Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Nested Phase-Specific Speed-Deficit Viability for Computer-Assisted Robotic Surgical Training Assessment

Ali Algarni

ahalgarni@kau.edu.sa

King Abdulaziz University

Abdulrahman Ali Alsolami

King Abdulaziz University

Research Article

Keywords: Computer-assisted surgery, robotic surgical training assessment, dVRK kinematics, phase-normalized motion analysis, surgical motion quality, operator-disjoint validation, kinematic performance metrics

Posted Date: September 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10580275/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Nested Phase-Specific Speed-Deficit Viability for Computer-Assisted Robotic Surgical Training Assessment

Ali Algarni[1,*], Abdulrahman Ali Alsolami[2]
[1]Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[1]Email: ahalgarni@kau.edu.sa
[2]Department of Mathematics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Email: aalsolami1@kau.edu.sa
*Corresponding Author email: ahalgarni@kau.edu.sa

Abstract: Objective motion-quality assessment in computer-assisted robotic surgical training requires metrics that preserve task-phase structure rather than reducing complete trajectories to global kinematic summaries. This paper introduces Nested Phase-Specific Speed-Deficit Viability (Nested PS-SDVI), an interpretable and leakage-controlled trajectory-level scoring method for dVRK kinematics. For each held-out operator, task-specific good-quality speed reference profiles are estimated from training operators only. Positive speed-deficit exposure is then measured within normalized task phases, while phase-bin resolution, reference quantile, and deficit summary are selected through inner user-level validation. The evaluated operator is excluded from reference construction, thresholding, score standardization, and candidate selection. Technical validation used 46 ROSMA dVRK trials from three operators across Pea-on-a-Peg, Post-and-Sleeve, and Wire-Chaser I tasks. Nested PS-SDVI achieved macro-AUC of 0.762 for poor-quality trial detection and macro-Spearman association of 0.570 with train-standardized within-task ROSMA score, outperforming fixed PS-SDVI variants and raw kinematic baselines. These results support phase-conditioned speed-deficit scoring as a compact assessment layer for computer-assisted robotic surgical training analysis.

Keywords: Computer-assisted surgery; robotic surgical training assessment; dVRK kinematics; phase-normalized motion analysis; surgical motion quality; operator-disjoint validation; kinematic performance metrics.

1. INTRODUCTION

Objective technical-skill assessment has long been central to surgical training. Early structured tools such as OSATS established standardized expert-based evaluation, while later motion-analysis work showed that surgical skill can be quantified through tool motion, haptic information, and statistical models of surgical activity (Martin et al., 1997; Rosen et al., 2001; Reiley et al., 2011; Vedula et al., 2017). In robotic surgery, GEARS and subsequent technical-skill assessment studies extended this direction toward robotic-console performance and task-level skill evaluation (Goh et al., 2012; Chen et al., 2019). Recent reviews confirm that objective and AI-assisted surgical-skill assessment remains an active problem, with persistent concerns regarding validation design, heterogeneity, interpretability, and generalizability (Lam et al., 2022; Boal et al., 2024; Kankanamge et al., 2025; Laughlin et al., 2025). Kankanamge et al. (2025), for example, reported substantial heterogeneity in the reliability of AI-based minimally invasive skill assessment despite the use of standardized objective metrics.

Public robotic-surgery datasets have enabled reproducible algorithmic evaluation. JIGSAWS provides da Vinci kinematics, video, gesture annotations, and skill labels, and Ahmidi et al. (2017) formalized benchmark segmentation and recognition evaluation on these data (Gao et al., 2014; Ahmidi et al., 2017). ROSMA later expanded the dataset landscape with dVRK recordings of three common training maneuvers, 207 trials from 12 users, 160 kinematic variables recorded at 50 Hz, and video recordings at 15 frames per second; ROSMAG40 further adds gesture annotations to a ROSMA subset (Rivas-Blanco et al., 2024). These datasets show that robotic surgical motion is not a homogeneous time series but a structured task trajectory with phase, gesture, and subtask organization.

Automated assessment methods broadly fall into handcrafted kinematic descriptors and learned representations. Kinematic studies use speed, jerk, path length, smoothness, end-effector motion, coordination, and related descriptors to quantify skill (Zia & Essa, 2018; Lefor et al., 2020). Deep and multimodal approaches learn from kinematic, video, physiological, or combined streams and have shown strong discriminative ability in skill classification or regression tasks (Wang & Fey, 2018; Fawaz et al., 2019; Nguyen et al., 2019; Funke et al., 2019; Lavanchy et al., 2021; Pan et al., 2023; Chen et al., 2025; Shafiei et al., 2026). Chen et al. (2025) used video and kinematic classifiers on JIGSAWS and ROSMA, while Shafiei et al. (2026) showed that multimodal deep learning using physiological signals can classify robot-assisted surgical skill in task-dependent ways.

1.1. Study Theme

Robot-assisted surgical training produces high-frequency kinematic trajectories that can support objective motion-quality assessment. Public datasets have enabled this direction: ROSMA provides dVRK-based recordings of common surgical training tasks with kinematic data recorded at 50 Hz, while JIGSAWS provides da Vinci kinematics, video, gesture annotations, and skill-related labels for surgical activity analysis (Gao et al., 2014; Rivas-Blanco et al., 2024). Existing assessment approaches commonly use global kinematic descriptors such as speed, path length, jerk, smoothness, bimanual coordination, completion time, or error counts. These quantities are informative, but they compress a temporally structured surgical task into trial-level summaries. Gesture-aware and phase-aware studies show that robotic surgical motion has structured temporal organization, yet many skill-assessment scores still rely on global descriptors or learned representations whose relation to task phase is indirect (Lin et al., 2006; Tao et al., 2012; Zappella et al., 2013; Ahmidi et al., 2017). This creates a specific methodological gap: an interpretable trial-level score is needed that preserves task phase, distinguishes phase-appropriate controlled motion from phase-inappropriate speed deficit, and prevents validation leakage during reference construction and model selection.

1.2. Novelty and Contributions

This paper makes four contributions.

i. Nested PS-SDVI defines motion-quality degradation as positive speed-deficit exposure relative to task-specific phase references, rather than as raw speed, jerk, or other global kinematic summaries. This makes the score sensitive to where the deficit occurs within the surgical task.

ii. The method separates outer held-out-operator evaluation from inner candidate selection. The evaluated operator is excluded from reference construction, poor-quality thresholding, score standardization, and model selection.

iii. The phase-bin count, reference quantile, and deficit summary are selected inside the training folds. This allows different robotic-surgery tasks to use different temporal resolutions and summary forms without tuning on the held-out operator.

iv. Nested PS-SDVI is evaluated against fixed PS-SDVI variants and raw kinematic baselines, including low-speed, jerk, asynchrony, and bimanual-closeness exposure. This comparison tests whether phase-normalized speed-deficit viability adds information beyond unnormalized motion descriptors.

The proposed framework is positioned as a computer-assisted surgical training analytics method: it converts dVRK kinematic trajectories into an interpretable phase-normalized motion-quality score that can be evaluated under operator-disjoint validation.

