MEASURE, CALIBRATE, REFUSE—AND VERIFY: AN EVIDENCE CONTRACT FOR MOBILE AQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployment-facing action quality assessment combines repeated observations, user adaptation, human review, physical measurements, and abstention, yet these components are often evaluated as if their evidence were independent. We formulate an *evidence contract*: every adaptive or human-assisted degree of freedom requires a matching source of evaluation independence. We instantiate the contract in a longitudinal audit of an eight-month mobile rehabilitation system and obtain four paired reversals. Same-user threshold fitting raised accuracy from 0.738 to 0.918, whereas nested leave-one-user-out calibration achieved 0.557. Pooled repetition agreement of 0.937 became 0.748 when users received equal weight. On the same 23 videos and frozen predictions, model agreement changed from 0.913 under model-visible review to 0.391 against blind consensus; 51 videos were retrospectively labeled, of which 38 entered frozen-rule scoring after exclusions, yielding error-detection accuracies of 0.167 and 0.357 for two exercises. Finally, within-user correlation of 0.807 coexisted with centimeter limits of agreement $[-7.1, +4.2]$, while refusal retained 0.137 user-level risk at 0.964 coverage. The resulting evaluation stack—user-level aggregation, nested adaptation, blind-first labeling, measurement agreement, and coverage–risk analysis—turns apparently reliable outputs into testable stage-wise claims. It applies beyond rehabilitation wherever systems personalize, expose machine suggestions to reviewers, or withhold predictions.

1 INTRODUCTION

Physical rehabilitation exercises must be performed correctly, yet home coaching systems observe users through a single phone camera, with scarce labels, changing viewpoints, and substantial anthropometric variation. Action quality assessment (AQA) increasingly combines learned scores, explicit rubrics, uncertainty, and human review (Pirsiavash et al., 2014; Parmar & Morris, 2019; Yu et al., 2021; Xu et al., 2022; Majeedi et al., 2024). Deployment, however, asks a different question from benchmark ranking: what evidence licenses a system to measure, adapt, issue a verdict, or abstain for a particular user?

The obstacle is not simply distribution shift. Each useful system capability creates a corresponding route for evaluation dependence. Repeated observations can let prolific users dominate pooled metrics; personalization can fit the person later used for testing; model-visible review can pull labels toward the prediction; correlation can reward association without physical agreement; and a reject option can preserve high coverage without concentrating errors in the refused set. Each mechanism can make a sensible component look reliable while leaving its deployment claim untested.

We unify these mechanisms in an *evidence contract*: every adaptive or human-assisted degree of freedom requires a matching source of evaluation independence. Repeated observations require user-level inference; user-specific parameters require new-user holdout or temporally separated calibration and test data; model-assisted review requires labels recorded before model exposure; physical-unit claims require agreement with a reference; and abstention requires risk reported across coverage. The contract converts a single end-to-end accuracy claim into a sequence of stage-wise claims, each paired with the separation or reference needed to test it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We instantiate this contract by auditing an eight-month mobile rehabilitation effort that evolved from synthetic regression into an interpretable *measure, calibrate, refuse* (MCR) pipeline. Four matched comparisons isolate distinct dependencies while holding predictions or inputs fixed where possible. Same-user threshold fitting raises accuracy from 0.738 to 0.918, but nested leave-one-user-out calibration reaches 0.557. Equal weighting by user changes pooled agreement from 0.937 to 0.748. On the same 23 videos, agreement changes from 0.913 under model-visible review to 0.391 against blind consensus. A within-user correlation of 0.807 coexists with limits of agreement $[-7.1, +4.2]$,cm, and 0.964 coverage retains 0.137 user-level selective risk. Together, these interventions reveal which dependency produced each apparent gain.

This work makes three contributions. **First**, it introduces a deployment-facing evidence contract linking five common system freedoms to five concrete evaluation checks. **Second**, it provides paired empirical audits—user balancing, nested calibration, blind-first re-labeling, metrological agreement, and coverage–risk analysis—that expose failures hidden by pooled accuracy and interpretable output. **Third**, it turns those diagnoses into system decisions: use global parameters until stable reference-measured bias supports personalization, preserve blind judgments before assistance, and design refusal scores against observed error modes. The current data are development and retrospective-audit evidence, not prospective clinical validation; the methodological contribution is the contract that determines what must be verified next.

2 Related work

Action and rehabilitation assessment. AQA regresses or ranks expert judgments from video (Pirsiavash et al., 2014; Parmar & Morris, 2019; Tang et al., 2020; Yu et al., 2021; Xu et al., 2022); rehabilitation systems often operate on skeleton sequences (Vakanski et al., 2018; Capecci et al., 2019; Li et al., 2024; Bruce et al., 2021) with GCN-based models (Yan et al., 2018; Yu et al., 2022; Bruce et al., 2024). Recent work such as RICA^2 explicitly models scoring rubrics and calibrates predictive uncertainty (Majeedi et al., 2024). Interpretable and rule-based assessment also predates our system (Zhao et al., 2017; Fieraru et al., 2021; Ienaga et al., 2022). Our concern is complementary: whether the measurement, adaptation, label, and refusal evidence remains independent enough to support a deployment claim.

Dependence in evaluation and annotation. In longitudinal health data, record-wise validation can strongly overestimate new-user performance; the split must approximate the intended use case (Saeb et al., 2017; Little et al., 2017). The same principle applies after model fitting: a reviewer who sees the prediction is no longer an independent reference. Controlled work on model-assisted annotation shows that suggestions shift label distributions and can inflate the apparent performance of the suggesting model (Schroeder et al., 2025). We test both dependencies directly through user-level aggregation, nested access control, and a matched anchored-versus-blind label comparison.

Measurement, calibration, and abstention. Correlation measures association rather than agreement between measurement methods; bias and limits of agreement instead expose absolute error relative to the intended decision scale (Bland & Altman, 1986). This metrological sense of calibration—correcting systematic measurement error against a reference—differs from confidence calibration (Guo et al., 2017) and RICA^2's alignment of predictive uncertainty with error (Majeedi et al., 2024). Rejection rules and selective classification formalize abstention (Chow, 1970; El-Yaniv et al., 2010), while conformal methods offer distribution-aware alternatives (Vovk et al., 2005). Our evidence contract connects these literatures by asking, at each stage, which reference, split, or coverage condition makes the claimed reliability identifiable.

3 Setting and framework

System. Given a single-camera repetition or static hold, the system returns a coached-error verdict, the underlying measured quantity, and an evidence frame, or refuses with `not_observable` or `not_calibrated`. Each rule $(m_d, \mathcal{B}_d)$ applies a deterministic pose-sequence feature m_d—side selection, visibility filtering, smoothing, and temporal aggregation—and a physical-unit band $\mathcal{B}_d$. MediaPipe supplies 33 joints per frame (Lugaresi et al., 2019; Bazarevsky et al., 2020); the mo-

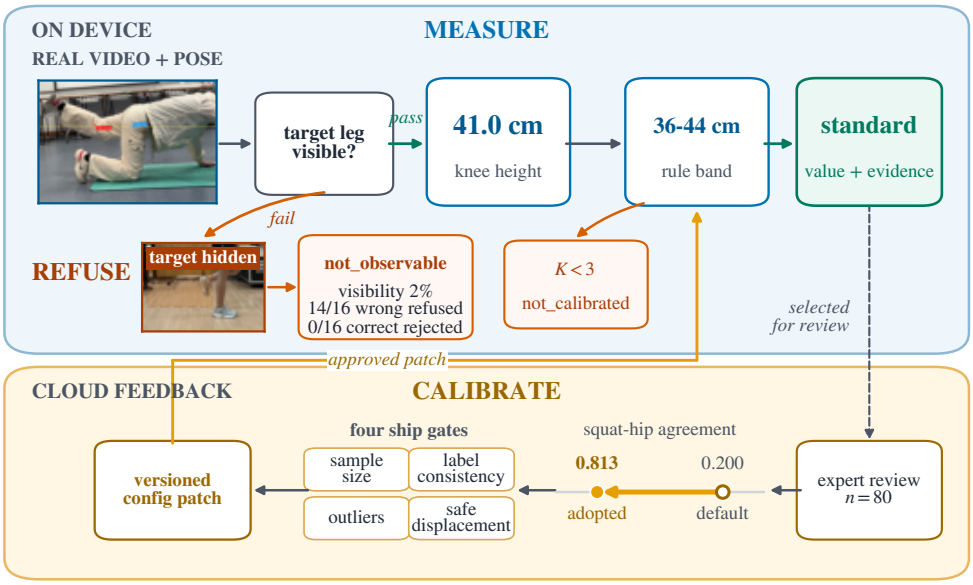


Figure 1: **The audited measure–calibrate–refuse pipeline.** The device exposes a physical-unit measurement and evidence or refuses; reviewed cases can propose versioned cloud-to-device patches. Values are audited development cases: the visibility guard was designed and evaluated on the same 32 videos, and the $0.200 \rightarrow 0.813$ squat-hip patch was selected on C3. Downstream analyses test each stage under its matched independence condition.

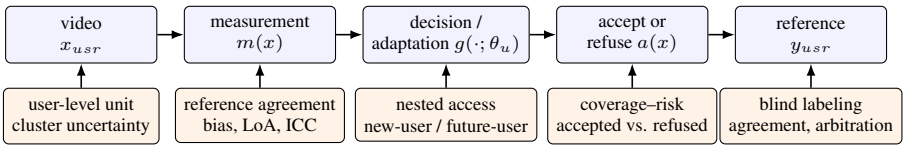


Figure 2: **Audit map.** An interpretable output exposes a chain of claims, not a certificate of validity. Each stage has a matched check and can fail while downstream accuracy appears high.

bile and cloud implementations share a versioned rule schema. This one-quantity–one-band design exposes intermediate measurements for audit, but does not guarantee their validity.

