DiMiRoS: A Digital Robotic Microscopic Dataset for Benchmarking High-Precision Surgery & Expert Dexterity Understanding

Anonymous authors
Paper under double-blind review

Abstract

For a century, what a microsurgeon saw was locked inside a pair of eyepieces. Digital exoscopic robotic microscopes open that lock: a camera and display make the visual feedback guiding every movement a signal that can be recorded and deliberately changed. This matters beyond surgery. Dexterity emerges as perception and action are calibrated through experience, yet recordings that could teach embodied AI expert manipulation remain scarce, scarcest where precision is extreme. Microsurgery is that extreme: bimanual, magnified, sub-millimetre, with almost no tolerance for error. We introduce DiMiRoS, the first surgical dataset recorded on a digital exoscopic robotic microscope: 308 end-to-end microvascular anastomoses by sixteen participants from novices to experts, captured from eleven synchronized views of the operative field, hands, body and operating room. Each participant operates in 2D and stereoscopic 3D with task and phantom unchanged, making visual feedback a controlled variable. Phase labels, object annotations, hand and body keypoints, and a blinded expert-derived evaluation ground benchmarks of suture quality and phase prediction. Reasoning VLMs overestimate quality, miss execution errors and confuse distinct actions performed with the same tools, while multi-view predictions do not reliably improve assessment. These failures expose a gap between seeing skilled movement and reasoning about its progress and consequences. DiMiRoS makes that gap measurable, opening a path towards embodied AI that learns how experts move and how vision shapes their movement.

1 Introduction

Human dexterity is among the most sophisticated forms of embodied intelligence. The human hand combines 27 degrees of freedom Sobinov & Bensmaia (2021); Lin et al. (2000), more than 30 muscles Lemelin & Diogo (2016), and dense sensory innervation, including over 17,000 tactile receptors in the palm Johansson & Vallbo (1979); Johansson & Flanagan (2008); Corniani & Saal (2020). Yet anatomy alone does not produce dexterity. Early movements combine reflexes with emerging voluntary and visual control Futagi et al. (2012); van der Meer et al. (1995); Pogetti et al. (2013), while neural maturation, sensory feedback, exploration, and interaction with the environment then refine motor commands into increasingly precise and adaptable goal-directed actions Needham et al. (2002); Forssberg et al. (1992; 1995). Dexterity emerges through *embodied learning*: perception and action are continuously calibrated through experience.

Robotics and machine learning pursue the same paradigm through embodied AI: systems that acquire skills by interacting with the physical world. Learning from skilled human behaviour requires recordings of expert manipulation, yet such datasets remain rare, particularly where precision matters most. Microsurgery provides a demanding testbed for studying human dexterity. Microvascular anastomosis (suturing) requires movements of extremely small amplitude, exceptional stability, and precise bimanual coordination under magnification, with minimal tolerance for error. An algorithm that models and discriminates expertise here would strongly validate its ability to encode complex human motor behaviour.

The optical interface through which microsurgeons exercise this dexterity is changing. Digital exoscopic microscopes move the operative field from eyepieces to a display, while robotic systems

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.

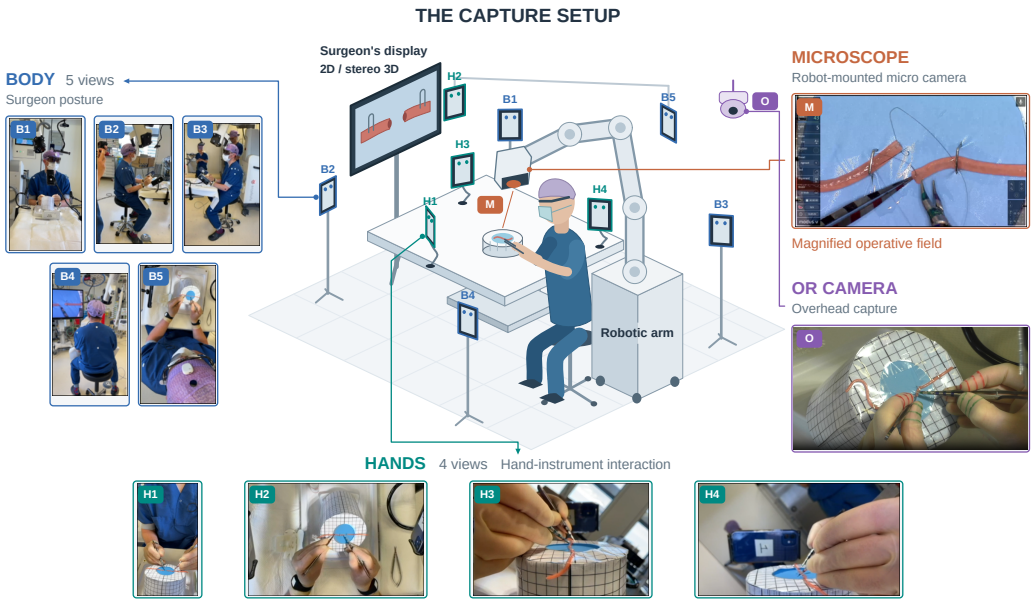

Figure 1: **DiMiRoS multi-view capture.** Multi-view recordings combine the microscopic operative view with surrounding views of instrument-hand interaction, surgeon posture, and operating-room context. This multi-scale setup links fine-grained surgical execution with broader operator behaviour, providing complementary visual evidence for modelling microsurgical dexterity.

track instruments and maintain focus Muhammad et al. (2019); Calvanese et al. (2024). Their 2D and stereoscopic 3D modes alter the visual feedback guiding each movement, opening a new possibility for computational modelling: what was locked inside the oculars becomes an explicit, recordable, and *manipulable* signal.

We introduce DiMiRoS, to our knowledge the first microsurgical dataset recorded on a digital exoscopic robotic microscope for a suturing task. It makes three contributions:

- **Controlled recordings across expertise and visual feedback.** Participants at three expertise levels perform the same standardized suturing task in 2D and stereoscopic 3D, enabling study of skilled motor behaviour across experience while varying visual feedback.
- **Multi-view representations of skilled behaviour.** Eleven synchronized views capture the operative field, instruments, hands, whole-body behaviour and operating-room context, supporting multiview representation learning, viewpoint-invariant behaviour modelling, and fine-grained spatial and action understanding.
- **Benchmarks of expert human dexterity.** Fine-grained spatial, temporal and expert-derived performance annotations ground suture quality and phase prediction from multiview video, testing reasoning VLMs on the platform where the next generation of microsurgeons will train.

2 Related Work

Surgical scene and workflow understanding. A large body of work in surgical video understanding models the structure of procedures through phases, actions, instruments, and scene semantics. OphNet (Hu et al., 2024) provides a large-scale benchmark for ophthalmic workflow recognition, while EgoSurgery-Phase (Fujii et al., 2024) studies phases in egocentric open-surgery video. Other datasets emphasize spatial understanding: CholecTrack20 (Nwoye et al., 2025) provides tool detection and tracking labels, and ProstaTD (Chen et al., 2026) combines instrument localization with temporally grounded actions. At the operating-room scale, 4D-OR (Özsoy et al., 2022), MM-OR (Özsoy et al., 2025b), and EgoExOR (Özsoy et al., 2025a) extend coverage to people, objects, activities, and their relations. The temporal and spatial columns of Table 1 reflect this rich supervision, while the performance column highlights a different question: how well was the task executed? DiMiRoS

Table 1: **Comparison with related surgical datasets.** DiMiRoS combines multi-view operator capture with temporal, spatial, and expert-derived quality annotations for microsurgical skill modelling.

Dataset	Setting	Domain	Task	Capture		Annotations			Performance Evaluation	Modalities
				Views	Operator	Expertise Levels	Temporal	Spatial		
4D-OR (2022)	P	OR	Total knee replacement	6	B	–	Phase	3D box + pose + mask	–	RGB-D
MM-OR (2025b)	P	OR	Total/partial knee replacement	8	B	–	Phase + action	Mask + graph	–	RGB-D + A + R
EgoExOR (2025a)	P	OR	Spine surgery, US-guided needling	>6	HB	–	Action	Hand pose + graph	–	RGB-D + G + A
EgoSurgery (2024)	R	Open	10 open-surgery procedures	1	H	–	Phase	–	–	RGB + G
SHands (2026)	P	Open	Incision and suturing	5	H	2	Gesture	–	Errors	RGB
OphNet (2024)	R	Micro	Cataract, glaucoma, corneal surgery	1	–	–	Phase	–	–	RGB
OphNet-3D (2025)	P	Micro	Cataract surgery	8	H	2	Phase	3D Hand mesh + tool pose	–	RGB-D
MISAW (2021)	P	Micro	Microvascular anastomosis	1	–	2	Phase + step + activity	–	–	RGB+Stereo + Kinematics
MSTA (2026)	P	Micro	Supermicrosurgical anastomosis	1	–	–	–	Segmentation	Quality	RGB
SutureBot (2025)	P	RAS	End-to-end suturing	3	–	–	Subtask + action	Goal locations	–	RGB + Stereo + Kinematics
ProstaTD (2026)	R	RAS	Robot-assisted prostatectomy	1	–	–	Action	Boxes	–	RGB
CholecTrack20 (2025)	R	Endo	Laparoscopic cholecystectomy	1	–	–	Phase	Boxes + tracks	–	RGB
DiMiRoS	**P**	**Micro**	**Microvascular anastomosis**	**11**	**HB**	**3**	**Phase**	**Segmentation + Boxes + 3D Hand mesh**	**Skill + Errors**	**RGB+A**

Setting: real clinical data (R), phantom/ex-vivo (P). *Operator:* hands (H), body (B). *Modalities:* stereo video (Stereo), gaze (G), audio (A), robot signals (R). US denotes ultrasound.

couples scene and workflow annotations with expert performance scores, linking what occurs in a procedure to its technical quality.

Operator-centric and microsurgical modelling. Closer to our setting are datasets of fine-grained hand–instrument interaction and microsurgical workflow. MISAW (Huaulmé et al., 2021) describes anastomosis through hierarchical phase, step, and activity labels. OphNet-3D (Hu et al., 2025) focuses on dynamic 3D reconstruction of ophthalmic hand–instrument interaction using multi-view capture. SHands (Ma et al., 2026) records surgical hand behaviour from multiple viewpoints and labels gestures and execution errors across experience levels. These datasets capture complementary aspects of skill, from workflow structure and hand–instrument geometry to gestures and errors, through the viewpoints and operator coverage compared in Table 1. DiMiRoS records the operative field, hands, body, and surrounding room through eleven views, allowing these behavioural signals to be studied together within one standardized microsurgical task.

Skill assessment and performance supervision. Skill datasets also provide supervision of execution quality. SHands (Ma et al., 2026) pairs gesture labels with fine-grained errors, connecting local execution patterns to incorrect technique. MSTA (Liang et al., 2026) combines high-resolution supermicrosurgical video with expert performance scores and segmentation labels. As the performance column of Table 1 shows, these targets complement phase and object recognition. Two demonstrations can follow the same sequence yet differ substantially in precision, efficiency, and error severity. DiMiRoS combines expert-derived suture quality scores and explicit error penalties with phase labels and object annotations across three expertise groups (Table 1, expertise column). Repeating the task under 2D and stereoscopic 3D visualization allows observable behaviour and task progress to be related to both experience and visual feedback.

3 DiMiRoS: Dataset and Task Description

DiMiRoS brings the surgeon's visual feedback, motor behaviour and task outcomes into one dataset. We describe the standardized task, the recordings and their annotations, and the expert assessment that connects execution to quality.

Table 2: **Dataset composition.** Trials by expertise and visualization mode. The video column counts microscope and selected hand-view recordings used for metadata extraction.

Cohort	# Participants	# Trials (2D)	# Trials (3D)	# Trials (total)	# Videos
Experienced	6	60	60	120	220
Semi-experienced	5	48	50	98	164
Inexperienced	5	50	40	90	117
Total	16	158	150	308	501

3.1 Task Description

Microvascular anastomosis (suturing), or suturing, reconnects two severed vessel ends through fine, coordinated movements under magnification. Participants suture a vessel phantom, placing four stitches at the 12, 3, 6 and 9 o'clock positions in a self-selected order. The protocol schedules ten trials in 2D followed by ten in stereoscopic 3D. Trials last seven minutes, with time to finish a knot already in progress when the limit is reached. The same assistant cuts the thread after each suture. A trial constitutes one demonstration, whose structural quality is assessed using the rubric in Section 3.6. Appendix A.2 gives the equipment and acquisition details.

3.2 Recording Environment and System Setup

Recordings took place at Erasmus MC under standardized lighting and operating conditions. The Modus V digital exoscopic robotic microscope streams the magnified operative field to an external display in 2D or stereoscopic 3D. Participants work on the same 3 mm vessel phantom in both modes, preserving the physical task while changing its visual presentation.

Eleven views connect this close view of manipulation to the surgeon's hands, whole-body behaviour and operating-room context (Fig. 1). Nine smartphone cameras cover the operative field, posture and overhead views, complemented by the microscope and room cameras. The microscope stream records the operative field presented to the surgeon, providing a direct link between visual feedback and skilled movement. Camera placements, frame rates and hardware specifications are in Appendix A.2.

3.3 Participants

The 16 participants span three expertise cohorts: **(1) Experienced** ($n = 6$), certified microsurgeons with more than five years of dedicated practice; **(2) Semi-experienced** ($n = 5$), plastic surgery residents with at least two years of microsurgical exposure; and **(3) Inexperienced** ($n = 5$), medical students without prior microsurgical training. The standardized task allows motor behaviour and quality to be compared across these levels of experience.

3.4 Dataset Composition and Statistics

DiMiRoS contains 308 scored trials: 158 in 2D and 150 in 3D (Table 2). Each trial is one attempt recorded from multiple views, which share the same expert score. Metadata extraction uses 209 microscope videos and 292 videos from one selected hand viewpoint, pairing fine operative detail with a consistent view of hand motion. Section 4 evaluates subsets of these recordings. Appendix A reports metadata coverage by view, expertise and visualization mode.

3.5 Data Processing and Annotations

We extract metadata at approximately 5 fps, retaining original frame indices and timestamps as well as manually annotated frames between samples. We extract object masks and surgical phases from microscope videos and hand geometry from the selected hand view (Fig. 2).

Object annotations. Two annotators used OpenLabeling (Cartucho et al., 2018) to produce 21,425 boxes in 1,341 frame annotations across 21 videos. These sparse annotations cover 45 classes of vessel and instrument parts, body landmarks and action events, and anchor mask propagation and phase extraction. Class counts and annotation conventions appear in Appendices A.4 and A.