2. METHODOLOGY

2.1. Problem Statement

Robotic-surgery training trajectories contain phase-dependent motion patterns that are not fully represented by global speed, jerk, or completion-time summaries. A low-speed segment may reflect poor execution in one part of a task but controlled manipulation in another. The proposed method therefore

evaluates motion quality through phase-specific speed-deficit viability, where each held-out trial is compared with a task-specific good-quality reference learned only from training operators.
Let

$$\mathcal{D} = \{(\mathcal{X}_i, y_i, \tau_i, u_i, g_i) : i = 1, \dots, N\} \tag{1}$$

denote the trial-level dataset. Here, $\mathcal{X}i = \{x_i(t_k)\}_{k=1}^{T_i}$ is the kinematic trajectory of trial i, $y_i \in \mathbb{R}$ is the ROSMA performance score, $\tau_i \in \mathcal{T}$ is the task, $u_i \in \mathcal{U}$ is the operator, and g_i contains auxiliary outcomes such as task time and errors. The score y_i is oriented so that larger values indicate poorer performance.
For each trial, let $v_i(t_k) \geq 0$ be the bimanual tool-speed signal extracted from Cartesian motion. The normalized phase coordinate is

$$\phi_i(t_k) = \frac{t_k - t_{i1}}{t_{iT_i} - t_{i1}} \in [0,1] \tag{2}$$

For $K \geq 2$, define the ordered phase partition

$$P_b = \left[\frac{b-1}{K}, \frac{b}{K}\right), \; b = 1, \dots, K-1, \qquad P_K = \left[\frac{K-1}{K}, 1\right]. \tag{3}$$

For an outer held-out operator u and task τ, the operator-disjoint training index set is

$$\mathcal{I}_{-u,\tau} = \{i : u_i \neq u, \; \tau_i = \tau\}. \tag{4}$$

Given reference quantile $q \in (0,1)$, the good-quality reference subset is

$$\mathcal{G}_{-u,\tau}(q) = \left\{i \in \mathcal{I} - u, \tau : y_i \leq Q_q\left(y_j : j \in \mathcal{I} - u, \tau\right)\right\}. \tag{5}$$

For phase bin P_b, define the reference speed center and scale by

$$m_{b,-u,\tau} = \text{median}\left\{v_i(t_k) : i \in \mathcal{G}_{-u,\tau}(q), \; \phi_i(t_k) \in P_b\right\}, \tag{6}$$

and $m_{b,-u,\tau} > 0$ as the corresponding robust scale. For a held-out trial i with $u_i = u$ and $\tau_i = \tau$, the positive phase-specific speed deficit is

$$d_i(t_k; K, q) = \left[\frac{m_{b,-u_i,\tau_i} - v_i(t_k)}{s_{b,-u_i,\tau_i}}\right]_+, \qquad \phi_i(t_k) \in P_b, \tag{7}$$

where $[z]_+ = \max(z, 0)$.
The trial-level Nested PS-SDVI score is

$$S_i = S\{d_i(t_k; K, q) : k = 1, \dots, T_i\}, \tag{8}$$

where S is selected from candidate summaries including mean deficit, high-deficit fraction, tail mean, 95th percentile, or a convex hybrid. The tuple (K, q, S) is selected by inner user-level validation using only the outer training operators.
The validation objective is

$$S_i \uparrow \quad \Leftrightarrow \; y_i^{(z)} \uparrow, \tag{9}$$

where $y_i^{(z)}$ is the train-standardized within-task ROSMA score. Poor-quality detection is evaluated by AUC, and continuous agreement is evaluated by Spearman association. Thus, the methodological problem is to construct a leakage-controlled, phase-normalized viability score that assigns larger values to lower-quality held-out robotic-surgery trials.

2.2. Phase-Specific Reference and Deficit Viability

This section makes the abstract score in Eq. (8) explicit. For trial i, the raw bimanual speed $v_i(t_k)$ is converted to a bounded-scale working signal

$$z_i(t_k) = \log 1 + \tilde{v}_i(t_k), \tag{10}$$

where $\tilde{v}_i(t_k)$ denotes the winsorized bimanual speed. The logarithmic transform reduces the influence of extreme frame-level velocity spikes while preserving the order of speed values.
For an outer held-out operator u, task τ, reference quantile q, and phase partition $\mathcal{P}_K$, define the reference sample in phase bin P_b as

$$\mathcal{V}_{b,-u,\tau}(q) = \left\{z_i(t_k) : i \in \mathcal{G}_{-u,\tau}(q), \; \phi_i(t_k) \in P_b\right\}. \tag{11}$$

The phase-specific reference center and scale are

$$m_{b,-u,\tau}(q) = \text{median}\mathcal{V}_{b,-u,\tau}(q), \qquad \text{s}_{b,-u,\tau}(q) = \max\left\{\text{MAD}\mathcal{V}\mathcal{V}_{b,-u,\tau}(q), \varepsilon_s\right\}, \tag{12}$$

where $\varepsilon_s > 0$ prevents division by a degenerate within-bin scale. If a phase bin has insufficient reference samples, the center and scale are replaced by the task-level training median and MAD computed from $\bigcup_{b=1}^{K} \mathcal{V}_{b,-u,\tau}(q)$

For a held-out trial i with $u_i = u$ and $\tau_i = \tau$, the phase-specific standardized speed-deficit field is

$$d_i(t_k; K, q) = \left[\frac{m_{b,-u_i,\tau_i}(q) - z_i(t_k)}{s_{b,-u_i,\tau_i}(q)}\right]_+, \qquad \phi_i(t_k) \in P_b. \tag{13}$$

Only positive deficits are retained. Speeds at or above the phase-specific good-quality reference contribute zero deficit.

Let $D_i(K, q) = \{d_i(t_k; K, q) : k = 1, \ldots, T_i\}$. The candidate trial-level summaries are

$$S_i^{\text{mean}} = T_i^{-1} \sum_{k=1}^{T_i} d_i(t_k; K, q), \qquad S_i^{\text{p95}} = Q_{0.95}\{D_i(K, q)\}, \tag{14}$$

$$S_i^{\text{frac}} = T_i^{-1} \sum_{k=1}^{T_i} b1d_i(t_k; K, q) > 1, S_i^{\text{tail}} = \frac{\sum_{k=1}^{T_i} d_i(t_k; K, q) 1d_i(t_k; K, q) \geq Q_\rho(D_i)}{\sum_{k=1}^{T_i} \text{b1}d_i(t_k; K, q) \geq Q_\rho(D_i)}, \tag{15}$$

where $Q_\rho(D_i)$ is the empirical upper-tail threshold, with $\rho = 0.80$ in the validation.

A convex hybrid summary is also allowed:

$$S_i^{\text{hyb}} = w_1 S_i^{\text{p95}} + w_2 S_i^{\text{mean}} + w_3 S_i^{\text{tail}}, \qquad w_r \geq 0, \quad \sum_{r=1}^{3} w_r = 1. \tag{16}$$

The complete candidate family is

$$\mathfrak{S} = \{S^{\text{mean}}, S^{\text{p95}}, S^{\text{frac}}, S^{\text{tail}}, S^{\text{hyb}}\}. \tag{17}$$

For a candidate $c = (K, q, S)$, the resulting score of trial i is

$$S_i(c) = S\{D_i(K, q)\}, \qquad c \in \mathcal{C} = \mathcal{K} \times \mathcal{Q} \times \mathfrak{S}. \tag{18}$$

Eqs. (10)–(18) define PS-SDVI as a phase-normalized positive-deficit functional. The score is nonnegative by construction, task-specific through τ_i, operator-disjoint through $\mathcal{G}_{-u_i,\tau_i}(q)$, and phase-aware through $\mathcal{P}_\mathcal{K}$. The next subsection defines how the candidate (c) is selected without using the outer held-out operator.

2.3. Nested Candidate Selection

Nested PS-SDVI selects the candidate $c = (K, q, S) \in \mathcal{C}$ in Eq. (18) without using the outer held-out operator. For an outer held-out operator u, define the outer training operators by $\mathcal{U} - u = \mathcal{U} \setminus \{u\}$. Candidate selection is performed by inner leave-one-operator validation over $\mathcal{U}_{-u}$.

For an inner held-out operator $r \in \mathcal{U}_{-u}$ and task τ, define the inner training and validation index sets as

$$\mathcal{I}_{-u,-r,\tau} = \{i : u_i \notin \{u, r\},\ \tau_i = \tau\}, \qquad \mathcal{I}_{r,\tau} = \{i : u_i = r,\ \tau_i = \tau\}. \tag{19}$$

For each candidate $c = (K, q, S)$, reference profiles are built only from $\mathcal{I}_{-u,-r,\tau}$. Scores $S_i(c)$ are then computed for $i \in \mathcal{I}_{r,\tau}$. The inner standardized score is

$$y_{i,-u,-r,\tau}^{(z)} = \frac{y_i - \text{median}\{y_j : j \in \mathcal{I}_{-u,-r,\tau}\}}{\text{MAD}\{y_j : j \in \mathcal{I}_{-u,-r,\tau}\}}, \qquad i \in \mathcal{I}_{r,\tau}. \tag{20}$$

The corresponding poor-quality label is

$$b_{i,-u,-r,\tau} = 1\{y_i \geq Q_{0.75}(y_j : j \in \mathcal{I}_{-u,-r,\tau})\}. \tag{21}$$

For candidate c, inner validation performance is computed by pooling inner validation trials across $r \in \mathcal{U}_{-u}$. Let

$$A_{u,\tau}(c) = \text{AUC}(\{b_{i,-u,-r,\tau}\}, \{S_i(c)\}) \tag{22}$$

and

$$R_{u,\tau}(c) = \text{Spearman}\left(\{S_i(c)\}, \left\{y_{i,-u,-r,\tau}^{(z)}\right\}\right). \tag{23}$$

The inner selection score is

$$\mathcal{L}_{u,\tau}(c) = \lambda_A\{2A_{u,\tau}(c) - 1\} + \lambda_R R_{u,\tau}(c) - \lambda_K \frac{K}{K_{max}}, \tag{24}$$

where $\lambda_A, \lambda_R \geq 0$, $\lambda_A + \lambda_R = 1$, and $\lambda_K \geq 0$ is a small complexity penalty for excessive phase discretization.