The object being audited. Let x_{usr} be repetition r from session s and user u. The system computes a quantity $m(x_{usr})$, optionally adapts parameters θ_u from a calibration set, emits $\hat{y}_{usr} = g(m(x_{usr}); \theta_u)$, and accepts when a guard $a(x_{usr}) = 1$. Its reference label is y_{usr}. This decomposition separates five claims that a single accuracy number conflates: whether m measures the intended construct; whether g transfers to a new user; whether adaptation generalizes beyond the samples that fitted θ_u; whether y was obtained independently of $\hat{y}$; and whether accepted predictions are safer than refused ones.

Estimands and independence. For loss ℓ, pooled repetition risk weights prolific users more heavily. We therefore make the user the primary unit,

$$R_{\text{user}} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{n_u} \sum_{s,r} \ell(\hat{y}_{usr}, y_{usr}). \tag{1}$$

For a guard setting τ, coverage is $C(\tau) = P[a_\tau(x) = 1]$ and selective risk is $R_{\text{sel}}(\tau) = E[\ell(\hat{y}, y) \mid a_\tau(x) = 1]$; a refusal mechanism is useful only if lowering coverage lowers the relevant risk, not merely if some bad views are rejected. For new-user evaluation, every fitted component of θ must use users outside the held-out fold. For within-user adaptation, θ_u may use only an earlier calibration prefix and is scored on later repetitions. Test-user fitting is reported only as an oracle ceiling.

Table 1: **Development readings and their audit counterparts.** All evidence is developmental or retrospective; none is a prospective population estimate.

Reading	Development evaluation	Stricter evaluation
C3 rule patches	macro $0.663 \rightarrow 0.758$	same four users informed design; nested thresholds, not independent validation
C4 agreement	0.937 pooled repetitions	0.748 user-balanced; 0.391 blind consensus on matched videos
Per-user calibration	$0.738 \rightarrow 0.918$ test-user oracle	0.557 nested LOSO; 0.806 within-user $K{=}10$ split
Blind labels (H1)	raw agreement $23/29$, $17/22$	$\kappa{=}0.45$, 0.51; rule accuracy 0.167, 0.357
Centimeter estimate	within-user $r{=}0.807$	bias -1.4 cm; LoA $[-7.1, +4.2]$ cm; ICC 0.59–0.64
Refusal guards	$14/16$ wrong views refused on design data	coverage 0.964 with per-user selective risk 0.137

Corpora. C1 contains 613 videos from 52 app users and was repeatedly consulted during synthetic-model development. C2 contains 61 laboratory videos from three subjects with grid-measured knee height. C3 contains 544 multi-exercise videos from four subjects and informed threshold, feature, and guard design. C4 contains 156 selected valid repetitions from 27 videos and 10 users whose reviewers saw the automatic verdict. H1 retrospectively re-labels 51 C4-pool videos with two blinded raters and arbitration; $45/51$ videos and $11/12$ users overlap C4, so only the labels are independent. A frozen prospective protocol over new users (C5) was not executed. Appendix A gives sampling and label details.

Audit protocol. We distinguish development analyses (C1–C4), retrospective audits (H1 and the C2 grid re-analysis), controlled simulation (EGPA), and a public benchmark (DeepRehabPile); none is a prospective clinical test. Five checks target the five claims in Figure 2: thresholds are evaluated with nested user holdout; repetition-level results are re-aggregated with users as the statistical unit; historical videos are re-labeled without model outputs; physical estimates are compared with a grid reference using bias and limits of agreement (LoA), rather than correlation alone; and refusals are evaluated through coverage–risk. User-level uncertainty uses cluster bootstrap intervals (2,000 resamples, seed 0). All protocol choices that could benefit from the evaluated outcomes are marked as development or retrospective. Table 1 pairs each original reading with its audit counterpart; full protocols and tables are in the appendix.

Executable sequence on historical data. The audit is designed to remain informative when new-user collection is unavailable. First, freeze predictions, rule versions, and a provenance ledger recording which users influenced each design choice. Second, recompute every headline metric at the deployment unit and bootstrap that unit rather than individual rows. Third, replay each fitted component under the strongest split supported by timestamps and user identifiers: other-user folds for transfer, or an earlier labeled prefix for within-user adaptation. Fourth, obtain labels for a sampled historical subset through a blind-first interface, retain rater disagreement, and reveal the model only after the independent decision. Fifth, evaluate physical measurements against any archived reference and report refusal as a coverage–risk curve. The output is a claim ledger with one of four statuses—development-only, contradicted, promising but non-independent, or supported—rather than a single replacement accuracy. Historical replay can falsify a deployment claim or expose its dependence structure; it cannot convert participating users into a prospective cohort.

4 Results

We first show why the system adopted an explicit measurement chain, then pair each favorable development reading with the independence check required by its deployment claim.

4.1 Synthetic shift motivates an auditable measurement chain

Synthetic-only regression did not transfer. Five compact regressors (0.022–0.395M parameters) were trained on increasingly varied synthetic pose data while C1 served as a repeatedly consulted real-data development set. No version mean beat guess-the-mean MAE 0.2233: normalization reduced real-data MAE from 0.416 ± 0.073 to 0.255 ± 0.014, but the best sweep reached only 0.234 ± 0.026. Within-source MSE was 0.0315 versus 0.0865 across sources ($+175\%$); absolute-

coordinate features dominated correlations, synthetic validation continued improving after real validation degraded, and a label-mean shift appeared across real splits. These observations diagnose shortcut and protocol sensitivity in this development path, not learning methods in general (Appendix A.2).

A common-input comparison favored geometry, not calibration. C1's manual scores were not deterministically generated from the MediaPipe landmarks used by later regressors: the synthetic-label formula itself obtained MAE 0.252 on those landmarks, while C1-trained linear maps reached 0.171–0.183. On C2, all methods instead used the same cached landmarks, grid-derived three-class target, and leave-one-user-out splits. Direct geometry reached 0.738 accuracy and macro-F1 0.571, above majority (0.656), logistic regression (0.525), ST-GCN (0.528 ± 0.184), and an MLP (0.580 ± 0.075). Measured-torso calibration scored 0.721 and nested threshold calibration 0.557 (Table 2). With only three users, the direct-geometry accuracy interval is 0.45–0.90; the result supports this representation on this pilot, not a general rules-over-learning claim (full per-class results: Appendix A.4, Table 8).

4.2 Dependencies behind favorable development metrics

Three mechanisms inflated later development readings. First, C3's four subjects supported threshold search, LOSO screening, and feature redesign; its $0.663 \rightarrow 0.758$ macro therefore describes the design corpus, not new users (Appendix A.3). Second, C4 pooled 156 repetitions although one user contributed 85 (54.5%) at 0.977 agreement, allowing that user to dominate the 0.937 result. Third, C4 reviewers saw the automatic label, measurement, and evidence frame, with the automatic answer pre-selected. On 23 videos later given independent blind labels, this anchored workflow agreed with the model on $21/23$ (0.913). Each mechanism is corrected below without changing the underlying predictions.

4.3 Matched independence checks reverse the readings

Calibration gains reverse as evaluation becomes stricter. Figure 3 separates four regimes: thresholds fitted and scored on the test user (M1, oracle), fitted and scored on all rows (M2, pooled), fitted on other users (M3, nested LOSO), and fitted on a user's first $K{=}10$ labeled repetitions before scoring later repetitions (M4, deployment-available). On C2, default accuracy 0.738 rises to 0.918 under the oracle but falls to 0.557 under nested LOSO; nested calibration hurts two of three users. The within-user $K{=}10$ split reaches 0.806, but is a same-user temporal estimate, not new-user evidence. C3's six dynamic rules show the same ordering, although less sharply. Features were designed on all of C3 and C2 has three users, so even nested values remain pilot estimates. The reversal identifies the source of the gain: access to the evaluated user's labels, not a threshold that transfers across users.

User-balanced aggregation removes single-user leverage. For online C4, agreement falls from 0.937 over pooled repetitions to 0.748 when each user receives equal weight (user-bootstrap 95% CI 0.49–0.97); leaving one online user out reduces the latter to 0.691. Figure 4 shows the same pattern for the full 10-user set. The predictions and labels are unchanged—only the estimand changes from "a random recorded repetition" to "a random observed user." The 0.189-point gap is therefore attributable to sample concentration, not model behavior.

Blinded labels reverse the anchored reading. Two raters independently re-labeled 51 historical videos without model outputs; a third adjudicated disagreements. These are independent labels, not independent users: $45/51$ videos and $11/12$ users overlap C4. After excluding missing outputs and invalid consensus labels, 38 videos remain. Under the pre-specified primary any-error aggregation, frozen-rule error-detection accuracy is 0.167 on bird-dog (24 videos) and 0.357 on kneeling leg raise (14); majority- and all-error sensitivity analyses do not change the conclusion (Appendix A.5). On 23 videos matched across workflows, model agreement is 0.913 under anchored review and 0.391 against blind consensus. This matched-video contrast holds both the inputs and frozen predictions fixed; only the label-production process changes. Seven of the 12 overturned bird-dog "agreements" are online videos where the anchored reviewer followed an automatic "over" verdict that blind con-

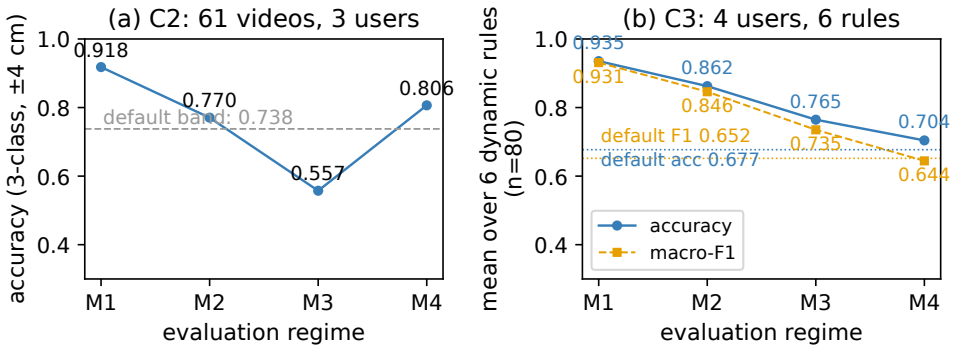


Figure 3: **Calibration depends on evaluation access.** M1: same-user oracle; M2: pooled in-sample fit; M3: nested new-user LOSO; M4: first $K{=}10$ repetitions to later repetitions of the same user. Oracle and pooled fits overstate new-user transfer; M4 measures a different, within-user deployment claim. Dashed lines mark defaults.

