

First In Vivo Feasibility Study of Humanoid Robots in Surgery

Corresponding Author: Mr Zekai Liang

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Comments;

The manuscript presents a humanoid tele-lap framework evaluated through benchtop workspace and precision testing, dry-lab user studies (ring transfer and FLS peg transfer comparing humanoid, manual lap, and robo platforms, and two in-vivo porcine chole. The staged evaluation design is appropriate and demonstrates technical feasibility of humanoid-assisted lap task execution.

Originality and Significance

The in-vivo teleoperated humanoid cholecystectomy is noteworthy.

However, the primary novelty lies in system integration & staged evaluation rather than conceptual breakthrough.

The manuscript should position its contribution more explicitly relative to prior teleoperated lap robotics literature and existing humanoid medical task teleoperation studies. Claims of being "first" should be precisely scoped and supported with appropriate references to avoid overstating novelty.

Data & Methodology (Validity, Data Quality, Presentation)

The staged approach (benchtop → dry-lab → in-vivo) is appropriate.

However, several aspects require strengthening for interpretability and reproducibility:

- * Provide a clearer ---> tool-tip error sources including extended kinematic chain effects, RCM calibration, latency, joint compliance/backlash, tracker error.

- * Expand description of RCM calibration, drift handling, and enforced safety constraints.

- * Clarify in-vivo procedural protocol (team roles especially that surgeon was assisting mostly, pause/recalibration criteria, thresholds to convert, sterile workflow, instrument exchange details).

- * Report latency characterization (mean and variance) and relate to performance observations.

Currently, safety characterization and robustness analysis are insufficiently detailed.

Appropriate Use of Statistics / Treatment of Uncertainties

Please;

- * Justify statistical test selection and state distribution assumptions.

- * Include effect sizes and confidence intervals, not only p-values.

- * Clarify normalization methods used in FLS scoring and whether they introduce bias.

Conclusions: Robustness, Validity, Reliability

Conclusions suggesting efficiency, scalability, ergonomic superiority, or workflow enhancement exceed the supporting evidence.

The platform remains console-operated, and human bedside assistance was required during cases. Claims regarding reduced congestion, preserved sightlines, improved team communication, or "free console setup" are speculative and unsupported by objective ergonomic or workflow metrics. The manuscript should more clearly separate feasibility

(demonstrated) from clinical readiness (not demonstrated).

Suggested Improvements for Revision

1. Strengthen statistical reporting (effect sizes, CI, justified test selection).
2. Provide detailed safety architecture and error budget.
3. Quantify latency and correlate with task performance.
4. Expand in-vivo procedural transparency for reproducibility.
5. Temper speculative workflow and cost/scalability claims unless supported by data.
6. Add performance metrics relevant to MIS safety (e.g., smoothness, overshoot, collisions, time out of view).

References:

The manuscript would benefit from clearer positioning relative to prior work in teleoperated laparoscopy, RCM control methodologies, and humanoid medical teleoperation studies.

Explicit comparative discussion is needed to avoid overstating novelty or implying parity with mature surgical robotic systems.

Clarity and Context

The manuscript is generally readable but the abstract and introduction are somewhat broad. Market and funding context can be reduced in favor of earlier emphasis on the surgical evaluation question and success criteria. The conclusion should align more closely with demonstrated feasibility rather than extend toward clinical readiness or system superiority claims.

Overall Recommendation: Major Revision

Technically compelling feasibility work. With strengthened methodological rigor, clearer safety characterization, more balanced claims, and improved statistical reporting, this could become a meaningful contribution to the field.

Referee #2

(Remarks to the Author)

General summary:

The manuscript demonstrates impressive translation of humanoid robots into in vivo surgery as a substitution for popular teleoperation embodiments. The main technical contribution is the integration of the humanoid into the teleoperation chain with a custom end-effector mount that enables control of non-robotic, wristed laparoscopic instruments through humanoid hands. The work is evaluated through benchtop studies and two in vivo porcine cholecystectomies.

Despite these impressive contributions and strong interest of this work to the community, the technical contributions of the work are limited and there are some key elements of the results that need to be addressed (i.e. limited expert users, limited analysis for benchtop comparisons, missing discussion points). Furthermore, the motivation for using the humanoid in laparoscopic surgery in place of a conventional fixed-based patient side manipulator is unclear.

Key results:

- This work demonstrates and evaluates a novel clinical application of humanoids in surgery, specifically to assist with laparoscopic surgical tasks. In general, humanoids have been proposed or investigated as surgical assistants operating at the patient bedside or as substitutes for nurses. This is the first manuscript that I'm aware of that explores the potential use of humanoids as part of the teleoperation chain for robotic surgery. It is also the first in-vivo study I have seen of humanoids in a clinical context.
- The manuscript provides an evaluation in both benchtop settings and in-vivo porcine surgeries. It is highly encouraging and outstanding to see this level of translation, particularly the in-vivo studies, given that humanoid robots have only recently emerged with the appropriate perception, control, and motion capabilities that they have now (as the authors point out in the introduction).
- Methodologically, the teleoperation framework that was used to map the master manipulator of the daVinci console to the humanoid is a useful technical contribution.
- The paper answers critical questions and includes thoughtful analysis about components like the robot workspace, user interactions with the system, task completion rates, potential interruptions to the surgical flow, etc.

Validity:

My biggest concern is about the motivation for the chosen deployment of the humanoid. The introduction describes how humanoids inherently fit environments designed to accommodate the human form-factor which I agree with, however, the motivation for replacing the patient side manipulators in the teleoperation chain with the humanoid is unclear to me. The paper states: "Laparoscopy is chosen as a representative benchmark due to its widespread clinical adoption and the availability of established robotic systems such as Intuitive's da Vinci Surgical System." - but purpose-built systems like the da Vinci can accommodate minimally invasive surgery very well and are not impeded by the limited mobility of the base. Additionally, Intuitive has achieved widespread adoption in clinical centers despite some of the limitations pointed out like storage demands, capital costs, etc.

One of the key advantages of a humanoid robot is the bipedal movement which is not taken advantage of in this study. The humanoid form-factor may be useful in open surgical procedures such as trauma procedures where the surgeon is moving around the operating room, however, I don't see the real benefit in minimally invasive surgery. With this, I'd like to see

stronger evidence to motivate the choice to use the humanoid as the patient side manipulator. What does this validation enable and where might the innovations presented here be most beneficial? Could this system be deployed in trauma settings where the surgeon is moving around a lot and the bipedal mobility would be a major asset?

I have some additional concerns about the current manuscript that should be addressed before publication:

- Related to my previous point, in the introduction there are some mentions of other potential use cases for humanoid robots. I recommend additional discussion about how the work presented in this manuscript may support these other use-cases where a humanoid form-factor may be more explicitly beneficial.
- Much of the discussion and motivation is centered around cost, and smaller physical footprint of the humanoid compared to a conventional fixed-based robot. While I agree that the footprint of one humanoid is smaller than the da Vinci, if you had 2 humanoids in place like you do in the supplemental videos to accommodate 3 laparoscopic instruments, I think the footprint comparison becomes weak. In fact, in many procedures it is common to have 4 – 5 ports, which would require up to 3 humanoids at which point, I would guess the footprint and storage demands are not significantly different.
- I would like to see an improved discussion of the actual costs of the humanoid robot in comparison to the da Vinci patient side manipulator. The sentence: “Economic constraints are a major barrier to broader robotic adoption, as high acquisition costs, ongoing maintenance contracts, proprietary consumables, and extensive training requirements have confined robotic surgery largely to well-resourced tertiary medical center” implies that the humanoid will require less acquisition costs, maintenance, less consumables, and less training. However, the current system still needs wristed laparoscopic tools and, in the teleoperation mode like it is shown here, would probably require the same amount of training. Surgical robot costs exceed manufacturing costs of the hardware many folds due to regulatory, liability, marketing, and profit overhead so directly comparing to commercial humanoid sales is unfair. Furthermore, is there reason to believe the maintenance of a humanoid would be different than that of a fixed-based system? More evidence is necessary to justify these claims in the discussion.
- I recommend changing the title to emphasize the key role of teleoperation in this study. The current title “Humanoid Robots in Surgery: First In-vivo Feasibility Study”, does not make it clear to the reader that the humanoid is teleoperated.
- User study: The balanced FLS score is interesting but I would have liked to see some comparison in task performance using the conventional criteria from the FLS program ([https://cpd.partners.org/sites/default/files/manual_skills_guidelines\[1\].pdf](https://cpd.partners.org/sites/default/files/manual_skills_guidelines[1].pdf)) which penalizes based on the pegs falling out of the field of view. If not, then perhaps a better explanation as to why a unique FLS metric was derived for this study. There are also some nice review papers that analyze different data-driven quantitative metrics for evaluating performance in robot-assisted laparoscopy (ie. <https://link.springer.com/article/10.1007/s00464-021-08792-5>). Given the variability in the results with just FLS score, weighted error, and time, I would be interested if some more analysis could be done with these validated metrics to better understand the gaps between manual, daVinci, and humanoid.
- In table 1, it is unclear why the group distinction (surgeons versus novices) is only provided for the o-ring transfer task.
- The user experience and system performance evaluation is also a bit confusing. My understanding is that the four survey participants were attending surgeons, clinical fellows and residents that were present during the in-vivo studies. Although each of these users were present, I would assume that they didn't each complete every task that was assessed (ie. setup and docking, operating from the surgeon's console, tool insertion). If there was systematic assessment of each of these, the description should be explained in more detail, if not, I think the metrics of the user who actually completed each part of the workflow should be separated as that would be more representative (i.e. I want to see the distinct survey results about port placement from the user who consistently helped with port placement).
- I also have some major questions about sterility. The da Vinci robot is typically draped apart from the tools. I imagine that the humanoid would need similar sterility (or could perhaps be sterilized like a human surgeon) and this could potentially block some of the sensors that enable the current motion and stability. Has any analysis been done to this effect? If not, maybe commenting on this and the potential challenges towards clinical translation would benefit the discussion.
- One of the biggest discussion points is the high number of redeployments during the in-vivo studies, characterized as instrument repositioning. I imagine this would be comparable to repositioning done by a human during manual laparoscopic surgery but some sort of comparison to the manual procedure and how frequently repositioning occurs would help with the interpretation of the results.
- Another question I have is about the harnesses that appear to be holding up the humanoids in the videos. I believe one of the biggest concerns with these systems would be the risk associated with a humanoid falling onto the patient during the operation if they were to become unstable. More discussion on the necessity of the harnesses and the stability of the system would help clarify this. Also, if the system requires a ceiling mounted harness, does the workspace argument still hold-up?
- The manuscript would benefit from more discussion about the functional dexterity differences between the conventional PSM and the humanoid. During the benchtop evaluation, did the users note feeling any major differences or limitations?
- In the NASA TLX survey, please consider including the frustration category. It's unclear why this was omitted.
- The radial error in the motion precision evaluation of over 10mm is massive and much larger than the allowable error in the FLS circular cutting task of 2mm. This needs to be addressed.
- Latency needs to be measured and reported and compared to SOTA teleop systems and performance impact discussed.

More minor concerns:

- Define acronyms like MTM on first occurrence.
- More details should be included about the motion precision evaluation with the Optitrack system (ie. how was the tool tip tracked, calibration error, etc.), a photo of the experimental setup would be useful. I wonder if the lever arm effect is contributing to the out-of-plane error. Nonetheless, an image that helps convey how the ground truth measurement was obtained would help.
- More details on the RCM calibration using the ArUco markers or a reference to a study that defines this process should be included.
- Minor errors: There are some typos through the manuscript (ie. on page 19, mis-spelling in MTM definition, typo in figure 2 caption (benchtop), missing capital on pg 21 “we parameterize”).
- Figure 1A should be improved to better visualize the MTMs.

- Figure 2 (bi and bii) are a bit hard to interpret at first glance, I wonder if an additional side view of the humanoid / patient would make it easier to see the hand and instrument workspace
- Workspace analysis: One small comment is that it would be useful to see the da Vinci workspace size as either a dashed line or additional bar chart on Figure 2(c) for quick reference.
- Revise this sentence, demonstrate what? "Quantitative assessment of surgical performance demonstrated that active console time, in which the surgeon was directly manipulating instruments using the humanoid robot, was seen in Table 2."
- Inconsistent capitalization in the NASA TLX bar chart labels.
- Typo "exchange6" on page 18

Originality and significance:

- As I pointed out in my first comment, this work is highly novel and innovative. I have not seen any other studies that use humanoids in surgery in the capacity described in this manuscript. On a subjective note, I am not convinced this deployment of a humanoid is particularly impactful given the widespread success of purpose-built laparoscopic robot systems and the limited benefits demonstrated in the results. The results here appear to further reinforce the idea that purpose-built robotic systems are better tools for teleoperation. This also ties back to my previous comment about clarifying the motivation for using the humanoid to replace the patient side manipulator.

