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

Anonymous Author 1*
anonymous1@example.com
Anonymous Institution 1
Anonymous City, Anonymous State, Anonymous Country

Anonymous Author 2*
anonymous2@example.com
Anonymous Institution 2
Anonymous City, Anonymous State, Anonymous Country

Abstract

Vision-language models (VLMs) are increasingly used in high-stakes workflows, yet reliable deployment depends on more than raw multimodal capability alone. In healthcare settings, trustworthy use additionally 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 a pilot dataset of Parkinson's disease, essential tremor, and control participants recorded with multiple consumer devices and viewpoints. Beyond strong clinician-aligned severity estimation, the main 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 suggest that trustworthy VLM use in healthcare benefits from structured intermediate representations, calibration, selective prediction, and abstaining assistance rather than unrestricted multimodal generation.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Neural networks*; • **Applied computing** → *Health care information systems.*

Keywords

VLMs, Abstention, Uncertainty, Calibration, Healthcare

ACM Reference Format:

Anonymous Author 1 and Anonymous Author 2. 2026. Toward Trustworthy Vision-Language Reporting for Tremor Assessment under Distribution Shift. In *Proceedings of Anonymous Workshop Submission.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

Anonymous Workshop Submission,

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 depends not only on downstream hallucination control, but also on 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.

This setting is particularly relevant to TrustVLM because visually grounded clinical reporting is useful only if it remains calibrated, bounded, and robust to real-world variation. Tremor-related hand dysfunction provides a useful test case for trustworthy VLM deployment. In Parkinson's disease (PD) and essential tremor (ET), functional burden often emerges during hand-object interaction rather than isolated limb motion. 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.

We therefore focus this paper on trustworthy VLM-assisted reporting rather than unrestricted end-to-end multimodal generation. 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 designed not as an autonomous clinical agent, but as a bounded decision-support interface that preserves structured evidence and makes failure more visible.

This work makes three contributions. First, we present a pipeline from structured evidence to VLM reporting for tremor assessment. Second, we show that **constrained reporting and abstention** reduce unsupported outputs relative to free-form and forced-answer baselines. Third, we demonstrate utility under realistic shift conditions, including device, viewpoint, and image-quality degradation.

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, not a full physiological simulator.

2.3 Trustworthy VLM Reporting

The reporting layer receives only structured evidence, not raw video. It outputs either: (1) a clinician-facing summary, (2) a patient-facing instruction set, or (3) an abstain/escalate message. We intentionally avoid an agentic or autonomous VLM design; in this setting, the goal is bounded reporting from structured evidence with explicit abstention, not open-ended clinical reasoning.

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 former removes schema constraints; the latter preserves structured evidence slots but 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 pilot dataset of 35 participants: 18 PD, 10 ET, and 7 controls, collected in a simulated home environment 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 rather than direct numerical equivalence.

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 compare against four baselines: (1) Landmark–Kinematics, (2) Temporal without Twin, (3) Hand-Only, and (4) Forced-Answer Language. We also evaluate a free-form language variant for hallucination analysis. Metrics include clinical validity (Spearman, ICC, MAE), functional relevance (task F1, risk AP), trustworthiness (ECE, answered-case accuracy, coverage), and language safety (unsupported-statement rate, abstention correctness).

Evaluation protocol. We treat trustworthy VLM-assisted reporting as a joint problem of evidence quality, predictive validity, and report reliability. Accordingly, we evaluate not only clinician agreement and task-level relevance, but also 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. Figure 1 shows the corresponding main-system comparison for the TrustVLM setting. The proposed model performs

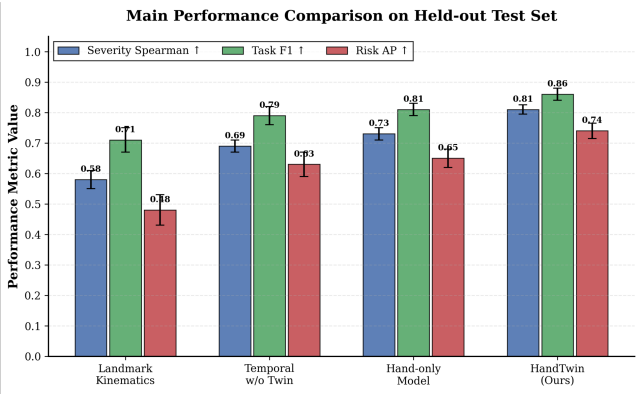

Figure 1: Main system performance across landmark-only, hand-only, and temporal baselines. Reliable VLM reporting depends on accurate upstream structured evidence extraction.