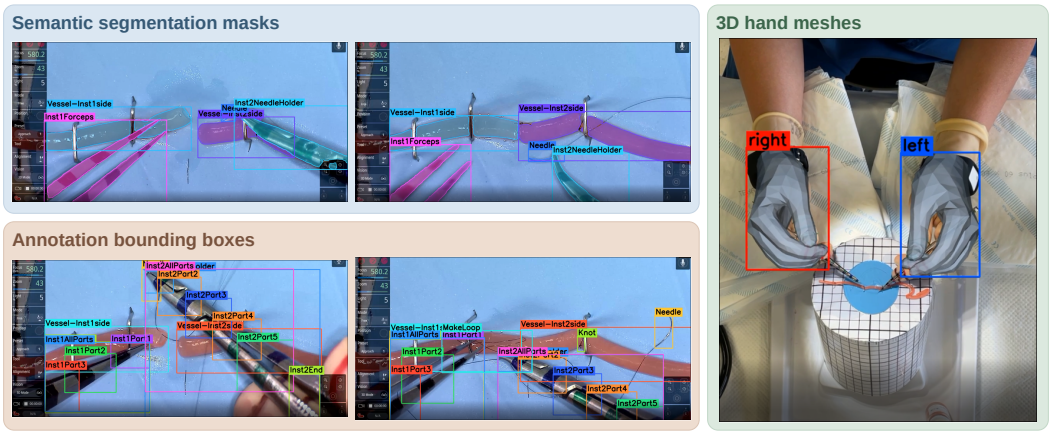

Figure 2: **DiMiRoS data and annotations.** Microscope frames show class-labelled SAM 3 masks (top left) and sparse manual object and instrument-part boxes (bottom left). HaMeR reconstructs 3D hand meshes from the selected hand viewpoint (right). The panels illustrate complementary view-specific metadata and do not assert cross-camera temporal alignment.

Hand keypoints and meshes. HaMeR (Pavlakos et al., 2024) reconstructs MANO hands (Romero et al., 2017) from the selected hand-view videos. The metadata includes 3D joints and meshes, handedness, and projected 2D keypoints, connecting hand configuration to its image evidence. Detection stages and stored parameters are detailed in Appendix A.3.

Semantic segmentation. SAM 3 (Carion et al., 2025) extracts microscope masks for vessels, instruments, needle, thread and hands using text prompts or manual boxes. Boxes initialize labelled masks that are tracked forward and backward and re-anchored at annotated frames. Text prompting supplies thread masks and labels in videos without boxes. Both approaches retain bounding boxes, confidence scores and label sources, with propagation and validation detailed in Appendix A.

Surgical phases. We derive eight phases from microscope object and action annotations: **(1) needle pick-up**, marked by the first `Needle` observation; **(2) needle pass A** through the first vessel end and **(3) needle pass B** through the opposite end, identified by successive `Insert` sequences; **(4) thread pull**, identified by `Pull`; **(5) first knot**, **(6) second knot** and **(7) third knot**, anchored by successive `MakeLoop` sequences with `Knot` completion cues; and **(8) thread cutting**, identified by `Cut`. Figure 3 illustrates this phase progression, and Appendix A.3 details how annotations are matched to video frames and how ambiguous cases are handled.

3.6 Anastomosis Quality Assessment

One experienced microsurgical trainer scored all 308 trials, blinded to participant expertise and visualization mode. After each trial, the anastomosis was examined under a separate microscope, first externally and then opened longitudinally to inspect the lumen and suture line. The assessor used an adapted Likert instrument (Atlan et al., 2018), providing one expert reference for structural quality across the dataset.

The trial suture quality score combines five skill criteria with penalties for major and minor faults (Table 3):

$$R = s_{\text{num}} + s_{\text{space}} + s_{\text{tight}} + s_{\text{bite}} + s_{\text{approx}} - 4\,n_{\text{major}} - 2\,n_{\text{minor}}, \tag{1}$$

Here, $s_{\text{num}} \in \{0, \ldots, 4\}$ counts placed sutures, while the remaining criteria score spacing, knot tightening, needle bite and vessel approximation from 0 to 5. The counts n_{major} and n_{minor} accumulate faults across the anastomosis. Each distinct fault on each placed suture counts separately: one stitch can carry several faults, including both major and minor defects. Repeated errors therefore incur repeated penalties, rather than collapsing into a trial-level flag.

Suture quality scores can reach 24, but negative values are retained to distinguish trials with repeated defects. The score measures the quality of the resulting repair, rather than progress through the task:

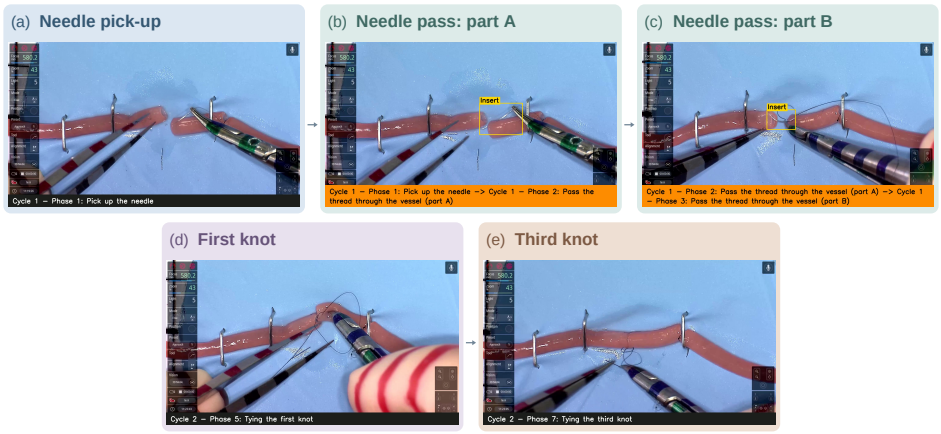

Figure 3: **Progression through selected microsurgical phases.** Microscope frames show (a) needle pick-up, (b–c) passage through both vessel ends, (d) first-knot tying, and (e) third-knot tying. Panels follow procedural order, with annotation-derived starts (orange) that may lag physical transitions.

Table 3: **Microsurgical quality rubric.** Each anastomosis receives skill scores and a penalty for every distinct fault. Higher totals indicate better structural quality.

Type	Evaluation criterion	Recorded per	Score
Skill	Number of sutures	anastomosis	0–4
	Appropriate suture spacing	anastomosis	0–5
	Adequate tightening of the knot	anastomosis	0–5
	Adequate needle bite from the edge	anastomosis	0–5
	Adequate vessel approximation	anastomosis	0–5
Penalty	Major fault: catching the back- or side wall, tearing the vessel wall, large luminal intimal flap, thread in lumen	suture, fault	-4 per fault
	Minor fault: small luminal intimal flap, partial-thickness stitch	suture, fault	-2 per fault
Trial score R (Eq. 1), maximum			24
minimum: none by construction, set by the number of recorded faults			—

zero can indicate either no completed suture or fault penalties that cancel the skill score. We release component scores alongside the total so these outcomes remain distinguishable. Together with the phase labels, they support separate study of what was completed and how well it was done. This phantom assessment measures structural quality rather than functional patency. Appendix A.5 gives the assessment rationale, fault-counting examples and interpretation of the scale.

3.7 Performance Across Experience and Visualization Mode

Visual feedback and expertise shape performance. Mean ground-truth suture quality assessment scores rise from 2D to 3D in all three cohorts: 11.73 to 17.77 for experienced, 12.50 to 13.92 for semi-experienced, and 5.96 to 9.30 for inexperienced participants (Fig. 4). However, the proportion of trials with a major error falls from 55% to 27% among experienced surgeons, but rises from 50% to 60% and from 38% to 70% in the other cohorts. Minor-error penalties instead decrease in 3D for experienced and inexperienced participants, but remain nearly unchanged for semi-experienced participants. Since higher total scores can coexist with more faults, these patterns motivate modelling technique and errors separately and relating visual feedback to experience. Appendix A.6 provides the component scores and full error and risk breakdowns.

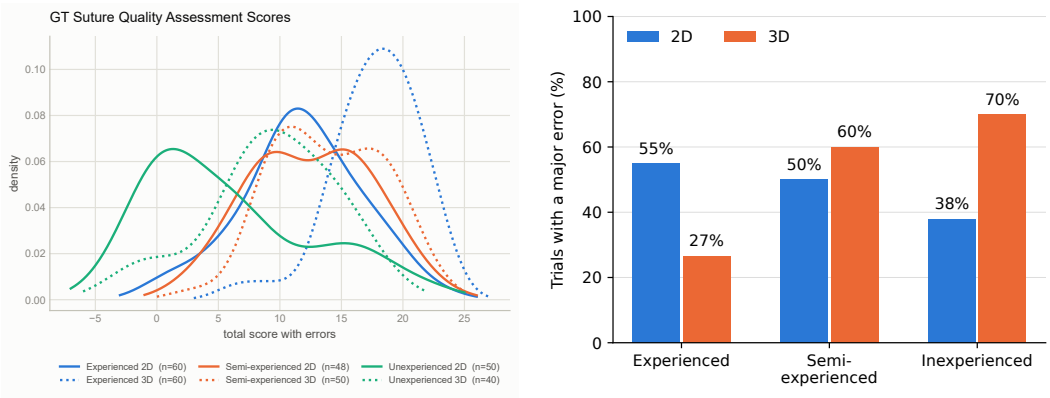

Figure 4: **Higher suture quality can coexist with more major errors.** Left: Kernel (gaussian) Density Estimation plot of suture quality scores across all 308 trials, by expertise and visualization. Right: percentage of trials with at least one recorded major error. The error pattern reverses across experience groups despite higher mean quality scores in 3D for every group.

4 Benchmark Tasks and Results

We test whether reasoning VLMs can follow skilled manipulation and judge its outcomes through two benchmarks of *expert human dexterity*. Suture quality prediction uses microscope and hand-view videos, whereas phase prediction uses microscope clips only. The microscope is the surgeon's primary visual feedback for correcting their own actions, while the hand view reveals how those actions are performed. Both use few-shot in-context learning baselines with Qwen3.5-9B (Qwen Team, 2026) and MedGemma-1.5-4B (Sellergren et al., 2026) with reasoning enabled.

4.1 Suture Quality Prediction

Task and data. Given microscope and hand-view videos, models predict the expert scorecard and total trial score R (Section 3.6). Accurate assessments could support microsurgical reward models for inference-time guidance and reinforcement fine-tuning of VLMs and vision-language-action (VLA) policies (Intelligence et al., 2025; Haworth et al., 2026). The 20 test trials balance 2D/3D visualization across expertise levels and the score range. Each query includes three same-mode expert-scored examples drawn from six disjoint trials covering all expertise–visualization groups (Appendix B.1).

Inference and evaluation. Models process each view at 1 fps in chronological chunks, updating a record of sutures, faults and scores before fusing both records with final microscope frames. Table 4 reports mean absolute error (MAE) against expert scores for single views, fusion and the rounded component-wise few-shot mean baseline. Appendices B and E.1 give inference details and the full few-shot example sequence.

Reasoning VLMs struggle to assess dexterous execution. Qwen and MedGemma have fused quality score MAE of 6.90 and 7.20, versus 3.90 for the few-shot mean, and both overestimate quality while poorly ranking trials (Appendix C.1). As a medical foundation model, MedGemma supports medical image interpretation and clinical reasoning, yet reports four sutures in every trial although only 12 of 20 have four, and misses all nine trials with major faults, of which Qwen detects only two. These uniform counts and missed faults expose a gap between medical image perception and reasoning about the progress and quality of skilled high-precision microsurgical actions. Appendix C details these errors in tracking what was completed and whether it was done correctly.

Multiple views require reliable integration of evidence. Fusion does not outperform the best single view for either model (Table 4). In Figure 5, Qwen's microscope and hand-view records disagree on suture count, and fusion selects three against an expert count of two. Its total nevertheless approaches the reference (18 versus 17) because errors in suture count and knot tightness partly offset underestimated spacing. This plausible total conceals an incorrect account of the procedure, showing

Table 4: **Suture quality prediction.** MAE ± sample standard deviation of absolute error (lower is better). Skill and penalty columns report errors in rubric scores and fault counts, respectively. Micro and hands use individual views, while fused combines their summaries.

		Skill					Penalty		Total
Source / view		**Sutures**	**Spacing**	**Bite**	**Suture tight.**	**Knot tight.**	**Major**	**Minor**	**R**
Few-shot mean		0.65 ±0.81	1.15 ±0.88	0.90 ±1.02	0.95 ±0.76	0.95 ±0.94	0.55 ±0.60	0.60 ±0.50	3.90 ±3.21
Qwen 3.5-9B	hands	1.40 ±1.39	1.70 ±1.66	1.65 ±1.35	1.60 ±1.50	1.90 ±1.86	0.55 ±0.69	0.50 ±0.61	7.65 ±5.02
	micro	1.15 ±1.31	1.55 ±1.50	1.05 ±1.05	1.30 ±1.08	1.00 ±0.92	0.60 ±0.60	0.45 ±0.51	6.85 ±4.79
	fused	0.95 ±1.23	1.45 ±1.54	0.95 ±1.05	1.25 ±1.07	0.90 ±0.91	0.60 ±0.60	0.45 ±0.51	6.90 ±4.53
MedGemma 1.5-4B	hands	0.95 ±1.28	1.35 ±1.46	0.85 ±1.14	0.95 ±0.89	1.00 ±0.92	0.55 ±0.69	0.45 ±0.51	7.00 ±5.02
	micro	0.90 ±1.25	1.45 ±1.43	0.95 ±1.10	1.15 ±0.99	1.00 ±0.97	1.10 ±2.65	0.45 ±0.51	8.75 ±10.00
	fused	0.95 ±1.28	1.30 ±1.49	1.00 ±1.17	1.10 ±1.02	1.15 ±0.99	0.55 ±0.69	0.45 ±0.51	7.20 ±4.95

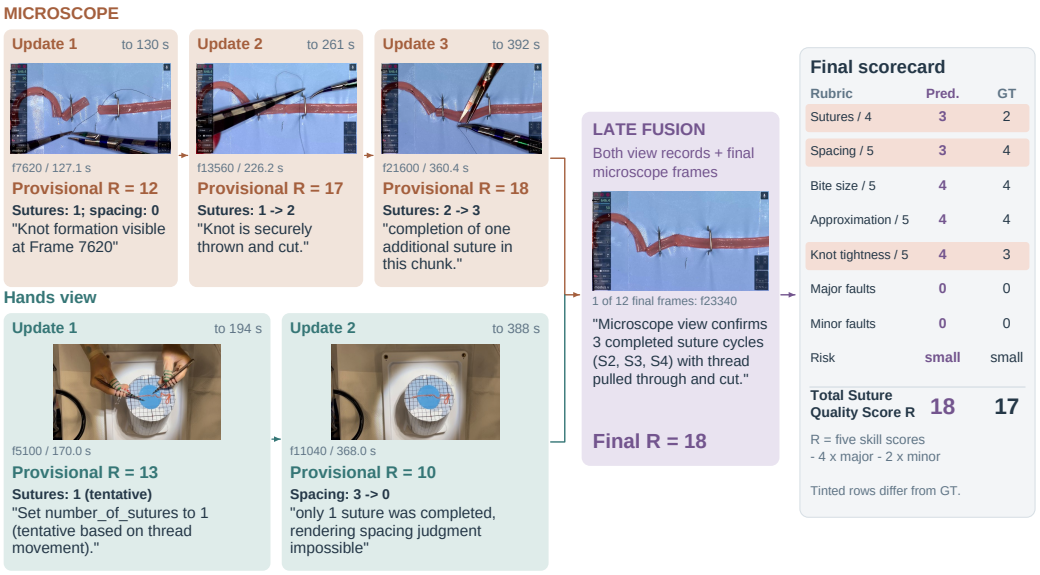

Figure 5: **A close total can conceal component errors.** Qwen fuses microscope and hand-view records into a suture quality score of 18 versus the expert's 17, despite the component errors marked by tinted rows. Quotations depict the model's reasoning output.

why multi-view assessment must resolve conflicting evidence while preserving the components that explain quality accurately.

