Toward Trustworthy Vision-Language Reporting for Tremor Assessment under Distribution Shift

Xinjun Li
xl863@cornell.edu
Cornell University
New York, NY, USA

Abstract

Vision-language models (VLMs) are increasingly used in high-stakes workflows, and reliable deployment requires calibration under distribution shift, selective abstention, and bounded reporting grounded in structured evidence. We present a VLM-assisted framework for tremor assessment from monocular RGB video, where modular hand-object perception and temporal modeling first extract structured clinical evidence, and a constrained reporting layer then generates clinician-facing or patient-facing outputs under uncertainty-aware abstention. A baseline-aware patient state supports longitudinal comparison against prior function. We evaluate the system on prospectively collected in-person recordings from Parkinson's disease, essential tremor, and control participants using multiple consumer devices and viewpoints. Alongside clinician-aligned severity estimation, the key result is that constrained VLM reporting with abstention substantially reduces unsupported outputs compared with free-form and forced-answer baselines, while remaining stable under moderate device and viewpoint shift. These findings position tremor assessment as a compact testbed for trustworthy medical VLM reporting, where structured intermediate evidence, calibrated uncertainty, selective abstention, and bounded language generation provide the basis for reliability under real-world visual shift.

CCS Concepts

• **Applied computing** → **Health informatics**; *Health care information systems*; • **Computing methodologies** → **Computer vision**; *Activity recognition and understanding*; *Neural networks.*

Keywords

Medical AI, vision-language models, movement disorders, tremor assessment, clinical reporting, video-based motor assessment, uncertainty-aware abstention, calibration

1 Introduction

Vision-language models (VLMs) are becoming attractive interfaces for high-stakes workflows because they can summarize, explain, and communicate structured multimodal information. However, in real-world settings, reliable use requires downstream hallucination control, calibration under distribution shift, selective prediction, and the ability to avoid unsupported outputs. These concerns are especially acute in healthcare, where visual evidence may be heterogeneous, labels may be imperfect, and overconfident errors can carry substantial clinical cost.

Visually grounded clinical reporting depends on calibrated perception, bounded language generation, and robustness to real-world variation. Tremor-related hand dysfunction provides a useful test case for trustworthy VLM deployment. Prior mixed-reality work has examined ergonomic hand-motion assistance and AR-guided rehabilitation for tremor disorder therapies, highlighting the importance of task-level interaction, accessibility, and functional support in this clinical domain [1]. We extend this line of inquiry from intervention and rehabilitation interfaces to trustworthy VLM-assisted reporting: how video-derived hand-object evidence can be structured, calibrated, and communicated under uncertainty. In Parkinson's disease (PD) and essential tremor (ET), functional burden often emerges during hand-object interaction, with isolated limb motion capturing only part of the clinical picture. A patient may complete a brief examination task yet still struggle to stabilize a cup, open a pill bottle, or transport an object without spill or release failure. Existing video-based systems can recover clinically meaningful tremor features, but their outputs must still be translated into bounded, trustworthy reports before they can be safely used in clinical workflows.

This paper centers on trustworthy VLM-assisted reporting as a controlled framework for converting video-derived evidence into bounded clinical language. Our design follows three principles: (1) **structured evidence first**, through modular hand-object perception and temporal modeling; (2) **uncertainty-aware reporting**, through calibration, selective prediction, and abstention; and (3) **robustness in the wild**, through evaluation across device shift, viewpoint shift, and difficult visual conditions.

Our method, *Abstain to Assist*, combines quality-controlled video acquisition, hand-object representation learning, temporal symptom modeling, a baseline-aware patient memory for longitudinal comparison, and a constrained reporting layer that can abstain when confidence is low. The resulting system is a bounded decision-support interface that preserves structured evidence and makes failure more visible.

This work makes three contributions. First, we formulate tremor assessment as a trustworthy VLM reporting problem under real-world visual shift, with unconstrained video-to-text generation serving as the failure mode the system controls. Second, we introduce an evidence-first reporting pipeline that combines hand-object perception, temporal symptom modeling, baseline-aware patient state, calibration, and abstention. Third, we evaluate the system beyond accuracy, using participant-level clinical agreement, functional risk metrics, calibration, coverage, unsupported-output analysis, and robustness under device and viewpoint shift.