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**

best on all primary metrics. The largest gains appear in tasks requiring sustained object stabilization, especially cup holding and pill-bottle manipulation. This supports explicit hand-object modeling over hand-only kinematics.

4.2 Trustworthy VLM Reporting

The main TrustVLM result is that constrained abstaining reporting is substantially more reliable than unconstrained alternatives. Table 2 isolates the trustworthiness contribution of the reporting layer by directly comparing free-form, forced-answer, and abstaining reporting modes. Relative to the forced-answer language variant, uncertainty-aware abstention improves answered-case accuracy from 81.4% to 89.2% at 73% coverage. Unsupported statements fall from 9.7% in the free-form variant to 3.2% under schema constraints and to 1.4% with abstention enabled. Figure 2 shows the corresponding accuracy–coverage trade-off.

These findings suggest that trust gains arise primarily from selective deferral of ambiguous cases rather than uniformly increasing confidence across all cases. This interpretation is consistent with recent work on selective VLM prediction and abstention-aware reasoning. The comparison also shows that schema constraints alone are helpful, but insufficient: the largest reliability gain appears when structured evidence is combined with an explicit reject option.

4.3 Robustness under Shift

Under cross-device evaluation, severity correlation drops by less than 0.04; under moderate viewpoint shift, by less than 0.06. Under

Table 2: Trustworthiness of alternative reporting modes.

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

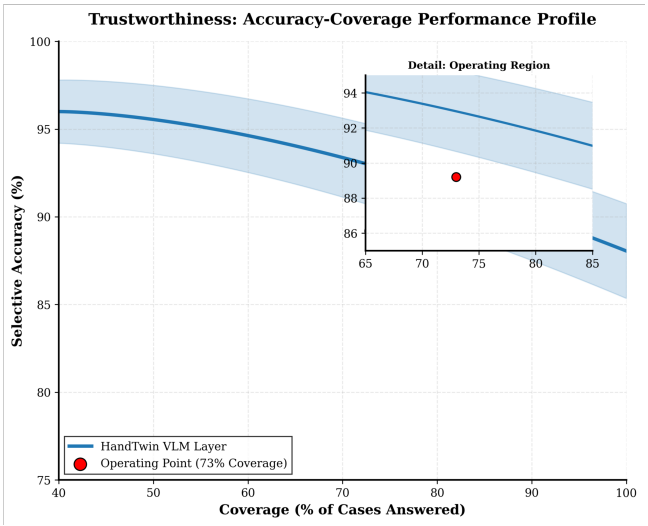

Figure 2: Accuracy–coverage trade-off for uncertainty-aware abstention in constrained VLM reporting. Higher reliability is obtained by selectively deferring ambiguous cases rather than forcing outputs for every query.

severe occlusion and extreme backlighting, severity correlation falls to 0.62 and abstention rises to 38%, 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, the baseline-aware patient state improves longitudinal tracking, with second-visit Spearman increasing from 0.61 to 0.79 and worsened-status detection improving from 0.54 to 0.73 F1. Figure 3 shows representative stable and worsening trajectories. In this paper, the role of the patient state is not to foreground "digital twin" as a concept, but to provide reliable longitudinal context for reporting. In the present TrustVLM framing, this module mainly serves to provide longitudinal context for reporting rather than to define a separate digital-twin contribution.

5 Discussion and Conclusion

The central message is not that tremor assessment is merely another clinical application, but that it provides a useful stress test for trustworthy VLM-assisted decision support in the wild. Our results suggest that reliable healthcare reporting from VLMs benefits

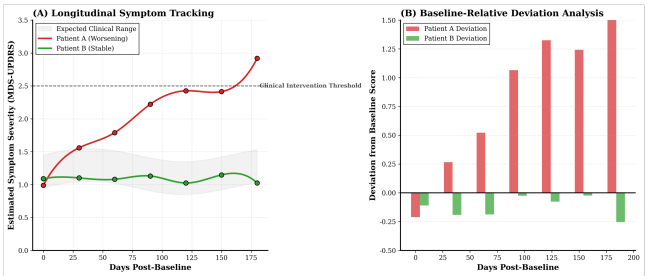

Figure 3: Representative stable and worsening longitudinal trajectories. The baseline-aware patient state provides longitudinal context that can be surfaced to the reporting layer without requiring unrestricted free-form reasoning.

from modular visual evidence extraction, calibration, participant-level evaluation, and abstaining assistance rather than unrestricted multimodal generation.

The system is intended for structured assessment support, not autonomous diagnosis or treatment planning. 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] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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