Data & methodology:

- In my previous comments, I have requested some additional details about the ground truth for the motion precision and some further analysis on the distinction between different robot platforms with more task-specific metrics from the literature. I think these would greatly strengthen the work.
- Please see my comments below about the statistics and uncertainties with respect to the number of expert surgeons in the benchtop user studies.
- The code used to collect this data is already available. I wonder if the authors have any intention to release the specific data (ie. recordings of the in-vivo procedures, benchtop analysis). The provided zenodo link offers a ros2 stack for setting up the humanoid teleoperation complete with urdf descriptions, setup files, and important mathematical definitions. I believe this would be sufficient material to reproduce the system.

Appropriate use of statistics and treatment of uncertainties:

- The benchtop user studies rely on only 2 expert surgeon users which makes me question if the p-values in the surgeon section of table 1 are truly meaningful. Also, there is no standard deviation reported for table 3. More information on the surgeon training (i.e. years of practice in robotic and laparoscopic surgery) would be helpful.
- Generally, I'm not confident that the quantitative results from the benchtop user studies are significant given the limited number of expert users. The novice group is sufficient, but surgeons trained in both manual and daVinci guided laparoscopy will likely have different experiences with the system. I think the study would strongly benefit from more surgeon users for the benchtop studies.

Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

- I think there are some points in the conclusions and data interpretation that can be improved as I have commented on earlier. In particular, the number of expert surgeon users limits the statistical power of the comparisons to manual and dVRK systems. Further, the patterns between benchtop tasks are not distinct, as the authors mention, each system appears to have different strengths and weaknesses with the current analysis. I believe that a more extensive analysis with validated evaluation metrics from the literature will glean more insight on these comparisons.
- Lastly, I think the survey results from the in-vivo work require further information (like standard deviations) or distinction between each of the participants and their specific role during the experiment.

Suggested improvements:

I have pointed out where I think the results are weak in my second comment about the manuscript validity but here is a list of the additional data and experiments that I feel would significantly strengthen the work:

- Further analysis of benchtop experiments with metrics previously validated for robot-assisted laparoscopy
- More details about the ground truth for motion precision evaluation
- Additional expert users for benchtop validations
- Detailed user breakdown for the in-vivo study surveys and standard deviations
- A better comparison of quantitative outcomes in the in-vivo study (like robot re-deployments versus tool repositioning) to manual laparoscopic procedures
- Sterility discussion
- Discussion or analysis of the dependency on the ceiling mounted harness / tethers that are used to support the humanoids

References:

- I believe this manuscript includes appropriate references to the literature, in particular, I really enjoyed how the introduction covers both research and commercial advances.

Clarity and context:

- The abstract and paper are well written, clear, and very accessible to a non-expert audience and I think well-suited for Nature

(Remarks on code availability)

From above:

The code used to collect this data is already available. I wonder if the authors have any intention to release the specific data (ie. recordings of the in-vivo procedures, benchtop analysis). The provided zenodo link offers a ros2 stack for setting up the humanoid teleoperation complete with urdf descriptions, setup files, and important mathematical definitions. I believe this would be sufficient material to reproduce the system.

Referee #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

all concerns were addressed appropriately.

(Remarks on code availability)

all concerns were addressed appropriately.

Referee #2

(Remarks to the Author)

We appreciate the authors detailed response to review letter. There have been some strong improvements to the manuscript. In particular:

- The motivation for this deployment choice for the humanoid is clearer in the manuscript now. I think the strongest argument, which is now highlighted in the introduction, is the multi-purpose aspect of these systems which could potentially make them more economical. One minor comment is that the example of fitting through existing OR door frames that is stated in the response letter (R2.3) is strong and should be in the introduction. Limiting the need to retrofit operating rooms or infrastructure is a major selling point for these humanoid systems.

- The new text in the manuscript that discusses sterility is good and I think essential when discussing the feasibility of humanoids in the OR.

- Responses to the feedback on the latency, error analysis, workspace visualization, RCM calibration, etc. are sufficient and the paper is much clearer now in these areas.

Some remaining concerns:

- R2.6 – No additional analysis was added, the response just indicates how the FLS score was normalized, and some additional statistics were included on the same metrics (time, weighted error). I still feel that additional metrics could be derived from this data to better understand how the humanoid vs. dVRK vs. manual performance gap truly materializes.

- R2.7 – The manuscript would benefit from more explicitly labelling the number of users in each part of the study (n = 12, n=9 in the captions where relevant). Without this, and without the now added sentence that states 9 users did the FLS peg transfer task, it is easy to assume that all 12 users did all the tasks. I think it is a key detail that the sample size changed and that the user groups are not the same from o-ring to FLS peg transfer task. I'm also not sure that grouping these results in Table 1 is the best way to present them given the cohort differences for each category.

- R2.8 – It's good to split the responses into who did which task in Table 3 and I understand this is a subjective assessment of the in vivo feasibility study. However, it's clearer now that these results are entirely dependent on one user's opinion in some categories. I think that's okay for clinical feasibility, but it underscores why the bench top analysis should be stronger.

- R2.11 – The comment about the harness as a safety precaution should be in the paper, not just the response letter.

- R2.13 – Deciding to omit the frustration, temporal demand and effort in a conventional NASA TLX survey is still unclear to me. I understand that the intention was not to do a full workload assessment, but this would have helped to better understand the differences between this system and conventional robots or manual laparoscopic tools.

- R2.14 - The discussion and response with respect to the large radial error and its impact on the system performance is currently insufficient in the paper. The description of why this gap arises in 2.1.2 (extended kinematic chain - resulting in discrepancies between commanded and executed motion & ArUco marker-based RCM localization error) is not detailed enough. I would like to see a more explicit description why this error is so high and how it impacts the overall system. An estimate or assessment of what can be attributed to the ArUco marker-based RCM localization error would help too. Given this seems to be one of the biggest technical limitations of the system now, it should also be explicitly addressed in the discussion when summarizing the performance gaps. Further, the closed-loop visual feedback compensation (described in the response letter) is not mentioned anywhere in the paper. If this was feedback from the users, it should be included.

- R2.27 – Thank you for the response. 12 and then 9 users, only 2 of which are surgeons, still does not feel like a sufficient cohort. The inconsistency with the user groups between tasks is also a limitation. The new statistics (confidence interval,

effect size, etc.) don't seem to really strengthen many of these results either. This statement "In the surgeon group, performance is generally more accurate and consistent across all platforms; surgeons are fastest on the dVRK, achieve their lowest error on the humanoid, and exhibit a similar timing trend as novices." - only has 2 data points supporting it. Furthermore, most of the confidence intervals for the o-ring transfer task and FLS peg transfer cross zero, some with very wide intervals which points to high uncertainty.

- It appears that there are only 3 significant claims from the dry lab evaluation (below - taken directly from the text):

1. The humanoid platform yields error levels comparable to the dVRK (5.49 ± 3.35), with no significant difference in weighted error between humanoid and dVRK operation ($\Delta = 0.49$, 95% CI $[-1.38, 2.36]$, $p = 0.575$).

2. Completion times differ significantly across platforms: the dVRK is faster than manual operation ($\Delta = -30.06$ s, 95% CI $[-49.35, -10.77]$, $p = 0.006$, $dz = 0.99$), while the humanoid platform requires longer completion times than the dVRK ($\Delta = 43.86$ s, 95% CI $[27.87, 59.85]$, $p < 0.001$, $dz = 1.74$).

3. Paired comparisons indicate that the da Vinci platform significantly outperforms both the humanoid and manual conditions in completion time and balanced FLS score (Table 1 E).

- These significant conclusions are contradictory between tasks which makes it difficult to draw conclusions. In the o-ring transfer task the significant finding is that there is no real difference between the dVRK and the humanoid, but for the FLS peg transfer task – the dVRK is significantly better. The timing differences are also inconsistent – in Figure 3 B(ii) it seems the humanoid is the slowest for o-ring (significantly), but falls in the middle for FLS peg transfer (Figure 3 Cii). Given the limited sample size, and the difference in cohorts and experience levels – I don't think these results support meaningful or rigorous conclusions.

Referee #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

no further comments.

Referee #2

(Remarks to the Author)

We appreciate the authors response letter and efforts to improve their submission. In the revised manuscript, the following changes were made which have substantially strengthened the work:

- Several additional experienced users were added. Task 1 evaluation now consists of $n = 18$ users (12 novices and 6 surgeons – previously only 12 users), and Task 2 now has $n = 13$ users (8 junior and 5 senior surgeons – previously only 9 users).

- Additional evaluation metrics were also included for the dry-lab evaluation studies (peg drops, collisions, failed pick-ups, etc.) that further underscore the study findings.

- The summary tables for task 1 and 2 have been clearly separated to better distinguish the results and different task cohorts.

- The tables for tasks 1 and 2 were separated. This now highlights that the dVRK was used for task 1 whereas the da Vinci Xi system was used for task 2.

- The significance claims in the paper have now been clarified and the potential contribution of the performance gap between the dVRK and the da Vinci Xi as robotic benchmarks are also discussed.

The authors adequately addressed all the comments and the manuscripts is now ready to publish.

Minor remaining comment: Be sure to define the error bars in the legend for Figure 3.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Referees

We thank the referees for their in-depth and critical assessment of our work. The enthusiasm for the paper is appreciated and we have done our best to address the referees' comments point by point in the following text.

Referee 1 [R1]

R1.1

Provide a clearer ---> tool-tip error sources including extended kinematic chain effects, RCM calibration, latency, joint compliance/backlash, tracker error.

Reply:

We thank the referee for this helpful comment. In the Motion accuracy evaluation section, we added more detailed discussion on the tool-tip error including extended kinematic chain measurement issues, RCM calibration procedure and inaccuracy. We also added a quantitative latency characterization and discussion of how these factors contribute to discrepancies between commanded and executed tool motion.

These additions appear in the revised manuscript page 5, 6.

Motion accuracy is a key determinant for safely and effectively manipulating delicate anatomical structures during laparoscopic surgery. Prior studies [49, 50] have identified common sources of error in conventional teleoperation systems, including communication latency, kinematic inaccuracies, and human-robot interaction dynamics. In the proposed humanoid laparoscopic framework, manual laparoscopic instruments are incorporated as an extended kinematic chain with more complex control loops.

Furthermore, because the commercial instruments used in this study are not publicly documented, the geometric parameters of the extended kinematic chain must be obtained through manual measurement and modeling, introducing potential calibration errors. To quantify and evaluate the motion accuracy of the system, we approach this from both system latency and command-execution tracking perspectives.

inaccuracies in the geometric measurements of the extended kinematic chain, resulting in discrepancies between commanded and executed motion. In addition, manipulation of wristed laparoscopic instruments requires both software-enforced RCM control and physical trocar constraints, making the kinematic chain more sensitive to the accuracy of ArUco marker-based RCM localization.

System latency: We characterize the system latency in terms of follower-leader latency and control loop latency. The follower-leader latency, defined as the delay between operator hand motion and robot end-effector motion, was measured as ~ 156 ms using a high frame rate (120 FPS) external camera. The control loop latency measured on the control PC is approximately 24 ms, within which the inverse kinematics computation requires approximately 11 ms per cycle. Recent work [51] reports that existing humanoid teleoperation systems commonly exhibit end-to-end latencies on the order of hundreds of milliseconds, suggesting that such platforms still remain under active development and that further improvements are needed. Prior studies indicate that latency below 150 ms is generally considered desirable for conventional surgical robotic systems [52].

R1.2

Expand description of RCM calibration, drift handling, and enforced safety constraints.

Reply:

We appreciate this insightful comment. In the Live porcine surgery section, we have extended explanations on the RCM calibration procedures during the operation. RCM drift handling and safety constraints are now further discussed, including necessary recalibration when large drifts happen and a temporal filter implemented to prevent unsafe disturbances. The total recalibration time is reported in Table 2.

These additions appear in the revised manuscript page 14, 15, 16.

figuration with four trocars (three 5 mm, one 12 mm). As shown in Fig. 4B(i), the ports are localized before the operation by attaching calibration boards with ArUco markers, which are detected by the humanoid's head camera. After port placement,

Due to the animal's respiration and slight robot base drift during the operation, the actual RCM positions were dynamic. A temporal filter on the estimated RCM positions was adopted to prevent unsafe disturbances. Procedures were paused for recalibration in cases of large RCM drift, degraded visualization, unexpected instrument behavior, or at the surgeon's discretion. During these pauses, instruments were held in a stable position and system functionality was verified before resuming the procedure. Instrument exchanges and robot redeployments for recalibration or repositioning typically required at least three minutes and were classified as major pauses in the workflow analysis.

Table 2 Quantitative metrics of humanoid robotic performance during live porcine cholecystectomies. Times reflect durations of active console control and segment-specific operative steps. Major pauses (> 3 min) indicate system recalibration or instrument exchanges; minor pauses (< 1 min) indicate brief camera or instrument adjustments. Deployments were defined as discrete episodes of robot recalibration or physical repositioning relative to trocar placement. Δ denotes the change in each metric between Case 1 and Case 2.

Metric	Case 1	Case 2	Δ (Change)
Active console time (mm:ss)	56:15	31:59	−24:16
Robot initiation → initiation of tissue dissection (mm:ss)	4:31	2:12	−2:19
Initiation of tissue dissection → critical view of safety (mm:ss)	15:42	6:19	−9:23
Critical view of safety → completion of lap clipping (mm:ss)	4:04	3:19	−0:45
Gallbladder mobilization off liver bed (mm:ss)	31:58	23:08	−8:50
Number of robot deployments (n)	8	4	−4
Number of major pauses (> 3 min, n)	2	3	+1
Number of minor pauses (< 1 min, n)	7	4	−3

R1.3

Clarify in-vivo procedural protocol (team roles especially that surgeon was assisting mostly, pause/recalibration criteria, thresholds to convert, sterile workflow, instrument exchange details).