Trust setting. We consider three failure modes that are especially relevant for VLM-assisted reporting in healthcare: (1) unsupported free-form generation from visually ambiguous or shifted inputs, (2) overconfident reporting under device, viewpoint, or image-quality

shift, and (3) loss of task-level clinical meaning when raw multimodal inputs are passed directly to a language model without structured intermediate evidence. Our design addresses these risks through structured evidence extraction, calibration, and abstention-aware reporting.

2 Method

2.1 Overview

Our framework consists of five components: (1) quality-controlled video acquisition, (2) hand and hand-object perception, (3) temporal symptom modeling, (4) baseline-aware patient state updating, and (5) constrained reporting with abstention. The design is intentionally modular: perceptual evidence is produced before reporting, and reporting cannot bypass uncertainty or evidence constraints. This makes the system easier to interpret and better aligned with trustworthy deployment requirements.

2.2 Structured Evidence Extraction

Given a monocular RGB sequence $X = \{x_t\}_{t=1}^{T}$, a hand detector and landmark estimator extract frame-level hand structure. For object-manipulation tasks, an object detector identifies the manipulated object and estimates contact, grip, and release state. We use a landmark-first representation because it exposes clinically interpretable intermediate structure and reduces dependence on purely black-box appearance features.

This choice matters because functional burden in tremor disorders often arises at the hand-object interface. A hand-only representation may capture oscillation, but it can miss clinically relevant instability during stabilization, transport, or release.

A temporal encoder f_θ processes landmark trajectories, motion spectra, and hand-object state features to estimate severity, instability, and object risk. The model predicts $\hat{y}$ severity, $\hat{r}$ task-level risk, and $\hat{u}$ predictive uncertainty.

To support longitudinal tracking, we maintain a patient-specific state:

$$z_i^{(n)} = U\big(z_i^{(n-1)}, f_\theta(X^{(n)}), c^{(n)}, m_i\big), \tag{1}$$

where $z_i^{(n)}$ is the updated state after encounter n, $c^{(n)}$ is the task type, and m_i denotes static metadata. We use the term "digital twin" in a restrained sense: a baseline-aware longitudinal patient state. Full physiological simulation remains outside the scope of this module.

2.3 Trustworthy VLM Reporting

The reporting layer operates through an evidence contract between the visual-temporal model and the language interface. It receives structured task evidence, quality indicators, temporal severity estimates, predictive uncertainty, and baseline-aware change signals. The contract specifies what has been observed, how reliable the evidence is, which uncertainty signals are active, and which report actions are permitted. It outputs a clinician-facing summary, a patient-facing instruction set, or an abstain/escalate message. This design prevents ambiguous visual input from being converted into unsupported clinical language.

The trustworthiness design is motivated by recent work on calibrated robust fine-tuning of vision-language models under ID/OOD shift, selective prediction for unreliable black-box VLM responses, and efforts to reduce unnecessary abstention in vision-language reasoning. For language-safety analysis, we compare this constrained reporting layer against a free-form variant and a forced-answer variant. The free-form variant removes schema constraints. The forced-answer variant preserves structured evidence slots and disables abstention.

2.4 Training Objective

The total objective is

$$L = \lambda_1 L_{sev} + \lambda_2 L_{risk} + \lambda_3 L_{task} + \lambda_4 L_{temp} + \lambda_5 L_{cal} + \lambda_6 L_{twin}, \tag{2}$$

where the loss terms supervise severity estimation, unsafe interaction events, task success, temporal consistency, confidence calibration, and longitudinal state consistency. This modular design exposes clinically meaningful intermediate variables and makes failure analysis, calibration, and abstention easier than in unrestricted end-to-end generation pipelines.