The selected candidate for outer held-out operator u and task τ is

$$\hat{c}_{u,\tau} \in \arg\max_{c \in \mathcal{C}} \mathcal{L}_{u,\tau}(c). \tag{25}$$

Ties are resolved by a fixed deterministic rule, applied identically across all folds.

After $\hat{c}_{u,\tau}$is selected, the final reference profile is rebuilt using all available outer-training trials $\mathcal{I}_{-u,\tau}$. The selected score for an outer held-out trial i with $u_i = u$ and $\tau_i = \tau$ is

$$\hat{S}_i = S_i(\hat{c}_{u,\tau}). \tag{26}$$

Thus, the outer held-out operator contributes only to the final evaluation of $\hat{S}_i$, not to candidate selection or reference estimation.

The fixed comparator used in validation is obtained by setting

$$c_0 = (K_0, q_0, S^{p95}), \qquad K_0 = 15, \quad q_0 = 0.35. \tag{27}$$

Raw kinematic comparators are computed from unnormalized low-speed, jerk, asynchrony, and closeness exposure channels. These baselines test whether the nested phase-specific construction adds information beyond global motion summaries.

2.4. Mathematical Properties

The following properties are stated for a fixed task τ, held-out operator u, and candidate $c = (K, q, S)$. Assume that the transformed speed signal $z_i(t_k)$ is finite and that each robust phase scale satisfies $s_{b,-u,\tau}(q) \geq \varepsilon_s > 0$. Then

$$0 \leq d_i(t_k; K, q) < \infty, \qquad k = 1, \dots, T_i. \tag{28}$$

4For a nonnegative vector $d = (d_1, \dots, d_T)$, let $d_{(1)} \leq \cdots \leq d_{(T)}$ denote its order statistics. The upper-tail mean in Eq. (15) can be written as

$$S^{\text{tail}}(d) = m_\rho^{-1} \sum_{\ell=T-m_\rho+1}^{T} d_{(\ell)}, \qquad m_\rho = \lceil (1-\rho)T \rceil. \tag{29}$$

Proposition 1 (Finite nonnegative scoring): For every $c \in \mathcal{C}$ and every held-out trial i

$$0 \leq S_i(c) < \infty. \tag{30}$$

Proof. Eq. (13) defines $d_i(t_k; K, q)$ as the positive part of a finite standardized deficit. Hence $D_i(K, q)$ contains finite nonnegative values. The mean, high-deficit fraction, empirical quantile, tail mean, and convex hybrid in Eqs. (14)– (16) are finite nonnegative functionals of $D_i(K, q)$. ■

Proposition 2 (Reference-consistency): If

$$z_i(t_k) \geq m_{b,-u_i,\tau_i}(q) \quad \text{for all } k \text{ such that } \phi_i(t_k) \in P_b, \tag{31}$$

then $S_i(c) = 0$ for all summaries in $\mathfrak{S}$.

Proof. Under Eq. (31), the numerator in Eq. (13) is nonpositive at every time point. Therefore $d_i(t_k; K, q) = 0$ for all k, and every summary in $\mathfrak{S}$ is zero. ■

Proposition 3 (Deficit monotonicity): Let $d, d' \in R_+^T$ satisfy $d_k \leq d'_k$ for all k. Then

$$S(d) \leq S(d') \qquad \text{for every } S \in \mathfrak{S}. \tag{32}$$

Proof. The result is immediate for the mean and high-deficit fraction. Empirical quantiles are monotone under component-wise ordering, so the 95-th percentile is monotone. Order statistics also preserve component-wise ordering, giving $d_{(\ell)} \leq d'_{(\ell)}$ for all ℓ. Eq. (29) gives monotonicity of the tail mean. A convex hybrid of monotone summaries is monotone. ■

Proposition 4 (Bounded perturbation stability): For two held-out speed signals evaluated against the same reference profile, suppose

$$\max_k |z(t_k) - z'(t_k)| \leq \delta. \tag{33}$$

Then

$$\max_k |d(t_k; K, q) - d'(t_k; K, q)| \le \delta/\varepsilon_s. \tag{34}$$

Consequently, S^{mean}, S^{p95}, S^{tail}, and any convex hybrid of these summaries change by at most δ/ε_s.

Proof. The positive-part map is 1-Lipschitz. Since every denominator in Eq. (13) is at least ε_s, Eq. (34) follows. The mean, empirical quantile, and upper-tail mean are 1-Lipschitz under the sup-norm on the deficit vector. Convex combinations preserve the same bound. ■

The high-deficit fraction is threshold based. If $|d(t_k) - 1| > \eta$ for all k and $\max_k |d(t_k) - d'(t_k)| < \eta$, then $S^{\text{frac}}(d) = S^{\text{frac}}(d')$.

Proposition 5 (Phase-reference specificity): For two phase bins $P_b \neq P_{b'}$, the same transformed speed value may receive different penalties:

$$z_i(t_k) = z_i(t_\ell), \quad \phi_i(t_k) \in P_b, \quad \phi_i(t_\ell) \in P_{b'} \quad \not\Rightarrow \quad d_i(t_k; K, q) = d_i(t_\ell; K, q). \tag{35}$$

The deficits depend not only on $z_i(t_k)$ but also on the phase-specific reference center and scale in Eq. (13). Thus, PS-SDVI penalizes speed relative to the expected phase of the same task, not as an unconditioned raw low-speed quantity.

Proposition 6 (Global low-speed exposure as a special case): If $K = 1$ and the robust scale is computed from the single global training reference, Eq. (13) reduces to

$$d_i(t_k; 1, q) = \left[\frac{m_{1,-u_i,\tau_i}(q) - z_i(t_k)}{s_{1,-u_i,\tau_i}(q)}\right]_+. \tag{36}$$

Thus, global low-speed exposure is a special one-phase case of PS-SDVI.

Proof. When $K = 1$, the partition contains only $P_1 = [0,1]$. Every time point is compared with the same global reference center and scale. Eq. (13) therefore becomes Eq. (36). ■

For leakage control, define the outer training information for operator u and task τ as

$$\mathcal{F}_{-u,\tau} = \sigma\{(X_i, y_i, g_i): u_i \neq u,\ \tau_i = \tau\}. \tag{37}$$

Theorem 1 (Held-out-operator separation): For every outer held-out operator u and task τ, the selected candidate $\hat{c}_{u,\tau}$, the final reference profile, the standardized outcome reference, and the poor-quality threshold are $\mathcal{F}_{-u,\tau}$-measurable.

Proof. The inner validation sets in Eq. (19) exclude the outer held-out operator u. Eqs. (20)–(24) compute score standardization, poor-quality labels, AUC, Spearman association, and the inner objective using only operators in $\mathcal{U}_{-u}$. Hence $\hat{c}_{u,\tau}$ is determined by $\mathcal{F}_{-u,\tau}$. The final reference profile and outer evaluation thresholds are rebuilt from $\mathcal{I}_{-u,\tau}$, which also excludes u. Therefore, all selection, thresholding, and reference-estimation operations are $\mathcal{F}_{-u,\tau}$-measurable. ■

Proposition 7 (Finite candidate well-posedness): If $\mathcal{K}$, $\mathcal{Q}$, and the hybrid-weight grid are finite, then $\mathcal{C}$ is finite. If at least one candidate has an estimable inner objective, then

$$\hat{c}_{u,\tau} \in \arg\max_{c \in \mathcal{C}} \mathcal{L}_{u,\tau}(c) \tag{38}$$

exists.