Reply:

We thank the referee for pointing this out. Team roles details are now added. Pause/recalibration criteria, and instrument exchange details are further discussed in the Live porcine surgery section (see also R1.2). We also included sterile workflow during the procedure as well as the limitation discussion. Additionally, we added clarification that the procedures were completed without conversion to conventional laparoscopy or open surgery.

These additions appear in the revised manuscript page 15, 20.

Procedures followed a standard laparoscopic workflow, with one console surgeon and one bedside assistant. To ensure consistency, a single senior surgeon performed all procedures from the robotic console using the humanoid platform, while bedside assistance was provided by either another senior surgeon or a clinical fellow experienced in laparoscopic surgery. The assistant managed camera control, tissue retraction, exposure, and bedside adjustments such as camera cleaning and instrument adjustments. Instrument adjustments and camera cleaning followed standard laparoscopic

Integrating humanoid robots into sterile operating rooms introduces additional challenges. In this study, sterility was maintained by placing gloves over the robot’s arms, a strategy that does not fully replicate sterile workflows required in human surgery, where all components contacting the operative field must be autoclaved and appropriately draped. Current commercial humanoid systems lack autoclavable components, and preserving sterility without impairing sensor function or motion calibration remains a critical barrier. Transparent sterile drapes may offer a potential solution; however, future work must establish robust protocols for fully sterile integration while maintaining sensor fidelity.

R1.4

Report latency characterization (mean and variance) and relate to performance observations.

Reply:

We agree latency characterization is important for the proposed system. The latency measurement details and quantitative results have been included in the Motion accuracy evaluation section (see also R1.1). In the Dry lab evaluation of the framework we have discussed that the performance aligned well with the benchtop experiments results.

These additions appear in the revised manuscript page 6, 10, 11.

the da Vinci system, and reduced reachability due to the humanoid's compact form factor. These feedback align with the quantitative benchtop analysis discussed in previous sections, where higher latency and motion error require greater operator skill to close the control loop. Additionally, the Unitree G1 humanoid has an arm span of approximately 450 mm, compared with 1.6–1.8 m for an adult human [54], a con-

R1.5

Justify statistical test selection and state distribution assumptions.

Reply:

We appreciate the referee for pointing this out, and we agree more statistical analysis is necessary for evaluating the performance. In the revised manuscript, we included paired t-test and Wilcoxon signed-rank test in table 1, decided by normal distribution checks.

These additions appear in the revised manuscript page 13, table 1.

Table 1 Participants’ performance across dry-lab tasks. (A) O-ring replacement summary. (B) O-ring paired comparison statistics. (C) O-ring error type occurrences. (D) FLS peg transfer summary. (E) FLS paired comparison statistics. All values are reported as mean $\pm$ std using per-participant means.

(A) O-ring replacement: weighted error and time (mean $\pm$ std)						
Group	Platform	Weighted Error		Time (s)		
Novices	Manual	7.91 $\pm$ 5.64		75.43 $\pm$ 32.66		
	Humanoid	5.49 $\pm$ 3.35		89.23 $\pm$ 38.91		
	dVRK	5.00 $\pm$ 2.83		45.37 $\pm$ 17.93		
Surgeons	Manual	4.25 $\pm$ 2.47		44.80 $\pm$ 9.30		
	Humanoid	2.44 $\pm$ 2.03		66.75 $\pm$ 15.92		
	dVRK	3.56 $\pm$ 3.80		39.16 $\pm$ 9.58		
(B) O-ring replacement: paired comparison statistics						
Metric	Comparison	Δ [95% CI]	p	d_z	Test	
Weighted Error	Humanoid – Manual	-2.42 [-5.30, 0.45]	0.091	0.54	t	
Weighted Error	dVRK – Manual	-2.92 [-6.44, 0.61]	0.096	0.53	t	
Weighted Error	Humanoid – dVRK	0.49 [-1.38, 2.36]	0.575	0.17	t	
Time (s)	Humanoid – Manual	13.80 [-9.56, 37.16]	0.220	0.38	t	
Time (s)	dVRK – Manual	-30.06 [-49.35, -10.77]	0.006	0.99	t	
Time (s)	Humanoid – dVRK	43.86 [27.87, 59.85]	4.88e-04	1.74	w	
(C) O-ring replacement: error type occurrences (mean $\pm$ std)						
Platform	Failed pick	Stretch (peg)	Stretch (handoff)	Drop	Collision	Straw disp.
dVRK	3.65 $\pm$ 2.14	0.88 $\pm$ 0.98	0.33 $\pm$ 0.94	0.38 $\pm$ 0.69	8.19 $\pm$ 5.20	0.64 $\pm$ 0.74
Humanoid	6.27 $\pm$ 5.22	3.35 $\pm$ 3.39	0.48 $\pm$ 0.65	1.22 $\pm$ 1.44	3.70 $\pm$ 4.03	0.34 $\pm$ 0.83
Manual	6.25 $\pm$ 4.86	6.16 $\pm$ 5.87	1.04 $\pm$ 1.17	0.74 $\pm$ 1.07	8.49 $\pm$ 6.22	0.31 $\pm$ 0.51
(D) FLS peg transfer: performance summary (mean $\pm$ std)						
Platform	Time (s)		Error	Balanced FLS (0–100)		
Manual	1204.4 $\pm$ 808.5		3.86 $\pm$ 2.97	57.18 $\pm$ 25.59		
Humanoid	713.8 $\pm$ 687.9		2.75 $\pm$ 3.24	72.79 $\pm$ 30.49		
da Vinci	146.7 $\pm$ 58.1		0.14 $\pm$ 0.38	98.21 $\pm$ 2.96		
(E) FLS peg transfer: paired comparison statistics						
Metric	Comparison	Δ [95% CI]	p	d_z	Test	
Time (s)	da Vinci – Humanoid	-694.50 [-1441.64, 52.64]	0.031	0.98	w	
Time (s)	da Vinci – Manual	-1079.33 [-1947.57, -211.10]	0.024	1.30	t	
Time (s)	Humanoid – Manual	-425.86 [-767.33, -84.38]	0.022	1.15	t	
Error	da Vinci – Humanoid	-3.33 [-6.63, -0.04]	0.048	1.06	t	
Error	da Vinci – Manual	-4.00 [-7.32, -0.68]	0.027	1.26	t	
Error	Humanoid – Manual	-0.71 [-4.94, 3.51]	0.694	0.16	t	
Balanced FLS	da Vinci – Humanoid	31.91 [0.62, 63.20]	0.047	1.07	t	
Balanced FLS	da Vinci – Manual	43.03 [16.16, 69.91]	0.009	1.68	t	
Balanced FLS	Humanoid – Manual	12.18 [-13.28, 37.64]	0.286	0.44	t	

Notes: Δ is the mean paired difference (first minus second) with a 95% Confidence Interval (CI). d_z is Cohen’s d for paired samples computed on paired differences. Test indicates paired two-sided t -test (t) or Wilcoxon signed-rank test (w), selected based on normality checks.

R1.6

Include effect sizes and confidence intervals, not only p-values.

Reply:

We acknowledge this limitation. In addition to the statistical test selection and state distribution assumptions, we also added effect sizes and confidence intervals in the table 1 (see also R1.5)

R1.7

Clarify normalization methods used in FLS scoring and whether they introduce bias.

Reply:

This is a helpful observation. In the revised manuscript, we have further clarified that the total time open exceeds the 300 s limit for platforms including humanoid and manual operation, which would result in a negative FLS score, and create an imbalance between how much time and errors contribute to the scores. The normalization is used to scale the score into 0-100 with interpretable overall performance.

These adjustments appear in the revised manuscript page 9, 10.

We evaluate task performance using completion time, peg drops, and the FLS score. The FLS peg transfer task conventionally has a 300 s time limit, and both completion time and errors are considered in the performance evaluation [53]. A commonly used scoring formulation is $300 - t - 10e$, where t is the completion time (s) and e is the number of peg drops. However, completion times for humanoid and manual operation frequently exceed the 300 s limit (Table 1), applying this metric would produce negative or otherwise unbalanced scores. To provide a single interpretable measure that reflects both speed and reliability, the balanced FLS score is computed by min-max normalizing time and error across all valid trials and combining them with equal weights, then rescaling to $[0, 100]$, where higher values indicate faster completion with fewer drops. The score can be calculated as

$$t' = \frac{t - t_{\min}}{t_{\max} - t_{\min}}, \quad e' = \frac{e - e_{\min}}{e_{\max} - e_{\min}} \quad (1)$$

$$S = 100 (1 - (w t' + (1 - w) e')), \quad w = 0.5 \quad (2)$$

where t is the completion time (s), S is the balanced FLS score, e is the number of drops in the trial, and $t_{\min}, t_{\max}, e_{\min}, e_{\max}$ are the minimum and maximum values of completion time and peg drops across all participants. $w = 0.5$ is the weighting factor that balances the relative contributions of time and drops.

R1.8

The manuscript would benefit from clearer positioning relative to prior work in teleoperated laparoscopy, RCM control methodologies, and humanoid medical teleoperation studies. Explicit comparative discussion is needed to avoid overstating novelty or implying parity with mature surgical robotic systems.

Reply:

We acknowledge that this is an important consideration. In the introduction, we have added references of prior works in teleoperated laparoscopy and RCM control, and compared them with the proposed system to better highlight our contributions. We have also referred to the recent Surgie study that performs dexterous medical tasks under controlled conditions.

These adjustments can be found in the revised manuscript page 2, 3.

profile claims of “superhuman surgeries” [24]. Prior work has extensively explored teleoperated robotic systems for minimally invasive surgery, most notably through purpose-built platforms such as the da Vinci Surgical System and related research platforms [25, 26]. To reduce reliance on mechanically regulated remote center-of-motion (RCM) mechanisms, general-purpose manipulators have also been employed to maintain laparoscopic port constraints through kinematic control strategies [27–30]. However, these systems still rely on robot-specific laparoscopic instruments [31, 32], including those used in Intuitive’s da Vinci Surgical System. In contrast, Humanoid robots, with human-like kinematics and dexterous upper limbs, offer a potential pathway to directly manipulate surgical tools designed for human use. Nevertheless, it

The recent Surgie study [40] demonstrated that general-purpose platforms can be teleoperated to perform dexterous medical tasks under controlled conditions. Although no surgical procedures were performed, the study suggests a plausible pathway from service-oriented hospital tasks toward minimally invasive surgical workflows. Yet a

R1.9

The manuscript is generally readable but the abstract and introduction are somewhat broad. Market and funding context can be reduced in favor of earlier emphasis on the surgical evaluation question and success criteria. The conclusion should align more closely with demonstrated feasibility rather than extend toward clinical readiness or system superiority claims.

Reply:

We appreciate the referee’s suggestion. To address this point, we reduced the discussion on the market and funding related context in the Abstract and Introduction, and put earlier emphasis on the surgical evaluation of humanoids. In the Discussion section, we further clarified the key limitations of the current system, and distinguished feasibility and potential improvements in the future.

These adjustments can be found on page 1, 2, 20, 21

ment. Healthcare is a particularly pressing domain, where staffing shortages and increasing care demand are widening the gap between clinical workload and the availability of skilled labor. While current automation has largely focused on digital and logistical tasks, much of hospital work remains embodied, requiring mobility, manipulation, and safe interaction in human-designed environments. Humanoid form factors offer unique potential, particularly when used to assist with surgical tasks. Traditionally, robotic systems for surgery are purpose-built

in growing industrial momentum and early commercialization efforts [18, 19]. These trends suggest that humanoids may soon be capable of performing labor-intensive human tasks at scale.

Healthcare represents a particularly critical application domain, where rising care demands and persistent workforce shortages place increasing strain on clinical systems [20–22]. While existing hospital robots primarily support logistics and administrative tasks, extending robotic assistance to clinical workflows requires embodied interaction, dexterous manipulation, and safe operation in human-centered environments. Early deployments such as hospital logistics robots demonstrate the feasibility of robotic assistance in healthcare settings [23]. However, most existing systems focus on routine support tasks rather than safety-critical clinical procedures.

Beyond support and logistics, a particularly high-impact and safety-critical question is whether humanoid robots can assist with surgical tasks. Indeed, the idea

3 Discussion

In this study, we present a comprehensive evaluation of contemporary humanoid technology for surgical applications and report the first in vivo use of a humanoid robotic surgical system to perform standard laparoscopic cholecystectomies in a porcine model. Before in vivo deployment, benchtop experiments and dry-lab user studies provided controlled evaluations of the platform’s technical capabilities and usability. These experiments demonstrated the feasibility of accurately manipulating manual laparoscopic instruments under externally enforced RCM constraints. At the same time, quantitative analyses revealed performance gaps compared with established surgical robotic systems, including control latency, reduced workspace reachability, and sensitivity to calibration drift. Future work should focus on improving system integration and control strategies to reduce user-perceived latency, enhancing the robustness of control under noisy or dynamic RCM conditions, and investigating human-scale humanoids to better match the workflow of surgical environments. These findings help identify key technical challenges that must be addressed before clinical deployment.