3 Experimental Setup

We evaluate on a prospectively collected in-person clinical dataset of 35 real participants: 18 with Parkinson's disease, 10 with essential tremor, and 7 controls. Video recordings were acquired during structured study visits using three consumer smartphones and three synchronized viewpoints. The analyzed dataset contains 1,470 single-view clips corresponding to 490 unique task performances, with 12 participants returning for a second visit.

Standardized tasks include resting posture, postural hold, finger tapping, hand open-close, and pronation-supination. ADL tasks include cup grasping, cup drinking, pill-bottle opening, book holding, spoon use, and phone pickup. PD labels use the relevant MDS-UPDRS Part III items; ET labels use a TETRAS-aligned rubric; pooled severity is built through clinician-supervised ordinal harmonization, which avoids treating the two scales as directly numerically equivalent.

We use diagnosis-stratified, patient-wise train/val/test splits (21/7/7 participants). Multi-view recordings from the same physical performance are treated as correlated observations. During evaluation, predictions are fused within performance and then aggregated at the participant level. This participant-level protocol is critical for trustworthy evaluation under heterogeneous recording conditions.

We evaluate the system along four axes: clinical validity, functional relevance, calibration and selective reporting, and language safety. Clinical validity is measured by Spearman correlation, ICC, and MAE. Functional relevance is measured by task F1 and risk AP. Calibration and selective reporting are measured by ECE, answered-case accuracy, and coverage. Language safety is measured by unsupported-statement rate and abstention correctness. We define a reporting case as one task-level assessment instance passed to the reporting layer after visual-temporal evidence extraction. An unsupported statement is a generated report claim that is not entailed by the structured evidence representation, including unsupported diagnosis, causal attribution, treatment recommendation, or longitudinal change.

Evaluation protocol. We treat trustworthy VLM-assisted reporting as a joint problem of evidence quality, predictive validity, and

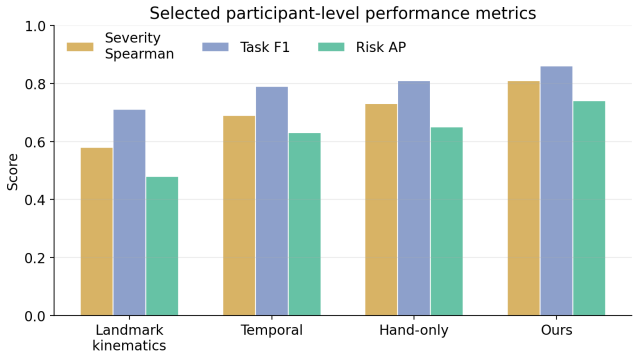

Figure 1: Selected participant-level performance metrics across landmark-kinematics, temporal, hand-only, and full-model baselines. Reliable VLM reporting depends on accurate upstream structured evidence extraction.

Table 1: Participant-level results after within-performance view fusion.

Method	Severity Spearman↑	MAE↓	ICC↑	Task F1↑	Risk AP↑	ECE↓
Landmark-kinematics	0.58	0.72	0.51	0.71	0.48	0.15
Temporal w/o twin	0.69	0.58	0.64	0.79	0.63	0.11
Hand-only	0.73	0.52	0.68	0.81	0.65	0.09
Ours	**0.81**	**0.41**	**0.76**	**0.86**	**0.74**	**0.06**

report reliability. Accordingly, we evaluate clinician agreement, task-level relevance, calibration, coverage under abstention, unsupported statement rate, and robustness under device/viewpoint shift. This protocol is intended to better reflect real-world deployment conditions than accuracy-only evaluation.

4 Results

4.1 Main System Performance

Table 1 reports participant-level results after within-performance view fusion. The full model achieves the strongest performance across the primary metrics, with severity Spearman 0.81, MAE 0.41, ICC 0.76, Task F1 0.86, Risk AP 0.74, and ECE 0.06. Figure 1 visualizes selected metrics from the same comparison. The largest gains appear in tasks requiring sustained object stabilization, especially cup holding and pill-bottle manipulation.