Proof. The candidate set $\mathcal{C} = \mathcal{K} \times \mathcal{Q} \times \mathfrak{S}$ is finite under the validation grid. The objective $\mathcal{L}_{u,\tau}(c)$ is real-valued for every estimable candidate. A real-valued function over a finite nonempty set attains a maximum. ■

Proposition 8 (Inner-objective dominance over included fixed candidates): Let $c_0 \in \mathcal{C}$ be any fixed comparator included in the candidate family, such as the fixed 15-bin p95 score in Eq. (27). Then

$$\mathcal{L}u, \tau(\hat{c}u, \tau) \ge \mathcal{L}_{u,\tau}(c_0). \tag{39}$$

Proof. The result follows directly from Eq. (38), since $\hat{c}_{u,\tau}$ maximizes $\mathcal{L}_{u,\tau}(c)$ over all $c \in \mathcal{C}$, and $c_0 \in \mathcal{C}$. ■

Propositions 1–8 establish the mathematical basis of Nested PS-SDVI within the proposed agenda. The score is finite, nonnegative, zero when no phase-specific speed deficit is present, monotone with increasing deficit exposure, stable under bounded perturbations, phase-specific by construction, and a strict generalization of global low-speed exposure. The nested selection procedure is well posed and separated from the held-out operator.

These properties establish that Nested PS-SDVI is a finite, nonnegative, phase-specific, leakage-controlled viability score whose selected candidate is well defined under finite-grid validation.

2.5. Dataset and Signal Construction

The empirical instantiation uses a ROSMA subset with 46 dVRK trials from three operators across Pea-on-a-Peg, Post-and-Sleeve, and Wire-Chaser I tasks. ROSMA contains dVRK kinematic recordings for three common surgical training tasks, with kinematic variables recorded at 50 Hz and task-evaluation annotations based on completion time and task-specific errors (Rivas-Blanco et al., 2024). The present study uses only the kinematic trial files and task-level score, time, and error annotations.

For trial i, let the left and right patient-side manipulator positions be

$$p_i^L(t_k) = \left(x_i^L(t_k), y_i^L(t_k), z_i^L(t_k)\right), \qquad p_i^R(t_k) = \left(x_i^R(t_k), y_i^R(t_k), z_i^R(t_k)\right). \tag{40}$$

A centered moving average is applied to each Cartesian coordinate before numerical differentiation. With sampling interval $\Delta t = 1/f_s$, velocities are computed by finite differences:

$$\dot{p}_i^a(t_k) = \frac{p_i^a(t_k) - p_i^a(t_{k-1})}{\Delta t}, \qquad a \in \{L, R\}. \tag{41}$$

The bimanual speed signal used by Nested PS-SDVI is

$$v_i(t_k) = \left\|\dot{p}_i^L(t_k)\right\|_2 + \left\|\dot{p}_i^R(t_k)\right\|_2. \tag{42}$$

This signal is winsorized at extreme empirical quantiles and transformed through Eq. (10). The normalized phase coordinate $\phi_i(t_k)$ is then obtained from Eq. (2), so each trial is represented by a phase-indexed speed sequence.

For baseline comparisons, three additional kinematic exposure channels are constructed. Let $\dddot{p}_i^a(t_k)$ denote the finite-difference jerk of arm a. Define

$$J_i(t_k) = \left\|\dddot{p}_i^L(t_k)\right\|_2 + \left\|\dddot{p}_i^R(t_k)\right\|_2, \tag{43}$$

$$A_i(t_k) = \left|\left\|\dot{p}_i^L(t_k)\right\|_2 - \left\|\dot{p}_i^R(t_k)\right\|_2\right|, \tag{44}$$

and

$$C_i(t_k) = \frac{1}{\left\|p_i^L(t_k) - p_i^R(t_k)\right\|_2 + \varepsilon_c}, \qquad \varepsilon_c > 0. \tag{45}$$

Here, $J_i(t_k)$ is bimanual jerk exposure, $A_i(t_k)$ is speed-asynchrony exposure, and $C_i(t_k)$ is inverse bimanual-closeness exposure. Each baseline channel is winsorized and transformed with $\log(1 + x)$, matching the preprocessing applied to the speed signal.

The task-level score y_i, completion time, and error count are aligned by trial identifier. Scores are oriented so that larger values indicate lower-quality execution. For an outer held-out operator u and task τ, the train-standardized within-task score is

$$y_{i,-u,\tau}^{(z)} = \frac{y_i - \text{median}\{y_j : j \in \mathcal{I}_{-u,\tau}\}}{\text{MAD}\{y_j : j \in \mathcal{I}_{-u,\tau}\}}, \qquad i : u_i = u,\ \tau_i = \tau. \tag{46}$$

The outer poor-quality label is

$$b_{i,-u,\tau} = \mathbf{1}\{y_i \geq Q_{0.75}(y_j : j \in \mathcal{I}_{-u,\tau})\}. \tag{47}$$

Both $y_{i,-u,\tau}^{(z)}$ and $b_{i,-u,\tau}$ are computed from outer-training operators only. Thus, the held-out operator contributes only to evaluation, not to score orientation, standardization, or threshold construction.

Raw baseline scores are obtained by applying the same trial-level p95 summary to low-speed, jerk, asynchrony, and closeness exposure channels. These baselines provide the empirical comparison needed to determine whether phase-normalized speed-deficit viability adds information beyond unnormalized kinematic summaries.

2.6. Validation Protocol and Algorithmic Pipeline

Nested PS-SDVI is evaluated under outer leave-one-operator-out validation. For each outer held-out operator u and task τ, the candidate $\hat{c}_{u,\tau}$ is selected by Eq. (25), the final reference profile is rebuilt from $\mathcal{I}_{-u,\tau}$, and the selected score $\hat{S}_i$ is computed for each held-out trial $i: u_i = u, \tau_i = \tau$. The held-out evaluation set is

$$\mathcal{H} = \{i: u_i = u,\ \tau_i = \tau,\ u \in \mathcal{U}, \tau \in \mathcal{T}\}. \tag{48}$$

For each task τ, poor-quality detection is evaluated by

$$\mathrm{AUC}_\tau = \mathrm{AUC}\left(\{b_{i,-u_i,\tau}\}_{i:\tau_i=\tau}, \{\hat{S}_i\}_{i:\tau_i=\tau}\right), \tag{49}$$

and continuous agreement is evaluated by

$$r_\tau = \mathrm{Spearman}\left(\{\hat{S}_i\}_{i:\tau_i=\tau}, \left\{y^{(z)}_{i,-u,\tau}\right\}_{i:\tau_i=\tau}\right). \tag{50}$$

The macro validation metrics are

$$\mathrm{AUC}_{\mathrm{macro}} = |\mathcal{T}|^{-1} \sum_{\tau \in \mathcal{T}} \mathrm{AUC}_\tau, \qquad r_{\mathrm{macro}} = |\mathcal{T}|^{-1} \sum_{\tau \in \mathcal{T}} r_\tau. \tag{51}$$

Cross-task robustness is reported by

$$\mathrm{AUC}_{\mathrm{min}} = \min_{\tau \in \mathcal{T}} \mathrm{AUC}_\tau. \tag{52}$$

Bootstrap confidence intervals are computed from held-out trial scores. In each bootstrap replicate, trials are sampled with replacement from $\mathcal{H}$, and the macro metrics in Eq. (51) are recomputed. The empirical 2.5-th and 97.5-th percentiles define the reported 95% intervals.

Raw baseline scores are evaluated under the same held-out folds. For a baseline exposure channel $E_i(t_k)$, the trial-level baseline score is

$$B_i^E = Q_{0.95}\{E_i(t_k): k = 1, \ldots, T_i\}. \tag{53}$$

The evaluated baseline channels are raw low-speed, jerk, speed asynchrony, and inverse bimanual closeness. Fixed PS-SDVI comparators are obtained by replacing $\hat{c}_{u,\tau}$ with prespecified candidates, including the fixed 15-bin 95-th-percentile score in Eq. (27). The complete computational workflow, including signal construction, nested candidate selection, held-out scoring, baseline evaluation, and bootstrap uncertainty estimation, is summarized in Algorithm 1.