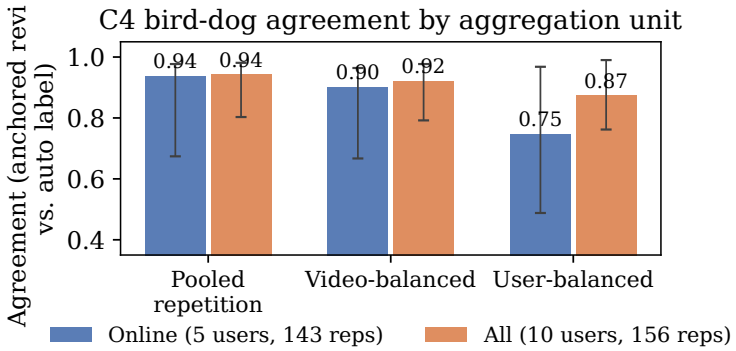


Figure 4: **C4 agreement by aggregation unit.** User balancing removes the influence of a single user who supplied 54.5% of valid repetitions; whiskers are user-clustered 95% CIs.

sensus read as standard. The kneeling misses trace to segmentation (2), pose-feature failure (3), and threshold mismatch (3), showing why calibration alone cannot repair them.

4.4 Stage-specific diagnostics localize the remaining errors

Correlation concealed measurement disagreement. Against C2's grid reference, the deployed centimeter estimator has mean bias $-1.4\,$cm, MAE $2.6\,$cm, RMSE $3.2\,$cm, and 95% limits of agreement $[-7.1, +4.2]\,$cm, with ICC$(2,1){=}0.59$ and ICC$(3,1){=}0.64$. The limits are wider than the $40 \pm 4\,$cm coaching decision band despite within-user correlation $r{=}0.807$. Per-user biases ($-1.2/-4.2/+0.9\,$cm) implicate pose estimation as the largest identifiable source, but three subjects make every index a pilot estimate (Appendix A.4, Figure 8).

Blinded raters exposed label uncertainty. Raw agreement was $23/29$ for bird-dog and $17/22$ for kneeling, but Cohen's κ was only 0.45 and 0.51 under strongly skewed class distributions; 11 videos required arbitration. Thus neither anchored agreement nor a single expert workflow provides an unquestioned reference.

Refusal missed the dominant errors. The original quad-stretch guard refused $14/16$ wrong-view videos and no right-view videos, but was designed and evaluated on those same 32 videos. Across 311 historical bird-dog and kneeling repetitions, only nine (2.9%) were refused. User-level coverage was 0.964 and accepted-set selective risk 0.137 (95% CI 0.02–0.20); threshold sweeps left risk

Table 2: **Strict evidence behind the two model-facing conclusions.** Left: common-input C2 comparison under user holdout; direct geometry is strongest on this three-user pilot, while calibration does not improve it. Right: frozen rules against blind H1 consensus (primary any-error aggregation), plus the matched-video label-workflow contrast. Complete per-class and sensitivity results are in Appendix A.4 and Appendix A.5.

C2: common input, LOSO

Method	Acc.	Macro-F1
Majority	0.656	0.264
Logistic	0.525	0.413
ST-GCN (5 seeds)	0.528±0.184	0.242
MLP (5 seeds)	0.580±0.075	0.316
Direct geometry	**0.738**	**0.571**
+ measured torso	0.721	0.545
+ nested thresholds	0.557	0.473

H1: blind consensus, video level

Action	Users	Videos	κ	Acc./F1
Bird-dog	8	24	0.45	0.167/0.161
Kneeling	8	14	0.51	0.357/0.354

Matched bird-dog videos

Label workflow	Videos	Agreement
Anchored review	23	0.913
Blind consensus	23	0.391

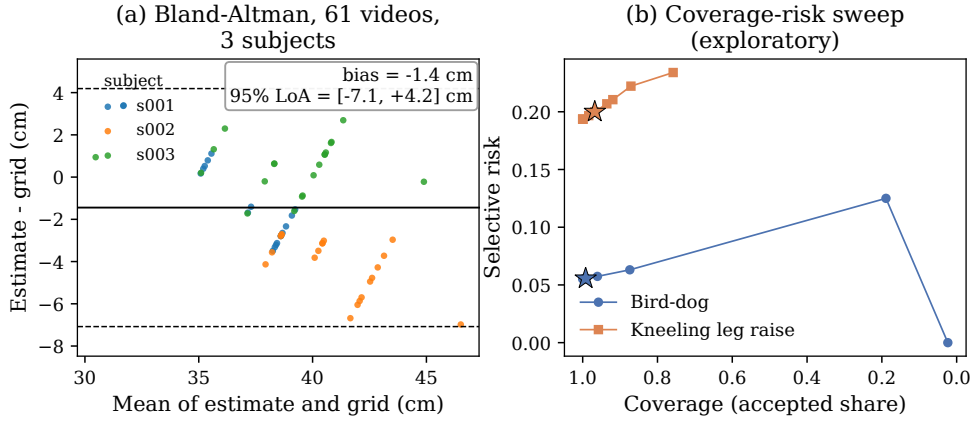


Figure 5: **Measurement and refusal require their own diagnostics.** Left: Bland–Altman differences against the C2 grid reference; correlation cannot reveal the subject-specific offsets or limits of agreement. Right: exploratory threshold sweeps on judgeable historical repetitions; stars mark frozen settings. Risk does not improve over the useful high-coverage region, while the apparent zero-risk bird-dog endpoint accepts almost no repetitions.

nearly flat until coverage collapsed. The guards captured a narrow observability failure, not the segmentation, pose, and boundary failures dominating the blinded audit (Appendix A.7).

4.5 Evidence-gated adaptation and refusal

Calibration requires evidence of stable bias. The C2 oracle suggests that user-specific offsets can be corrected, but the gain disappears under new-user evaluation. On DeepRehabPile's 14 regression datasets (161 subjects), per-user fitting never beats memorizing the user's historical mean ($29/161$ wins, p=1); on UI-PRMD it beats the LOUO global fit for only $37/60$ users. Conversely, a 60-user simulation recovers gains when stable per-user bias is injected and adds at most 0.009 MAE without it (Appendix A.8). These testbeds establish neither necessity nor sufficiency. They support a conservative rule: retain global parameters unless bias is measured against a reference and shown to persist across sessions.

Interpretability enables stage-wise verification. The one-quantity–one-band representation localized errors to segmentation, measurement, threshold, or observability. The blind labels, nested evaluation, and coverage analysis then identified which exposed component lacked supporting evidence. MCR's value is therefore structural: it makes each component separately testable under the contract.

Table 3: **Claim ladder for deployment-facing assessment.** "Disposition" describes only the evidence in this study.

Claim	Required separation or reference	Diagnostic	Disposition here
Construct is measured	physical or expert reference independent of prediction	bias, LoA, ICC; failure cases	wider than decision band
Calibration transfers	users held out from all fitting choices	user-level risk and uncertainty	fails C2 new-user test
Personalization helps	earlier labeled prefix, later same-user test	paired gain over global/default	within-user pilot
Verdict is correct	blind labels with disagreement preserved	matched agreement; reliability	fails H1 blind test
Refusal reduces risk	accepted and refused cases across thresholds	coverage–risk curve	risk remains flat
Clinical benefit	prospective target population and outcomes	functional/safety endpoint	prospective test required

4.6 THE EVIDENCE CONTRACT

The four reversals are not four estimates of one latent "true accuracy." They intervene on different dependencies: user weighting changes the target population; nested calibration removes access to the evaluated user; blinding removes access of the reference process to the prediction; measurement agreement replaces rank association with absolute error; and coverage–risk conditions performance on acceptance. Their convergence is the result: favorable development readings relied on at least one dependency incompatible with the deployment claim.

Table 3 turns the findings into a claim ladder. A stage may advance only when the evidence in its row is available; passing a later-looking software check cannot compensate for a failed earlier claim. In particular, a calibrated threshold cannot repair a quantity that is not measuring the intended geometry, and a refusal guard cannot confer safety when its score is unrelated to the dominant errors.

This ladder turns the audit into a release rule. The system currently supports an auditable research instrument: intermediate quantities localize failures, and historical deployment supplies realistic stress cases. Advancing it to a clinical assessor requires the references and separations in the final column, rather than another pooled metric. The contract therefore converts the 0.937 agreement and 0.918 oracle result from endpoints into diagnostics of what the next evaluation must isolate.

5 DISCUSSION

Independence is stage-specific. The results support one organizing principle rather than a collection of reporting preferences. User balancing targets dependence among repeated observations; nesting removes outcome access from adaptation; blinding separates the reference from the prediction; agreement analysis supplies an external physical scale; and coverage–risk tests whether abstention selects easier cases. None substitutes for another. A clean user split cannot repair anchored labels, and a calibrated confidence score cannot validate the quantity being measured.

The contract changes system design. The synthetic regressors exposed source shortcuts, while geometry made the three-user grid pilot easier to diagnose. The useful division is not rules versus learning: learning remains essential for perception and segmentation, where the blind audit found several failures. A stronger architecture can learn perception, expose the measured quantity, adapt only after stable reference-measured bias, and train refusal against observed error modes. Each capability then ships only with its matched evaluation.

Label interfaces are evaluation infrastructure. Moderate inter-rater agreement and the anchored-to-blind reversal show that the interface can change the apparent model. A blind-first workflow should record an independent judgment and uncertainty before revealing the prediction; assisted review can follow, but must remain a separate quantity. Preserving disagreement and arbitration then makes label uncertainty visible instead of folding it into model agreement.