4.2 Suture Phase Prediction

Task and data. Using microscope video only, models identify the phase at a marked time. This tests the subtask recognition needed to adapt hierarchical VLA approaches (Shi et al., 2025; Peschl et al., 2025; Intelligence et al., 2025) to microsurgical dexterity, where policies must select and execute successive steps of a precise, long-horizon procedure. We group the eight phases into five types: pick-up, insertion, pull, knot and cut. The 77 test clips come from 11 videos across both visualization modes. Five disjoint few-shot examples provide one per type and span both modes and all expertise levels (Appendix B.1).

Inference and evaluation. Queries present clips sampled at 5 fps up to five seconds before and after the marked start, together with the five examples, and request a phase with a frame-linked explanation (Appendices B and E.2). Table 5 reports per-phase recall and its mean, balanced accuracy, alongside overall accuracy and macro-F1. Recall measures the fraction of a phase's clips identified correctly,

Table 5: **Suture phase prediction.** Per-phase recall and aggregate metrics (higher is better). Uniform chance is an analytic reference.

	Per-phase recall					Overall		
Model	**Pick-up**	**Insert**	**Pull**	**Knot**	**Cut**	**B.Acc.**	**Macro-F1**	**Accuracy**
Chance (uniform)	0.20	0.20	0.20	0.20	0.20	0.20	–	0.20
Qwen 3.5-9B	0.38	0.87	0.00	0.14	0.33	0.35	0.33	0.42
MedGemma 1.5-4B	0.31	0.65	0.00	0.04	0.00	0.20	0.15	0.26

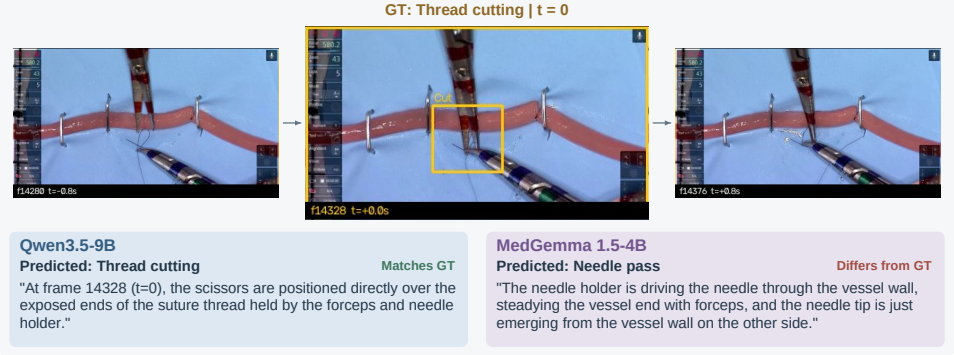

Figure 6: **Recognizing the tools does not settle the action.** At the marked cutting event ($t = 0$, yellow), Qwen predicts cutting and MedGemma incorrectly predicts needle passage. Their explanations appear below.

whereas precision measures how often its predictions are correct. Macro-F1 averages their harmonic mean across phases, while accuracy counts correct predictions across all clips (Appendix D.1).

VLMs confuse persistent tools with changing actions. Qwen achieves 0.35 balanced accuracy, while MedGemma matches the 0.20 chance level despite its medical specialization. Both concentrate on needle actions, predicting pick-up or insertion in 66 and 65 of 77 clips, although only 36 belong to those types. Neither identifies a pull, and knot and cut recall remain low (Appendix D). Figure 6 illustrates this difficulty, where Qwen correctly identifies cutting, while MedGemma offers a specific account of a needle but an incorrect explanation of the needle passing through the vessel.

The same needle holder serves insertion and knot tying, so recognizing the action requires tracking its relation to the vessel and thread as the goal changes. These confusions expose difficulty following task state, which DiMiRoS makes measurable through phase and quality targets and geometric annotations for studying how fine motion relates to progress and execution quality.

5 Conclusion and Future Directions

DiMiRoS connects multiview recordings across expertise and visual feedback with spatial, temporal, and performance annotations to study expert human dexterity. It supports multiview representation learning, viewpoint-invariant behaviour modelling, and spatial and action understanding. The benchmarks reveal the challenge: VLMs miss completed actions and faults, confuse phases as instrument–object relations change, and do not reliably improve assessment by combining views. Learning from skilled behaviour requires tracking both task progress and execution quality.

Future work will utilize the multi-view capture setup to reconstruct the full microsurgical scene in 3D meshes for inference with 3D VLMs and VLAs. The phase and quality annotations could also help microsurgical VLM backbones and hierarchical VLA policies to learn long-horizon surgical task decomposition into a curriculum for effective sub-goal conditioned action-policies, and for precise execution steering through reward-guidance from expert-derived feedback. These directions connect embodied learning from expert human dexterity to the platform on which the next generation of microsurgeons will train.

AI USE STATEMENT

We used generative AI tools to implement methods, assist with translation, improve manuscript language and presentation, provide feedback on research methodology and experiments, support qualitative and quantitative analysis, and interpret results. We did not use these tools to generate synthetic datasets, clean or reformat datasets, or propose or refine hypotheses. The development of theoretical models or conceptual frameworks, formulation of mathematical claims, and assistance with proofs were not applicable to this work.

We reviewed all AI-assisted work. Methodological suggestions were checked against relevant literature for overlap with existing work, and AI-assisted analyses were manually reviewed for accuracy and consistency with the underlying data and research materials. All AI-generated code was reviewed, tested and verified for correctness by at least one author. We take responsibility for the final content, including any text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The Medical Ethics Committee (METC) approval has been received. Written informed consent was obtained from all participants prior to participation.

REPRODUCIBILITY STATEMENT

The appendices provide the information needed to reproduce metadata extraction and both benchmarks. Appendix A documents acquisition, annotation conventions, coordinate systems, mask propagation, phase extraction and the expert assessment. Appendix B specifies sample selection, exemplar allocation, temporal sampling, model settings and output processing. Appendices C and D provide detailed results and evaluation definitions, while Appendix E reproduces the complete prompts and in-context sequences. Code, benchmark configurations, and a few sample videos of the data, including their semantic segmentation masks, annotations and hand meshes overlayed, are part of the supplementary material.

REFERENCES

Michael Atlan, Alexandre G. Lellouch, Josette Legagneux, Marc Chaouat, Alain C. Masquelet, and Didier Letourneur. A new synthetic model for microvascular anastomosis training? a randomized comparative study between silicone and polyvinyl alcohol gelatin tubes. *Journal of Surgical Education*, 75(1):182–187, 2018. doi: 10.1016/j.jsurg.2017.06.008.

Francesco Calvanese, Anna Maria Auricchio, Ville Vasankari, et al. Digital 3D exoscope is safe and effective in surgery for intradural extramedullary tumors: a comparative series. *World Neurosurgery*, 184:e1–e8, 2024. doi: 10.1016/j.wneu.2024.01.136.

Teodor Z. Carciumaru, Vlad Esanu, Clemens M. F. Dirven, Dalibor Vasilic, and Victor Volovici. Reliable but not interchangeable: What validated instruments measure when they grade an anastomosis. *Journal of Clinical Medicine*, 15(15):5779, 2026. doi: 10.3390/jcm15155779.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL `https://arxiv.org/abs/2511.16719`.

J. Cartucho, R. Ventura, and M. Veloso. Robust object recognition through symbiotic deep learning in mobile robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2336–2341, 2018.

Yiliang Chen, Zhixi Li, Cheng Xu, Alex Qinyang Liu, Ruize Cui, Xuemiao Xu, Jeremy Yuen-Chun Teoh, Shengfeng He, and Jing Qin. ProstaTD: Bridging surgical triplet from classification to fully supervised detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Giulia Corniani and Hannes P Saal. Tactile innervation densities across the whole body. *Journal of neurophysiology*, 124(4):1229–1240, 2020.

H. Forssberg, H. Kinoshita, A. C. Eliasson, R. S. Johansson, G. Westling, and A. M. Gordon. Development of human precision grip. II. anticipatory control of isometric forces targeted for object's weight. *Experimental Brain Research*, 90(2):393–398, 1992. doi: 10.1007/BF00227253.

H. Forssberg, A. C. Eliasson, H. Kinoshita, G. Westling, and R. S. Johansson. Development of human precision grip. IV. tactile adaptation of isometric finger forces to the frictional condition. *Experimental Brain Research*, 104(2):323–330, 1995. doi: 10.1007/BF00242017.

Ryo Fujii, Masashi Hatano, Hideo Saito, and Hiroki Kajita. EgoSurgery-Phase: A dataset of surgical phase recognition from egocentric open surgery videos. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15006 of *Lecture Notes in Computer Science*, pp. 187–196. Springer, 2024. doi: 10.1007/978-3-031-72089-5_18.

Yasuyuki Futagi, Yasuhisa Toribe, and Yasuhiro Suzuki. The grasp reflex and Moro reflex in infants: Hierarchy of primitive reflex responses. *International Journal of Pediatrics*, 2012:191562, 2012. doi: 10.1155/2012/191562.

Ali M. Ghanem, Yasser Al Omran, Bishara Shatta, Edward Kim, and Simon Myers. Anastomosis lapse index (ALI): A validated end product assessment tool for simulation microsurgery training. *Journal of Reconstructive Microsurgery*, 32(3):233–241, 2016. doi: 10.1055/s-0035-1568157.

Marjan J. B. Govaerts, Lambert W. T. Schuwirth, Cees P. M. Van der Vleuten, and Arno M. M. Muijtjens. Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education*, 16(2):151–165, 2011. doi: 10.1007/s10459-010-9250-7.

Jesse Haworth, Juo-Tung Chen, Nigel Nelson, Ji Woong Kim, Masoud Moghani, Chelsea Finn, and Axel Krieger. Suturebot: A precision framework & benchmark for autonomous end-to-end suturing. *arXiv preprint arXiv:2510.20965*, 2025.

Jesse Haworth, Juo-Tung Chen, Nigel Nelson, Ji Woong Kim, Masoud Moghani, Chelsea Finn, and Axel Krieger. Suturebot: A precision framework & benchmark for autonomous end-to-end suturing. *Advances in Neural Information Processing Systems*, 38, 2026.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, Jun Cheng, Chi Liu, Kaijing Zhou, and Zongyuan Ge. OphNet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024. doi: 10.1007/978-3-031-73235-5_27.

Ming Hu, Zhengdi Yu, Feilong Tang, Kaiwen Chen, Yulong Li, Imran Razzak, Junjun He, Tolga Birdal, Kaijing Zhou, and Zongyuan Ge. Towards dynamic 3D reconstruction of hand-instrument interaction in ophthalmic surgery. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, Pablo Arbeláez, Wolfgang Reiter, Manoru Mitsuishi, Kanako Harada, and Pierre Jannin. MIcro-Surgical Anastomose Workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021. doi: 10.1016/j.cmpb.2021.106452.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi^*_{0.6}$: a VLA That Learns From Experience. *arXiv preprint arXiv:2511.14759*, 2025.

R Johansson and J Flanagan. 6.05-tactile sensory control of object manipulation in humans. *Senses Compr. Ref. Acad. Press NY NY*, 67:86, 2008.

Roland S. Johansson and Ake B. Vallbo. Tactile sensibility in the human hand: relative and absolute densities of four types of mechanoreceptive units in glabrous skin. *The Journal of physiology*, 286 (1):283–300, 1979.

Pierre Lemelin and Rui Diogo. Anatomy, function, and evolution of the primate hand musculature. In *The evolution of the primate hand: Anatomical, developmental, functional, and paleontological evidence*, pp. 155–193. Springer, 2016.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2203.16527.

Jing Liang, Jingyuan Fan, Aihemaitijiang Yusufu, Jianxi Hou, Dewei Zhao, and Qingtang Zhu. A large 4 k dataset of supermicrosurgical anastomosis at the 0.2 mm scale. *Scientific Data*, 2026.

John Lin, Ying Wu, and Thomas S Huang. Modeling the constraints of human hand motion. In *Proceedings workshop on human motion*, pp. 121–126. IEEE, 2000.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. URL https://arxiv.org/abs/1405.0312.

Le Ma, Thiago Freitas dos Santos, Nadia Magnenat-Thalmann, and Katarzyna Wac. SHANDS: A multi-view dataset and benchmark for surgical hand-gesture and error recognition toward medical training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Sajjad Muhammad, Martin Lehecka, and Mika Niemelä. Preliminary experience with a digital robotic exoscope in cranial and spinal surgery: a review of the Synaptive Modus V system. *Acta Neurochirurgica*, 161(10):2175–2180, 2019. doi: 10.1007/s00701-019-03953-x.

Amy Needham, Tracy Barrett, and Karen Peterman. A pick-me-up for infants' exploratory skills: Early simulated experiences reaching for objects using 'sticky mittens' enhances young infants' object exploration skills. *Infant Behavior and Development*, 25(3):279–295, 2002. doi: 10.1016/S0163-6383(02)00097-8.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L. Lavanchy, and Nicolas Padoy. CholecTrack20: A multi-perspective tracking dataset for surgical tools. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8942–8952, 2025. doi: 10.1109/CVPR52734.2025.00836.

Ege Özsoy, Evin Pınar Örnek, Ulrich Eck, Tobias Czempiel, Federico Tombari, and Nassir Navab. 4D-OR: Semantic scene graphs for OR domain modeling. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 13437 of *Lecture Notes in Computer Science*, pp. 475–485. Springer, 2022. doi: 10.1007/978-3-031-16449-1_45.