Algorithm 1. Nested PS-SDVI validation

Input: trial trajectories $\mathcal{X}_i$, oriented ROSMA scores y_i, task labels τ_i, operator labels u_i, candidate grid $\mathcal{C}$, sampling frequency f_s.

Output: selected held-out scores $\hat{S}_i$, task-level AUC values, task-level Spearman values, macro metrics, and baseline comparisons.

i. Extract left and right Cartesian tool positions for each trial.
ii. Smooth Cartesian coordinates and compute $v_i(t_k), J_i(t_k), A_i(t_k)$, and $C_i(t_k)$ by Eqs. (42)–(45).
iii. For each outer held-out operator u and task τ:
    a. Construct $\mathcal{I}_{-u,\tau}$ by Eq. (4).
    b. For each candidate $c \in \mathcal{C}$, perform inner user-level validation using Eqs. (19)–(24).
    c. Select $\hat{c}_{u,\tau}$ by Eq. (25).
    d. Rebuild the task-specific reference profile from $\mathcal{I}_{-u,\tau}$.
    e. Compute $\hat{S}_i = S_i(\hat{c}_{u,\tau})$ for all held-out trials $i: u_i = u, \tau_i = \tau$.
    f. Compute $y^{(z)}_{i,-u,\tau}$ and $b_{i,-u,\tau}$ by Eqs. (46)–(47).
iv. Pool held-out scores across outer folds.
v. Compute task-level and macro validation metrics using Eqs. (49)–(52).
vi. Compute baseline scores using Eq. (53) and evaluate them under the same held-out folds.

Estimate bootstrap confidence intervals for macro-AUC and macro-Spearman.

The computational order is

$$\mathcal{O}\left(|\mathcal{U}|, |\mathcal{T}|, |\mathcal{C}|, N_{\mathrm{tp}}\right), \tag{54}$$

where $N_{\mathrm{tp}} = \sum_i T_i$ is the total number of trajectory time points. The procedure is finite because the candidate grid, operator set, task set, and observed trajectories are finite. It also preserves the single validation agenda: every reported score is computed for held-out operators after reference construction and candidate selection have been completed without using their trials.

3. RESULTS

3.1. Validation Design and Fold Configuration

The evaluation used 46 ROSMA trials from three operators across Pea-on-a-Peg, Post-and-Sleeve, and Wire-Chaser I tasks. The data structure supported eight outer held-out-operator folds: three folds for Pea-on-a-Peg, three folds for Post-and-Sleeve, and two folds for Wire-Chaser I because X03 trials were not available for that task. Figure 1 summarizes the trial coverage across operators and tasks.

Table 1 reports the fold-level configuration selected by the nested validation procedure. In each outer fold, the held-out operator was excluded from reference-profile construction, poor-quality threshold construction, score standardization, and candidate selection. The selected configurations varied across folds, with phase-bin counts ranging from 8 to 18 and score summaries including high-fraction, mean, 95th percentile, hybrid, and tail-mean forms. This confirms that the final score was not imposed as a fixed global statistic; it was selected within the training data for each held-out evaluation setting.

3.2. Distribution of Nested PS-SDVI Exposure

Figure 2 shows the distribution of selected Nested PS-SDVI exposure across the three ROSMA tasks. Wire-Chaser I produced the highest exposure distribution, whereas Pea-on-a-Peg and Post-and-Sleeve showed lower median values and broader overlap. This pattern is consistent with the purpose of PS-SDVI: the score measures phase-specific speed-deficit exposure relative to a good-quality task reference, not raw speed alone.

The separation across tasks indicates that the proposed score preserves task structure. Larger values represent stronger deviation from the learned reference speed profile, and therefore correspond to greater motion-quality burden under the phase-normalized viability formulation.

3.3. Baseline Comparison for Poor-Quality Detection

Nested PS-SDVI achieved the strongest poor-quality detection performance among all evaluated methods. Table 2 and Figure 3A show that Nested PS-SDVI reached a macro-AUC of 0.762, with a bootstrap 95% confidence interval of [0.522, 0.844]. The fixed 15-bin PS-SDVI p95 comparator achieved a macro-AUC of 0.738, and the selected-reference PS-SDVI p95 score achieved 0.715.

Raw kinematic baselines were substantially weaker. Raw closeness p95 reached a macro-AUC of 0.521, raw low-speed p95 reached 0.509, raw jerk p95 reached 0.433, and raw asynchrony p95 reached 0.416. Thus, the proposed nested phase-normalized score provided the strongest discrimination of poor-quality trials under held-out-operator validation.

3.4. Association with Within-Task ROSMA Score

The continuous association results showed the same ordering. Nested PS-SDVI achieved the highest macro-Spearman association with train-standardized within-task ROSMA score, reaching 0.570 with a bootstrap 95% confidence interval of [0.119, 0.653] (Table 2; Figure 3B). The fixed PS-SDVI p95 and fixed PS-SDVI tail-mean comparators achieved lower macro-Spearman values of approximately 0.384. Raw closeness and raw low-speed showed weak positive associations, while raw jerk and raw asynchrony showed negative macro associations.

Figure 5 further supports this result at the trial level. Higher train-standardized ROSMA scores, representing poorer within-task performance, were associated with higher selected Nested PS-SDVI scores. This relationship is expected under the proposed formulation: trials with poorer motion quality should exhibit stronger or more persistent deficits relative to the task-specific good-quality reference profile.

3.5. Cross-Task Robustness

The minimum task AUC for Nested PS-SDVI was 0.597. This value is important because it shows that the macro-AUC was not produced by a single favorable task alone. The fixed 15-bin PS-SDVI p95 score had a slightly higher minimum task AUC of 0.653, but its macro-Spearman association was lower than that of Nested PS-SDVI. This indicates a clear trade-off: the fixed p95 score is competitive for binary poor-quality detection, whereas nested candidate selection better captures continuous variation in ROSMA performance.

The combined results support the use of nested selection as the primary formulation. It improves the association with within-task performance while retaining stronger poor-quality detection than the raw kinematic baselines.

3.6. Task-Specific Candidate Selection

Table 3 and Figure 4 summarize the nested candidate-selection behavior of NPS-SDVI. Table 3 reports the median selected configuration by task, while Figure 4 gives the fold-level distribution of the selected phase-bin counts and score summaries.

Figure 4A shows that the selected phase resolution was task dependent. Pea-on-a-Peg selected 12 phase bins in two folds and 15 phase bins in one-fold. Post-and-Sleeve selected 8 phase bins in two folds and 18 phase bins in one-fold. Wire-Chaser I selected 15 phase bins in both available folds. These selections indicate that the nested procedure did not impose a single temporal resolution across all surgical tasks. Instead, the number of phase bins was adapted within the training folds according to the task-specific speed-deficit structure.

Figure 4B shows the corresponding selected score summaries. Pea-on-a-Peg selected three different summaries across folds: high-fraction, hybrid, and mean. Post-and-Sleeve selected high-fraction, mean, and tail mean. Wire-Chaser I selected the 95th-percentile summary in both folds. This pattern supports the design of NPS-SDVI as a task-adaptive viability score. The selected score was not fixed globally as a p95 or mean statistic; it was chosen through inner user-level validation before being applied to the outer held-out operator.

Together, Table 3 and Figure 4 show that task heterogeneity affects both the temporal resolution and the deficit summary used by NPS-SDVI. This is consistent with the central methodological premise: different robotic-surgery training tasks impose different movement rhythms and speed-control demands, so phase-specific speed-deficit viability should be selected within the task rather than imposed as a universal raw kinematic threshold.

Nested PS-SDVI provided the strongest overall validation profile among the evaluated motion-quality scores. It achieved the highest macro-AUC for poor-quality detection and the highest macro-Spearman association with train-standardized within-task ROSMA score. Its advantage over raw low-speed, jerk, asynchrony, and closeness baselines indicates that the useful signal is not explained by a single unnormalized kinematic exposure. Instead, performance improves when speed-deficit exposure is measured relative to task-specific phase references and selected under a leakage-controlled nested validation design.