Integrating humanoid robots into sterile operating rooms introduces additional challenges. In this study, sterility was maintained by placing gloves over the robot’s arms, a strategy that does not fully replicate sterile workflows required in human surgery, where all components contacting the operative field must be autoclaved and appropriately draped. Current commercial humanoid systems lack autoclavable components, and preserving sterility without impairing sensor function or motion calibration remains a critical barrier. Transparent sterile drapes may offer a potential solution; however, future work must establish robust protocols for fully sterile integration while maintaining sensor fidelity.

This study establishes an early, quantitative benchmark for humanoid surgical performance in vivo. Active console efficiency and precise instrument manipulation were achievable even in these preliminary deployments, demonstrating feasibility while identifying areas for technical and workflow refinement. With further development, humanoid systems may provide scalable alternatives to conventional robotic platforms and expand robotic capabilities beyond traditional operating rooms. At our institution, close clinician–engineer collaboration has enabled rapid, iterative testing in realistic surgical environments, laying the foundation for continued refinement. Future investigations will evaluate performance across a broader range of procedures, extended case durations, and more autonomous tasks to assess the capacity of humanoid surgical robots to address remaining clinical, economic, and structural gaps in contemporary robotic surgery.

R1.10

The manuscript should position its contribution more explicitly relative to prior teleoperated lap robotics literature and existing humanoid medical task teleoperation studies. Claims of being 'first' should be precisely scoped and supported with appropriate references to avoid overstating novelty.

Reply:

We appreciate this recommendation. As the referee suggested in R1.8, we included references of previous literature in teleoperated robotic laparoscopy, and we referred to the most recent Surgie study that conducted non-surgical teleoperation tasks. To best scope the novelty of our work, we further clarified in the revised manuscript that it's the first humanoid-based in-vivo porcine study.

These changes appear on page 3.

ies and further validate the framework in the first humanoid-based in-vivo porcine study in which the teleoperated humanoid completes two cholecystectomies. Together,

R1.11

Claims regarding reduced congestion, preserved sightlines, improved team communication, or “free console setup” are speculative and unsupported by objective ergonomic or workflow metrics. The manuscript should more clearly separate feasibility (demonstrated) from clinical readiness (not demonstrated).

Reply:

We agree that additional clarification is needed. As the referee suggested in R1.9, in the discussion section we separated the system's feasibility and clinical readiness. Claims such as reduced congestion, preserved sightlines, improved team communication are further clarified as potentials once the system becomes clinically ready in the future.

These adjustments can be found on page 21.

Once clinically ready, humanoid platforms may help address several limitations inherent to contemporary robotic surgery. High acquisition costs, proprietary consumables, ongoing maintenance contracts, and extensive training requirements have historically confined robotic surgery to well-resourced tertiary centers [61, 62]. By providing a versatile humanoid platform at a lower cost, these systems may improve accessibility and broaden the availability of robotic-assisted surgical care. Additionally, the humanoid robot's reduced footprint and positional adaptability may reduce crowding at the operating table, potentially providing bedside assistants greater freedom to perform tasks such as suction, retraction, and instrument exchange. The absence of large bedside consoles or obstructive robotic arms may also help preserve sightlines, which could facilitate team situational awareness, learner engagement, and intraoperative communication.

Referee 2 [R2]

R2.1

My biggest concern is about the motivation for the chosen deployment of the humanoid. The introduction describes how humanoids inherently fit environments designed to accommodate the human form-factor which I agree with, however, the motivation for replacing the patient side manipulators in the teleoperation chain with the humanoid is unclear to me. The paper states: "Laparoscopy is chosen as a representative benchmark due to its widespread clinical adoption and the availability of established robotic systems such as Intuitive's da Vinci Surgical System." - but purpose-built systems like the da Vinci can accommodate minimally invasive surgery very well and are not impeded by the limited mobility of the base. Additionally, Intuitive has achieved widespread adoption in clinical centers despite some of the limitations pointed out like storage demands, capital costs, etc.

Reply:

We appreciate the referee's comments and have worked to have clearer motivation in the examination of humanoids. As also pointed out by referee 1 in R1.9, we have further clarified the distinction between demonstrated feasibility and clinical readiness in the revised manuscript. In the introduction, we also acknowledged the exceptional performance and widespread adoption of specialized surgical robots. Additionally, this study is an exploration of humanoids' potential in surgical applications, which does not aim to replace the current surgical robotic systems.

While we acknowledge the existing performance gap, current purpose-built systems still face limitations. Recent studies note that robot-assisted surgery often requires operating rooms larger than those originally designed for manual procedures, which remains the case in most clinical settings. This clarification and references are included in the introduction section.

These changes can be found on page 3

While purpose-built surgical platforms such as Intuitive Surgical's da Vinci Surgical System have demonstrated exceptional performance and widespread clinical adoption through years of industrial refinement, studies note that most existing operating rooms (ORs) were originally designed for manual surgery, and robotic-assisted surgery (RAS) often requires additional space and coordination for equipment positioning, robot docking, and intraoperative maneuvering [33, 34]. Humanoid robots, by contrast, share a similar form factor with humans, potentially offering practical deployment advantages in operating rooms. Their human-like morphology allows them to operate within existing surgical infrastructure and interact directly with standard manual medical equipment. Although the comparatively lower price tags of current

R2.2

Related to my previous point, in the introduction there are some mentions of other potential use cases for humanoid robots. I recommend additional discussion about how the work presented in this manuscript may support these other use-cases where a humanoid form-factor may be more explicitly beneficial.

Reply:

This is an insightful comment. In the introduction, we have added more potential use cases discussion with the corresponding references,

These changes can be found on page 3.

offer the potential for versatile task deployment. Beyond minimally invasive surgery, humanoid robots capable of dexterous manipulation may support a broader range of clinical activities, including patient interaction [35, 36], bedside assistance [37], and hospital logistics [38, 39]. Such multi-purpose capabilities may improve resource utilization and reduce capital, storage, and operational demands in resource-constrained healthcare environments.

R2.3

Much of the discussion and motivation is centered around cost, and smaller physical footprint of the humanoid compared to a conventional fixed-based robot. While I agree that the footprint of one humanoid is smaller than the da Vinci, if you had 2 humanoids in place like you do in the supplemental videos to accommodate 3 laparoscopic instruments, I think the footprint comparison becomes weak. In fact, in many procedures it is common to have 4 – 5 ports, which would require up to 3 humanoids at which point, I would guess the footprint and storage demands are not significantly different.

Reply:

We acknowledge that this is an important consideration. As discussed in R2.1, the mobility and setup of current purpose-built surgical robots still present practical limitations and impose greater requirements on operating room (OR) size and layout. Beyond footprint within the OR itself, deployment logistics can also pose challenges. For example, many operating rooms—particularly in older hospitals or outpatient surgical centers—may not

accommodate large robotic systems such as the da Vinci through standard doorways. In such scenarios, the smaller footprint and inherent mobility of humanoid platforms may offer practical advantages for deployment.

R2.4

I would like to see an improved discussion of the actual costs of the humanoid robot in comparison to the da Vinci patient side manipulator. The sentence: “Economic constraints are a major barrier to broader robotic adoption, as high acquisition costs, ongoing maintenance contracts, proprietary consumables, and extensive training requirements have confined robotic surgery largely to well-resourced tertiary medical center” implies that the humanoid will require less acquisition costs, maintenance, fewer consumables, and less training. However, the current system still needs wristed laparoscopic tools and, in the teleoperation mode, as shown here, would probably require the same amount of training. Surgical robot costs exceed manufacturing costs of the hardware manyfold due to regulatory, liability, marketing, and profit overhead, so directly comparing to commercial humanoid sales is unfair. Furthermore, is there reason to believe the maintenance of a humanoid would be different than that of a fixed-based system? More evidence is necessary to justify these claims in the discussion.

Reply:

We appreciate this suggestion and agree that direct price comparisons between humanoids and established surgical robotic systems may be inappropriate given their different levels of market maturity and regulatory development. Accordingly, we have reduced the emphasis on direct price comparisons in the revised manuscript. Nevertheless, humanoid platforms may offer potential advantages in addressing economic constraints due to their versatility and multi-purpose capabilities, as discussed in R2.3. Unlike purpose-built surgical robots designed for specific procedures, humanoids may support a broader range of clinical tasks and manipulate conventional non-robotic instruments without requiring specialized tooling. Their mobility and general-purpose manipulation capabilities may also enable assistance with setup, instrument exchange, or other perioperative tasks. While these possibilities remain speculative, they highlight potential advantages of general-purpose humanoid platforms beyond single-task surgical automation.

These clarifications can be found on page 3

manual medical equipment. Although the comparatively lower price tags of current humanoid platforms may partly reflect differences in regulatory, liability, and commercialization pathways compared with established surgical robots, humanoid systems offer the potential for versatile task deployment. Beyond minimally invasive surgery, humanoid robots capable of dexterous manipulation may support a broader range of clinical activities, including patient interaction [35, 36], bedside assistance [37], and hospital logistics [38, 39]. Such multi-purpose capabilities may improve resource utilization and reduce capital, storage, and operational demands in resource-constrained healthcare environments.

R2.5

I recommend changing the title to emphasize the key role of teleoperation in this study. The current title “Humanoid Robots in Surgery: First In-vivo Feasibility Study”, does not make it clear to the reader that the humanoid is teleoperated.

Reply:

We appreciate this recommendation. In the revised manuscript, we have clarified that the system is teleoperated by positioning this work explicitly as a teleoperated feasibility study and by expanding comparisons with prior teleoperated laparoscopy research (see also R1.8). In addition, we have emphasized the role of teleoperation in both the Abstract and Introduction to ensure this aspect is clear to readers.

R2.6

User study: The balanced FLS score is interesting but I would have liked to see some comparison in task performance using the conventional criteria from the FLS program ([https://cpd.partners.org/sites/default/files/manual_skills_guidelines\[1\].pdf](https://cpd.partners.org/sites/default/files/manual_skills_guidelines[1].pdf)) which penalizes based on the pegs falling out of the field of view. If not, then perhaps a better explanation as to why a unique FLS metric was derived for this study. There are also some nice review papers that analyze different data-driven quantitative metrics for evaluating performance in robot-assisted laparoscopy (ie. <https://link.springer.com/article/10.1007/s00464-021-08792-5>). Given the variability in the results with just FLS score, weighted error, and time, I would be interested if some more analysis could be done with these validated metrics to better understand the gaps between manual, daVinci, and humanoid.

Reply:

This is a helpful suggestion, which was also raised by Referee 1 (R1.7). To address this comment, we clarified that the conventional FLS scoring formulation may produce negative or otherwise unbalanced scores in our experimental setting. The normalized score used in this study assigns equal weights (0.5) to completion time and errors, providing a more interpretable scale for performance comparison while ensuring balanced contributions from both factors (see also R1.7).

R2.7

In table 1, it is unclear why the group distinction (surgeons versus novices) is only provided for the o-ring transfer task.

Reply:

We appreciate this helpful observation. The participants' information was unintentionally omitted during the preparation of the initial manuscript. We have now added the participants' background for the FLS task in the revised version.

These changes can be found on page 9

titative evaluation protocol. Nine participants with professional medical training took part, including seven surgical residents and two medical students. As illustrated in

R2.8

The user experience and system performance evaluation is also a bit confusing. My understanding is that the four survey participants were attending surgeons, clinical fellows and residents that were present during the in-vivo studies. Although each of these users were present, I would assume that they didn't each complete every task that was assessed (ie. setup and docking, operating from the surgeon's console, tool insertion). If there was systematic assessment of each of these, the description should be explained in more detail, if not, I think the metrics of the user who actually completed each part of the workflow should be separated as that would be more representative (i.e. I want to see the distinct survey results about port placement from the user who consistently helped with port placement).

Reply:

This is a very insightful suggestion from the referee. We acknowledge that separating the users' response based on their roles can present more informative metrics. To address this, we have clarified participants' roles and corresponding questions they answered in Table 3. This is also noted in the text.

These adjustments appear on page 17, 19

cal trainees involved in system setup, bedside assistance, or console operation. Each participant first indicated which workflow components they performed, and survey responses were collected only for those tasks. Consequently, reported scores reflect

Table 3 Post-procedure survey results evaluating the humanoid robotic surgical system across setup, bedside assistance, and console operation. Survey responses ($n = 4$) were collected from attending surgeons and surgical trainees involved in system setup (1), bedside assistance (2), or console operation (1). Participants completed only the questions corresponding to the role they performed. Ratings were provided on a Likert scale from 1 (poor/low readiness) to 5 (excellent/high readiness).