Ege Özsoy, Arda Mamur, Felix Tristram, Chantal Pellegrini, Magdalena Wysocki, Benjamin Busam, and Nassir Navab. Egoexor: An ego-exo-centric operating room dataset for surgical activity understanding. *arXiv preprint arXiv:2505.24287*, 2025a.

Ege Özsoy, Chantal Pellegrini, Tobias Czempiel, Felix Tristram, Kun Yuan, David Bani-Harouni, Ulrich Eck, Benjamin Busam, Matthias Keicher, and Nassir Navab. MM-OR: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19378–19389, 2025b. doi: 10.1109/CVPR52734.2025.01805.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024. URL https://geopavlakos.github.io/hamer/.

Markus Peschl, Pietro Mazzaglia, and Daniel Dijkman. From code to action: Hierarchical learning of diffusion-VLM policies. *arXiv preprint arXiv:2509.24917*, 2025. URL https://arxiv.org/abs/2509.24917.

Lívia S. Pogetti, Rosana M. de Souza, Eloísa Tudella, and Luis A. Teixeira. Early infant's use of visual feedback in voluntary reaching for a spatial target. *Frontiers in Psychology*, 4:520, 2013. doi: 10.3389/fpsyg.2013.00520.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. URL https://mano.is.tue.mpg.de/.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Lucy Xiaoyang Shi, Brian Ichter, Michael Robert Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, Adrian Li-Bell, Danny Driess, Lachy Groom, Sergey Levine, and Chelsea Finn. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54919–54933, 2025. URL https://proceedings.mlr.press/v267/shi25d.html.

Anton R Sobinov and Sliman J Bensmaia. The neural mechanisms of manual dexterity. *Nature Reviews Neuroscience*, 22(12):741–757, 2021.

Piotr Stogowski, Filip Flićiński, Jan Białek, Filip Dąbrowski, Maciej Piotrowski, and Tomasz Mazurek. Microsurgical anastomosis rating scale (MARS10): A final product scoring system for initial microsurgical training. *Plastic Surgery*, 29(4):243–249, 2021. doi: 10.1177/2292550320969649.

A. L. H. van der Meer, F. R. van der Weel, and D. N. Lee. The functional significance of arm movements in neonates. *Science*, 267:693–695, 1995. doi: 10.1126/science.7839147.

Victor Volovici et al. Microvascular errors of technique: a systematic review. *Acta Neurochirurgica*, 168:65, 2026. doi: 10.1007/s00701-026-06810-w.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38571–38584, 2022. URL https://arxiv.org/abs/2204.12484.

APPENDIX ROADMAP

The appendix provides the dataset documentation and experimental details that complement the main paper:

- **Appendix A: Dataset.** Hosting and access, recording setup, metadata extraction, annotation coverage, quality assessment, and performance across expertise and visualization modes.
- **Appendix B: Benchmark protocol.** Sample selection, few-shot allocation, baselines, and inference settings.
- **Appendix C: Suture quality results.** Calibration, ranking, errors by visualization mode, and confusion matrices.
- **Appendix D: Suture phase results.** Metric definitions, per-phase performance, and confusion patterns.
- **Appendix E: Benchmark prompts.** Complete prompts and in-context sequences for both tasks.

A DATASET: COMPOSITION, RECORDING SETUP, METADATA AND COVERAGE

A.1 HOSTING, LICENSING AND ORGANIZATION

DiMiRoS comprises synchronized video from eleven viewpoints, with 308 scored trials performed by 16 participants across three expertise levels and two visualization modes (Table 2). The release will include the videos, spatial and temporal annotations, expert assessments, and benchmark resources described in this paper, organized as shown in Figure 7. The dataset will be made publicly accessible under restricted access defined by an End-User License Agreement (EULA), with hosting details and access instructions provided at release.

```
dimiros_dataset/
├── surgeon_01/
│   ├── session_01/
│   │   ├── 2d/
│   │   │   ├── microscope/
│   │   │   │   ├── <modality> S<s>S<n> Camera microscope.mp4
│   │   │   │   ├── manifest.json
│   │   │   │   ├── frames/
│   │   │   │   │   └── <frame_idx>.jpg
│   │   │   │   ├── annotations/
│   │   │   │   │   ├── boxes.jsonl
│   │   │   │   │   └── link_report.json
│   │   │   │   ├── phases/
│   │   │   │   │   └── phases.json
│   │   │   │   ├── semseg/
│   │   │   │   │   ├── instance_png/
│   │   │   │   │   │   └── <frame_idx>.png
│   │   │   │   │   ├── masks.jsonl
│   │   │   │   │   ├── propagated.jsonl
│   │   │   │   │   ├── summary.json
│   │   │   │   │   ├── propagation_summary.json
│   │   │   │   │   ├── propagation_holdout_eval.json
│   │   │   │   │   └── eval_vs_gt.json
│   │   │   │   ├── rewards/
│   │   │   │   │   └── scores.json
│   │   │   │   └── overlays/
│   │   │   │       ├── annotations.mp4
│   │   │   │       ├── phases.mp4
│   │   │   │       ├── semseg.mp4
│   │   │   │       └── propagated.mp4
│   │   │   ├── or_view/
│   │   │   │   ├── <modality> S<s>S<n> Camera OR.mp4
│   │   │   │   ├── manifest.json
│   │   │   │   ├── frames/
│   │   │   │   │   └── <frame_idx>.jpg
│   │   │   │   ├── annotations/
│   │   │   │   │   ├── boxes.jsonl
│   │   │   │   │   └── link_report.json
│   │   │   │   ├── rewards/
│   │   │   │   │   └── scores.json
│   │   │   │   └── overlays/
│   │   │   │       └── annotations.mp4
│   │   │   ├── cam_view_1/ … cam_view_7/, cam_view_9/
│   │   │   │   ├── <modality> S<s>S<n> Camera <cam_num>.MOV
│   │   │   │   ├── manifest.json
│   │   │   │   ├── frames/
│   │   │   │   │   └── <frame_idx>.jpg
│   │   │   │   ├── annotations/
│   │   │   │   │   ├── boxes.jsonl
│   │   │   │   │   └── link_report.json
│   │   │   │   ├── rewards/
│   │   │   │   │   └── scores.json
│   │   │   │   └── overlays/
│   │   │   │       └── annotations.mp4
│   │   │   └── cam_view_8/
│   │   │       ├── <modality> S<s>S<n> Camera 8.MOV
│   │   │       ├── manifest.json
│   │   │       ├── frames/
│   │   │       │   └── <frame_idx>.jpg
│   │   │       ├── hands/
│   │   │       │   ├── meshes/
│   │   │       │   │   └── <frame_idx>.npz
│   │   │       │   ├── meshes_smoothed/
│   │   │       │   │   └── <frame_idx>.npz
│   │   │       │   ├── obj/
│   │   │       │   │   └── <frame_idx>_<hand_id>_<L|R>.obj
│   │   │       │   ├── hands.jsonl
│   │   │       │   ├── hands_smoothed.jsonl
│   │   │       │   ├── summary.json
│   │   │       │   └── smoothing_report.json
│   │   │       ├── rewards/
│   │   │       │   └── scores.json
│   │   │       └── overlays/
│   │   │           └── hands.mp4
│   │   └── 3d/
│   │       └── (same layout as 2d/: microscope/, or_view/, cam_view_[1-9]/)
│   └── session_02/ … session_10/
│       └── (same layout as session_01/)
└── surgeon_02/ … surgeon_16/
    └── (same layout as surgeon_01/)
```

Figure 7: **DiMiRoS dataset organization.** Recordings and associated metadata are organized by surgeon, session, visualization mode, and camera view.

Table 6: Camera configuration of DiMiRoS. Each trial is recorded from 11 viewpoints at a resolution of 1080×1920.

View group	Device	# Views	Placement / target	FPS
Operative field	iPhone 12	3	Left, right, anterior of phantom	30
Surgeon posture	iPhone 12	4	Left, right, anterior, posterior of surgeon	30
Overhead	iPhone 12	2	Phantom and hands; upper body	30
Microscope	Modus V camera	1	Standard optical axis	59.94
OR overview	Panasonic AW-HE40	1	Top-down oblique view of setup	60
Total		11		

A.2 Acquisition and Equipment

The procedure uses 8-0 Ethilon thread and the Ethicon 2808G, BV130-5 needle (J&J Med Tech). The protocol schedules ten trials in 2D followed by ten in 3D, with the scored trials available in each cohort and mode reported in Table 2.

Recording environment. All sessions were recorded in the surgical operating room complex of Erasmus MC under standardized conditions (controlled lighting, minimal background noise). Participants performed a phantom-based microsurgical task, a controlled and reproducible setting that approximates both skills-lab training and a real clinical microsurgical procedure.

Digital robotic microscope. Participants operated with the Modus V digital exoscopic robotic microscope (Synaptive Medical Inc., Canada): a camera on a robotic arm streams the operative field to a 55-inch Ultra HD 4K monitor, and a $12.5\times$ optical zoom provides the magnification for fine microsurgical manipulation. The Modus V supports 2D and stereoscopic 3D visualization, the display configurations of contemporary digital operating rooms.

Surgical phantom. Participants performed end-to-end microvascular anastomosis on a simulated blood vessel phantom of 3 mm diameter (ASY-1065, Synaptive Medical Inc.), secured horizontally on a pedestal inclined 45° towards the surgeon over a blue surgical background for visual contrast. The setup was identical for both visualization modes, enabling systematic comparison of 2D and 3D.

Multi-view camera configuration. Each trial was recorded by 11 cameras (Table 6) to capture both fine-grained manipulation in the operative field and the surgeon's whole-body motor behaviour: nine smartphone cameras (iPhone 12), the Modus V microscope camera, and an overhead operating-room (OR) camera (Panasonic AW-HE40). The 11 views form four groups: i. **Operative field (3 views):** smartphone cameras left, right, and anterior of the phantom (the anterior one facing the surgeon), all focused on the operative field, ii. **Surgeon posture (4 views):** smartphone cameras at operating-table height left, right, anterior, and posterior of the surgeon, capturing body posture, iii. **Overhead (2 views):** smartphone cameras above the surgical field, one framing the phantom and the surgeon's hands, the other the surgeon's upper body, iv. **Microscope and OR (2 views):** the microscope camera records the operative field along its standard optical axis, recording the field presented to the surgeon, while the OR camera gives a top-down oblique overview of the surgical setup.

A.3 Metadata Extraction and Coverage

Recording selection and counts. Table 7 lists the microscope and hand-view recordings selected for metadata extraction, distinguishing scored trials from the individual views that share each trial's score. Annotation totals include labelled frame records from both annotators, whose labels for the same source frame are combined when extracting temporal metadata.

Hand reconstruction. ViTDet (Li et al., 2022) detects the person and ViTPose (Xu et al., 2022) locates the hands before HaMeR reconstruction. Each MANO estimate has 778 vertices, 3D joints, handedness, pose and shape parameters, and camera parameters. Projected keypoints and vertices connect the reconstruction to the source image.

Storage and coordinate conventions. Image metadata uses a common 1920×1080 coordinate system. Masks are stored using COCO run-length encoding (Lin et al., 2014). Each video has a manifest containing source identity, native resolution and frame rate, transformation to the common

Table 7: **Videos selected for metadata extraction.** Counts by visualization mode and experience. Hand-view videos provide hand geometry, while microscope videos provide object masks and, where annotated, boxes and phases.

Mode	Cohort	Microscope	Hand view	Total
2D	Experienced	58	52	110
2D	Semi-experienced	28	47	75
2D	Inexperienced	24	43	67
3D	Experienced	50	60	110
3D	Semi-experienced	39	50	89
3D	Inexperienced	10	40	50
Total		209	292	501

image coordinates, and extraction summaries. Mask and box records use JSONL indexed by native frame number, with timestamps expressed in the video's own timebase. Mesh arrays are compressed per-frame NPZ files, with a lightweight 2D hand index in JSONL. The hand records retain vertices, joints, MANO pose and shape, full-frame camera translation, focal length, principal point, and handedness. Rescaling a display changes 2D projections and intrinsics, not the stored 3D coordinates. Masks use nearest-neighbour resizing after COCO-RLE decoding. Source images can be regenerated from native indices, so storing a JPEG for every sampled frame is optional. These conventions preserve frame identities when metadata from different extraction steps is combined.

Manual annotation conventions. The two annotators jointly labelled 100 frames to agree on conventions, discussed ambiguous cases with a supervising researcher, then divided the remaining frames. Boxes cover the visible extent of objects, including blurred and partially occluded objects. Action cues are sparse relative to instrument boxes: cutting, for example, appears in 38 annotated frame records. All manually annotated frames are retained alongside the regular 5 fps samples.

Annotation matching and vocabulary. We match annotated images to their source videos before extracting masks or phases. Candidate matches are checked using five images from each candidate video and a median normalized cross-correlation threshold of 0.90. Accepted matches score 0.999–1.000, compared with 0.08–0.77 for incorrect candidates. Matching also recovers the original frame indices. Annotations contribute to temporal metadata after this verification, yielding 11 microscope videos for the phase benchmark.

The 50 standardized class names used for metadata extraction are given in Table 8, after reconciling naming variants within the nominal 54-class annotation scheme. Class IDs are resolved per archive before aliases such as `Instrument1Forceps`→`Inst1Forceps` are applied. Hand aliases follow the annotation's instrument-side convention rather than an independent determination of handedness. The annotation files use five class-list variants, so numerical class IDs are interpreted using the class list associated with each file. Action classes denote annotated regions of an event and are retained separately from physical-object track labels.

Text-prompted masks and label propagation. The microscope concept list is `silicone tube`, `suture needle`, `suture thread`, `tweezers`, `needle holder`, and `gloved hand`. SAM 3 returns masks, bounding boxes, and scores for each prompt (Carion et al., 2025), with overlapping masks used to suppress duplicate detections across prompts. Each label maps to the common vocabulary and retains its source phrase, but these framewise predictions do not assign persistent identities across the video. The propagation route instead tracks seven physical identities: `Vessel-Inst1side`, `Vessel-Inst2side`, `Needle`, `Inst1Forceps`, `Inst2NeedleHolder`, `Hand-Inst1`, and `Hand-Inst2`. Manual boxes supply the initial labels, which SAM 3 video tracking carries between annotated frames. We combine predictions from the nearest annotated frames and periodically restart tracking from manual boxes. Annotated microscope videos are processed by both routes: free-text masks provide a comparison without manual seeds and the only available thread masks.