Together, the results support Nested PS-SDVI as an interpretable phase-normalized framework for robotic surgical motion-quality assessment under held-out-operator validation. The evidence establishes the value of nested phase-specific speed-deficit viability in the present ROSMA setting and provides a structured basis for broader validation on larger robotic-surgery datasets.

Table 1. Nested PS-SDVI fold configuration and held-out performance.

Held-Out Operator	ROSMA Task	Train/Test Trials	Phase Bins	Ref. Quantile	Score Summary
X01	Pea-on-a-Peg	12/6	15	0.45	High-fraction
X01	Post-and-Sleeve	10/6	8	0.25	High-fraction
X01	Wire-Chaser I	6/6	15	0.35	95th percentile

X02	Pea-on-a-Peg	12/6	12	0.45	Mean
X02	Post-and-Sleeve	10/6	8	0.25	Mean
X02	Wire-Chaser I	6/6	15	0.35	95th percentile
X03	Pea-on-a-Peg	12/6	12	0.45	Hybrid summary
X03	Post-and-Sleeve	12/4	18	0.25	Tail mean

Table 2. Cross-user validation of nested PS-SDVI against motion-quality baselines.

Method	Macro AUC	AUC 95% CI	Macro Spearman	Spearman 95% CI	Minimum Task AUC
Nested PS-SDVI	0.762	[0.522, 0.844]	0.570	[0.119, 0.653]	0.597
Fixed PS-SDVI p95, 15 bins	0.738	[0.450, 0.817]	0.384	[-0.018, 0.519]	0.653
PS-SDVI p95, selected reference	0.715	[0.441, 0.792]	0.323	[-0.024, 0.489]	0.583
Fixed PS-SDVI tail mean, 15 bins	0.674	[0.483, 0.795]	0.384	[0.049, 0.571]	0.438
PS-SDVI tail mean, selected reference	0.644	[0.485, 0.794]	0.341	[0.040, 0.554]	0.438
Fixed PS-SDVI mean, 15 bins	0.544	[0.427, 0.748]	0.301	[0.086, 0.554]	0.188
Raw closeness p95	0.521	[0.288, 0.671]	0.012	[-0.241, 0.378]	0.188
Raw low-speed p95	0.509	[0.189, 0.564]	0.063	[-0.264, 0.333]	0.125
Raw jerk p95	0.433	[0.304, 0.669]	-0.416	[-0.617, -0.113]	0.042
Raw asynchrony p95	0.416	[0.182, 0.500]	-0.467	[-0.760, -0.317]	0.083

Table 3. Task-level nested configuration selected by NPS-SDVI.

ROSMA task	Held-Out Folds	Median Phase Bins	Median Reference Quantile	Selected Score Summary
Pea-on-a-Peg	3	12	0.45	High-fraction; hybrid; mean
Post-and-Sleeve	3	8	0.25	High-fraction; mean; tail mean
Wire-Chaser I	2	15	0.35	95th percentile

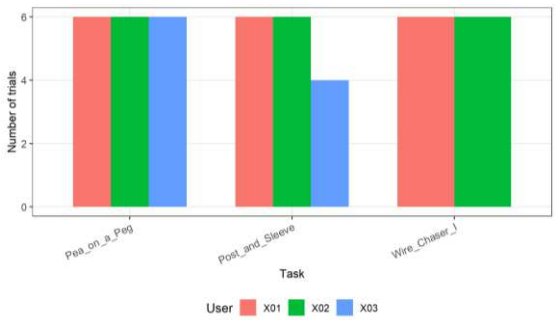


Figure 1. ROSMA trial coverage across operators and tasks.

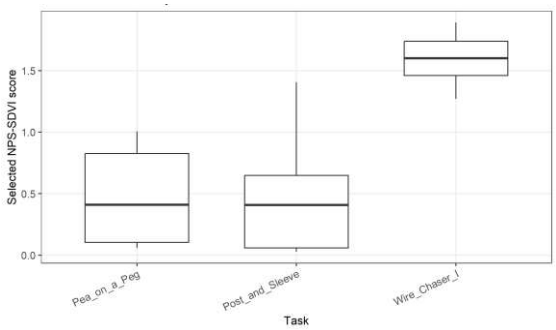


Figure 2. Task-wise distribution of nested PS-SDVI exposure.

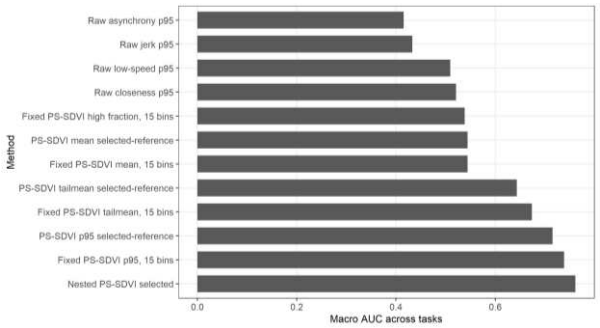


Figure 3A. Macro poor-quality detection.

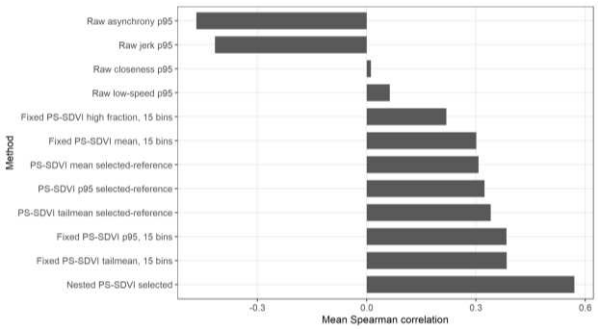


Figure 3B. Macro association with within-task score.

Figure 3. Cross-user validation of nested PS-SDVI against motion-quality baselines.

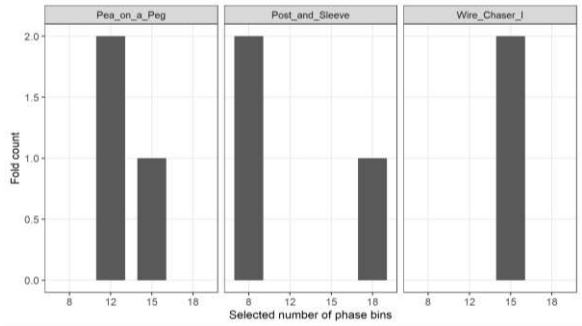


Figure 4A. Selected phase-bin counts.

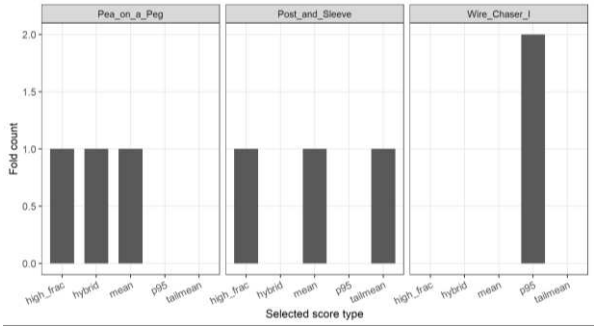


Figure 4B. Selected score summaries.

Figure 4. Task-specific candidate selection under nested validation.