Diagnosis allocates the next experiment. The calibration reversal prioritizes a pre/post within-user study over further cross-user threshold search. The measurement analysis prioritizes pose accuracy and reference quality; the blind taxonomy points to segmentation and feature repair; and the

Table 4: **Robustness checks for the two reference-dependent conclusions.** Left: on the same 38 H1 videos, changing the definition of a video-level error raises apparent accuracy only by shrinking the positive class; error recall falls from 0.429 to 0.143. Right: replacing the deployed $50\,\text{cm}$ torso default with measured torso lengths leaves C2 agreement with the grid essentially unchanged. Thus neither label aggregation nor anthropometric personalization rescues the audited claim.

H1: sensitivity to video aggregation

Error definition	Acc.	Macro-F1	Prec.	Recall
Any (primary)	0.237	0.232	0.222	0.429
Majority	0.421	0.394	0.250	0.286
All	0.605	0.474	0.400	0.143

C2: measurement arm, 61 videos

Torso	Bias	MAE	RMSE	LoA	ICC(3,1)
Default	-1.4	2.6	3.2	$[-7.1, 4.2]$	0.64
Measured	-1.4	2.6	3.2	$[-7.2, 4.4]$	0.63

flat coverage–risk curve motivates learning observability from those failures rather than tuning the present guards. An exposed measurement chain therefore converts a weak aggregate result into a specific experimental program.

Prospective validation follows the same contract. A frozen new-user evaluation can estimate user-balanced blind accuracy; repeated reference measurements can test whether LoA fits the decision band; a pre/post within-user protocol can test persistent adaptation gains; and a redesigned observability score can be judged by decreasing risk as coverage falls. The evidence contract fixes these estimands before new data arrive, making the prospective study a direct test rather than another round of metric selection.

Robustness checks isolate the bottleneck. On C2, measured torso lengths alter bias and RMSE by less than $0.05\,\text{cm}$; the residual is dominated by pose-estimation error. Hip height matched the reference by construction and off-plane videos were re-shot, so neither deployment source is estimated. H1's eight kneeling misses divide into segmentation (2), pose-feature failure (3), threshold mismatch (3), and observability (0). Thus personalization, calibration, and the present guard each leave observed failure modes unresolved.

Retrospective audits are powerful falsification tools. Historical data permit controlled interventions: re-weight the same predictions by user, relabel the same videos blind, and replay the same calibration routine without held-out access. These pairs isolate why a headline moved. They do not estimate future-user performance because capture, design, and record availability already encode selection. Retrospective checks are therefore efficient gates for prospective validation: they eliminate unsupported claims while reserving new data for hypotheses that survive.

Scope beyond rehabilitation. The same dependency pattern arises in personalized education, wellness sensing, human-in-the-loop moderation, and any system receiving many records from few people. Whenever an output is adapted, shown to a reviewer, or filtered by a reject option, the unit, access pattern, reference, and coverage should be specified together. The pose features are domain-specific; the evidence contract is not.

6 Conclusion

Deployment-facing assessment needs an evidence contract: user-level inference for repeated observations, nested or temporal separation for adaptation, blind references for assisted review, agreement analysis for physical measurements, and risk for abstention coverage. Applying this contract reversed four favorable readings and localized their dependencies. The transferable result is to measure, calibrate, and refuse only to the extent that each stage has been independently verified.

Limitations. C1 was repeatedly consulted; C3's four subjects informed its thresholds, features, and guards; C2 has three subjects; and C4 is selected and dominated by one user. H1 blinds labels on old videos, not new users. Interventions are retrospective, and EGPA tests injected assumptions. Reliability is moderate where measured; clinical, functional, safety, and adherence outcomes were not evaluated. All measurements use MediaPipe and may vary across devices, viewpoints, pose backends, and populations. Without the frozen prospective C5 protocol, results remain development or retrospective audits, not population accuracy or clinical effectiveness.

AI USE STATEMENT

Generative AI tools were used under author supervision for feedback on hypothesis framing and evaluation methodology; implementation and debugging of rule, calibration, analysis, and figure-generation code; checks on statistical interpretation; and drafting, editing, and compression of the manuscript. The GLM vision API was additionally evaluated as an experimental comparator in Appendix A.12. Models did not supply participant data or substitute for expert labels. Reported results come from the documented project pipelines and dated run reports; the authors inspected the generated code, verified manuscript numbers against those artifacts, checked cited sources, made all scientific decisions, and take responsibility for the paper.

ETHICS STATEMENT

Participants were adults recruited through an open university-wide call; no patients, minors, or vulnerable populations were enrolled. Participation was voluntary and compensated under terms disclosed before enrollment. All participants provided documented informed consent that explicitly covered research analysis and publication. No health records or diagnostic data were collected. Under the applicable institutional policy and local regulations, this data-collection activity did not require review by an institutional ethics committee; consequently, no approval identifier exists. Underlying videos are not released, and only aggregate statistics over de-identified measurements are reported. The still frames in Figure 1 were separately permitted for publication and pixelated to obscure identity and surroundings. The system is decision-support rather than a medical device, and this study makes no claim of clinical benefit.

REPRODUCIBILITY STATEMENT

The underlying C1–C4 videos are proprietary consented program data and cannot be released; the production rule engine, mobile app, and calibration service are likewise not included. To make the reported analyses independently checkable, the supplementary material (an anonymous, self-contained `submission/` bundle) contains: de-identified pose landmarks and derived measurements for the C2 grid-truth corpus (61 videos, 3 subjects), the C4 reflow labels with pseudonymized user/record identifiers, the 51-video blinded dual-rater consensus labels, the EGPA simulator and its 60-user closed-loop results, and the evaluation scripts behind the audits that are reproducible from the bundle. Running `run_all.sh` regenerates the main tables, the Bland–Altman figure, the pooled-vs-user figures, and the calibration-strictness figure from the bundled data (CPU only, no network); seeds, bootstrap settings, pinned dependency versions, and the per-result expected numbers are documented in its `README.md`, and every shipped file is covered by `MANIFEST.sha256`. Results that are *not* reproducible from the bundle are stated as such there: the C1 five-version regression runs (training logs and weights were not archived), the C3 11-rule patched/LOSO table (the MediaPipe environment behind the 544-video cache is deprecated; the cached per-rule measurements are included and the four calibration modes are recomputable), the VLM study (requires API credentials), the refusal-guard stress test (requires internal C3 landmarks), and the historical coverage–risk audit of the production guards (its script and the per-repetition landmarks are internal; the bundled `labels.csv` carries the underlying reflow labels). Released derived data are de-identified (real names, absolute paths, and production identifiers removed); public release remains subject to organizational privacy review.

REFERENCES

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. arxiv 2020. *arXiv preprint arXiv:2006.10204*, 2020.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

XB Bruce, Yan Liu, Keith CC Chan, Qintai Yang, and Xiaoying Wang. Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer's progression. *Pattern Recognition*, 119:108095, 2021.

XB Bruce, Yan Liu, Keith CC Chan, and Chang Wen Chen. Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6471–6485, 2024.

Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on neural systems and rehabilitation engineering*, 27(7):1436–1448, 2019.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9914–9923. IEEE, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Naoto Ienaga, Shuhei Takahata, Kei Terayama, Daiki Enomoto, Hiroyuki Ishihara, Haruka Noda, and Hiromichi Hagihara. Development and verification of postural control assessment using deep-learning-based pose estimators: Towards clinical applications. *Occupational Therapy International*, 2022(1):6952999, 2022.

Ali Ismail-Fawaz, Maxime Devanne, Stefano Berretti, Jonathan Weber, and Germain Forestier. A standardized benchmark for skeleton-based rehabilitation assessment using deep learning. In *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2026.

Jianwei Li, Jun Xue, Rui Cao, Xiaoxia Du, Siyu Mo, Kehao Ran, and Zeyan Zhang. Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3184–3193. IEEE, 2024.

Max A. Little, Gael Varoquaux, Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. Using and understanding cross-validation strategies: Perspectives on saeb et al. *GigaScience*, 6(5), 2017. doi: 10.1093/gigascience/gix020.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. RICA^2: Rubric-informed, calibrated assessment of actions. In *European Conference on Computer Vision*, pp. 143–161. Springer, 2024. doi: 10.1007/978-3-031-73036-8_9.

Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 304–313. IEEE, 2019.

Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *European conference on computer vision*, pp. 556–571. Springer, 2014.

Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5), 2017. doi: 10.1093/gigascience/gix019.

Hope Schroeder, Deb Roy, and Jad Kabbara. Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25771–25795. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1323.

Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9836–9845. IEEE, 2020.

Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1):2, 2018.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2939–2948. IEEE, 2022.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

BX Yu, Yan Liu, Xiang Zhang, Gong Chen, and KC Chan. Egcn: An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3681–3687, 2022. doi: 10.24963/ijcai.2022/511.

Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7899–7908. IEEE, 2021.

Wenbing Zhao, M Ann Reinthal, Deborah D Espy, and Xiong Luo. Rule-based human motion tracking for rehabilitation exercises: realtime assessment, feedback, and guidance. *IEEE Access*, 5:21382–21394, 2017.

A Corpora and label protocols

A.1 Corpora and label protocols

C1 (scored corpus). 613 videos, 52 app users, two exercises (bird-dog and kneeling leg raise). The continuous label is a normalized leg-lift height in $[0, 1]$ produced by an early annotation campaign (manual peak-frame judgment); its per-video measurement protocol was not archived. The labels are not a deterministic function of the landmarks the regressors consumed: recomputing the synthetic generator's ankle–hip formula $\Delta Y = (y_{\text{ankle}} - y_{\text{hip}})/L_{\text{torso}}$, clipped to $[-0.3, 1.0]$ and mapped to $[0, 1]$, from the MediaPipe landmarks fed to V4–V5 yields MAE 0.252 against the labels—worse than the guess-the-mean baseline (0.2233, dataset mean 0.5269)—with correlation 0.61 and a systematic per-exercise offset. Three-class labels ($<$0.6 insufficient / 0.6–1.0 standard / 1.0 over) derive from the same quantity. User distribution is long-tailed (top user 111 videos; 14 users with 1–3 videos). C1 was repeatedly consulted to compare V1–V5 and select diagnostic changes; it is a development set, not a final held-out test set.