To assess propagation, we withhold annotated frames from the prompts and compare the predicted bounding boxes with the withheld boxes using intersection over union (IoU). We also compare text-prompted detections at class and object-group levels. This evaluates object localization against the available bounding-box annotations.

Table 8: **Standardized annotation vocabulary.** Class names grouped by their common label. Action names are grouped under thread but remain distinct from physical thread tracks. Original class lists contain 53–55 entries.

Group	Annotation names
vessel	`Vessel-Inst1side`, `Vessel-Inst2side`, `Vessel-AfterKnot`
needle	`Needle`
thread	`Insert`, `MakeLoop`, `Pull`, `Knot`, `Cut`
forceps	`Inst1Forceps`, `Inst1Open`, `Inst1Part1`, `Inst1Part2`, `Inst1Part3`, `Inst1Part4`, `Inst1AllParts`, `Inst1End`
needle_holder	`Inst2NeedleHolder`, `Inst2Open`, `Inst2Part1`, `Inst2Part2`, `Inst2Part3`, `Inst2Part4`, `Inst2Part5`, `Inst2AllParts`, `Inst2End`
other	`Inst3`, `Inst3Open`, `Inst3AllParts`, `Inst3End`, `GridPlane`, `BaseCylander`, `WristSensor`, `BlackPoint`, `WhitePoint`, `CoveredPoint`, `ArmPoint`, `FeetPoint`, `HeadRectangle`, `ShoulderRectangle`, `BackRectangle`, `ShoulderPoints`
hand	`Hand-Inst1`, `Hand-Inst2`, `Thumb-Inst1`, `Thumb-Inst2`, `Index-Inst1`, `Index-Inst2`, `Middle-Inst1`, `Middle-Inst2`

Phase extraction and uncertainty. We combine both annotators' labels by source frame and divide a trial into suture cycles at consecutive `Cut` annotations. Within each cycle, phase evidence follows the surgical sequence. The first `Needle` observation marks pick-up. Two separate sequences of `Insert` labels identify passage through the first and second vessel ends, while inseparable sequences are recorded as a combined passage. A `Pull` annotation between insertion and knot tying supports the pulling phase. Otherwise, an interval inferred between these phases is marked as inferred.

The first, second, and third knots are identified from sequences of `MakeLoop` annotations separated by at least three seconds, up to three knots per cycle. A `Knot` annotation marks completion before the next loop. Starts inferred from a knot annotation without a preceding loop are marked as uncertain. Missing evidence does not establish completion. The first cycle may also include a separate SAM-derived record of needle visibility, stored separately from the human annotations that supply the phase benchmark's starts.

For visualization, phases extend to the next phase's start, with a three-second buffer after the final phase. The metadata stores observed endpoints separately from interpolated intervals, together with the source annotations, cycle index, and uncertainty flags. Missed cuts can merge cycles, and sparse annotations can merge phases. These labels therefore provide temporal supervision derived from object and action boxes rather than exhaustive frame-by-frame phase labels. Phase times follow the microscope video, while hand-view videos retain their own timestamps.

A.4 Annotation distribution by object group

Figure 9 groups the 45 classes shown in Fig. 8. Together they contribute 21,425 boxes. This observed vocabulary is distinct from the larger standardized vocabulary listed in Table 8.

A.5 Anastomosis Quality Assessment

Assessor and inspection procedure. A microsurgical trainer from Erasmus MC SkillsLab scored all 308 trials using the adapted instrument (Atlan et al., 2018). The assessor has more than eleven years of teaching experience and has assessed more than five thousand anastomoses on living animal models. Coded trial identifiers concealed expertise cohort and visualization mode. Each repair was inspected under a separate surgical microscope, first from the outside and then after a longitudinal cut exposed the lumen and inner suture line. This inspection reveals defects that may not be visible from the exterior alone.

Rationale for the expert reference. The rubric records technical features associated with anastomotic quality: suture placement and spacing, knot tightening, needle bite, vessel approximation, and distinct major and minor faults. Related work connects such errors to reduced patency (Volovici et al., 2026), supports end-product assessment in simulation through the Anastomosis Lapse Index (Ghanem et al., 2016), and relates final-product ratings to patency (Stogowski et al., 2021). These studies

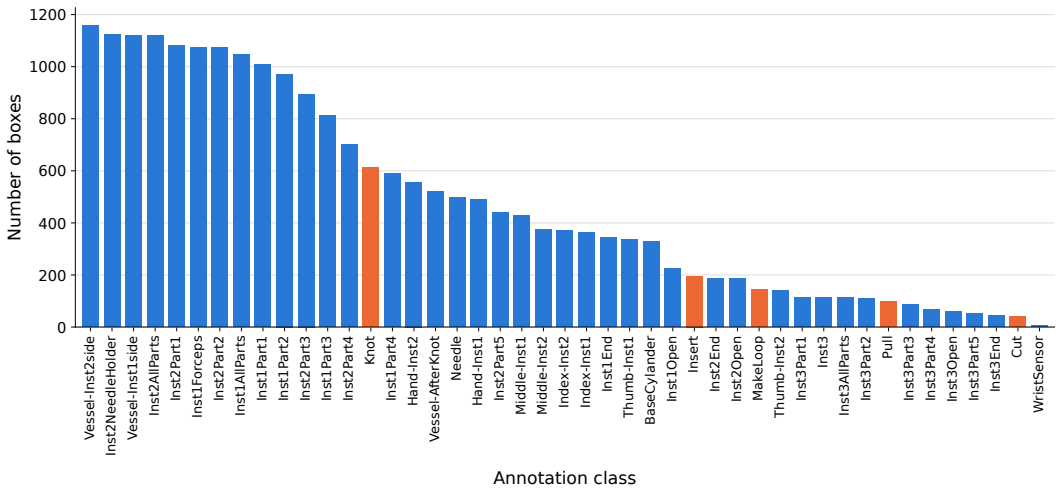

Figure 8: **Distribution of all object and action annotations.** Box counts across all 45 observed classes, with objects and body parts in blue and actions in orange.

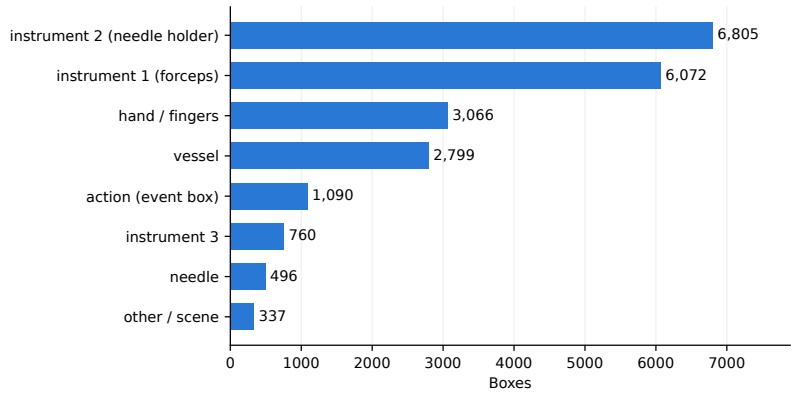

Figure 9: **All manual boxes by object group.** Counts from all 42 annotation archives, including unlinked records. Groups cover the complete 45-class observed vocabulary, including action boxes that supply temporal evidence rather than physical object tracks.

motivate structural assessment without establishing a score-to-patency relationship for DiMiRoS, whose phantom does not provide a functional patency measurement.

Experience influences how raters interpret performance (Govaerts et al., 2011). A single assessor provides a common expert reference throughout the dataset, and blinding keeps cohort and visualization labels from informing that judgement without establishing that assessor bias is absent or constant. Validated assessment instruments can still exhibit differences between raters (Carciumaru et al., 2026). DiMiRoS therefore reports an expert reference rather than a consensus label, and inter-rater agreement cannot be estimated from these scores.

Counting distinct faults. Equation 1 combines the five trial-level criteria with all recorded faults. Every placed suture is checked against both fault categories. A back-wall catch and a vessel-wall tear on the same stitch count as two major faults, just as a major and a minor defect on one stitch both count. The same defect on several stitches is counted once on each affected stitch. Thus four back-wall catches incur four deductions, rather than one trial-level penalty. Opening the lumen allows these defects to be attributed to individual sutures.

The maximum skill sum is 24, but penalties are not capped at that sum. The rubric consequently has no fixed lower bound: a negative value records penalties exceeding the technique score. A high value

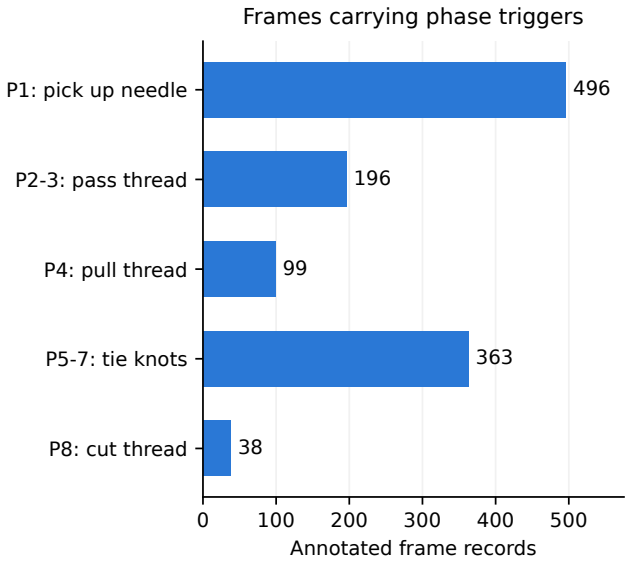

Figure 10: **Annotated frames carrying phase cues.** Bars count frame records with the relevant object or action labels. A frame can contain several cues, so these counts describe annotation coverage rather than the number of phase starts.

indicates few recorded defects and strong technique ratings under this assessment, without predicting functional success in a living vessel.

Zero scores and incomplete attempts. A zero total can describe different behavioural outcomes: an attempt with no suture completed within the time limit, or a repair whose penalties cancel its skill score. Because faults are charged to placed sutures, damage during an unsuccessful pass without a placed suture is not penalized by this instrument. An attempt can therefore score zero despite visible damage, while its video still records needle handling, unsuccessful passes and repositioning.

Zero is therefore neither the bottom of the scale nor an absence of behaviour. An attempt with no placed suture can score above a completed repair carrying many defects. This follows from the instrument's emphasis on the resulting repair rather than effort or elapsed task progress.

Using the scores for learning. Component scores distinguish attempts with no completed suture from trials whose penalties cancel their skill scores. Suture counts and phase labels describe task progress, while total and component scores describe execution quality and its sources. Per-suture faults can also be related to the corresponding placement phases. Retaining negative values preserves variation among trials with repeated defects, which clipping at zero would merge with unsuccessful attempts.

A.6 Performance by Expertise and Visualization

Figures 11 and 12 expand the main-paper comparison into technique components, penalties and assessor-rated risk. Most technique-component means increase in 3D, but major-error frequency rises for semi-experienced and inexperienced participants while falling for experienced surgeons. Minor-error penalties differ: mean penalty magnitude falls from 0.93 to 0.60 for experienced participants, remains near 0.67–0.68 for semi-experienced participants, and falls from 0.43 to 0.35 for inexperienced participants. These descriptive comparisons retain the protocol's fixed 2D-then-3D order, so visualization and accumulated practice are not separated.

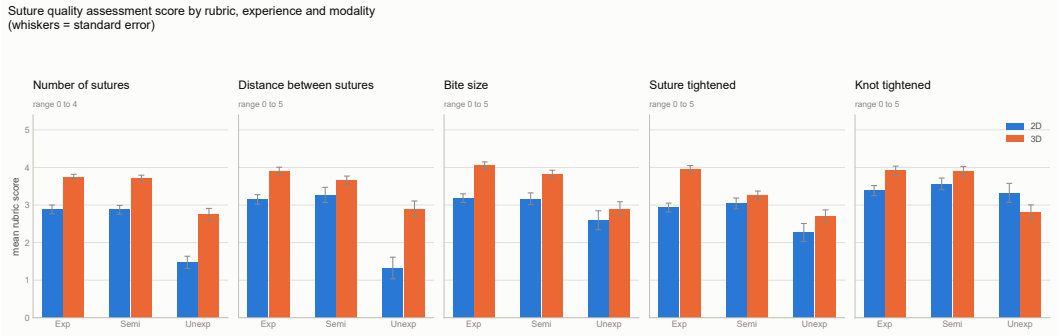

Figure 11: **Technique components expose different aspects of dexterity.** The five rubric components are shown separately by cohort and visualization mode, with whiskers showing standard errors across trials. The panels share a plotting scale, while the completed-suture count is bounded by four.



Figure 12: **Technique scores and errors need not improve together.** Penalties, major-error frequency, and assessor-rated thrombosis/leakage risk by cohort and visualization mode. Each panel retains its own measurement scale. These dataset statistics motivate separate evaluation of technical quality and faults.

B Benchmark Protocol and Reproducibility

B.1 Benchmark Data

The quality test set balances visualization mode and uses a 4/3/3 experience allocation within each mode, with six exemplars covering all mode–experience combinations. Phase exemplars balance the five types, while the 77 eligible test clips retain their observed class imbalance. The phase split is not balanced by visualization mode or experience.

Example selection and clip boundaries. Quality exemplars are 2D S1S3 (experienced), 2D S2S2 (semi-experienced), 2D S8S2 (inexperienced), 3D S1S1 (experienced), 3D S4S2 (semi-experienced), and 3D S16S5 (inexperienced). The fixed test allocation is in Table 9, with participant IDs and expert scores recorded in the benchmark manifest. Exemplars were selected near each group's median expert score, while test trials were selected across score quantiles. Thus the benchmark covers a range of quality within each experience and visualization group. Models use fixed in-context examples without supervised training or validation.

For phase prediction, the fixed exemplars are pick-up at 2D S9S4 frame 17698, insertion at 2D S12S2 frame 4967, pull at 2D S10S1 frame 14145, knot at 3D S3S2 frame 16884, and cut at 3D S2S7 frame 7363. Starts of every type, including excluded starts, constrain clip trimming. Another type before the anchor moves the clip start beyond it, while another type after the anchor moves the clip end to the preceding annotated frame. Same-type starts do not force trimming. Two test clips consequently end at the annotated anchor, with no post-start frames. Exemplar–test overlap is excluded, but test clips

Table 9: **Quality-benchmark allocation.** Each trial has two input videos (microscope and selected hand view). Exemplars are disjoint from test trials, and each prediction receives three same-mode exemplars.

	Exemplar trials		Test trials	
Cohort	**2D**	**3D**	**2D**	**3D**
Experienced	1	1	4	4
Semi-experienced	1	1	3	3
Inexperienced	1	1	3	3
Total trials	3	3	10	10
Input videos	6	6	20	20

Table 10: **Phase-benchmark construction.** Counts are phase starts or clips rather than trials. Exclusions are counted before exemplar selection, and the final column gives the number of test clips for each class metric.