Survey Metric	Respondents (n)	Mean Effectiveness	Mean Clinical Readiness
Setup, port placement, docking	1	4	3
Ease of bedside tool insertion & attachment	1	3	–
Bedside assistance with system	2	3	2.5
Ergonomics of bedside assistance	2	3	–
Surgeon console effectiveness	1	3	3
Surgeon console intuitiveness	1	4	–
Fidelity of motion of surgeon console	1	3	–
Overall clinical readiness	4	–	2.5

R2.9

I also have some major questions about sterility. The da Vinci robot is typically draped apart from the tools. I imagine that the humanoid would need similar sterility (or could perhaps be sterilized like a human surgeon) and this could potentially block some of the sensors that enable the current motion and stability. Has any analysis been done to this effect? If not, maybe commenting on this and the potential challenges towards clinical translation would benefit the discussion.

Reply:

We thank the referee for pointing out the lack of surgery details such as sterility in the original manuscript. In the revised version, we have included the sterility procedure in the study, while acknowledging the current sterility challenges in the Discussion section. Additionally, humanoid systems can scrub in similarly to human operators; however, pre- and post-operative handwashing is currently unnecessary because the robot hand module can be detached.

These adjustments appear on page 20

Integrating humanoid robots into sterile operating rooms introduces additional challenges. In this study, sterility was maintained by placing gloves over the robot's arms, a strategy that does not fully replicate sterile workflows required in human surgery, where all components contacting the operative field must be autoclaved and appropriately draped. Current commercial humanoid systems lack autoclavable components, and preserving sterility without impairing sensor function or motion calibration remains a critical barrier. Transparent sterile drapes may offer a potential solution; however, future work must establish robust protocols for fully sterile integration while maintaining sensor fidelity.

R2.10

One of the biggest discussion points is the high number of redeployments during the in-vivo studies, characterized as instrument repositioning. I imagine this would be comparable to repositioning done by a human during manual laparoscopic surgery but some sort of comparison to the manual procedure and how frequently repositioning occurs would help with the interpretation of the results.

Reply:

We appreciate this comment. The redeployments recorded in our study were primarily system recalibration events caused by animal respiration and slight robot base drift, which resulted in dynamic changes of the estimated RCM position. We have clarified this process and its classification in the revised manuscript. Brief adjustments and camera cleaning events (<1 min) were categorized as micro-pauses and are comparable to routine adjustments performed by human assistants during manual laparoscopy. Longer interruptions (>3 min) corresponded to full system recalibration or physical repositioning of the robot and reflect current system limitations rather than procedural requirements.

These changes can be found on page 15, 17

Due to the animal's respiration and slight robot base drift during the operation, the actual RCM positions were dynamic. A temporal filter on the estimated RCM positions was adopted to prevent unsafe disturbances. Procedures were paused for recalibration in cases of large RCM drift, degraded visualization, unexpected instrument behavior, or at the surgeon's discretion. During these pauses, instruments were held in a stable position, and system functionality was verified before resuming the procedure. Instrument exchanges and robot redeployments for recalibration or repositioning typically required at least three minutes and were classified as major pauses in the workflow analysis.

procedures. Brief instrument adjustments and camera cleaning events lasting less than 1 minute were classified as micro-pauses. Similar to those routinely performed by human assistants in laparoscopic surgery [55–57], these reflect normal, expected pauses in the surgical workflow rather than limitations of the humanoid platform. In contrast, major pauses lasting more than three minutes corresponded to full redeployments of the robot for system recalibration or physical repositioning of the robot base or arms relative to the trocars. Unlike routine micro-pauses, these longer interruptions disrupt procedural flow and highlight areas where the technology requires improvement before clinical implementation. Importantly, these redeployments in each case did not affect performance during active console use, and both procedures maintained excellent execution of critical operative tasks (Table 2).

R2.11

Another question I have is about the harnesses that appear to be holding up the humanoids in the videos. I believe one of the biggest concerns with these systems would be the risk associated with a humanoid falling onto the patient during the operation if they were to become unstable. More discussion on the necessity of the harnesses and the stability of the system would help clarify this. Also, if the system requires a ceiling mounted harness, does the workspace argument still hold-up?

Reply:

We apologize for the lack of clarity on the harness. Since the system in this study is still experimental and not clinically ready, we used robot harness only as a precautionary safeguard, which is not required for any system functionality.

R2.12

The manuscript would benefit from more discussion about the functional dexterity differences between the conventional PSM and the humanoid. During the benchtop evaluation, did the users note feeling any major differences or limitations?

Reply:

Benchtop experiments provide a broad understanding of the humanoid's ability relative to conventional PSM, however we agree with the referee that it does not provide a functional dexterity evaluation specific to a user operating. Therefore, we included additional statistical results from our user studies where users performed complex surgical tasks hence providing a functional evaluation of the dexterity (see also R1.5).

R2.13

In the NASA TLX survey, please consider including the frustration category. It's unclear why this was omitted.

Reply:

We thank the referee for the suggestion. We did not administer the full NASA-TLX survey in these initial in-vivo studies. Instead, we used a modified, pilot survey inspired by NASA-TLX to capture the humanoid system's technical performance, including mental demand, physical demand, motion accuracy, visual feedback quality, and overall task performance. Frustration and other standard NASA-TLX categories (e.g., temporal demand, effort) were not assessed in this pilot study. Our goal was to evaluate system-specific workflow and operational issues rather than perform a comprehensive workload assessment. This clarification has been added in the Dry lab evaluation of the framework section.

The adjustments can be found on page 8

Post-experiment survey: Upon completing all required trials, participants fill out a post-experiment questionnaire to assess their experience with each platform. A Likert-scale survey was used to capture user experience, focusing on mental demand, physical demand, motion accuracy, visual feedback, and overall performance for specific workflow components. These pilot data provide targeted insights into the humanoid platform's usability and task-specific performance.

R2.14

The radial error in the motion precision evaluation of over 10mm is massive and much larger than the allowable error in the FLS circular cutting task of 2mm. This needs to be addressed.

Reply:

We thank the referee for pointing this out. The relatively large motion error primarily arises from multiple sources, including extended kinematic chain measurement inaccuracies and limitations in the current RCM calibration procedure (see also R1.1). We acknowledge this as a limitation of the present system. As discussed in the dry-lab user study section, although human operators can partially compensate for such inaccuracies through closed-loop visual feedback during teleoperation, the observed task performance remains consistent with the benchtop motion precision evaluation and typically requires greater user effort and skill.

The edits appear on page 5, 6, 10, 11

factor. This feedback aligns with the quantitative benchtop analysis discussed in previous sections, where higher latency and motion error require greater operator skill to close the control loop. Additionally, the Unitree G1 humanoid has an arm span

R2.15

Latency needs to be measured and reported and compared to SOTA teleop systems and performance impact discussed.

Reply:

This is a very helpful observation, which is also pointed out by referee 1 in R1.4. To address this, we have characterized the system latency using an external high frame rate camera, which has been included in the Benchtop experiments section (see also R1,1).

R2.16

Define acronyms like MTM on first occurrence.

Reply:

We appreciate the referee for pointing this out. We have defined MTM acronym at the first occurrence in the introduction. The adjustment can be found on page 3.

form laparoscopic manipulation using manual wristed instruments via a Master Tool Manipulator (MTM). Leveraging stereo vision and manual laparoscopic tools, the pro-

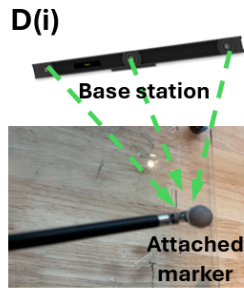
R2.17

More details should be included about the motion precision evaluation with the Optitrack system (ie. how was the tool tip tracked, calibration error, etc.), a photo of the experimental setup would be useful. I wonder if the lever arm effect is contributing to the out-of-plane error. Nonetheless, an image that helps convey how the ground truth measurement was obtained would help.

Reply:

This is a valuable suggestion. To address this concern, we have included more details on the Optitrack system as well as a demonstration photo of the experiment setup in Fig 2 D(i). Additionally, we have further discussed how the evaluation was conducted given the tracked trajectories.

These adjustments appear on page 6, 7.



jectories while maintaining the RCM constraint. The executed tool-tip trajectory is tracked using an OptiTrack motion-capture system (NaturalPoint) by attaching a reflective marker to the tool tip, as shown in Fig. 2 D(i).

Since aligning the OptiTrack base-station coordinate frame with the robot base frame would introduce additional physical measurement errors, direct trajectory comparison in Cartesian space becomes unreliable. Instead, precision is measured using RMS geometric residuals relative to fitted motion primitives. The fitted trajectories are obtained via Principal Component Analysis (PCA): perpendicular deviations are reported for lines, while radial deviation and out-of-plane deviation are reported for circles. The detailed derivation is illustrated in the Method section. The predefined primitives include 100 mm straight lines and planar circles with a diameter of 80 mm. The recorded qualitative results are shown in Fig. 2D(ii, iii).

R2.18

More details on the RCM calibration using the ArUco markers or a reference to a study that defines this process should be included.

Reply:

We appreciate this suggestion. In this study, the humanoid's head camera was used to detect ArUco markers to estimate the physical position of the ports relative to the robot, a widely used approach for object localization in robotics. We have added this clarification in the revised manuscript.

These updates can be found on page 4, 14

motion to robot wrist movements. The physical RCM is localized using ArUco markers detected by the robot's front head camera. Detailed system design and technical figuration with four trocars (three 5 mm, one 12 mm). As shown in Fig. 4B(i), the ports are localized before the operation by attaching calibration boards with ArUco markers, which are detected by the humanoid's head camera. After port placement,

R2.19

Minor errors: There are some typos through the manuscript (ie. on page 19, mis-spelling in MTM definition, typo in figure 2 caption (bencht op), missing capital on pg 21 "we parameterize".)

Reply:

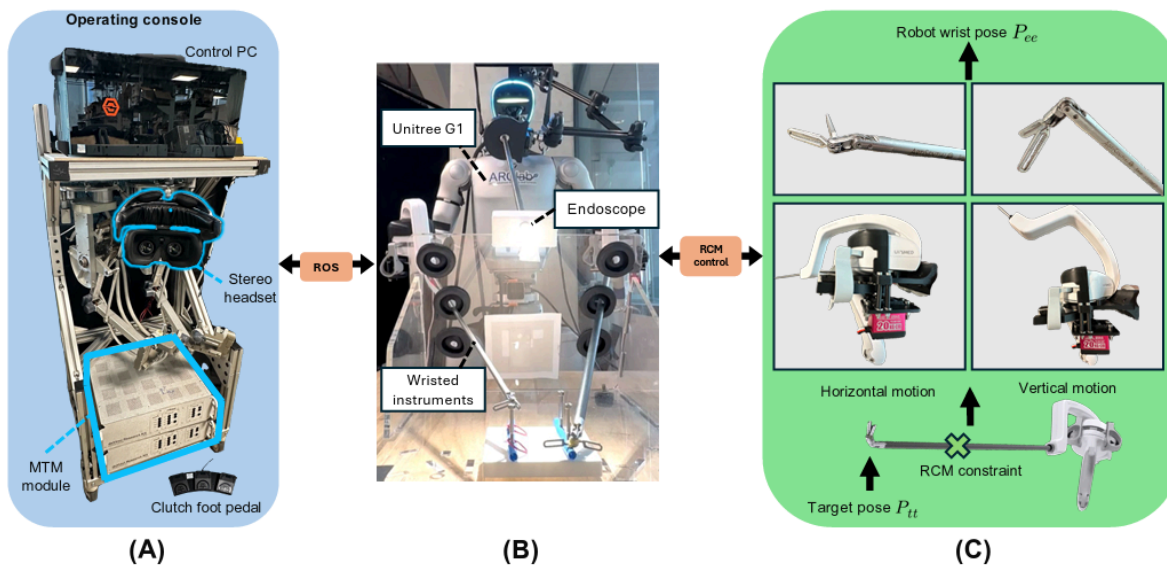
This is a very helpful observation. We have fixed the mentioned typos in the revised manuscript.

R2.20

Figure 1A should be improved to better visualize the MTMs.

Reply:

We have revised the figure to improve the visualization of the MTMs. In the updated version, the MTMs are highlighted and shown at a larger scale.

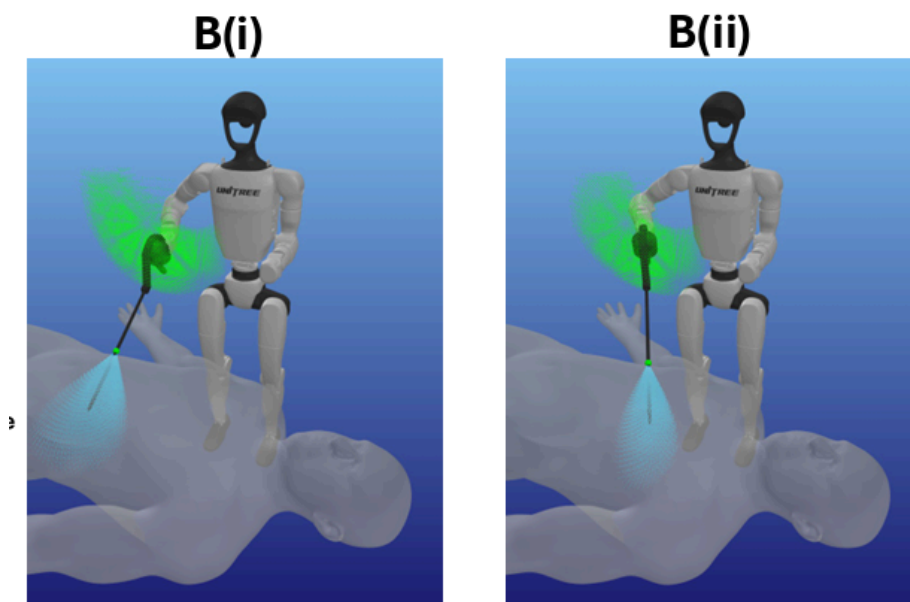


R2.21

Figure 2 (bi and bii) are a bit hard to interpret at first glance, I wonder if an additional side view of the humanoid / patient would make it easier to see the hand and instrument workspace

Reply:

To address this, we reprocessed the figure from side views with more realistic models.