C2 (grid-measured calibration corpus). 61 laboratory videos, 3 subjects, with peak knee height measured against a physical grid in the camera plane, plus per-subject torso length and hip height. The regression target in the LOUO report is knee height divided by torso length. The direct 0.0337 reconstruction uses world-3D pose, each subject's measured torso length, and a fixed $40\,$cm grid-derived hip height; it is not calibration-free in the sense of requiring no subject or protocol information. Labels for the three-class task follow the coach's $40 \pm 4\,$cm coaching decision band. Pose-estimated peak knee height is biased -1.2, -4.2, and $+0.9\,$cm vs. grid truth for the three subjects (mean $-1.4\,$cm; Appendix A.4).

C3 (multi-exercise form corpus). 4 subjects $\times$ 136 videos $=$ 544 per viewpoint, captured synchronously at four viewpoints ($0°/90°/180°/270°$; the side ($90°$) batch is analyzed in Table 6). Ten exercises (six dynamic: squat, glute bridge, Bulgarian split squat, kneeling leg lift, single-leg bridge, side leg raise; four static: balance hold, side plank, single-leg stand, quad stretch), each video conditioned on a coached error dimension and amplitude tier (dynamic: 20 videos/subject; static: 4). Portrait captures were physically rotated to landscape (ffprobe-verified 208/208) before processing. Ground truth is the per-video expert amplitude judgment (visual-only), collapsed into each rule's label space (e.g., "hip-flexion insufficient" counts as "standard" for the knee-flexion rule).

C4 (live reflow). $\sim$2,500 app records (user, exercise, error-flagged clips) were pulled through the production API. The reported bird-dog result uses 156 valid repetitions (143 online, 13 offline) from 27 videos and 10 users (5 online, 5 offline) after expert review. The sample was selected by the production review pipeline and should be interpreted descriptively, not as a prevalence-weighted population estimate. Its "human" labels are *anchored*: the review interface shows the auto label and evidence frames and pre-selects the auto answer before the reviewer confirms or corrects, so agreement between auto and human is an upper bound on true agreement (Appendix A.6).

H1 (blinded re-analysis of the reflow corpus). 51 reflowed videos (29 bird-dog, 22 kneeling leg raise) were re-labeled by two raters working independently, blind to all machine outputs and to each other; disagreements were arbitrated by a third rater. Because $45/51$ videos and $11/12$ users overlap with C4, H1 is a *blinded retrospective re-analysis* of old videos from design-participating users—independent labels, not independent user samples. The frozen prospective protocol over new users (C5, v1.1) is defined in the project archive but was not executed and remains future work. Evidence-status labels used throughout: C1–C4 are *development analysis* (data or labels participated in design), H1 and the C2 grid re-analysis are *retrospective audit*, EGPA is *controlled simulation*, and DeepRehabPile is *public benchmark*; no result is a prospective test.

Public and controlled complements. DeepRehabPile (Ismail-Fawaz et al., 2026): 14 regression datasets, 161 subjects; per-user splits seed-controlled; seven arms as in Section 4.5. EGPA: 60 users $\times$ 20 reps, seed 2026; measurement chain identical to the deployed rule (self-test: zero-perturbation reproduction to 0.05 cm).

A.2 The five sim-to-real versions (Act I detail)

Table 5: The five regression versions. Real data never entered training gradients, but C1 (613 videos) was repeatedly consulted during development (MAE; guess-mean 0.2233). V1–V3 predate real-set connection in comparable terms.

Version	Architecture / data	Headline result and diagnosis
V1	3-branch (vision/image/geo) dual-head, 0.395M	Immediate overfit; attention peaks at frame 0 (initial-pose shortcut); dropping the image branch best
V2	2-branch, size ablation 0.022–0.248M	All sizes within noise of guess-mean; **source-fingerprint wall**: within-source CV 0.0315 vs. cross-source 0.0865 (+175%)
V3	MediaPipe 99-dim features, synthetic corpus	Linear probe R^2=1.0000 on synthetic (single-feature shortcut); real-data collapse to guess-mean
V4	+ hip-centered/torso-normalized inputs; 4-level $\times$ 16-dimension synthetic perturbations	0.416 $\rightarrow$ 0.255 $\pm$ 0.014 (5/5 seeds); best sweep 0.234 $\pm$ 0.026; 3/18 runs below baseline
V5	+ real-val checkpoint selection; 5-seed ensembling	Selection arm 0.2597 vs. synthetic-val 0.2463; ensemble gain +0.0013; no support for the model-selection account in this comparison

Why this learning line was stopped. C2 was used as a measurement-validity probe before a planned two-stage recipe (synthetic pre-training + real fine-tuning). For the normalized target, the best LOUO geometry obtained MAE 0.0500 against a fold-specific mean baseline of 0.0564 and a global-mean baseline of 0.0498; it won in two of three folds. In absolute centimeters the deployed estimator showed mean bias -1.4 cm, MAE 2.6 cm, RMSE 3.2 cm, and 95% LoA $[-7.1, +4.2]$ cm (ICC 0.59–0.64; three-subject pilot, Appendix A.4); within-user correlation (0.807) was a secondary tracking indicator. A direct reconstruction obtained 0.0337, but used world-3D pose, measured torso length, and a fixed 40 cm hip-height constant. The $\sim$0.0385 quantity is a grid-quantization estimate, not a strict noise floor. A three-class oracle also differed sharply between absolute-cm (0.984) and normalized (0.492) conventions. These exploratory results suggested that task definition and

measurement deserved priority, but they neither prove that the regression residual was exclusively protocol-bound nor rule out stronger learned alternatives.

A.3 RULE-LEVEL C3 DETAIL

The C3 rule evaluation is reported here in full; the main text keeps only the summary rows of Table 1. Thresholds start at literature-derived engineering defaults, and C3 subsequently informed threshold and feature choices, so all patch and decomposition results below are exploratory and retrospective rather than estimates on an independent test corpus.

Table 6: **Eleven C3 rules plus one separate C4 production result.** "Majority" = most-common-class baseline. The C3 macro is the unweighted mean of the 11 C3 rule accuracies; the bird-dog C4 row (n=143 online valid repetitions of 156 total) is excluded, and its 0.937 is an anchored pre-audit pooled agreement figure (user-balanced 0.748; blind consensus on matched videos 0.391), not a production-accuracy estimate (Section 4.3). LOSO evaluates patched bands across the same four C3 subjects and is a development gate, not independent validation. † withheld pending a labeling-convention decision; $*$ static rules, n=16 videos; dynamic rules n=80 videos; the C4 row n=143 online repetitions; quad stretch: §4.4.

Exercise / rule	Feature (unit)	n	Majority	Default	Patched	LOSO	Band change
Kneeling leg lift	thigh elevation p95 (deg)	80	0.500	0.825	**0.963**	0.900	$-10 \rightarrow -3$
Squat, knee flexion	knee angle p95 (deg)	80	0.750	0.913	0.913	0.913	none (gate)
Squat, hip flexion	hip angle p95 (deg)	80	0.500	0.200	**0.813**	pass	$[70, 120] \rightarrow [117.9, 158.3]$
Glute bridge	hip-ext. deviation p05	80	0.350	0.750	0.750	0.488	none (gate)
Single-leg bridge	hip-ext. deviation p05	80	0.500	0.500	**0.800**	0.725	$15 \rightarrow -0.7$
Bulgarian split squat	front-knee flexion p95	80	0.500	0.875	0.875	0.825	none (gate)
Side leg raise	thigh elevation p95 (deg)	80	0.500	0.600	0.900†	pass	$25 \rightarrow 51.6$
Bird-dog lift (production)	knee height (cm)	143	0.400	**0.937** (anchored reflow)		—	40 ± 4 (coaching decision band)
Balance hold$*$	free-leg elevation	16	0.500	0.813	—	—	default kept
Side plank$*$	hip sag (normalized)	16	0.500	0.688	—	—	default kept
Single-leg stand$*$	foot-lift height	16	0.500	0.688	—	—	default kept
Quad stretch	knee flexion p95	16	0.500	0.438	—	—	selection failure

Cold start (Table 6). With defaults, seven of eleven rules beat their majority baselines; two fail outright. Squat-hip scores 0.200—*below* the always-standard baseline 0.500—because its literature band $[70, 120]^\circ$ contradicts the observed C3 convention: parallel-squat bottom hip flexion measures 120–156°, so $39/40$ standard repetitions were labeled excessive. Quad stretch fails because side selection grabs the support leg under occlusion (§4.4). The unweighted mean of all 11 C3 rule accuracies is 0.663.

Calibration with LOSO gates. Band search under a ship-only-if-LOSO-not-worse gate adopted three patches: squat-hip $0.200 \rightarrow 0.813$, kneeling lift $0.825 \rightarrow 0.963$ (LOSO 0.900), and single-leg bridge $0.500 \rightarrow 0.800$. The squat-knee candidate gained in sample ($0.913 \rightarrow 0.925$, one correctly re-classified video at 129.8°) but its fitted band tied the default band under the strictly nested gate (0.9125 vs. 0.9125), so $130 \rightarrow 129$ was withheld as within-fold noise and the default kept. Keeping the other eight rules at their defaults gives an unweighted 11-rule macro of 0.758. The glute-bridge candidate reached 0.775 in sample but LOSO was 0.488, so it was withheld; Bulgarian was also blocked. These gates reduce obvious subject-specific overfitting but—because the same four subjects supported search, evaluation, and feature design—constitute nested threshold selection, not nested validation. The side-raise candidate ($0.600 \rightarrow 0.900$) was withheld because its 51.6° threshold conflicts with the 30–45° collection convention. The separate C4 production row and its audit are covered in Appendix A.6 and Section 4.3.