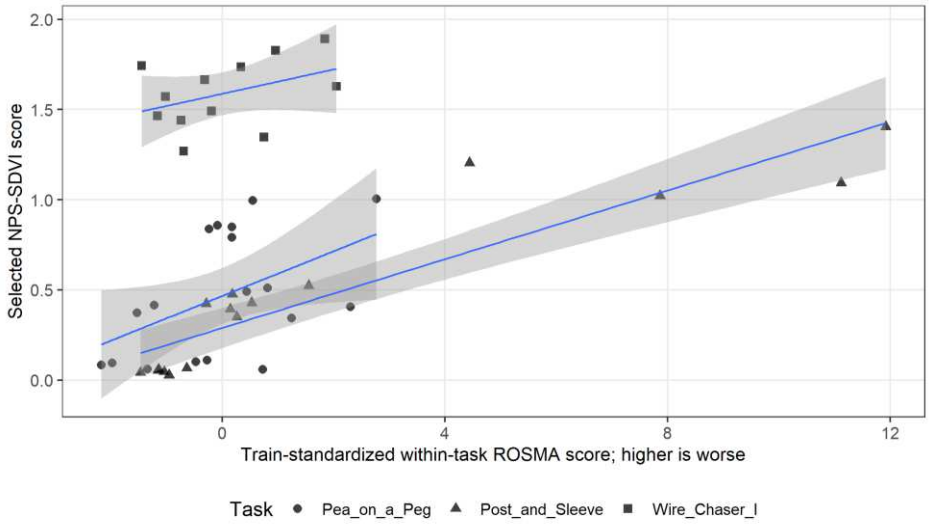


Figure 5. Relationship between nested PS-SDVI and train-standardized ROSMA score.

4. DISCUSSION

4.1. Main Finding

This study developed Nested PS-SDVI to assess robotic surgical motion quality from phase-normalized speed-deficit exposure. The main finding is that the nested score produced the strongest overall validation profile among all evaluated motion-quality measures. It achieved the highest macro-AUC for poor-quality trial detection and the highest macro-Spearman association with train-standardized within-task ROSMA score (Table 2; Figure 3A, B). This joint result is important because the method did not improve only a binary endpoint. It also preserved continuous agreement with within-task performance variation.

The comparison with raw kinematic baselines clarifies the contribution. Raw low-speed, jerk, asynchrony, and closeness summaries were weaker than Nested PS-SDVI. This indicates that the relevant signal is not captured by a single unconditioned kinematic magnitude. The improvement appears when speed deficit is evaluated relative to a task-specific phase reference and when phase resolution and score summary are selected inside a held-out-operator validation design.

4.2. Why Phase-Specific Deficit Matters

The central premise of PS-SDVI is that speed has different meaning at different phases of a robotic surgical task. A low-speed segment can indicate careful manipulation in one phase and inefficient execution in another. A global speed statistic cannot make this distinction because it removes the temporal context in which the movement occurred.

Nested PS-SDVI addresses this problem by comparing each time point with a good-quality reference speed profile for the same task and phase. The score therefore measures deficit relative to expected task-phase behavior, not raw slowness alone. This distinction is supported by the baseline comparison in Table 2. Raw low-speed exposure had weak macro performance, whereas Nested PS-SDVI produced the strongest macro-AUC and macro-Spearman. The result supports the methodological claim that phase-conditioned deficit measurement is more informative than unconditioned low-speed exposure.

4.3. Task-Specific Exposure Structure

The task-wise exposure pattern in Figure 2 shows that Wire-Chaser I had the highest selected NPS-SDVI distribution, while Pea-on-a-Peg and Post-and-Sleeve had lower and more overlapping distributions. This pattern should be interpreted as task-specific deficit burden within the present validation setting, not as a universal task-difficulty ranking.

The important point is that the score retained task structure. Because the reference profile is learned within each task, a high score reflects deviation from that task's good-quality phase profile. This gives the

metric an interpretable meaning: higher Nested PS-SDVI values correspond to stronger or more persistent speed-deficit exposure relative to the expected motion pattern for that task.

4.4. Leakage-Controlled Candidate Selection

Nested candidate selection is not a secondary implementation detail; it is part of the methodological contribution. The method selects the phase-bin count, reference quantile, and score summary inside the training operators before evaluation on the outer held-out operator. This design prevents the evaluated operator from influencing reference construction, thresholding, standardization, or candidate selection. Table 3 and Figure 4 show why these matters. Figure 4A shows that selected phase resolution differed across tasks: Pea-on-a-Peg mainly selected 12 phase bins, Post-and-Sleeve mainly selected 8 bins, and Wire-Chaser I selected 15 bins. Figure 4B shows that the selected score summaries also differed across tasks. Pea-on-a-Peg used high-fraction, hybrid, and mean summaries; Post-and-Sleeve used high-fraction, mean, and tail mean; Wire-Chaser I used the 95th-percentile summary.

These selections support the central design logic. Robotic-surgery training tasks do not share identical temporal structure. A fixed global summary may be adequate for some tasks, but it cannot represent all phase-dependent movement patterns equally well. Nested PS-SDVI handles this heterogeneity by selecting the temporal resolution and deficit summary within the training fold, then applying the selected candidate to the held-out operator.

4.5. Added Value of Nested Selection over Fixed PS-SDVI

The fixed 15-bin PS-SDVI p95 comparator was competitive for poor-quality detection, with a macro-AUC close to the nested score. However, it had a lower macro-Spearman association with train-standardized within-task ROSMA score (Table 2). This difference is informative. The fixed p95 summary captures strong upper-tail deficits and can perform well for binary discrimination, but it is less flexible in representing continuous variation across tasks.

Nested PS-SDVI improved the continuous association while retaining the best macro-AUC. This suggests that candidate selection contributes most strongly when the goal is not only to separate poor-quality trials but also to preserve the ordered relationship between motion-deficit exposure and within-task performance. The nested formulation is therefore better aligned with motion-quality assessment than a fixed universal p95 rule.

4.6. Relationship to Raw Kinematic Measures

The raw kinematic baselines provide a useful negative control. Jerk, asynchrony, closeness, and raw low-speed exposure are common motion descriptors, but their trial-level p95 forms did not reproduce the performance of Nested PS-SDVI. In the present validation, raw jerk and raw asynchrony showed negative macro-Spearman associations, while raw closeness and raw low-speed were close to uninformative for continuous score association.

This does not imply that these signals are irrelevant to surgical motion quality. Rather, it shows that unnormalized global summaries are insufficient for the present held-out-operator setting. The difference between raw low-speed and PS-SDVI is especially important. PS-SDVI includes low-speed exposure as a special case when phase structure is removed, but its stronger performance shows that phase-specific referencing adds information beyond the global low-speed baseline.

4.7. Methodological Implications

The proposed framework has three methodological implications for robotic surgical motion analysis. First, motion-quality scores should preserve task phase when the task contains structured sub-movements. Second, reference profiles should be learned from good-quality trials within the same task rather than from all available trials pooled together. Third, validation must separate model selection from held-out evaluation, especially when datasets contain few operators.

Nested PS-SDVI satisfies these requirements in a transparent way. Its score is interpretable because it is built from positive speed deficits. Its validation is controlled because the held-out operator does not contribute to reference construction or candidate selection. Its task adaptation is explicit because the selected phase bins and summaries can be inspected fold by fold.
In this sense, Nested PS-SDVI contributes a trajectory-level robotic assessment layer: it converts raw dVRK motion into a phase-conditioned viability signal that can be compared across held-out operators without using their data during reference construction or candidate selection.

5. Limitations and Future Work

The validation used 46 trials from three operators across three ROSMA tasks, with incomplete Wire-Chaser I coverage for X03. This sample size limits the precision of the bootstrap intervals and constrains the breadth of cross-operator generalization. The results therefore characterize Nested PS-SDVI within the present held-out-operator ROSMA setting. Larger multi-operator validation is needed to assess stability across broader training populations and task executions.
The method uses normalized task phase rather than manually annotated gestures. This design allows PS-SDVI to operate when gesture labels are unavailable, but gesture-aware or subtask-aware references may improve alignment for tasks with heterogeneous internal structure. Future work should evaluate Nested PS-SDVI on larger ROSMA cohorts and external robotic-surgery datasets, and should compare normalized-phase references with gesture-level or subtask-level reference profiles.
Nested PS-SDVI provides an interpretable phase-normalized viability score for robotic surgical motion-quality assessment. Its advantage over raw kinematic baselines shows that lower-quality trials are better characterized by task-phase-specific speed-deficit exposure than by unconditioned speed, jerk, asynchrony, or closeness summaries. The nested candidate-selection results further show that phase resolution and score summary should be adapted within training folds rather than imposed globally. Together, the findings support Nested PS-SDVI as a coherent scoring framework for held-out-operator robotic-surgery training analysis.