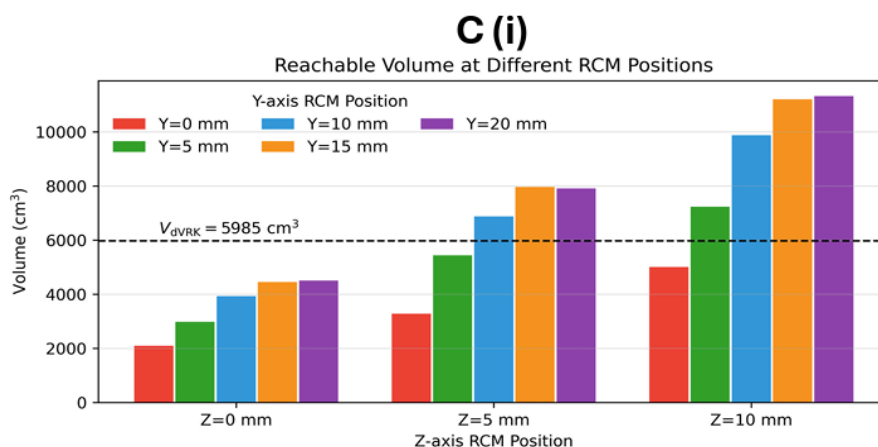


R2.22

Workspace analysis: One small comment is that it would be useful to see the da Vinci workspace size as either a dashed line or additional bar chart on Figure 2(ci) for quick reference.

Reply:

We appreciate the referee for pointing this out. We have included the da Vinci workspace reference in the chart.



R2.23

Revise this sentence, demonstrate what? “Quantitative assessment of surgical performance demonstrated that active console time, in which the surgeon was directly manipulating instruments using the humanoid robot, was seen in Table 2.”

Reply:

We have adjusted this sentence on page 15

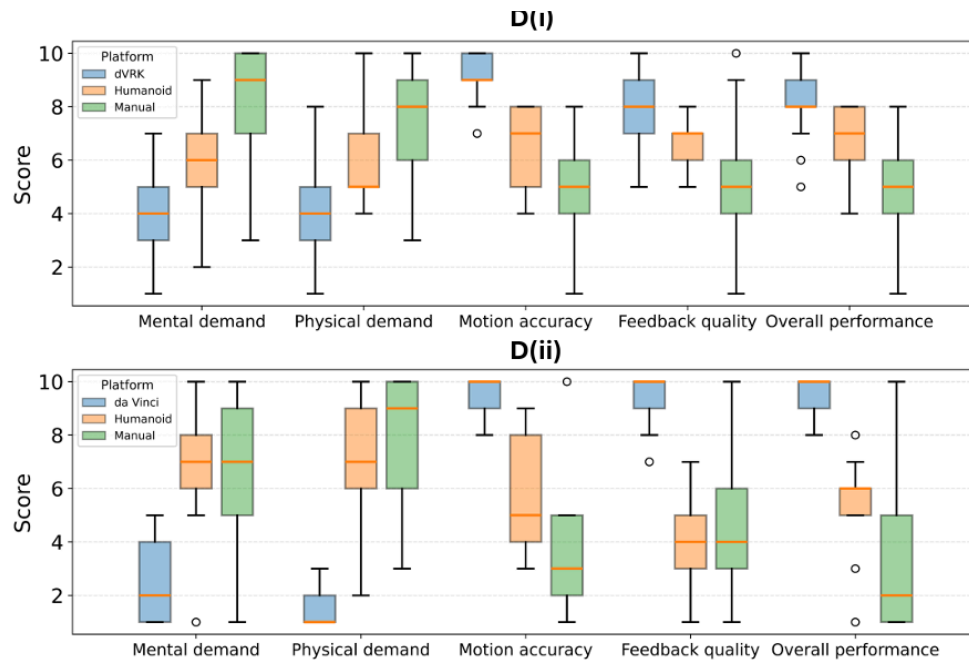
Quantitative assessment of surgical performance, including the duration of active console time when the surgeon directly manipulated instruments with the humanoid robot, is summarized in Table 2. Segment-specific analysis of active console times

R2.24

Inconsistent capitalization in the NASA TLX bar chart labels.

Reply:

We updated the capitalization in Fig 3 D(i, ii)



R2.25

Typo “exchange6” on page 18

Reply:

The typo has been fixed on page 21.

R2.26

The benchtop user studies rely on only 2 expert surgeon users which makes me question if the p-values in the surgeon section of table 1 are truly meaningful. Also, there is no standard deviation reported for table 3. More information on the surgeon training (i.e. years of practice in robotic and laparoscopic surgery) would be helpful.

Reply:

We acknowledge this concern regarding only 2 expert surgeons in the first task, and p-value may not be informative. Therefore, in the revised manuscript, we removed the p-values for the two surgeons, and additional statistical details have been reported in Table 1 (see also R1.5). In the in-vivo study, only four participants completed the survey, with one serving as the primary operating surgeon and the others in supporting roles. Due to this small sample size, we did not report standard deviations, as they would be statistically unreliable and potentially misleading. Instead, we now report the number of responses for each survey category, noting that not all participants answered every question. This clarification allows readers to interpret the results in the context of participant roles, particularly highlighting that only the primary surgeon provided responses for questions related to direct console operation (see also R2.8).

R2.27

Generally, I'm not confident that the quantitative results from the benchtop user studies are significant given the limited number of expert users. The novice group is sufficient, but surgeons trained in both manual and daVinci guided laparoscopy will likely have different experiences with the system. I think the study would strongly benefit from more surgeon users for the benchtop studies. Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

Reply:

We appreciate this observation. To further validate the system with participants with stronger clinical backgrounds, we invited seven surgical residents and two medical students for the second task (see also R2.7), as clarified on page 9. Additional statistical analyses, including confidence intervals and tests of distributional assumptions, have been added to Table 1 (see also R1.5) to provide more interpretable and informative results.

To further validate the proposed framework on participants with stronger medical backgrounds, the classic Fundamentals of Laparoscopic Surgery (FLS) peg transfer task is included in the user study for its diverse wrist motions and well-defined quantitative evaluation protocol. Nine participants with professional medical training took part, including seven surgical residents and two medical students. As illustrated in

R2.28

I think there are some points in the conclusions and data interpretation that can be improved as I have commented on earlier. In particular, the number of expert surgeon users limits the statistical power of the comparisons to manual and dVRK systems. Further, the patterns between benchtop tasks are not distinct, as the authors mention, each system appears to have different strengths and weaknesses with the current analysis. I believe that a more extensive analysis with validated evaluation metrics from the literature will glean more insight on these comparisons.

Reply:

This comment is very insightful. As discussed in R2.27, we added more extensive statistical data to the user study data, including those from more professional backgrounds. In addition, we have included more data interpretation to the observed performance across platforms, while acknowledging that there still exists a gap between the developed system and established surgical robots due to commercial maturity and technology refinements.

These edits appear on page 10, 13.

The results are reported in Table 1(C, D) and Fig. 3C(i, ii). The humanoid platform shows intermediate performance, completing the task faster than manual operation on average (713.8 ± 687.9 s) with lower errors (2.75 ± 3.24), yielding a moderate balanced FLS score (72.79 ± 30.49). Overall, the results demonstrate a performance separation similar to that observed in the first task, with the da Vinci platform consistently achieving the highest performance, manual operation the lowest, and the humanoid platform falling in between.

The results are reported in Table 1(D, E) and Fig. 3C(i, ii). The humanoid platform shows intermediate performance, completing the task faster than manual operation on average (713.8 ± 687.9 s) with fewer peg drops (2.75 ± 3.24), yielding a moderate balanced FLS score (72.79 ± 30.49). Paired comparisons indicate that the da Vinci platform significantly outperforms both the humanoid and manual conditions in completion time and balanced FLS score (Table 1 E). Overall, the results demonstrate a performance separation similar to that observed in the first task, with the da Vinci platform achieving the highest performance, manual operation the lowest, and the humanoid platform falling in between. The larger performance gap between the da Vinci and humanoid platforms is likely attributable to the greater commercial maturity of the da Vinci Xi system compared with the research-oriented dVRK platform used in the first task.

The results from both tasks collectively demonstrate the current performance of the humanoid platform for laparoscopic-style manipulation. In the O-ring replacement task, novice error rates on the humanoid are comparable to the dVRK and substantially lower than manual operation, while expert surgeons achieve their lowest errors on the humanoid, indicating strong support for precise manipulation with sufficient expertise. In the FLS peg transfer task, the humanoid shows intermediate performance, outperforming manual operation in both time and error but remaining behind the da Vinci system.

R2.29

Lastly, I think the survey results from the in-vivo work require further information (like standard deviations) or distinction between each of the participants and their specific role during the experiment.

Reply:

We appreciate this comment. As discussed in R2.26, the in-vivo survey included only four participants; therefore, standard deviations were not reported, and the number of responses for each survey category and participant roles have now been clarified in the revised manuscript (see also R2.8).

Response to Referees

We thank the referees for their careful assessment of the revised manuscript. Their comments provide valuable insights that have greatly helped improve the paper quality. We appreciate their enthusiasm for the work, and we have done our best to address the referees' remaining concerns point by point in the following text.

Referee 1 [R1]

No remaining comments to address.

Referee 2 [R2]

R2.1

R2.6 – No additional analysis was added, the response just indicates how the FLS score was normalized, and some additional statistics were included on the same metrics (time, weighted error). I still feel that additional metrics could be derived from this data to better understand how the humanoid vs. dVRK vs. manual performance gap truly materializes.

Reply:

We thank the referee for pointing this out. To better reflect performance across platforms with a more comprehensive analysis, we revisited the operational data and expanded our evaluation metrics to include common errors such as unintentional collisions and failed peg pick-up attempts. Furthermore, since we have invited several additional participants to this study, we also report group-wise performance summaries to better characterize the effect of expertise level on performance.

These additions appear in the revised manuscript page 10, 15.

To comprehensively evaluate efficiency and safety, task performance is quantified using completion time, operational errors, and the FLS score. Operational errors are

defined as peg drops, unintentional collisions, and failed peg pickups. The occurrences of each error type are reported in Table 2(B). The FLS peg transfer task convention-

Table 2 FLS peg transfer performance across platforms and expertise levels ($n = 13$). (A) Performance summary. (B) Error type occurrences report. (C) Paired comparison statistics across expertise levels. In this task, participants include 8 junior surgeons and 5 senior surgeons