Type	**Raw starts**	**Ambiguous/weak**	**Exemplars**	**Overlap excluded**	**Test clips**
Pick-up	14	0	1	0	13
Insertion	29	5	1	0	23
Pull	6	1	1	0	4
Knot	34	5	1	0	28
Cut	14	3	1	1	9
Total	97	14	5	1	77

may overlap one another. The final phase test set contains 57 2D and 20 3D clips. The class counts in Table 10 retain the observed distribution, and the phase test set is not balanced by experience.

B.2 Benchmark Baselines

Sampling and image preparation. Quality videos are sampled at 1 fps with widths of 640 pixels for the microscope and 448 pixels for the hand view. Images carry frame indices and timestamps. Phase queries span up to five seconds before and after the annotated start at 5 fps, shortened at other action types as described in Appendix B.1, resulting in 26–51 frames per clip. Qwen receives 832-pixel-wide phase images and MedGemma receives 896-pixel-wide images, with identical frame indices. The two vessel passes share the insertion label, and the three knots share the knot label.

Prompt construction. Each quality request is a system message plus one user message. The system describes the task, scoring rubric and output schema. The user prefix contains three expert-graded reference trials from the matching visualization mode. For each reference, ten uniformly spaced microscope indices and four end-state indices are requested and deduplicated (the final uniform index also belongs to the end-state sample), yielding 13 microscope frames in the saved benchmark examples, plus four hand-view frames. Test trials are shown more densely at 1 fps. Each chunk carries explicit frame/time labels and receives only the previous compact record (approximately 250 words) and full scorecard as memory. Scores can be revised on every turn. The final turn receives the two per-view summaries and 12 sampled frames from the final 15 seconds of microscope video, without previous per-chunk reasoning traces.

For phase prediction, the shared prefix contains five type descriptions and one labelled clip per type, each with 11 uniformly spaced frames including the annotated start. The test clip has up to 51 images and asks for the type at the marked start. Labels, object boxes, fine-phase ordinals, and the sample-specific preceding annotation gap are withheld. The expected answer is `PHASE: <type>` followed by `EVIDENCE: <text>`. We use the last parseable `PHASE:` line and map supported names and synonyms to the five types, applying the same parsing rule to answers with multiple labels.

Inference and output handling. Both models use temperature 1.0 and top-$p = 0.95$. Qwen uses top-$k = 20$, presence penalty 1.5, native thinking, and a 12,288-token combined reasoning/answer limit. MedGemma uses top-$k = 64$ and a 4,096-token thinking budget, after which the client closes an unfinished thought block and allows up to 1,536 answer tokens. Both use the 65,536-token server context. Quality inputs are identical frame files before model-specific image processing, whereas phase inputs share frame indices and temporal sampling but differ in width (832/896 pixels). MedGemma inputs are letterboxed before its square image encoding. Complete text templates and model-specific continuation prefixes appear in Appendices E.1 and E.2.

Quality responses use labelled score lines, a running record, observations, and edits (fusion uses scores, total, and rationale). A turn requires at least five of eight score fields to parse. On failure it is retried once with another seed before carrying forward the previous state and logging the failure if neither attempt succeeds. Output processing enforces valid score ranges and caps fault counts at 16, although the assessment rubric itself has no fixed maximum fault count. Missing fields, repairs, forced answers, raw responses, model reasoning, image indices, token counts, and per-turn latencies are retained. Totals are recalculated from the normalized components. For phase prediction, failure after one retry remains an unparsed prediction and is scored wrong.

C Additional Suture Quality Results

C.1 Calibration and Ranking

Table 11 reports total-score error, signed bias and Spearman rank correlation across the 20 quality-test trials. Model rows use fused predictions. Both models overestimate the expert score on average, while their rank correlations show difficulty ordering trials by quality. Risk MAE is reported separately from the reward. Qwen and MedGemma have signed biases of $+4.60$ and $+7.00$, and Spearman correlations of -0.21 and 0.24, respectively.

Table 11: **Quality-score calibration and ranking.** Results across all 20 test trials. Lower MAE and RMSE are better, bias is prediction minus expert, and ρ is Spearman correlation. Risk MAE has $n = 19$ recognized expert labels and is auxiliary to R.

Source	R MAE	R RMSE	R bias	R ρ	Risk MAE
Few-shot mean	3.90	5.00	-0.90	0.45	0.79
Qwen3.5-9B	6.90	8.19	4.60	-0.21	1.00
MedGemma-1.5-4B	7.20	8.67	7.00	0.24	1.05

C.2 Component Errors and Confusion Patterns

MedGemma predicts four completed sutures in every test trial, although only 12 of the 20 trials have four, and predicts no major faults in any trial, missing all nine affected trials. Qwen predicts four sutures in 16 trials, detects major faults in two of the nine affected trials, and reports them in three of the 11 remaining trials. Fusion leaves Qwen's total-reward MAE almost unchanged from microscope-only prediction (6.90 versus 6.85). MedGemma improves from 8.75 to 7.20 with fusion, while its hand-only MAE is 7.00.

Tables 12 and 13 split the same 20 quality-test trials by visualization mode under the same inference settings and exemplar rule. Both fused models have larger total-score error in 2D (Qwen 7.90, MedGemma 9.70) than in 3D (5.90, 4.70), while the few-shot mean remains lower in both groups (4.40, 3.40). Because the targets, participant composition, and exemplar scorecards differ by mode, these differences are not isolated measurements of the visual benefit of 3D. The ordinal confusion matrices in Figs. 13 and 14 show trial counts and reveal how predictions concentrate at a few score values.

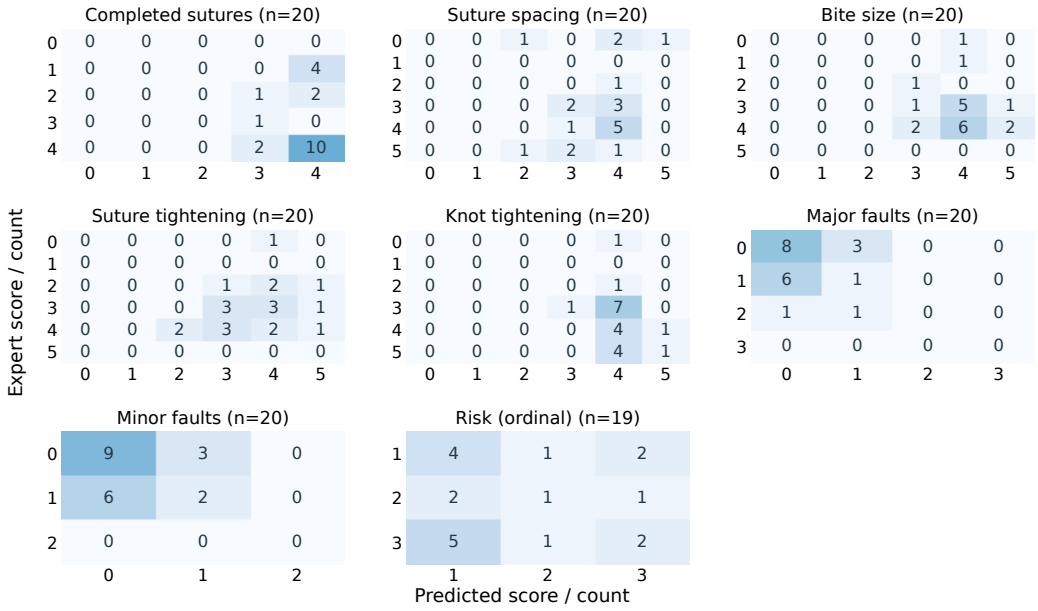

Figure 13: **Qwen3.5-9B fused ordinal predictions.** Expert labels form rows and predictions form columns. Values count trials, with $n = 20$ for the seven components of R and $n = 19$ for risk, whose remaining expert label is missing or unrecognized. Score levels with no reference trials remain visible to show the full scale. Fault panels show counts, and risk uses 1/2/3 for low/middle/high.

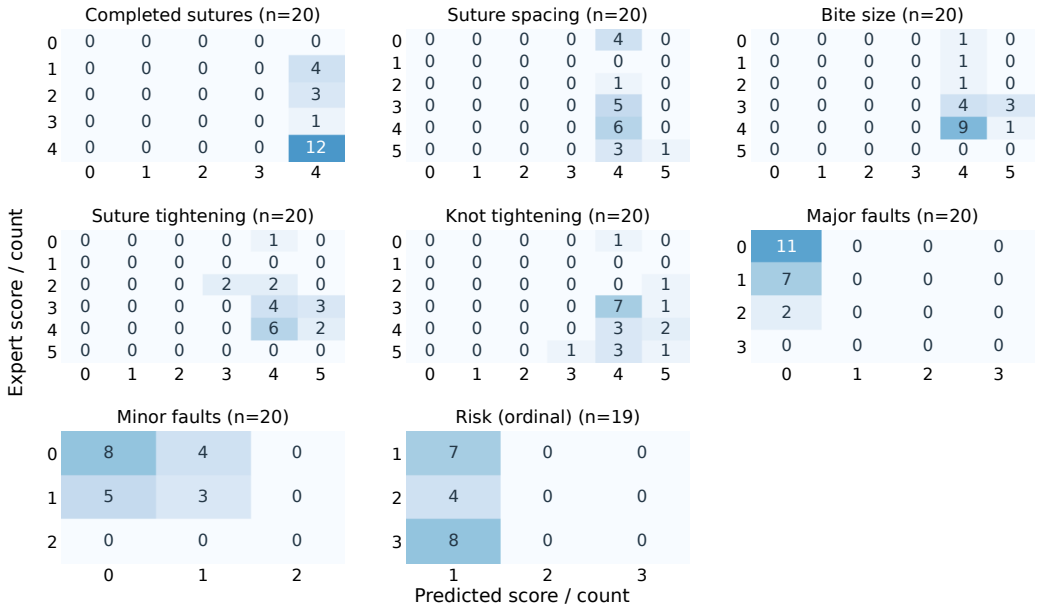

Figure 14: **MedGemma-1.5-4B fused ordinal predictions.** Same conventions and test trials as Fig. 13. Range-corrected predictions are shown, as in the result tables. Constant suture-count and major-fault predictions expose failure modes that an aggregate total alone obscures.

Table 12: **Suture quality prediction in 2D.** MAE ± sample standard deviation of absolute error across 10 2D trials (lower is better). Skill and penalty columns report errors in rubric scores and fault counts, respectively. Micro and hands use individual views, while fused combines their summaries.

		Skill					Penalty		Total
Source / view		**Sutures**	**Spacing**	**Bite**	**Suture tight.**	**Knot tight.**	**Major**	**Minor**	**R**
Few-shot mean		1.10 ±0.74	1.70 ±0.67	1.10 ±1.37	1.00 ±0.82	1.30 ±1.06	0.60 ±0.52	0.60 ±0.52	4.40 ±3.95
Qwen 3.5-9B	hands	2.00 ±1.56	1.90 ±1.91	2.40 ±1.51	2.50 ±1.51	2.80 ±1.99	0.60 ±0.70	0.40 ±0.52	9.80 ±5.01
	micro	1.70 ±1.34	2.10 ±1.66	1.20 ±1.40	1.30 ±1.34	1.20 ±1.14	0.50 ±0.53	0.30 ±0.48	8.10 ±5.80
	fused	1.60 ±1.35	2.00 ±1.76	1.20 ±1.40	1.30 ±1.34	1.20 ±1.14	0.50 ±0.53	0.40 ±0.52	7.90 ±5.82
MedGemma 1.5-4B	hands	1.70 ±1.34	2.10 ±1.66	1.20 ±1.40	1.00 ±1.05	1.40 ±1.07	0.60 ±0.70	0.40 ±0.52	9.00 ±5.52
	micro	1.60 ±1.35	2.30 ±1.49	1.30 ±1.34	1.30 ±1.16	1.30 ±1.25	1.70 ±3.68	0.40 ±0.52	12.60 ±12.76
	fused	1.70 ±1.34	2.00 ±1.76	1.40 ±1.43	1.20 ±1.23	1.60 ±1.17	0.60 ±0.70	0.30 ±0.48	9.70 ±4.81

Table 13: **Suture quality prediction in 3D.** MAE ± sample standard deviation of absolute error across 10 3D trials (lower is better). Skill and penalty columns report errors in rubric scores and fault counts, respectively. Micro and hands use individual views, while fused combines their summaries.

		Skill					Penalty		Total
Source / view		**Sutures**	**Spacing**	**Bite**	**Suture tight.**	**Knot tight.**	**Major**	**Minor**	**R**
Few-shot mean		0.20 ±0.63	0.60 ±0.70	0.70 ±0.48	0.90 ±0.74	0.60 ±0.70	0.50 ±0.71	0.60 ±0.52	3.40 ±2.37
Qwen 3.5-9B	hands	0.80 ±0.92	1.50 ±1.43	0.90 ±0.57	0.70 ±0.82	1.00 ±1.25	0.50 ±0.71	0.60 ±0.70	5.50 ±4.22
	micro	0.60 ±1.07	1.00 ±1.15	0.90 ±0.57	1.30 ±0.82	0.80 ±0.63	0.70 ±0.67	0.60 ±0.52	5.60 ±3.37
	fused	0.30 ±0.67	0.90 ±1.10	0.70 ±0.48	1.20 ±0.79	0.60 ±0.52	0.70 ±0.67	0.50 ±0.53	5.90 ±2.69
MedGemma 1.5-4B	hands	0.20 ±0.63	0.60 ±0.70	0.50 ±0.71	0.90 ±0.74	0.60 ±0.52	0.50 ±0.71	0.50 ±0.53	5.00 ±3.71
	micro	0.20 ±0.63	0.60 ±0.70	0.60 ±0.70	1.00 ±0.82	0.70 ±0.48	0.50 ±0.71	0.50 ±0.53	4.90 ±3.93
	fused	0.20 ±0.63	0.60 ±0.70	0.60 ±0.70	1.00 ±0.82	0.70 ±0.48	0.50 ±0.71	0.60 ±0.52	4.70 ±3.83

D Additional Suture Phase Results

D.1 Metrics

We evaluate five phase types on $N = 77$ test clips. For a phase c, let TP_c be the number of correctly labelled clips, FP_c the clips incorrectly assigned to c, and FN_c the clips of c assigned another label. Its support is $TP_c + FN_c$, and the number of predictions of that phase is $TP_c + FP_c$.