4.2 Trustworthy VLM Reporting

The central reporting-safety result is the language-safety comparison in Table 2. Table 1 establishes the reliability of the upstream visual-temporal evidence; Table 2 isolates how the reporting layer controls evidence fidelity under uncertainty.

Free-form reporting leaves evidence gaps open and reaches a 9.7% unsupported-statement rate. Schema-constrained forced answering lowers the rate to 3.2% by preserving evidence slots. Constrained reporting with abstention changes the output policy: the system answers evidence-supported cases and defers ambiguous cases. This yields 89.2% answered-case accuracy at 73% coverage and reduces unsupported statements to 1.4%.

Table 2: Trustworthiness of alternative reporting modes.

Reporting mode	Answered-case Acc.↑	Coverage↑	Unsupported Rate↓
Free-form	N/A	100%	9.7%
Forced-answer	81.4%	100%	3.2%
Constrained + abstain	89.2%	73%	1.4%

Note. Answered-case accuracy is defined for structured reporting modes with discrete report decisions. Free-form outputs are evaluated by unsupported-statement rate.

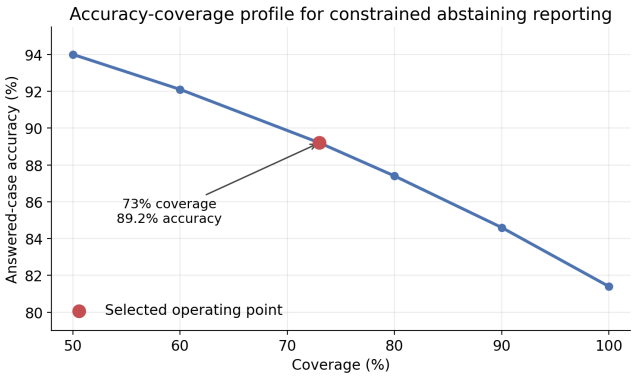

Figure 2: Accuracy-coverage profile for constrained abstaining reporting. The selected operating point reaches 89.2% answered-case accuracy at 73% coverage.

4.3 Robustness under Shift

Relative to the in-domain participant-level evaluation, severity Spearman drops by less than 0.04 under cross-device evaluation and by less than 0.06 under moderate viewpoint shift. Under severe occlusion and extreme backlighting, severity Spearman falls to 0.62 and abstention rises to 38% of evaluated reporting cases, indicating that the system becomes more conservative as evidence quality degrades. This behavior is desirable for trustworthy VLM-assisted deployment in the wild. Importantly, the increase in abstention under severe shift reduces the likelihood that degraded visual evidence is converted into unsupported natural-language output.

4.4 Baseline-Aware Longitudinal Tracking

Among the repeat-visit participants, baseline-aware patient-state updating improves longitudinal tracking. Second-visit Spearman increases from 0.61 to 0.79 relative to the temporal model without patient-state updating, and worsened-status detection improves from 0.54 to 0.73 F1. Figure 3 shows representative stable and worsening trajectories. The patient state provides longitudinal context for the reporting layer, allowing stable and worsening trajectories to be surfaced through constrained language generation.

5 Discussion and Conclusion

Tremor assessment provides a useful stress test for trustworthy VLM-assisted decision support in the wild. Reliable healthcare reporting from VLMs benefits from modular visual evidence extraction, calibration, participant-level evaluation, and abstaining assistance. Unrestricted multimodal generation defines the failure mode evaluated by the reporting baselines.

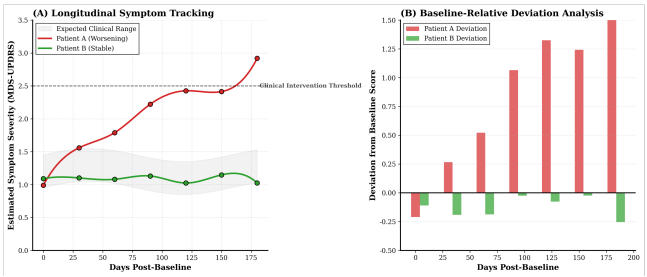

Figure 3: Representative stable and worsening longitudinal trajectories. The baseline-aware patient state provides longitudinal context that can be surfaced to the reporting layer through constrained reporting.