(A) FLS peg transfer: performance summary (mean $\pm$ std)					
Group	Platform	Time (s) $\downarrow$	Total Error $\downarrow$	FLS Score(0–100) $\uparrow$	
Overall	Manual	877.8 $\pm$ 774.3	17.45 $\pm$ 16.80	70.47 $\pm$ 27.53	
	Humanoid	560.1 $\pm$ 595.5	6.25 $\pm$ 6.36	85.39 $\pm$ 16.25	
	da Vinci Xi	118.2 $\pm$ 56.7	1.33 $\pm$ 1.23	97.67 $\pm$ 1.54	
Junior surgeons	Manual	1234.2 $\pm$ 881.4	24.83 $\pm$ 19.63	57.86 $\pm$ 31.47	
	Humanoid	765.1 $\pm$ 726.3	8.71 $\pm$ 7.52	79.54 $\pm$ 19.67	
	da Vinci Xi	146.7 $\pm$ 58.1	1.57 $\pm$ 1.27	96.94 $\pm$ 1.47	
Senior surgeons	Manual	450.2 $\pm$ 329.5	8.60 $\pm$ 6.62	85.61 $\pm$ 11.49	
	Humanoid	273.0 $\pm$ 86.6	2.80 $\pm$ 0.84	93.57 $\pm$ 1.73	
	da Vinci Xi	78.4 $\pm$ 19.6	1.00 $\pm$ 1.22	98.69 $\pm$ 1.04	
(B) Error type occurrences					
Platform	Drops	Collision	Failed Pick Up		
Manual	3.82 $\pm$ 3.76	8.27 $\pm$ 9.12	5.36 $\pm$ 5.61		
Humanoid	2.17 $\pm$ 2.76	2.08 $\pm$ 1.98	2.42 $\pm$ 2.39		
da Vinci Xi	0.17 $\pm$ 0.39	1.00 $\pm$ 1.04	0.17 $\pm$ 0.39		
(C) Paired statistics (overall)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time (s)	da Vinci Xi – Humanoid	−467.27 [−850.44, −84.10]	9.77e−04	0.82	w
Time (s)	da Vinci Xi– Manual	−757.73 [−1242.53, −272.93]	9.77e−04	1.05	w
Time (s)	Humanoid – Manual	−290.45 [−526.92, −53.98]	0.021	0.83	t
Error	da Vinci Xi – Humanoid	−5.18 [−9.67, −0.70]	0.004	0.78	w
Error	da Vinci Xi – Manual	−16.09 [−27.44, −4.74]	9.77e−04	0.95	w
Error	Humanoid – Manual	−10.91 [−18.03, −3.78]	9.77e−04	1.03	w
FLS Score	da Vinci Xi – Humanoid	12.97 [2.22, 23.72]	9.77e−04	0.81	w
FLS Score	da Vinci Xi – Manual	27.14 [9.21, 45.06]	9.77e−04	1.02	w
FLS Score	Humanoid – Manual	14.17 [5.77, 22.56]	0.004	1.13	t
(C2) Paired statistics (junior surgeons)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time	da Vinci Xi – Humanoid	−694.50 [−1441.64, 52.64]	0.031	0.98	w
Time	da Vinci Xi – Manual	−1079.33 [−1947.57, −211.10]	0.024	1.30	t
Time	Humanoid – Manual	−384.83 [−790.54, 20.87]	0.059	1.00	t
Error	da Vinci Xi – Humanoid	−8.00 [−16.55, 0.55]	0.061	0.98	t
Error	da Vinci Xi – Manual	−23.17 [−44.07, −2.27]	0.031	1.16	w
Error	Humanoid – Manual	−15.17 [−27.99, −2.34]	0.031	1.24	w
Balanced FLS	da Vinci Xi – Humanoid	19.52 [−1.30, 40.33]	0.031	0.98	w
Balanced FLS	da Vinci Xi – Manual	38.85 [6.51, 71.19]	0.027	1.26	t
Balanced FLS	Humanoid – Manual	19.33 [6.21, 32.46]	0.013	1.55	t
(C3) Paired statistics (senior surgeons)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time	da Vinci Xi – Humanoid	−194.60 [−314.32, −74.88]	0.062	2.02	w
Time	da Vinci Xi – Manual	−371.80 [−774.06, 30.46]	0.062	1.15	w
Time	Humanoid – Manual	−177.20 [−556.94, 202.54]	0.265	0.58	t
Error	da Vinci Xi – Humanoid	−1.80 [−3.64, 0.04]	0.053	1.21	t
Error	da Vinci Xi – Manual	−7.60 [−16.58, 1.38]	0.079	1.05	t
Error	Humanoid – Manual	−5.80 [−13.11, 1.51]	0.092	0.98	t
Balanced FLS	da Vinci Xi – Humanoid	5.12 [1.82, 8.42]	0.013	1.93	t
Balanced FLS	da Vinci Xi – Manual	13.08 [−1.60, 27.76]	0.062	1.11	w
Balanced FLS	Humanoid – Manual	7.96 [−4.86, 20.79]	0.160	0.77	t

Notes: Δ is the mean paired difference with 95% CI. d_z is Cohen’s d for paired samples.
Test: paired t -test (t) or Wilcoxon signed-rank (w).

R2.2

R2.7 – The manuscript would benefit from more explicitly labelling the number of users in each part of the study (n = 12, n=9 in the captions where relevant). Without this, and without the now added sentence that states 9 users did the FLS peg transfer task, it is easy to assume that all 12 users did all the tasks. I think it is a key detail that the sample size changed and that the user groups are not the same from o-ring to FLS peg transfer task. I'm also not sure that grouping these results in Table 1 is the best way to present them given the cohort differences for each category.

Reply:

We thank the referee for raising this concern. We have included the explicit participant numbers and their information in the captions of both tasks. And the tables for both tasks have been separated for better clarity.

These additions appear in the revised manuscript page 14, 15.

Table 1 O-ring transfer performance across platforms and expertise levels.(n = 18)
(A) Performance summary. (B) Error type occurrences. (C) Paired comparison statistics. In this task, participants include 12 novices and 6 surgeons.

(A) O-ring transfer: weighted error and time							
Group	Platform		Weighted Error ↓		Time (s) ↓		
Overall	Manual		7.03 ± 4.88		64.22 ± 31.82		
	Humanoid		4.53 ± 3.14		74.67 ± 39.22		
	dVRK		4.59 ± 2.53		43.01 ± 15.25		
Novices	Manual		7.91 ± 5.64		75.43 ± 32.66		
	Humanoid		5.49 ± 3.35		89.23 ± 38.91		
	dVRK		5.00 ± 2.83		45.37 ± 17.93		
Surgeons	Manual		5.26 ± 2.34		41.79 ± 13.79		
	Humanoid		2.61 ± 1.44		45.54 ± 19.25		
	dVRK		3.77 ± 1.72		38.30 ± 6.57		
(B) Error type occurrences							
Platform	Failed pick	Stretch (peg)	Stretch (handoff)	Drop	Collision	Straw disp.	
Manual	6.11 ± 4.30	5.17 ± 4.95	0.83 ± 1.04	0.83 ± 1.10	7.17 ± 5.70	0.28 ± 0.46	
Humanoid	5.44 ± 4.63	2.39 ± 2.59	0.39 ± 0.61	1.11 ± 1.23	3.33 ± 3.24	0.33 ± 0.77	
dVRK	3.17 ± 2.01	0.78 ± 0.88	0.22 ± 0.73	0.28 ± 0.57	7.44 ± 4.59	0.67 ± 0.69	
(C) Paired statistics (overall)							
Metric	Comparison		Δ [95% CI]		p	d _z	Test
Weighted error	Humanoid – Manual		−2.50 [−4.42, −0.58]		0.014	0.65	t
Weighted error	dVRK – Manual		−2.44 [−4.76, −0.12]		0.040	0.52	t
Weighted error	Humanoid – dVRK		−0.06 [−1.35, 1.23]		0.922	0.02	t
Time	Humanoid – Manual		10.45 [−5.86, 26.75]		0.194	0.32	t
Time	dVRK – Manual		−21.21 [−35.37, −7.04]		0.004	0.74	w
Time	Humanoid – dVRK		31.65 [17.22, 46.09]		2.41e−04	1.09	t
(C2) Paired statistics (Novices)							
Metric	Comparison		Δ [95% CI]		p	d _z	Test
Weighted error	Humanoid – Manual		−2.42 [−5.30, 0.45]		0.091	0.54	t
Weighted error	dVRK – Manual		−2.92 [−6.44, 0.61]		0.096	0.53	t
Weighted error	Humanoid – dVRK		0.49 [−1.38, 2.36]		0.575	0.17	t
Time	Humanoid – Manual		13.80 [−9.56, 37.16]		0.220	0.38	t
Time	dVRK – Manual		−30.06 [−49.35, −10.77]		0.006	0.99	t
Time	Humanoid – dVRK		43.86 [27.87, 59.85]		4.88e−04	1.74	w
(C3) Paired statistics (Surgeons)							
Metric	Comparison		Δ [95% CI]		p	d _z	Test
Weighted error	Humanoid – Manual		−2.65 [−5.08, −0.22]		0.038	1.15	t
Weighted error	dVRK – Manual		−1.49 [−3.75, 0.78]		0.094	0.69	w
Weighted error	Humanoid – dVRK		−1.16 [−2.51, 0.18]		0.076	0.91	t
Time	Humanoid – Manual		3.74 [−21.98, 29.46]		0.724	0.15	t
Time	dVRK – Manual		−3.49 [−16.98, 9.99]		0.535	0.27	t
Time	Humanoid – dVRK		7.23 [−13.68, 28.15]		0.688	0.36	w

Notes: Δ is mean paired difference with 95% CI. d_z is Cohen's d . Test: paired t -test (t) or Wilcoxon (w).

Table 2 FLS peg transfer performance across platforms and expertise levels ($n = 13$). (A) Performance summary. (B) Error type occurrences report. (C) Paired comparison statistics across expertise levels. In this task, participants include 8 junior surgeons and 5 senior surgeons

(A) FLS peg transfer: performance summary (mean $\pm$ std)					
Group	Platform	Time (s) $\downarrow$	Total Error $\downarrow$	FLS Score(0–100) $\uparrow$	
Overall	Manual	877.8 $\pm$ 774.3	17.45 $\pm$ 16.80	70.47 $\pm$ 27.53	
	Humanoid	560.1 $\pm$ 595.5	6.25 $\pm$ 6.36	85.39 $\pm$ 16.25	
	da Vinci Xi	118.2 $\pm$ 56.7	1.33 $\pm$ 1.23	97.67 $\pm$ 1.54	
Junior surgeons	Manual	1234.2 $\pm$ 881.4	24.83 $\pm$ 19.63	57.86 $\pm$ 31.47	
	Humanoid	765.1 $\pm$ 726.3	8.71 $\pm$ 7.52	79.54 $\pm$ 19.67	
	da Vinci Xi	146.7 $\pm$ 58.1	1.57 $\pm$ 1.27	96.94 $\pm$ 1.47	
Senior surgeons	Manual	450.2 $\pm$ 329.5	8.60 $\pm$ 6.62	85.61 $\pm$ 11.49	
	Humanoid	273.0 $\pm$ 86.6	2.80 $\pm$ 0.84	93.57 $\pm$ 1.73	
	da Vinci Xi	78.4 $\pm$ 19.6	1.00 $\pm$ 1.22	98.69 $\pm$ 1.04	
(B) Error type occurrences					
Platform	Drops	Collision	Failed Pick Up		
Manual	3.82 $\pm$ 3.76	8.27 $\pm$ 9.12	5.36 $\pm$ 5.61		
Humanoid	2.17 $\pm$ 2.76	2.08 $\pm$ 1.98	2.42 $\pm$ 2.39		
da Vinci Xi	0.17 $\pm$ 0.39	1.00 $\pm$ 1.04	0.17 $\pm$ 0.39		
(C) Paired statistics (overall)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time (s)	da Vinci Xi – Humanoid	−467.27 [−850.44, −84.10]	9.77e−04	0.82	w
Time (s)	da Vinci Xi– Manual	−757.73 [−1242.53, −272.93]	9.77e−04	1.05	w
Time (s)	Humanoid – Manual	−290.45 [−526.92, −53.98]	0.021	0.83	t
Error	da Vinci Xi – Humanoid	−5.18 [−9.67, −0.70]	0.004	0.78	w
Error	da Vinci Xi – Manual	−16.09 [−27.44, −4.74]	9.77e−04	0.95	w
Error	Humanoid – Manual	−10.91 [−18.03, −3.78]	9.77e−04	1.03	w
FLS Score	da Vinci Xi – Humanoid	12.97 [2.22, 23.72]	9.77e−04	0.81	w
FLS Score	da Vinci Xi – Manual	27.14 [9.21, 45.06]	9.77e−04	1.02	w
FLS Score	Humanoid – Manual	14.17 [5.77, 22.56]	0.004	1.13	t
(C2) Paired statistics (junior surgeons)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time	da Vinci Xi – Humanoid	−694.50 [−1441.64, 52.64]	0.031	0.98	w
Time	da Vinci Xi – Manual	−1079.33 [−1947.57, −211.10]	0.024	1.30	t
Time	Humanoid – Manual	−384.83 [−790.54, 20.87]	0.059	1.00	t
Error	da Vinci Xi – Humanoid	−8.00 [−16.55, 0.55]	0.061	0.98	t
Error	da Vinci Xi – Manual	−23.17 [−44.07, −2.27]	0.031	1.16	w
Error	Humanoid – Manual	−15.17 [−27.99, −2.34]	0.031	1.24	w
Balanced FLS	da Vinci Xi – Humanoid	19.52 [−1.30, 40.33]	0.031	0.98	w
Balanced FLS	da Vinci Xi – Manual	38.85 [6.51, 71.19]	0.027	1.26	t
Balanced FLS	Humanoid – Manual	19.33 [6.21, 32.46]	0.013	1.55	t
(C3) Paired statistics (senior surgeons)					
Metric	Comparison	Δ [95% CI]	p	d_z	Test
Time	da Vinci Xi – Humanoid	−194.60 [−314.32, −74.88]	0.062	2.02	w
Time	da Vinci Xi – Manual	−371.80 [−774.06, 30.46]	0.062	1.15	w
Time	Humanoid – Manual	−177.20 [−556.94, 202.54]	0.265	0.58	t
Error	da Vinci Xi – Humanoid	−1.80 [−3.64, 0.04]	0.053	1.21	t
Error	da Vinci Xi – Manual	−7.60 [−16.58, 1.38]	0.079	1.05	t
Error	Humanoid – Manual	−5.80 [−13.11, 1.51]	0.092	0.98	t
Balanced FLS	da Vinci Xi – Humanoid	5.12 [1.82, 8.42]	0.013	1.93	t
Balanced FLS	da Vinci Xi – Manual	13.08 [−1.60, 27.76]	0.062	1.11	w
Balanced FLS	Humanoid – Manual	7.96 [−4.86, 20.79]	0.160	0.77	t

Notes: Δ is the mean paired difference with 95% CI. d_z is Cohen’s d for paired samples.
Test: paired t -test (t) or Wilcoxon signed-rank (w).