Recall and precision. Recall measures the fraction of a phase's clips that the model identifies, while precision measures how often predictions of that phase are correct:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \qquad \text{Precision}_c = \frac{TP_c}{TP_c + FP_c}.$$

Predicting one phase too often can produce high recall, but the resulting false positives reduce precision.

F1 and macro-F1. F1 combines precision and recall through their harmonic mean:

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \qquad \text{Macro-F1} = \frac{1}{5}\sum_{c=1}^{5} F1_c.$$

F1 is zero for a phase with no correct predictions, and precision is set to zero when a supported phase is never predicted. Macro-F1 gives each phase equal weight, regardless of its frequency.

Accuracy and balanced accuracy. Overall accuracy is the proportion of all clips labelled correctly. Balanced accuracy instead averages per-phase recall:

$$\text{Accuracy} = \frac{\sum_c TP_c}{N}, \qquad \text{Balanced accuracy} = \frac{1}{5}\sum_{c=1}^{5} \text{Recall}_c.$$

The distinction matters because the test set contains 28 knot clips but only four pull clips: overall accuracy gives each clip equal weight, whereas balanced accuracy gives each phase equal weight. All five types occur in the full test set, but subset analyses average only over represented types.

Worked example. Qwen correctly identifies 20 of 23 insertion clips and predicts insertion 47 times. Its insertion recall is $20/23 = 0.87$, precision is $20/47 = 0.43$, and F1 is $40/(23 + 47) = 0.57$. Across all types, it correctly labels 32 clips, giving accuracy $32/77 = 0.42$, while the means of its five recalls and F1 values give balanced accuracy 0.35 and macro-F1 0.33, respectively.

Uncertainty and output handling. The 95% confidence intervals use 2,000 bootstrap resamples of the test clips, recomputing each metric after sampling clips with replacement. Unparseable answers count as incorrect, although all responses in the reported runs were parsed. Uniform prediction across five labels gives expected recall and accuracy of $1/5 = 0.20$, with macro-F1 omitted for this analytic chance reference.

D.2 Detailed results

Qwen achieves balanced accuracy 0.35, macro-F1 0.33, and accuracy 0.42, compared with 0.20, 0.15, and 0.26 for MedGemma. Qwen recognizes 20 of 23 insertions, four of 28 knots, and three of nine cuts, whereas MedGemma recognizes 15 insertions, one knot, and no cuts. Neither identifies a pull. They predict pick-up or insertion in 66 and 65 of the 77 clips, respectively, although only 36 clips belong to those types. Insertion precision is 0.43 for Qwen and 0.33 for MedGemma: frequent needle predictions raise insertion recall while obscuring actions involving the thread.

Figure 15 shows the confusion counts. Table 14 reports precision, recall, and F1 for each type, while Table 15 separates the eight fine phases under the five-type prediction task.

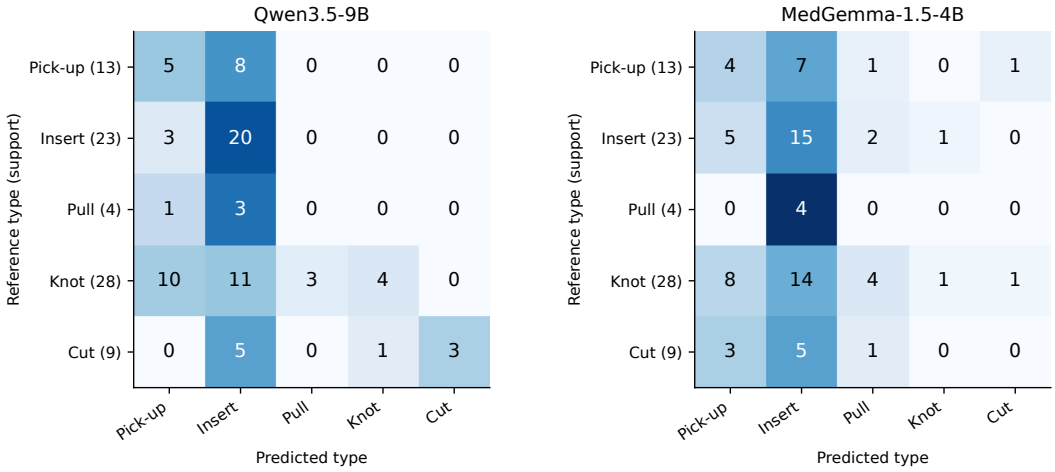

Figure 15: **Phase confusions concentrate on needle actions.** Rows are reference labels and columns are predictions, with numbers showing clip counts and colors normalized by recall within each row. Both models overpredict pick-up and insertion, while knot and pull recall remain low.

The same action remains difficult across its repetitions. Qwen's insertion recall is high for both vessel ends (0.86/0.89), while knot recall remains low across the first, second, and third knots (0.15/0.20/0.00). MedGemma's knot recalls are 0.00/0.00/0.20. The failures therefore extend across the knot sequence, rather than being confined to a particular knot. These results complement the quality benchmark's errors in counting completed sutures: both tasks require a model to follow how repeated motions change the state of the work.

Table 14: **Per-type precision (P), recall (R), and F1.** Support and Pred. count reference and predicted clips.

Type	Support	Pred.	P	R	F1
Qwen 3.5-9B					
Pick-up	13	19	0.26	0.38	0.31
Insert	23	47	0.43	0.87	0.57
Pull	4	3	0.00	0.00	0.00
Knot	28	5	0.80	0.14	0.24
Cut	9	3	1.00	0.33	0.50
MedGemma 1.5-4B					
Pick-up	13	20	0.20	0.31	0.24
Insert	23	45	0.33	0.65	0.44
Pull	4	8	0.00	0.00	0.00
Knot	28	2	0.50	0.04	0.07
Cut	9	2	0.00	0.00	0.00

Table 15: **Accuracy by fine phase.** Clips are grouped by their eight fine phases under the five-type prediction task.

Fine phase	Support	Qwen 3.5-9B	MedGemma 1.5-4B
P1: needle pick-up	13	0.38	0.31
P2: needle pass A	14	0.86	0.64
P3: needle pass B	9	0.89	0.67
P4: thread pull	4	0.00	0.00
P5: first knot	13	0.15	0.00
P6: second knot	10	0.20	0.00
P7: third knot	5	0.00	0.20
P8: thread cutting	9	0.33	0.00

E Complete Benchmark Prompts

This appendix gives the full instruction text and message assembly for each benchmark. Brace-delimited slots denote runtime values; repetitions and image payloads have explicit placeholders. These are templates, not model answers. Phase text comes from the saved run headers; quality text is expanded from the benchmark's prompt and rubric modules. Dynamic ledgers are model outputs and vary by trial and turn.

Both benchmarks rebuild a system message and one multimodal user message for every request; exemplars are reference material in the user message. Qwen uses native server thinking. MedGemma appends the native assistant continuation shown below. The optional prose-based thinking prompts in the code are not used by these evaluated server/native configurations.

E.1 Suture quality prediction: prompts and full in-context sequence

Each request has the quality system prompt, followed by the same-mode three-exemplar prefix and either the current view's chunk request or the final fusion request. Each example supplies 13 microscope frames and four hand-view frames in the saved artifacts, then its complete expert scorecard: 51 exemplar image presentations per request. The six frozen references are listed below; a test trial receives only its three same-mode examples. Exemplar trials are disjoint from test trials, but participants can recur: the literal prompt's phrase *other participants* should not be read as a participant-disjoint guarantee.

At 1 fps, the two test views are processed separately in context-sized chronological chunks. The first chunk has no prior state; later chunks receive only that view's preceding ledger and scorecard. The final chunk uses the last-chunk instruction in place of the provisional-update instruction. Fusion receives both final states and 12 microscope frames sampled from the last 15 seconds, again with the system and exemplar prefix. Missing views use the fallback text below. Decoding, retries, and normalization are detailed in Appendix B.

Quality system prompt (shared by both models)

```
You are an expert assessor of microsurgical skill. You grade microvascular
    anastomoses the way the expert grader of this study does.

## The task being performed
The participant performs an end-to-end anastomosis on a silicone vessel phantom
    with an 8-0 nylon suture under a surgical microscope. The participant's goal
```

```
     is four interrupted sutures at the 12, 3, 6 and 9 o'clock positions around
    the vessel circumference, in a self-selected order. You grade what was
    actually achieved, not what was intended. Each suture cycle is: pick up the
    needle, pass it through the first vessel end, pass it through the second
    vessel end, pull the thread through, tie three knot throws, and an assistant
     cuts the thread. A trial lasts at most 7 minutes. The participant works
    either with a 2D monitor or a stereoscopic 3D display; the recording you see
     is a 2D video in both cases.

## How the grade is given
After the trial, the expert inspects the finished anastomosis under a microscope,
     first from outside and then after cutting it open lengthwise to see the
    lumen. You cannot see that inspection. You see the recording of the
    procedure, and must infer the quality of the finished product from it: where
     and how each needle pass enters the wall, whether the needle catches the
    back wall, how the thread is pulled, how each knot is laid down, whether the
     wall tears, and how the anastomosis looks at the end.

## Rubric
Skill criteria, each rated once for the whole anastomosis:
- `number_of_sutures` (Number of sutures), integer 0-4: How many sutures were
    completed by the end of the trial: passed through both vessel ends, tied and
     cut. A suture that was started but not tied does not count.
- `distance_between_sutures` (Appropriate suture spacing), integer 0-5: Whether
    the completed sutures are evenly distributed around the circumference, at
    roughly equal angular spacing. Spacing cannot be judged with fewer than two
    sutures, and the grader then scores 0.
- `bite_size` (Adequate needle bite from the edge), integer 0-5: Whether each
    needle pass enters and exits at an appropriate, consistent distance from the
     cut edge, symmetric on both vessel ends: neither so close that the wall
    tears out nor so far that the edge folds in.
- `suture_tightened` (Suture tightened / vessel approximation), integer 0-5:
    Whether each suture is pulled tight enough to bring the two vessel ends into
     edge-to-edge contact (approximation), without gaps between the ends and
    without over-tightening that puckers or strangles the wall.
- `knot_tightened` (Adequate tightening of the knot), integer 0-5: Whether each
    knot is fully laid down and secure: throws seated flat on the vessel, no
    loose or air knots, no slipping.

Penalties, counted per placed suture and per distinct fault:
- `n_major_faults`, integer >= 0. Major fault, -4 each: catching the back or side
     wall with the needle, tearing the vessel wall, a large luminal intimal flap,
     thread in the lumen.
- `n_minor_faults`, integer >= 0. Minor fault, -2 each: a small luminal intimal
    flap, a partial-thickness stitch (needle not through the full wall).

- `risk_of_thrombosis_or_leakage`: one of 'small', 'moderate', 'large'. The
    grader's overall judgement of whether this anastomosis would clot or leak.

Trial score R = number_of_sutures + distance_between_sutures + bite_size +
    suture_tightened + knot_tightened - 4 * n_major_faults - 2 * n_minor_faults.
      Maximum 24; no lower bound.

How the grader applies it:
- With fewer than two completed sutures, suture spacing is scored 0.
- With no completed suture, every skill criterion is 0 and no fault is charged,
    since faults are recorded per placed suture.
- Faults are counted per suture and per distinct fault: one suture can carry a
    major and a minor fault, and two different major faults on one suture count
    twice. Across the dataset a trial has 0-3 major and 0-2 minor faults.
- Risk of thrombosis/leakage is 'large' in almost every trial with a major fault,
     and 'small' or 'moderate' otherwise.

## What you will see
```

```
Two views of the same trial, each sampled at 1 frame per second, each frame
    labelled with its native frame index and time:
- MICROSCOPE view: the magnified operative field, showing the two vessel ends,
    the needle, the thread and the instrument tips. This is where bites, knots
    and wall damage are visible.
- CAMERA 8 view: an overhead camera over the table, showing the surgeon's hands
    and instruments. The vessel itself is too small to judge here; this view
    shows handling, pace, how many suture cycles are performed and when the
    trial stops.
The two views are not time-synchronised with each other, so the same time value
    does not mean the same moment in both.

The video is far too long to see at once. It is shown to you in consecutive
    chunks, one view at a time, and you will not see a chunk again. After each
    chunk you keep a compact written record (the ledger) of everything the final
     grade depends on, per suture, and your current scorecard. The ledger is your
     only memory of earlier chunks, so put into it what you will need later and
    nothing else.

## Answer format
Answer in exactly the labelled format given at the end of each request, nothing
    else. One integer per score, no ranges, no extra text on a score line.
    Summarise what happens as events (a needle pass, a knot, a tear) with the
    frame indices where they occur; never describe or list frames one by one, in
     your reasoning or in your answer.
```

Few-shot reference block template

```
## Reference trials graded by the expert
Below are 3 trials from other participants with the expert's scorecards. Each is
    a SPARSE summary: a handful of microscope frames spread over the procedure
    plus the last seconds, where the finished anastomosis is in view, and a few
    camera-8 frames. Use them to calibrate how what you see maps to scores; the
    trial you assess will be shown far more densely.

### Reference trial {n} ({2D or 3D} display)
microscope frames, of {duration_s rounded to integer}s:
[frame {native_index} | t={time_s:.1f}s]<image>
{repeat for all microscope exemplar frames}
cam08 frames, of {duration_s rounded to integer}s:
[frame {native_index} | t={time_s:.1f}s]<image>
{repeat for all hand-view exemplar frames}

Expert scorecard:
{expert scorecard as key: value lines}
{repeat the reference-trial block for the other two same-mode exemplars}
```

Frozen reference scorecards and frame indices. The following score values and native indices are saved with the exemplars. Each scorecard fills the reference template above. Timestamps use the source video's frame-to-time mapping, rounded to one decimal place. Participant IDs identify the examples for reproducibility; they are not additional cohort labels supplied for a test trial.