Error decomposition: threshold / individual / feature. For every rule we computed in-sample ceilings: a *global oracle* (best single band), a *per-person oracle* (best band per subject), and the residual, $1 - \text{acc} = G_1 + G_2 + G_3$ (threshold / individual / feature-limited). On C3 (Figure 6), G_1 is large for side raise (0.300), single-leg stand (0.250), and side plank (0.187); G_2 is largest for glute bridge (0.225) and balance hold (0.125); G_3 is near zero for eight rules but larger for squat hip (0.190), quad stretch ($\sim$0.25), and single-leg bridge (0.150). The decomposition suggested candidate interventions, which we checked retrospectively on the same corpus.

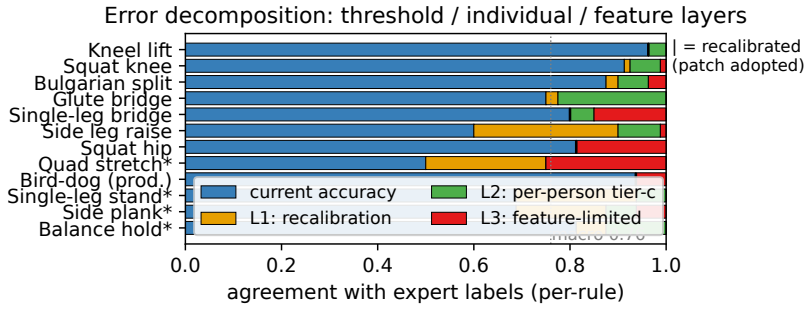


Figure 6: **Layered decomposition of residual error per rule.** Blue: current band; orange (L1): recoverable by global re-threshold; green (L2): additionally by per-subject thresholds; red (L3): unrecoverable by any threshold on the current feature. Filled/open diamonds: shipped/not-shipped patches; $*$: static rules (n=16); quad stretch pools L1/L2 to its global oracle ($\sim$0.75).

Retrospective feature intervention (squat hip). The bottom hip-flexion angle cannot express "leaning forward too much": after patching, its *standard* and *excessive* class medians differ by only 2.6° (ranges 119.6–155.9° vs. 104.3–163.7°), and for one subject the excessive interval lies *entirely inside* their standard interval—per-person calibration gains exactly zero. The trunk-lean angle (hip-center→shoulder-center vs. vertical) does express it: standard repetitions occupy 31–50° (median 41.2), excessive 54–82° (median 69.8), a near-separated margin (Figure 7a). The features are complementary—hip angle for *insufficient* ($<$ 118°), trunk lean for *excessive*—and a two-line combined rule reaches **0.963** (77/80), +15 points over the single-feature rule.

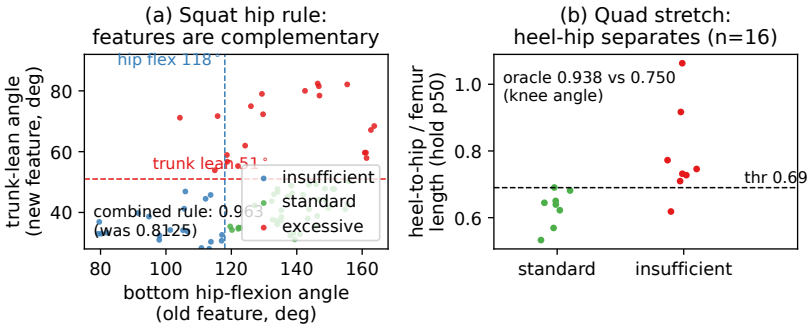


Figure 7: **Feature-layer interventions (C3 measurements, retrospective).** (a) Squat: hip flexion (x-axis) separates *insufficient* from *standard* but not *excessive* (red) from *standard* (green); trunk lean (y-axis) completes the job (dashed: combined rule, 0.963). (b) Quad stretch (n=16): heel-to-hip/femur separates the classes (0.938) where knee angle does not (0.750).

Retrospective feature intervention (quad stretch). On the correct-viewpoint subset, the knee-flexion feature's class medians differ by only $\sim$7° (oracle 0.75): coaches judge how close the heel comes to the buttock. A heel-to-hip distance normalized by femur length separates the classes at threshold 0.69 with oracle 0.938 (standard 0.53–0.69; insufficient 0.62–1.06; one overlap)—a change of *measured quantity*, not of threshold (Figure 7b).

Retrospective L1/L2 routing checks. Where G_3 is small, cheaper fixes sufficed (Appendix A.10): repairing the side-raise side-selection defect and re-thresholding (25° → 45–52°) yields 0.875–0.913 (per-person oracle 1.0); the single-leg-bridge support-leg repair (choose the *more extended* side; 22/80 flips) lifts 0.80 → 0.85; the Bulgarian front-leg repair adds +0.013—marginal, as its small G_1+G_2 predicted. The static rules' n=16 is a sample-size limitation: per-person oracles 0.94–1.0, so collection, not modeling, is the bottleneck.

A.4 MEASUREMENT-VALIDITY DETAIL (C2)

Indices and definitions. We treat the grid reading and the estimator as two raters of the same 61 videos (three subjects). Bias and 95% limits of agreement follow Bland–Altman (bias ± 1.96 SD of the paired differences). ICC$(2,1)$ is the two-way random-effects, single-measurement, absolute-agreement coefficient; ICC$(3,1)$ is the two-way mixed-effects, single-measurement, consistency coefficient. Because the grid is a fixed reference instrument, ICC$(3,1)$ is the more appropriate index; both are reported. Videos are repeated measures within three subjects: the subject-level 95% CI of the mean bias spans $[-7.9, +5.0]$ cm, so every index here is a pilot estimate, not a population statement, and correlation is reported only as a secondary tracking indicator.

Table 7: C2 measurement validity (61 videos, 3 subjects): the deployed default-torso estimator vs. the measured-torso variant. Grid truth is read in 5 cm steps; per-user bias is listed as s001/s002/s003.

Arm	bias (cm)	MAE	RMSE	LoA	ICC$(2,1)$	ICC$(3,1)$	user bias
Default torso 50 cm (deployed)	-1.4	2.6	3.2	$[-7.1, +4.2]$	0.59	0.64	$-1.2/-4.2/+0.9$
Measured torso (privileged)	-1.4	2.6	3.2	$[-7.2, +4.4]$	0.58	0.63	$-0.9/-4.3/+0.9$

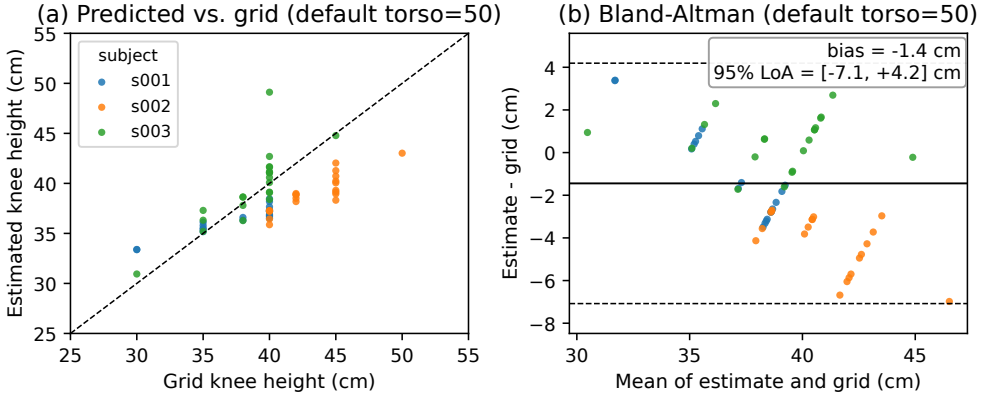


Figure 8: **C2 measurement validity (deployed default-torso estimator).** (a) Estimated vs. grid knee height with the identity line; (b) Bland–Altman plot with bias and 95% limits of agreement. Colors separate the three subjects.

Fair LOSO comparison. The uniform protocol comparison behind §4.1:

Table 8: **Fair C2 comparison (61 videos, 3 users, LOSO).** All arms share the cached MediaPipe landmarks, the same three-class labels, and leave-one-user-out splitting; learned baselines use five seeds with all seeds retained, normalization fit on training folds only, and no class weighting. Direct geometry uses the pre-fixed 40 ± 4 cm coaching band; nested threshold calibration fits bands on the other users only.

Method	Acc.	Macro-F1	Per-class recall (std/over/under)		
Majority class	0.656	0.264	1.000	0.000	0.000
Direct geometry (default anthropometry)	**0.738**	**0.571**	0.925	0.091	0.700
+ measured-torso calibration	0.721	0.545	0.925	0.091	0.600
+ nested threshold calibration (LOSO)	0.557	0.473	0.675	0.091	0.600
Logistic on lift (LOSO)	0.525	0.413	0.675	0.091	0.400
ST-GCN (5 seeds)	0.528 ± 0.184	0.242	error-class recall 0 in all seeds		
MLP (5 seeds)	0.580 ± 0.075	0.316	under recall 0.6 in 3/5 seeds; over 0		

Error-source decomposition. We bound each candidate source against C2: **(i) torso default**—replacing the default 50 cm torso with the measured lengths (46.7/52.0/55.3 cm) changes the aggregate error by < 0.05 cm (Δbias -0.04 cm, ΔRMSE -0.05 cm): the torso factor multiplies only the lift above the hip, and the three deviations largely cancel; users far from 50 cm remain unmeasured, and EGPA injects this source explicitly. **(ii) hip-height default**—both arms use the grid-derived

40 cm hip height, which equals the true hip height in all 61 rows, so this source is zero by construction in C2 and unidentifiable; deployed users without a measured hip height would add it. **(iii) pose estimation**—the measured-torso residual (bias -1.4 cm, 95% CI $[-2.2, -0.7]$ cm; per-user $-0.9/-4.3/+0.9$ cm) is the dominant identifiable source. **(iv) viewpoint**—C2 used one fixed side camera (off-plane views were grounds for re-shooting), so viewpoint error is not estimable here; wrong-view videos are handled by the refusal guard on other data (Section 4.4). **(v) segmentation**—peak-frame selection error is 0.084 s on average (0.449 s max); because the lift curve is flat at its extremum, timing error enters the centimeter value only at second order (sub-cm), and no rep-level split failure occurred in C2. The reference itself is quantized to 5 cm grid steps, which widens the LoA and deflates the ICC and correlation relative to a finer reference. C2 therefore identifies a measurement problem—bias -1.4 cm, LoA wider than the coaching decision band, primary source pose estimation—rather than establishing measurement validity.