R2.3

R2.8 – It's good to split the responses into who did which task in Table 3 and I understand this is a subjective assessment of the in vivo feasibility study. However, it's clearer now that these results are entirely dependent on one user's opinion in some categories. I think that's okay for clinical feasibility, but it underscores why the bench top analysis should be stronger.

Reply:

Thank you for the comment. We agree that the subjective nature of certain assessments reflects reliance on individual surgeon judgment. Early studies of the surgical robotic system primarily established feasibility and safety through in vivo clinical deployment without extensive benchtop evaluation (e.g. <https://www.nature.com/articles/35096636>, <https://pmc.ncbi.nlm.nih.gov/articles/PMC1422462/>), in part due to inherently operator-dependent and heterogeneous nature of surgical practice. While benchtop analysis enables controlled comparisons, it may not fully capture the variability and contextual factors present in clinical settings. Ultimately, the adoption of such systems depends on surgeons' assessment of safety, usability, and potential clinical benefit within their existing workflows. Accordingly, we place great emphasis on clinical feasibility as a key indicator of real-world applicability.

R2.4

R2.11 – The comment about the harness as a safety precaution should be in the paper, not just the response letter.

Reply:

We appreciate this comment. We have added the clarification for harness usage in the manuscript.

The adjustment can be found in the revised manuscript page 17.

field. In this study, a harness was used as a precautionary safeguard, which is not required for system functionality. Deployment required close coordination between sur-

R2.5

R2.13 – Deciding to omit the frustration, temporal demand and effort in a conventional NASA TLX survey is still unclear to me. I understand that the intention was not to do a full workload assessment, but this would have helped to better understand the differences between this system and conventional robots or manual laparoscopic tools.

Reply:

We agree that a task-oriented assessment is helpful for understanding system performance. Therefore, We have added a full TLX workload assessment alongside the system evaluation, and asked the five senior surgeons invited during this revision to complete it. The results align well with quantitative performance evaluation.

The adjustment can be found in the revised manuscript page 16

Table 3 Task workload assessment by senior surgeons. Subjective workload assessment (mean $\pm$ standard deviation) across platforms based on post-study questionnaires.

Metric	da Vinci Xi	Humanoid	Manual
Mental demand ↓	2.00 $\pm$ 1.22	5.40 $\pm$ 1.34	8.80 $\pm$ 0.84
Physical demand ↓	1.60 $\pm$ 0.89	4.40 $\pm$ 2.61	8.20 $\pm$ 0.84
Temporal demand ↓	1.60 $\pm$ 0.89	4.00 $\pm$ 1.58	7.40 $\pm$ 3.13
Overall performance ↑	9.20 $\pm$ 1.30	6.00 $\pm$ 1.22	3.40 $\pm$ 1.14
Effort ↓	2.20 $\pm$ 1.30	5.60 $\pm$ 1.82	8.80 $\pm$ 1.10
Frustration ↓	1.40 $\pm$ 0.89	4.20 $\pm$ 0.84	8.60 $\pm$ 1.34

R2.6

R2.14 - The discussion and response with respect to the large radial error and its impact on the system performance is currently insufficient in the paper. The description of why this gap arises in 2.1.2 (extended kinematic chain - resulting in discrepancies between commanded and executed motion & ArUco marker-based RCM localization error) is not detailed enough. I would like to see a more explicit description why this error is so high and how it impacts the overall system. An estimate or assessment of what can be attributed to the ArUco marker-based RCM localization error would help too. Given this seems to be one of the biggest technical limitations of the system now, it should also be explicitly addressed in the discussion when summarizing the performance gaps. Further, the closed-loop visual feedback compensation (described in the response letter) is not mentioned anywhere in the paper. If this was feedback from the users, it should be included.

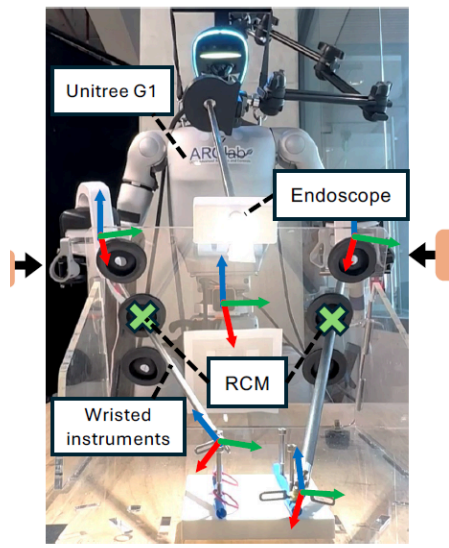
Reply:

We acknowledge this limitation. In the revised manuscript, we have added a more detailed discussion of the root causes of the command–execution errors, including a kinematic analysis and an improved visualization of the kinematic chain in the system figure. We have also incorporated this limitation into the summary. In addition, the feedback quality from the response letter refers to the perceived responsiveness from the system, which is caused by the system latency. This is further clarified in the user study section.

The adjustment can be found in the revised manuscript page 4, 6, 7, 12, 24

the order of ~ 1 mm under calibrated conditions [53]. From the experiments, this gap primarily arises from inaccuracies in the formulated kinematic chain for wristed laparoscopic instrument control. As shown in Fig. 1(B, C) and Fig. 5(A), the tool-tip pose of the manual instrument, observed in the endoscopic view, is used as the control input and is fed into the extended inverse kinematics solver to compute the corresponding robot wrist pose. This extended kinematic chain incorporates the instrument's geometric parameters, such as shaft lengths and link offsets. However, since these commercial

tools do not provide publicly available manufacturing specifications, all geometric parameters must be manually measured, introducing discrepancies between the modeled kinematic chain and the ground truth. Furthermore, RCM control is highly sensitive to trocar localization accuracy, particularly in terms of orientation. Due to the lever-like coupling between the robot wrist and the tool tip, even small errors in vision-based ArUco marker detection can be amplified along the kinematic chain. Together, these factors contribute to increased command-execution errors, particularly when compared to integrated, purpose-built surgical robotic systems.



and expand robotic capabilities beyond traditional operating rooms. In the future, improved system integration for humanoid manipulation of surgical instruments will be critical for enhancing control accuracy and operational safety. This includes the use of open-source laparoscopic instruments with well characterized geometries, as well as more accurate RCM localization techniques. Future work should also focus on

tion. Key limitations include reduced responsiveness in control feedback introduced by latency, a less intuitive control interface compared with the da Vinci Surgical System,

R2.7

R2.27 – Thank you for the response. 12 and then 9 users, only 2 of which are surgeons, still does not feel like a sufficient cohort. The inconsistency with the user groups between tasks is also a limitation. The new statistics (confidence interval, effect size, etc.) don't seem to really strengthen

many of these results either. This statement “In the surgeon group, performance is generally more accurate and consistent across all platforms; surgeons are fastest on the dVRK, achieve their lowest error on the humanoid, and exhibit a similar timing trend as novices.” - only has 2 data points supporting it. Furthermore, most of the confidence intervals for the o-ring transfer task and FLS peg transfer cross zero, some with very wide intervals which points to high uncertainty.

Reply:

We acknowledge the limitation in participant numbers. To address this, we invited several additional surgeons to participate in both tasks. As a result, the group was expanded to n=18 (12 novices, 6 surgeons) for the first task and n=13 (8 junior surgeons, 5 senior surgeons) for the second task, accompanied by a more comprehensive performance analysis (see also R2.2). The reported results demonstrate more robustness and reliability.

The adjustment can be found in the revised manuscript page 14, 15

R2.8

It appears that there are only 3 significant claims from the dry lab evaluation (below - taken directly from the text):

1. The humanoid platform yields error levels comparable to the dVRK (5.49 ± 3.35), with no significant difference in weighted error between humanoid and dVRK operation ($\Delta = 0.49$, 95% CI $[-1.38, 2.36]$, $p = 0.575$).
2. Completion times differ significantly across platforms: the dVRK is faster than manual operation ($\Delta = -30.06$ s, 95% CI $[-49.35, -10.77]$, $p = 0.006$, $dz = 0.99$), while the humanoid platform requires longer completion times than the dVRK ($\Delta = 43.86$ s, 95% CI $[27.87, 59.85]$, $p < 0.001$, $dz = 1.74$).
3. Paired comparisons indicate that the da Vinci platform significantly outperforms both the humanoid and manual conditions in completion time and balanced FLS score (Table 1 E).

These significant conclusions are contradictory between tasks which makes it difficult to draw conclusions. In the o-ring transfer task the significant finding is that there is no real difference between the dVRK and the humanoid, but for the FLS peg transfer task – the dVRK is significantly better. The timing differences are also inconsistent – in Figure 3 B(ii) it seems the humanoid is the slowest for o-ring (significantly), but falls in the middle for FLS peg transfer (Figure 3 Cii). Given the

limited sample size, and the difference in cohorts and experience levels – I don't think these results support meaningful or rigorous conclusions.

Reply:

Thank you for the comment, and we believe we can clear up any confusion, through greater clarity distinguishing the dVRK (used in the first task) and the da Vinci Xi (used in the second task) platform. The dVRK platform is the research platform that comprises a modified, first-generation da Vinci system, and is used across over 50 research labs around the world working on surgical robotics research, whereas the da Vinci Xi is a modern, clinically deployed platform. Naturally these two system performances will differ and the Xi will exhibit better performance due to the platform's maturity. To reduce potential confusion, we have added explicit labeling of the da Vinci Xi in the updated performance analysis table (see also R2.2). In addition to clarifying the performance presentation, we have expanded the discussion to highlight how the greater maturity of the da Vinci Xi may contribute to the observed performance gap relative to the dVRK used in the first task. Furthermore, following the referee's suggestion to include more participants, the results yield clearer and more robust conclusions (see also R2.2).

The adjustment can be found in the revised manuscript page xxx

distributions are shown in Fig. 3 B(i, ii). Overall, manual operation achieves moderate completion time (64.22 ± 31.82 s) but exhibits the highest weighted error and variability (7.03 ± 4.88). In contrast, both the humanoid platform (4.53 ± 3.14) and the dVRK (4.59 ± 2.53) achieve lower weighted errors, with comparable performance between the two, while the dVRK provides the fastest completion time (43.01 ± 15.25 s).

In the novice group, the dVRK achieves both the fastest completion time and the lowest error, while the humanoid platform demonstrates intermediate accuracy but requires the longest time. In the surgeon group, performance is generally more accurate and consistent across all platforms; surgeons achieve their lowest error on the humanoid platform while remaining fastest on the dVRK, following a similar timing trend as observed in novices. Paired comparisons show significant reductions in weighted error for both the humanoid platform and the dVRK compared to manual operation. Completion times differ substantially across platforms: the dVRK is faster than both manual operation and the humanoid platform.

Overall, these results highlight the performance improvements of the humanoid platform over manual operation while also underscoring the need for further development to achieve more responsive and precise control.

The results are reported in Table 2 and Fig. 3C(i, ii). Overall, a clear performance hierarchy is observed across platforms. The humanoid platform demonstrates reduced completion time and lower total errors relative to manual operation, resulting in an improved FLS score (85.39 ± 16.25 vs. 70.47 ± 27.53). However, it remains notably slower and less accurate than the da Vinci Xi system, which achieves the best overall performance with the shortest completion time (118.2 ± 56.7 s), lowest error (1.33 ± 1.23), and highest FLS score (97.67 ± 1.54). The trend is consistent across experience levels, with senior surgeons achieving better overall performance across all three platforms, likely due to greater familiarity and experience with both robotic and manual laparoscopic operations. Paired comparisons confirm that the da Vinci Xi significantly outperforms both the humanoid and manual conditions across all metrics, while the humanoid platform also shows significant improvements over manual operation. In addition, the five participating senior surgeons completed a task-oriented workload assessment questionnaire across platforms, as summarized in Table 3. The subjective feedback aligns with the quantitative performance metrics.

Overall, these results reinforce a consistent performance separation, with the da Vinci Xi platform achieving the highest performance, manual operation the lowest, and the humanoid platform falling in between. Compared to the first task, The larger performance gap between the da Vinci Xi and humanoid platforms is likely attributable to the greater commercial maturity of the da Vinci Xi system compared with the research-oriented dVRK platform used in the first task.

Response to the remaining issues raised by referees

We thank the referees for their careful assessment of the revised manuscript. We appreciate their enthusiasm for the work, and we have made the following adjustments to address the referees' remaining concerns point by point.

Referee 1 [R1]

No remaining comments to address.

Referee 2 [R2]

Minor remaining comment: Be sure to define the error bars in the legend for Figure 3.

Adjustment:

We thank the referee for pointing this out. We have added definitions for the error bars for Figure 3. This change can be found in the legend for Figure 3.

Figure 3. Operation performance, and post-study survey results are reported for user study tasks. (A)(i-iii) User study conditions. **(B)** O-ring replacement performance results ($n = 18$ independent participants). **(C)** FLS peg transfer performance results ($n = 13$ independent participants). **(D)** Post study survey results of the ring replacement **(i)** ($n = 18$ independent participants) and FLS peg transfer **(ii)** ($n = 13$ independent participants). For all box-and-whisker plots, the center line indicates the median, box bounds indicate the interquartile range (IQR), whiskers extend to the most extreme data points within $1.5 \times$ IQR, and dots indicate individual trials outside this range.