The system is intended for structured assessment support. Autonomous diagnosis and treatment planning remain outside its operating scope. Its main strengths come from bounded use: preserving structured evidence, correctly fusing correlated multi-view observations, deferring low-confidence cases, and reducing unsupported outputs under real-world shift conditions. Its limitations remain those expected for monocular real-world video, including occlusion, low-contrast scenes, sparse longitudinal follow-up, and the difficulty of harmonizing labels across heterogeneous tremor disorders. More broadly, the proposed evaluation protocol can serve as a compact benchmark template for trustworthy VLM-assisted reporting in healthcare, combining agreement, robustness under shift, calibration, coverage, and unsupported-output analysis within a single deployment-oriented setting.

References

[1] X. Li and Z. Lei. Ergonomic hand motion assistance and AR rehabilitation: Bridging the gap in tremor disorder therapies through mixed reality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 322–325. IEEE, 2024. doi:10.1109/VRW62533.2024.00064.

[2] 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*, 23(15):2129–2170, 2008.

[3] R. J. Elble. The Essential Tremor Rating Assessment Scale. *Journal of Neurology & Neuromedicine*, 1(4):34–38, 2016.

[4] P. M. Pecoraro, V. Riccio, A. L. Guida, et al. Computer Vision Technologies in Movement Disorders: A Systematic Review. *Movement Disorders Clinical Practice*, 12(9):1229–1243, 2025.

[5] R. Martínez-García-Peña, L. H. Koens, G. Azzopardi, and M. A. J. Tijssen. Video-Based Data-Driven Models for Diagnosing Movement Disorders: Review and Future Directions. *Movement Disorders*, 40(10):2046–2066, 2025.

[6] L. Kenny, Z. Azizi, K. Moore, M. Alcock, S. Heywood, A. Johnson, K. McGrath, M. J. Foley, B. Sweeney, S. O'Sullivan, et al. Inter-rater reliability of hand motor function assessment in Parkinson's disease: Impact of clinician training. *Clinical Parkinsonism & Related Disorders*, 11:100278, 2024.

[7] D. Deng, J. L. Ostrem, V. Nguyen, D. D. Cummins, J. Sun, A. Pathak, S. Little, and R. Abbasi-Asl. Interpretable video-based tracking and quantification of parkinsonism clinical motor states. *npj Parkinson's Disease*, 10:122, 2024.

[8] D. L. Guarín. Video-based quantification of hand postural tremor without external references: Integrating postural tremor quantification into visionMD. *npj Parkinson's Disease*, 11:351, 2025.

[9] R. Wolke, M. Schmidt, F. Nolte, et al. Validity of tremor analysis using smartphone-compatible computer vision frameworks. *Scientific Reports*, 2025.

[10] C. Oh, H. Lim, M. Kim, D. Han, S. Yun, J. Choo, A. Hauptmann, Z.-Q. Cheng, and K. Song. Towards Calibrated Robust Fine-Tuning of Vision-Language Models. *Advances in Neural Information Processing Systems*, 2024.

[11] Z. Khan and Y. Fu. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[12] T. Srinivasan, J. Hessel, T. Gupta, B. Y. Lin, Y. Choi, J. Thomason, and K. R. Chandu. Selective "Selective Prediction": Reducing Unnecessary Abstention in Vision-Language Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[13] Z. Gu, J. Chen, F. Liu, C. Yin, and P. Zhang. MedVH: Toward Systematic Evaluation of Hallucination for Large Vision Language Models in the Medical Context. *Advanced Intelligent Systems*, 2025.

[14] P. Wienholt, S. Caselitz, R. Siepmann, P. Bruners, K. Bressem, C. Kuhl, J. N. Kather, S. Nebelung, and D. Truhn. Hallucination filtering in radiology vision-language models using discrete semantic entropy. *European Radiology*, 2026.

[15] International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. IMDRF/AIML WG/N88 FINAL:2025, 2025.