A.5 Blind re-analysis detail (H1)

Primary table.

Table 9: **The frozen rules against blind consensus (primary: any-error, video level).** The error classes are the positive class for precision/recall; 95% CIs are user-level cluster bootstrap (n=2000, seed 0). The last row restates the pre-audit C4 figure at its own (repetition-level) unit; the rows are placed together to expose the anchoring and unit-of-inference differences, not to be subtracted.

Action	Users	Videos	κ	Acc.	Macro-F1	Prec.	Recall
Bird-dog	8	24	0.45	0.167	0.161	0.136	0.750
Kneeling leg raise	8	14	0.51	0.357	0.354	0.600	0.300
C4 bird-dog, anchored review (rep-level, 143 online reps)	5	22	—	0.937	—	—	—

Sensitivity aggregations and per-user results. Video-level error detection against blind consensus, error class as positive (pre-registered: any-error primary; majority-error and all-error as sensitivity); the combined rows pool both actions over the same 38 videos and are the source of the 0.24–0.61 range quoted in the abstract:

Aggregation	Action	Accuracy	Macro-F1	Precision	Recall	n
any-error (primary)	bird-dog	0.167	0.161	0.136	0.750	24
any-error (primary)	kneeling	0.357	0.354	0.600	0.300	14
majority-error	bird-dog	0.417	0.378	0.143	0.500	24
majority-error	kneeling	0.429	0.417	1.000	0.200	14
all-error	bird-dog	0.750	0.550	0.250	0.250	24
all-error	kneeling	0.357	0.326	1.000	0.100	14
any-error (primary)	combined	0.237	0.232	0.222	0.429	38
majority-error	combined	0.421	0.394	0.250	0.286	38
all-error	combined	0.605	0.474	0.400	0.143	38

Per-user any-error accuracy (videos): bird-dog—uid350 0.00 (3), uid351 0.50 (2), uid364 0.07 (14), uid374 0.00 (1), uid92703 0.00 (1), uid92704 1.00 (1), uid92705 1.00 (1), uid92706 0.00 (1); kneeling—uid350 0.50 (2), uid351 0.00 (2), uid364 0.20 (5), uid374 0.00 (1), uid92701 1.00 (1), uid92702 0.00 (1), uid92703 1.00 (1), uid92704 1.00 (1). The high-volume user's standard videos are polluted by single-repetition false alarms ($13/14$), and most users miss all of their kneeling error videos.

Kneeling-raise miss taxonomy. Eight error videos missed under majority-error aggregation (seven under any-error; video 2505 is detected by a single triggering repetition). One primary cause per case, from logged evidence:

Case	Measured range	Primary cause	Secondary	Evidence
2218 (uid374)	$+9.3$–89.4°	feature failure	segmentation	impossible "elevations" for a near-horizontal thigh
2316 (uid351)	-2.6–$+7.5^\circ$	threshold failure	segmentation	human boundary far above rule threshold
2318 (uid350)	-0.8–$+63.7^\circ$	feature failure	segmentation	old anchored labels followed the auto output
2322 (uid351)	-0.6–$+32.8^\circ$	feature failure	segmentation	anchored labels uncorrelated with measurements
2505 (uid364)	-10.9–$+23.1^\circ$	threshold failure	—	one triggering rep only; boundary mismatch
2506 (uid364)	$+3.0^\circ$ (1 segment)	segmentation	—	whole video compressed into one segment
2509 (uid364)	-4.7–$+11.5^\circ$	threshold failure	segmentation	-4.7° crosses frozen -3° but production -10°
927007 (uid92702)	$+74.9$–$+87.6^\circ$	segmentation	feature failure	manual re-split finds a -20.8° too-low segment

Distribution of primary causes: segmentation 2, feature failure 3, threshold failure 3, label ambiguity 0, observability 0; feature failure and segmentation co-occur three times and amplify each other. Three additional error videos have no auto output (coverage gaps, excluded from the accuracy denominators): one total segmentation failure and two offline records that never ran the pipeline, one of which also shows a label-ambiguity signal (blind consensus "too-low" vs. human re-segmentation reading "standard").

Unit-of-inference demonstration. On the same 20 online bird-dog videos and the same rule outputs: repetition-level anchored agreement $123/131 = 0.939$; video-level any-error detection against blind consensus 0.100 (TP 1, FP 18, FN 0, TN 1). The two figures differ only in statistical unit and label independence.

A.6 C4 AGGREGATION DETAIL

Three aggregation views (valid repetitions after expert review). Pooled repetition agreement vs. video-balanced vs. user-balanced agreement; CIs are user-level cluster bootstrap (n=2000, seed 0):

Set	n_{reps}	Pooled rep	Video-balanced	User-balanced	Pooled 95% CI
All (10 users, 27 videos)	156	0.942	0.921	0.874	0.803–0.981
Online (5 users, 22 videos)	143	0.937	0.903	0.748	0.674–0.977
Offline (5 users, 5 videos)	13	1.000	1.000	1.000	degenerate (n=13, all agree)

Single-user dominance. uid364 supplies $85/156 = 54.5\%$ of valid repetitions (59.4% of the online set) with user-level agreement 0.977, pulling the pooled figure toward 0.94. Leave-one-out on the online set: pooled $0.937 \to 0.879$, video-balanced $0.903 \to 0.792$, user-balanced $0.748 \to 0.691$. User-level agreement ranges from 0.00 (uid374) to 1.00 (uid350 and all five offline users, whose n=2–3 repetitions are nearly uninformative); user-balanced CI on the online set is 0.49–0.97.

Anchoring mechanism. The review interface displays the auto label, the measurement, and the evidence frames and pre-selects the auto answer before the human submits (recorded as *confirmed* or *corrected*). The human label is therefore a confirmation-or-correction of the machine output: agreement with the auto label is an upper bound on agreement with any independent reference, not a correctness estimate. This is why the paper reports the 0.937 figure as *anchored, pooled repetition-level agreement* rather than accuracy.

Same-video correspondence (23 matched bird-dog videos). Anchored video-level agreement $21/23 = 0.913$ (CI 0.75–1.00) vs. blinded consensus agreement $9/23 = 0.391$ (CI 0.00–0.56). Per user: uid350 $1.00 \to 0.00$ (3 videos), uid351 $1.00 \to 0.50$ (2), uid364 $0.923 \to 0.615$ (13), uid374 $0.00 \to 0.00$ (1), and the four offline users $1.00 \to 0.00$ (1 video each). Seven of the twelve overturned "agreements" are online videos where the anchored reviewer followed the auto "over" verdict that consensus read as standard.

A.7 REFUSAL AND OBSERVABILITY AUDIT DETAIL

Coverage–risk on historical reflow data (311 repetitions, 11 users, 45 videos; 218 judgeable). Primary aggregation is per-user macro; pooled and per-video are sensitivity views. Selective risk = accepted-set error among judgeable repetitions; refused-set error = error rate among refused-but-judgeable repetitions.

Set	Aggregation	Coverage	Selective risk	Refused-set error
Bird-dog (194 reps)	per-user	0.964	0.110	1.000
Bird-dog	pooled	0.964	0.052	1.000
Kneeling (117 reps)	per-user	0.994	0.159	0.000
Kneeling	pooled	0.983	0.133	0.000
Total (311 reps)	**per-user (primary)**	**0.964**	**0.137**	0.500
Total	pooled	0.971	0.074	0.333

User-level bootstrap 95% CIs (total): coverage 0.90–1.00, selective risk 0.02–0.20. Only $9/311$ (2.9%) repetitions were refused. **List A (refused but judgeable, 3):** one bird-dog "over" error the guard gave up on, plus two kneeling standard repetitions with insufficient frames. **List B (accepted but wrong, 16):** bird-dog "over" read as standard 6, kneeling standard read as too-low 5, kneeling

too-low read as standard 3, bird-dog standard read as under 1, bird-dog under read as standard 1. **List C (truly unobservable, 6):** refused and human-invalid, all bird-dog videos with insufficient knee-visibility frames. Forcing answers on the refused set would have produced no additional errors—the guards lost coverage without saving any mistake.

Exploratory threshold sweep (281 judgeable repetitions with cached landmarks). Bird-dog knee-visibility threshold: selective risk ≈ 0.056 from θ=1 (frozen) through θ=20, rising only as coverage collapses (0.063 at θ=40 with coverage 0.874). Kneeling min-valid-frames: risk $\approx$ 0.19–0.23 across the whole sweep, frozen point θ=10 inside the flat region (coverage 0.968, risk 0.200). The frozen thresholds cannot trade coverage for safety on this data; both sweeps are exploratory.

C3 four-viewpoint stress test (exploratory). The quad-stretch guards were designed on the 90° and 270° batches, which are also the only two viewpoints with cached landmarks; the 0° and 180° batches have labels but no landmarks, so per-viewpoint coverage and error rates cannot be completed—the stress test itself is under-specified by the available data.