2d_S1S3

```
2d_S1S3 (experienced)
microscope frames:
0, 2520, 5100, 7620, 10140, 12720, 15240, 17760, 20340, 22020, 22320, 22560,
    22860
cam08 frames:
0, 3780, 7530, 11310
```

```
Expert scorecard:
number_of_sutures: 2
distance_between_sutures: 4
bite_size: 4
suture_tightened: 3
knot_tightened: 2
n_major_faults: 0
n_minor_faults: 1
risk_of_thrombosis_or_leakage: small
total_score_with_errors: 13
```

2d_S2S2

```
2d_S2S2 (semi_experienced)
microscope frames:
0, 2520, 4980, 7500, 9960, 12480, 14940, 17460, 19920, 21600, 21900, 22140, 22440
cam08 frames:
0, 3600, 7200, 10800

Expert scorecard:
number_of_sutures: 3
distance_between_sutures: 3
bite_size: 4
suture_tightened: 3
knot_tightened: 4
n_major_faults: 1
n_minor_faults: 1
risk_of_thrombosis_or_leakage: large
total_score_with_errors: 11
```

2d_S8S2

```
2d_S8S2 (unexperienced)
microscope frames:
0, 2460, 4920, 7380, 9840, 12240, 14700, 17160, 19620, 21240, 21540, 21780, 22080
cam08 frames:
0, 3540, 7080, 10620

Expert scorecard:
number_of_sutures: 1
distance_between_sutures: 0
bite_size: 3
suture_tightened: 2
knot_tightened: 4
n_major_faults: 1
n_minor_faults: 0
risk_of_thrombosis_or_leakage: large
total_score_with_errors: 6
```

3d_S1S1

```
3d_S1S1 (experienced)
microscope frames:
0, 2340, 4680, 6960, 9300, 11640, 13980, 16260, 18600, 20100, 20400, 20640, 20940
cam08 frames:
0, 3330, 6690, 10020

Expert scorecard:
```

```
number_of_sutures: 4
distance_between_sutures: 4
bite_size: 4
suture_tightened: 4
knot_tightened: 4
n_major_faults: 0
n_minor_faults: 1
risk_of_thrombosis_or_leakage: small
total_score_with_errors: 18
```

3d_S4S2

```
3d_S4S2 (semi_experienced)
microscope frames:
0, 2220, 4440, 6600, 8820, 11040, 13260, 15420, 17640, 19020, 19320, 19560, 19860
cam08 frames:
0, 3120, 6240, 9360

Expert scorecard:
number_of_sutures: 4
distance_between_sutures: 3
bite_size: 4
suture_tightened: 3
knot_tightened: 4
n_major_faults: 1
n_minor_faults: 0
risk_of_thrombosis_or_leakage: moderate
total_score_with_errors: 14
```

3d_S16S5

```
3d_S16S5 (unexperienced)
microscope frames:
0, 2220, 4440, 6600, 8820, 11040, 13260, 15420, 17640, 19020, 19320, 19560, 19860
cam08 frames:
0, 2940, 5850, 8790

Expert scorecard:
number_of_sutures: 4
distance_between_sutures: 4
bite_size: 2
suture_tightened: 4
knot_tightened: 1
n_major_faults: 0
n_minor_faults: 1
risk_of_thrombosis_or_leakage: moderate
total_score_with_errors: 13
```

View description substitutions

```
microscope:
MICROSCOPE view: the magnified operative field, showing the two vessel ends, the
    needle, the thread and the instrument tips. This is where bites, knots and
    wall damage are visible.

cam08:
CAMERA 8 view: an overhead camera over the table, showing the surgeon's hands and
     instruments. The vessel itself is too small to judge here; this view shows
```

```
    handling, pace, how many suture cycles are performed and when the trial
    stops.
```

State text for first and subsequent chunks

```
First chunk:
(this is the first chunk of this view: no record yet)

Later chunks:
Ledger:
{previous ledger, or (empty)}

Current scorecard:
{previous scores as key: value lines}
```

Per-view chunk request and answer schema

```
## Trial under assessment ({2D or 3D} display)
{view description}
Chunk {chunk_no} of ~{n_chunks_est} of this view: frames {first}-{last}, t={
    first_time:.1f}-{last_time:.1f}s of {duration:.1f}s ({n_frames} frames at 1
    per second).

### Your record from the earlier chunks of this view
{state text}

### New frames
[frame {native_index} | t={time_s:.1f}s]<image>
{repeat for every frame in this chunk}

### Instruction
Update your record with these frames. Keep a score provisional where the evidence
     is not in yet (e.g. sutures not yet placed), and revise any score freely
    when new frames change your judgement.
Answer in this format:
SCORES:
number_of_sutures: <integer 0-4>
distance_between_sutures: <integer 0-5>
bite_size: <integer 0-5>
suture_tightened: <integer 0-5>
knot_tightened: <integer 0-5>
n_major_faults: <integer count>
n_minor_faults: <integer count>
risk_of_thrombosis_or_leakage: <small|moderate|large>
LEDGER: <your full updated record, replacing the previous one: per suture (
    position, status, bites, tightening, knot, faults seen with frame indices)
    plus overall notes; at most ~250 words>
OBSERVATIONS: <at most 3 sentences: the events in THESE frames that bear on the
    grade, with frame indices>
EDITS: <which scores you changed and why, or none>
```

Replacement instruction for the last chunk

```
This is the LAST chunk of this view: the video ends here. Your scores now are
    your final grade from this view alone.
```

Final fusion request and answer schema

```
## Final grade for the trial under assessment ({2D or 3D} display)
You have now watched both views in full. Your records from each:

### microscope
{final microscope state text}

### cam08
{final camera-8 state text}

### The last {span_s rounded to integer}s of the microscope view again, the
    finished anastomosis:
[frame {native_index} | t={time_s:.1f}s]<image>
{repeat for the 12 selected tail frames}

### Instruction
Combine the two views into the expert's final scorecard. The microscope view is
    the primary evidence for bites, tightening, knots and wall faults; camera 8
    mainly supports the count of completed sutures. Where they disagree, say why
     you trust one. Then compute total_score_with_errors with the formula of the
     rubric.
Answer in this format:
SCORES:
number_of_sutures: <integer 0-4>
distance_between_sutures: <integer 0-5>
bite_size: <integer 0-5>
suture_tightened: <integer 0-5>
knot_tightened: <integer 0-5>
n_major_faults: <integer count>
n_minor_faults: <integer count>
risk_of_thrombosis_or_leakage: <small|moderate|large>
TOTAL: <total_score_with_errors, integer>
RATIONALE: <per criterion, the evidence behind the score and which view it came
    from>
```

Fallback when all chunks of a view fail

```
(no usable record: every chunk of this view failed, so judge from the other view)
```

MedGemma native thought and forced-answer continuation

```
Initial assistant prefix for MedGemma:
<unused94>thought

If the thinking budget is exhausted, continue this assistant message:
<unused94>thought
{generated thought}
<unused95>SCORES:
```

E.2 SUTURE PHASE PREDICTION: PROMPTS AND FULL IN-CONTEXT SEQUENCE

Each request has the phase system prompt and one user message containing the five-exemplar prefix followed by one test-clip query. All five examples repeat in fixed order: pick-up, insertion, pull, knot, cut. Each contributes 11 frames from -5 to $+5$ seconds including the anchor: 55 exemplar image presentations plus 26–51 test images per request. No exemplar evidence narrative is supplied, only the type label. The examples span three 2D and two 3D videos. Exemplars and test clips can come from the same recording if their windows do not overlap; the prefix's phrase *other recordings* is not a recording-disjoint split guarantee.

The logged prefix below includes every exemplar frame index, relative timestamp, anchor marker, and answer label; only images are placeholders. Both models share this text and frame selection, with their respective 832- and 896-pixel image widths. The query uses the actual trimmed extent. No preceding annotation gap, fine-phase number, object boxes, or test label is included.

Phase system prompt (shared by both models)

```
You are an expert in microsurgery who recognises the surgical phases of
    microvascular anastomosis.

## The task being recorded
The participant performs an end-to-end anastomosis on a silicone vessel phantom
    (3 mm, two vessel ends held horizontally on a blue background) with an 8-0
    nylon suture under a surgical microscope, placing several interrupted
    sutures one after another. The view is the microscope's: the magnified
    operative field with the two vessel ends, the curved needle, the thread, the
     forceps and the needle holder.

## Phase types
Each suture runs through these phases, in this order: pickup -> insert -> insert
    -> pull -> knot -> knot -> knot -> cut, then the next suture.
- pickup (needle pick-up): The needle holder grasps the curved needle (from the
    blue background, or re-grasps it after the previous suture was cut) and
    loads and orients it for the first bite, often helped by the forceps. It
    ends when the needle tip reaches the vessel wall.
- insert (needle pass (insert)): A needle pass: the forceps steady or evert the
    edge of one vessel end and the needle holder drives the curved needle
    through that wall with a rotation of the wrist, until the tip comes out on
    the other side and is retrieved. Each suture has two of them, one through
    each vessel end; both are this type.
- pull (thread pull): The needle has come out of the second vessel end; the
    needle holder or forceps draws the long thread through both walls, away from
     the vessel, until only a short tail is left and the two ends come together.
     No loop is being wrapped yet.
- knot (knot tying): A knot throw: the long thread is wrapped around an
    instrument tip, the short tail is grasped and pulled through the loop, and
    the throw is tightened down onto the vessel. Each suture has three throws;
    all are this type.
- cut (thread cutting): The assistant brings scissors into the field and cuts
    both thread ends just above the finished knot.

Telling apart the types that look alike:
- pull vs knot: drawing long thread through with no loop is pull; as soon as the
    thread is wrapped around an instrument it is knot.
- insert vs pickup: handling and loading the needle away from the wall is pickup;
     once the tip goes into a vessel wall it is insert.
- cut vs pickup: the needle can stay in view after a cut; the scissors cutting
    the thread is cut, the needle holder taking the needle for the next suture
    is pickup.

## What you will see
A short clip in which ONE phase starts, at the frame marked ANNOTATED START (t=0)
    , usually in the middle of the clip. Frames are labelled with their frame
    index and their time relative to the annotated start: t=0 is the first
    annotated frame in which the new phase is under way. Annotators labelled only
     some frames, one every few seconds, so the phase may have begun somewhat
    before t=0. Frames before t=0 mostly show the phase that is ending; frames
    after it the phase that has begun.

Before the clip you get reference examples, one or more per type, each with its
    annotated type.

## Answer
```

```
Name the type of phase that BEGINS AT THE FRAME MARKED ANNOTATED START (t=0). The
     clip starts a few seconds earlier, so its first frames usually show the
    phase that is ENDING: do not answer with what is visible at the beginning of
     the clip. Compare the frames just before and just after t=0; the answer is
    the phase under way after t=0. Answer in exactly the format given at the end
     of the request, nothing else. Describe what happens as events with frame
    indices; never describe or list frames one by one, in your reasoning or in
    your answer.
```

Complete frozen five-example prefix

```
## Reference examples
Below are 5 annotated clips from other recordings, one or more per type. Each is
    a SPARSE summary of its clip: a few frames spread over the same time window;
     the clip you classify will be shown densely (5 frames per second).

### Example 1: a pickup (needle pick-up) begins at t=0
[frame 17398 | t=-5.0s]<image>
[frame 17458 | t=-4.0s]<image>
[frame 17518 | t=-3.0s]<image>
[frame 17578 | t=-2.0s]<image>
[frame 17638 | t=-1.0s]<image>
[frame 17698 | t=+0.0s | ANNOTATED START]<image>
[frame 17758 | t=+1.0s]<image>
[frame 17818 | t=+2.0s]<image>
[frame 17878 | t=+3.0s]<image>
[frame 17938 | t=+4.0s]<image>
[frame 17998 | t=+5.0s]<image>
Annotated answer:
PHASE: pickup

### Example 2: a insert (needle pass (insert)) begins at t=0
[frame 4667 | t=-5.0s]<image>
[frame 4727 | t=-4.0s]<image>
[frame 4787 | t=-3.0s]<image>
[frame 4847 | t=-2.0s]<image>
[frame 4907 | t=-1.0s]<image>
[frame 4967 | t=+0.0s | ANNOTATED START]<image>
[frame 5027 | t=+1.0s]<image>
[frame 5087 | t=+2.0s]<image>
[frame 5147 | t=+3.0s]<image>
[frame 5207 | t=+4.0s]<image>
[frame 5267 | t=+5.0s]<image>
Annotated answer:
PHASE: insert

### Example 3: a pull (thread pull) begins at t=0
[frame 13845 | t=-5.0s]<image>
[frame 13905 | t=-4.0s]<image>
[frame 13965 | t=-3.0s]<image>
[frame 14025 | t=-2.0s]<image>
[frame 14085 | t=-1.0s]<image>
[frame 14145 | t=+0.0s | ANNOTATED START]<image>
[frame 14205 | t=+1.0s]<image>
[frame 14265 | t=+2.0s]<image>
[frame 14325 | t=+3.0s]<image>
[frame 14385 | t=+4.0s]<image>
[frame 14445 | t=+5.0s]<image>
Annotated answer:
PHASE: pull

### Example 4: a knot (knot tying) begins at t=0
[frame 16584 | t=-5.0s]<image>
```

```
[frame 16644 | t=-4.0s]<image>
[frame 16704 | t=-3.0s]<image>
[frame 16764 | t=-2.0s]<image>
[frame 16824 | t=-1.0s]<image>
[frame 16884 | t=+0.0s | ANNOTATED START]<image>
[frame 16944 | t=+1.0s]<image>
[frame 17004 | t=+2.0s]<image>
[frame 17064 | t=+3.0s]<image>
[frame 17124 | t=+4.0s]<image>
[frame 17184 | t=+5.0s]<image>
Annotated answer:
PHASE: knot

### Example 5: a cut (thread cutting) begins at t=0
[frame 7063 | t=-5.0s]<image>
[frame 7123 | t=-4.0s]<image>
[frame 7183 | t=-3.0s]<image>
[frame 7243 | t=-2.0s]<image>
[frame 7303 | t=-1.0s]<image>
[frame 7363 | t=+0.0s | ANNOTATED START]<image>
[frame 7423 | t=+1.0s]<image>
[frame 7483 | t=+2.0s]<image>
[frame 7543 | t=+3.0s]<image>
[frame 7603 | t=+4.0s]<image>
[frame 7663 | t=+5.0s]<image>
Annotated answer:
PHASE: cut
```

Test-clip query and answer schema

```
## Clip to classify ({2D or 3D} display)
{n_shown} frames at 5 per second, t={negative_pre_s:+.1f}s to {post_s:+.1f}s
    around the annotated start at frame {anchor_index}.

### Frames
[frame {native_index} | t={relative_time_s:+.1f}s]<image>
{repeat for every selected test frame; at the anchor use:}
[frame {anchor_index} | t=+0.0s | ANNOTATED START]<image>

### Instruction
Which type of phase begins at frame {anchor_index}, the frame marked ANNOTATED
    START (t=0)? Not what the clip shows at its beginning: compare the frames
    just before and just after t=0, and name the phase under way after t=0.
    Answer in this format:
PHASE: <one of: pickup, insert, pull, knot, cut>
EVIDENCE: <at most 3 sentences: what changes around t=0, what begins after it,
    and what rules out the other types, with frame indices>
```

MedGemma native thought and forced-answer continuation

```
Initial assistant prefix for MedGemma:
<unused94>thought

If the thinking budget is exhausted, continue this assistant message:
<unused94>thought
{generated thought}
<unused95>PHASE:
```

Retries. A retry repeats the benchmark messages with another seed. When a quality-prediction request exceeds the context budget, the chunk is shortened and its frame and count fields are rebuilt.