6. CONCLUSION

This paper introduced Nested Phase-Specific Speed-Deficit Viability (Nested PS-SDVI) as a phase-normalized scoring method for robotic surgical motion-quality assessment. The method addresses a specific limitation of global kinematic summaries: the same speed value can have different meanings at different stages of a surgical task. Nested PS-SDVI resolves this by learning task-specific good-quality reference speed profiles and measuring positive speed-deficit exposure within normalized task phases. The nested design is central to the contribution. Phase-bin resolution, reference quantile, and score summary is selected inside training operators, while the evaluated operator remains excluded from reference construction, thresholding, score standardization, and candidate selection. This makes the score both task-adaptive and leakage-controlled, which is essential for cross-user robotic-surgery validation.
The empirical results support the proposed formulation. Nested PS-SDVI achieved the strongest validation profile among the evaluated methods, with the highest macro-AUC for poor-quality detection and the highest macro-Spearman association with train-standardized within-task ROSMA score. Its advantage over raw low-speed, jerk, asynchrony, and closeness baselines shows that motion quality is better represented by phase-conditioned speed-deficit exposure than by unconditioned global kinematic measures.
Overall, Nested PS-SDVI provides a mathematically defined, interpretable, and validation-controlled trajectory-level assessment layer for robotic surgical training analysis. The findings support phase-normalized viability scoring as a useful direction for objective robotic motion assessment and provide a clear basis for larger multi-operator validation and gesture-aware extensions.

Declarations
Ethics approval and consent to participate

Not applicable. This study involved secondary analysis of publicly available, de-identified robotic-surgery training data from the ROSMA dataset. No new participants were recruited, no clinical intervention was performed, and no identifiable personal information was accessed.

Consent for publication

Not applicable. The manuscript contains no identifiable personal information, images, videos, or participant-specific details.

Availability of data and materials

The dataset analysed during the current study is publicly available in the ROSMA repository at https://github.com/irivas-uma/rosma.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no funds, grants, or other financial support for the research, authorship, or publication of this article.

Authors' contributions

A.A. contributed to conceptualization, methodology, software, formal analysis, validation, visualization, and preparation of the original manuscript. A.A.A. contributed to methodological refinement, interpretation of the findings, supervision, and critical review and editing of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the public availability of the ROSMA robotic-surgery dataset, which enabled the empirical validation reported in this study.

References

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Béjar Haro, B., Zappella, L., Khudanpur, S., Vidal, R., & Hager, G. D. (2017). A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering, 64*(9), 2025–2041. https://doi.org/10.1109/TBME.2016.2647680

Boal, M. W., Anastasiou, D., Tesfai, F., Ghamrawi, W., Mazomenos, E., Curtis, N., & Francis, N. K. (2024). Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: A systematic review. *British Journal of Surgery, 111*(1), znad331. https://doi.org/10.1093/bjs/znad331

Chen, J., Cheng, N., Cacciamani, G. E., Oh, P. J., Lin-Brande, M., Remulla, D., Gill, I. S., & Hung, A. J. (2019). Objective assessment of robotic surgical technical skill: A systematic review. *The Journal of Urology, 201*(3), 461–469. https://doi.org/10.1016/j.juro.2018.06.078

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery, 14*, 1611–1617. https://doi.org/10.1007/s11548-019-02039-4

Funke, I., Mees, S. T., Weitz, J., & Speidel, S. (2019). Video-based surgical skill assessment using 3D convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery, 14*(7), 1217–1225. https://doi.org/10.1007/s11548-019-01995-1

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D. D., Chen, C. C. G., Vidal, R., Khudanpur, S., & Hager, G. D. (2014). JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A surgical activity dataset for human motion modeling. *M2CAI Workshop, MICCAI*.

Goh, A. C., Goldfarb, D. W., Sander, J. C., Miles, B. J., & Dunkin, B. J. (2012). Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of Urology, 187*(1), 247–252. https://doi.org/10.1016/j.juro.2011.09.032

Kankanamge, D., Wijeweera, C., Ong, Z., Preda, T., Carney, T., Wilson, M., & Preda, V. (2025). Artificial intelligence based assessment of minimally invasive surgical skills using standardised objective metrics:

A narrative review. *The American Journal of Surgery, 241*, Article 116074. https://doi.org/10.1016/j.amjsurg.2024.116074
Lam, K., Chen, J., Wang, Z., Iqbal, F. M., Darzi, A., Lo, B., Purkayastha, S., & Kinross, J. M. (2022). Machine learning for technical skill assessment in surgery: A systematic review. *npj Digital Medicine, 5*, Article 24. https://doi.org/10.1038/s41746-022-00566-0
Lavanchy, J. L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., & Beldi, G. (2021). Automation of surgical skill assessment using a three-stage machine learning algorithm. *Scientific Reports, 11*, Article 5197. https://doi.org/10.1038/s41598-021-84295-6
Lefor, A. K., Harada, K., Dosis, A., & Mitsuishi, M. (2020). Motion analysis of the JHU-ISI Gesture and Skill Assessment Working Set using Robotics Video and Motion Assessment Software. *International Journal of Computer Assisted Radiology and Surgery, 15*, 2017–2025. https://doi.org/10.1007/s11548-020-02259-z
Lin, H. C., Shafran, I., Yuh, D., & Hager, G. D. (2006). Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery, 11*(5), 220–230. https://doi.org/10.3109/10929080600989189
Martin, J. A., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., & Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery, 84*(2), 273–278. https://doi.org/10.1046/j.1365-2168.1997.02502.x
Nguyen, X. A., Ljuhar, D., Pacilli, M., Nataraja, R. M., & Chauhan, S. (2019). Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer Methods and Programs in Biomedicine, 177*, 1–8. https://doi.org/10.1016/j.cmpb.2019.05.008
Pan, M., Wang, S., Li, J., Li, J., Yang, X., & Liang, K. (2023). An automated skill assessment framework based on visual motion signals and a deep neural network in robot-assisted minimally invasive surgery. *Sensors, 23*(9), Article 4496. https://doi.org/10.3390/s23094496
Reiley, C. E., Lin, H. C., Yuh, D. D., & Hager, G. D. (2011). Review of methods for objective surgical skill evaluation. *Surgical Endoscopy, 25*(2), 356–366. https://doi.org/10.1007/s00464-010-1190-z
Rivas-Blanco, I., López-Casado, C., Herrera-López, J. M., Cabrera-Villa, J., & Pérez-del-Pulgar, C. J. (2024). Instrument detection and descriptive gesture segmentation on a robotic surgical maneuvers dataset. *Applied Sciences, 14*(9), Article 3701. https://doi.org/10.3390/app14093701
Rosen, J., Solazzo, M., Hannaford, B., & Sinanan, M. (2001). Objective laparoscopic skills assessments of surgical residents using hidden Markov models based on haptic information and tool/tissue interactions. *Studies in Health Technology and Informatics, 81*, 417–423.
Tao, L., Elhamifar, E., Khudanpur, S., Hager, G. D., & Vidal, R. (2012). Sparse hidden Markov models for surgical gesture classification and skill evaluation. In *Information Processing in Computer-Assisted Interventions* (pp. 167–177). Springer. https://doi.org/10.1007/978-3-642-30618-1_17
Trejos, A. L., Patel, R. V., Malthaner, R. A., & Schlachta, C. M. (2014). Development of force-based metrics for skills assessment in minimally invasive surgery. *Surgical Endoscopy, 28*, 2106–2119. https://doi.org/10.1007/s00464-014-3442-9
Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective assessment of surgical technical skill and competency in the operating room. *Annual Review of Biomedical Engineering, 19*, 301–325. https://doi.org/10.1146/annurev-bioeng-071516-044435
Wang, Z., & Fey, A. M. (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery, 13*, 1959–1970. https://doi.org/10.1007/s11548-018-1860-1
Zappella, L., Béjar, B., Hager, G., & Vidal, R. (2013). Surgical gesture classification from video and kinematic data. *Medical Image Analysis, 17*(7), 732–745. https://doi.org/10.1016/j.media.2013.04.007
Zia, A., & Essa, I. (2018). Automated surgical skill assessment in RMIS training. *International Journal of Computer Assisted Radiology and Surgery, 13*, 731–739. https://doi.org/10.1007/s11548-018-1735-5