ICCV. pp. 5842–5850 (2017)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
15. Islam, M.S., Adnan, T., Freyberg, J., Lee, S., Abdelkader, A., Pawlik, M., Schwartz, C., Jaffe, K., Schneider, R.B., Dorsey, R., et al.: Accessible, at-home detection of parkinson's disease via multi-task video analysis. In: AAAI. vol. 39, pp. 28125–28133 (2025)
16. Islam, M.S., Rahman, W., Abdelkader, A., Lee, S., Yang, P.T., Purks, J.L., Adams, J.L., Schneider, R.B., Dorsey, E.R., Hoque, E.: Using ai to measure parkinson's disease severity at home. npj Digital Medicine **6**(1), 156 (2023)
17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. Preprint arXiv:1705.06950 (2017)
18. Kilic-Berkmen, G., Pirio Richardson, S., Perlmutter, J.S., Hallett, M., Klein, C., Wagle-Shukla, A., Malaty, I.A., Reich, S.G., Berman, B.D., Feuerstein, J., et al.: Current guidelines for classifying and diagnosing cervical dystonia: empirical evidence and recommendations. Movement Disorders Clinical Practice **9**(2), 183–190 (2022)
19. Langevin, R., Ali, M.R., Sen, T., Snyder, C., Myers, T., Dorsey, E.R., Hoque, M.E.: The park framework for automated analysis of parkinson's disease characteristics. ACM IMWUT **3**(2), 1–22 (2019)
20. LemaÃŽtre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research **18**(17), 1–5 (2017)
21. Leone, C., Feys, P., Moumdjian, L., D'Amico, E., Zappia, M., Patti, F.: Cognitive-motor dual-task interference: a systematic review of neural correlates. Neuroscience & Biobehavioral Reviews **75**, 348–360 (2017)
22. Li, A., Li, C.: Detecting parkinson's disease through gait measures using machine learning. Diagnostics **12**(10), 2404 (2022)
23. Novotny, M., Rusz, J., Cmejla, R., Ruzicka, E.: Automatic evaluation of articulatory disorders in parkinson's disease. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(9), 1366–1378 (2014)
24. Pearson, C., Hartzman, A., Munevar, D., Feeney, M., Dolhun, R., Todaro, V., Rosenfeld, S., Willis, A., Beck, J.C.: Care access and utilization among medicare beneficiaries living with parkinson's disease. npj Parkinson's Disease **9**(1), 108 (2023)

25. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
26. Sarasso, E., Gardoni, A., Zenere, L., Emedoli, D., Balestrino, R., Grassi, A., Basaia, S., Tripodi, C., Canu, E., Malcangi, M., et al.: Neural correlates of bradykinesia in parkinson's disease: a kinematic and functional mri study. npj Parkinson's Disease **10**(1), 167 (2024)
27. Schilder, J.C., Overmars, S.S., Marinus, J., van Hilten, J.J., Koehler, P.J.: The terminology of akinesia, bradykinesia and hypokinesia: past, present and future. Parkinsonism & related disorders **37**, 27–35 (2017)
28. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS **35**, 10078–10093 (2022)
29. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: IEEE CVPR. pp. 14549–14560 (2023)
30. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., Hornung, R., Schroff, F., Yang, M.H., Ross, D.A., Wang, H., Adam, H., Sirotenko, M., Liu, T., Gong, B.: Videoprism: a foundational visual encoder for video understanding. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)