Viewpoint	n	Guard triggered	Coverage	Accepted-set error
90° (c90)	16	7 (G1:5, G2:2)	0.562	0.444
270° (c270)	16	7 (G1:6, G2:1)	0.562	0.556
0°, 180°	labels only	—	unmeasurable	unmeasurable

The kneeling-raise frozen guard on the c90 batch (80 videos, same development corpus): coverage 1.000, accepted-set error 0.037, zero refusals (exploratory). The quad-stretch guard remains a development case study re-tested on its own design data (Section 4.4); it is not evidence of generalization.

A.8 EGPA DETAIL

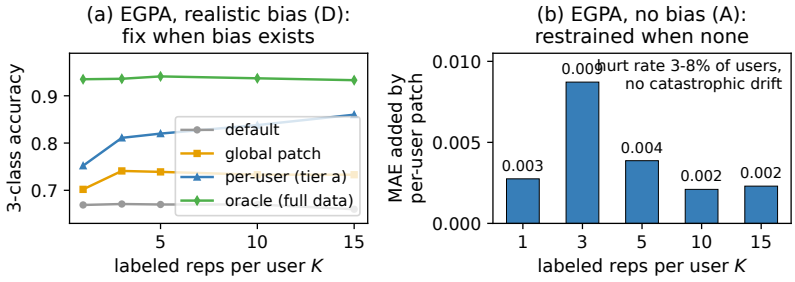


Figure 9: **Controlled personalization sanity check (60 synthetic users).** (a) Under injected per-user pose bias, missing anthropometrics, and perturbations, per-user calibration approaches the simulation oracle as labeled repetitions accumulate. (b) Without injected bias, the same procedure adds at most 0.009 MAE and slightly hurts 3–8% of simulated users. This controlled construction tests implementation behavior, not external validity.

The virtual-user measurement chain mirrors the deployed rule: lift $= \max_t(y_{\text{hip}} - y_{\text{knee}})$; $\hat{y} = h_{\text{hip}} + \text{lift} \cdot L^{\text{cm}}_{\text{torso}}/L^{\text{px}}_{\text{torso}}$. Per user, three bias factors are sampled independently: multiplicative torso default error (default 50 cm vs. true $\mathcal{N}(52, 5)$), additive hip-height default error (default 40 vs. $\mathcal{N}(41, 3)$), and a per-user pose bias $b_u \sim |\mathcal{N}(2.5, 1.2)|$ cm (magnitude calibrated to the measured MediaPipe-vs-grid offset on C2). Execution perturbations (low leg, knee-bend pseudo-height, pelvic lift, trunk sway, hold wobble) alter the true value or the measurement chain mechanistically. Expert labels apply the coach's 40 ± 4 cm convention (with 5% label noise) to the *true* knee height. Arms: default / global (pooled same-K fit) / per-user (tier a) / shrunk (40 ± 6 band, tier-c proxy) / oracle (full-data fit); $K \in \{1, 3, 5, 10, 15\}$. Four scenarios cross no bias, shared bias, dispersed per-user bias, and all factors. In the constructed per-user-bias scenarios, per-user tiers beat the default at every $K \geq 3$ ($p \leq 2 \times 10^{-5}$, Wilcoxon; all factors at K=10: accuracy 0.670 $\rightarrow$ 0.838, MAE 0.163 $\rightarrow$ 0.073). With only shared bias, the global patch (0.880) beats per-user calibration (0.853); without bias, calibration adds at most 0.009 MAE and slightly hurts 3–8% of simulated users. Full tables (accuracy, MAE, target-recovery error, and per-scenario Wilcoxon tests over 60 users) are in the project archive; headline values appear in Figure 9.

A.9 CALIBRATION TIERS

Three tiers re-fit a band from labeled data. Tier **a**: grid search over (ℓ, u) minimizing disagreement. Tier **b**: an LLM reads rule semantics and user history and proposes parameters directly. Tier **c** (deployed): the LLM proposes only a *search range*, clamped to device validity bounds, and the optimizer searches within it (fallback: tier a).

Pilot readings (30 segments, three subjects, test-user-oracle regime). Expert default 0.7667; pooled global re-fit 0.7333; tier-a per-user grid 0.8333; tier-b raw-LLM 0.7667; tier-c LLM-range + optimizer 0.9000. Every tier was fit and scored on the same three users' segments, so these figures—including tier-c's 0.9000—are same-user, in-sample, exploratory ceilings, not expected deployment accuracy; the strictness gradient of Section 4.5 reverses the apparent gain. Tier c exists because tier a over-fits small per-user samples—one subject's tier-a fit shifted the target 6 cm and *lost* holdout accuracy—while the coarse range prior suppresses exactly this failure.

A.10 SIDE-SELECTION REPAIRS AND L1/L2 ROUTING DETAIL

All four interventions below operated on cached landmarks of C3 (no re-extraction), re-computed measurements with the repaired selector, and re-ran threshold analysis; each repair was verified against rendered review frames of the failure cases it was designed to fix.

Side leg raise (selection + threshold). The legacy selector (larger knee-y range) picks the *support* limb during static holds—the raised leg is the still one. Selecting the side with larger hip-to-knee elevation p95 flips $2/80$ videos, including the diagnosed failure (a true elevated leg at $87.3°$ previously measured as the floor leg at $20.9°$). With the repair, class ranges are standard 50.3–$89.5°$ (median 63.7) vs. too-low 18.3–$59.9°$ (median 32.3); thresholds: $25°$ (legacy) $= 0.600$; $45° = 0.875$; $51.6°$ $= 0.9125$ (= global oracle); per-person oracle 1.0. Residual difficulty concentrates in one subject (0.55 at $45°$).

Single-leg bridge (support-side repair). The legacy selector (lower-foot side) fails when the free foot hangs below the support foot. Choosing the side with smaller hip-extension deviation p05 (the more extended side) flips $22/80$ selections (27.5%) and fixes the diagnosed failure ($+55.9$ on the wrong limb vs. -18.9 on the true support side). Accuracy $0.80 \rightarrow 0.85$; global oracle 0.85; per-person oracle 0.8875—the $\pm 6°$ class-overlap band caps single-threshold performance, as the decomposition predicted.

Bulgarian split squat (front-leg repair). Selecting the front leg at each side's own flexion-peak frame (rather than by whole-video ankle-y median, which standing/rest segments pollute) changes measurements by $> 10°$ on only 2/80 videos but removes the extreme outlier caused by picking the bench leg ($136.6° \rightarrow 67.8°$ on the diagnosed failure). $0.875 \rightarrow 0.8875$; per-person oracle 0.975; the insufficient class's long right tail (up to $124.3°$) caps the global threshold at 0.91.

Squat hip (feature definition detail). Legacy feature: bottom hip-flexion angle $= 180° - \angle(\text{shoulder}, \text{hip}, \text{knee})$, p95. New feature: trunk-lean = angle between the hip-center→shoulder-center line and vertical (orientation-free, $\arctan 2(|dx|, |dy|)$), p95. Class ranges: standard 31.2–$50.3°$ (median 41.2); excessive 53.9–$82.4°$ (median 69.8); insufficient 28.1–$46.9°$ (overlaps standard—the legacy feature covers that direction). Combined rule: hip-flexion $< 118° \Rightarrow$ insufficient; trunk-lean $> 51° \Rightarrow$ excessive; else standard; errors: 3 excessive mis-labeled insufficient, accuracy $77/80 = 0.9625$.

A.11 GUARD PARAMETERS

Quad-stretch occlusion guards (identical constants on device and cloud): `guard_min_winner_ratio=` 0.25 (winner side's share of visibility-filtered frames); `guard_min_stretch_flex=` $90°$ (visible-side knee flexion below this with the other knee invisible triggers refusal); `guard_max_loser_knee_vis=` 0.5. The guards are rule-level parameters delivered through the same patch channel as thresholds; on-device and cloud validity checkers share bounds.

A.12 VISION–LANGUAGE MODELS AS A SECOND OPINION (PILOT)

Could a vision–language model replace the measurement chain? In a single-model, single-exercise pilot—the GLM vision endpoint on the bird-dog lift over C2's 61 grid-measured videos—the VLM reaches 0.607 accuracy / 4.89 cm MAE against the rule's 0.721 / 2.62 cm, so the measured pipeline remains the primary judge (protocol and full results in Appendix A.13). Two findings bear on observability: disagreement between the VLM and the rule is a usable review-prioritization signal rather than an accept/reject policy, and the VLM's self-reported confidence spans 0.82–0.95 with zero observed correlation to correctness—supporting our choice of deployment-specific refusal triggers over predicted confidence (Section 4.4). Broader VLM claims require additional models and independently designed protocols.

A.13 VLM PROTOCOL DETAIL

Model: GLM vision production endpoint (glm-5v-turbo), temperature 0.1, one request per video, 12 uniformly sampled frames at 768 px. The prompt supplies the user's true torso length (physical scale bar) and the identical 40 ± 4 cm decision rule; the model returns peak frame, estimated cm height, class, and self-reported confidence. Cost: mean $5{,}928$ prompt + $3{,}147$ completion tokens per video; $\sim$554k tokens and RMB 6.9 ($\approx$USD 1) for the 61-video run.

Results. The VLM reaches 0.607 accuracy / 4.89 cm MAE vs. the rule's 0.721 / 2.62 cm. On the 24 disagreement videos, the VLM is right on 7, the rule on 14, and both are wrong on 3; on the remaining 37 agreements, 30 are correct (inferred from the two marginal accuracies), so agreement alone is only $30/37 = 81.1\%$ accurate, and under the idealized assumption that a human resolves every disagreement correctly, reviewing disagreements would yield at most $54/61 = 88.5\%$. Disagreement is thus a useful review-prioritization signal in this pilot, not an automatic accept/reject policy; the approximate cost is $\$0.02$/video. Self-reported confidence spans 0.82–0.95 with zero observed correlation to correctness (the worst class, accuracy 0.36, carries mean confidence 0.89). The step-by-step measurement variant was stopped after 7 videos ($1/7$ correct, 10.3 cm MAE, estimates oscillating 28.5–55.3 cm) under a pre-declared stopping rule—too small to localize the failure confidently.