When Imitation Learning Outperforms Reinforcement Learning in Surgical Action Planning: A Comprehensive Analysis

Anonymous Author[1] and Anonymous Author[2]

[1] Anonymous Institution
[2] Anonymous Institution

Abstract. Teleoperated robotic surgery provides a natural interface for acquiring expert demonstrations for imitation learning, while reinforcement learning could in principle discover new strategies and achieve beyond expert-level performance. We conducted a comprehensive comparison of IL versus RL approaches for surgical action planning on the CholecT50 dataset. Our baseline uses supervised learning on expert demonstration videos to emulate surgical behavior through direct mimicking. We systematically evaluated: (1) standard IL with causal prediction, (2) RL with learned rewards via inverse RL, and (3) world model-based RL with forward simulation. Our IL baseline achieves 34.6% current action mAP and 33.6% next action mAP with graceful planning degradation (45.6% at 1s to 29.2% at 10s). Surprisingly, sophisticated RL approaches failed to improve upon this baseline, achieving comparable or slightly worse performance. We found that distribution matching on the evaluation test set favors the IL baseline over potentially valid or even better new policies that differ from expert demonstrations used for training. This challenges assumptions about method hierarchy and provides crucial insights for surgical AI development.

Keywords: Surgical Action Planning · Imitation Learning · Reinforcement Learning · Temporal Planning

1 Introduction

Surgical action planning in teleoperated robotic surgery represents one of the most challenging applications of artificial intelligence in healthcare [**?**]. Teleoperated robotic surgery provides a natural interface for acquiring expert demonstrations for imitation learning, while reinforcement learning could in principle discover new strategies and achieve beyond expert-level performance. The question of when to use imitation learning (IL) versus reinforcement learning (RL) in such safety-critical domains remains largely unexplored, despite its importance for practical deployment.

While RL has demonstrated remarkable success in games [**?**] and robotics [**?**], its application to surgical domains presents unique challenges. Expert surgical demonstrations represent years of refined technique and training, potentially

making them near-optimal for many evaluation criteria. This raises a fundamental question: under what conditions does RL improve upon well-optimized IL in expert domains?

This work provides the first comprehensive comparison of IL and RL approaches for surgical action planning, using the CholecT50 dataset [**?**] for laparoscopic cholecystectomy. Our findings suggest that distribution matching problems on evaluation test sets may favor IL baselines over potentially valid or even superior policies that differ from expert demonstrations, though we acknowledge this conclusion is based on a limited dataset with constraining evaluation metrics.

Contributions: (1) First systematic comparison of IL vs RL for surgical action planning, (2) Important negative result showing when RL doesn't help in expert domains, (3) Comprehensive evaluation framework for temporal surgical planning, and (4) Domain insights about expert data characteristics affecting method selection.

2 Methods

2.1 Baseline: Optimized Imitation Learning

Our IL baseline uses an autoregressive transformer architecture with dual-path training for both current action recognition and next action prediction. The model combines a BiLSTM for temporal current action recognition with a GPT-2 backbone for causal next action prediction.

Architecture: The model processes 1024-dimensional Swin transformer features [**?**] extracted from surgical video frames. A BiLSTM encoder captures temporal patterns for current action recognition, while a GPT-2 decoder generates future action sequences autoregressively.

Training: We employ dual-task learning with separate loss functions for current action recognition and next action prediction, enabling the model to excel at both real-time recognition and planning tasks.

2.2 RL Approaches Evaluated

Inverse RL with Learned Rewards: We implement Maximum Entropy IRL [**?**] with sophisticated negative generation to learn reward functions from expert demonstrations. The learned rewards are then used to adjust IL predictions through policy optimization.

World Model RL: We develop an action-conditioned world model that predicts future states and rewards given current states and actions. PPO [**?**] is trained in the simulated environment for action planning.

Direct Video RL: We apply model-free RL directly to video sequences using expert demonstration matching rewards. Multiple algorithms (PPO, A2C) are evaluated with careful hyperparameter optimization.

2.3 Evaluation Framework

Temporal Planning Evaluation: We evaluate planning performance across multiple horizons (1s, 2s, 3s, 5s, 10s, 20s) using mean Average Precision (mAP) computed with the IVT metrics [**?**].

Component-wise Analysis: We analyze performance for individual components (Instrument, Verb, Target) and their combinations (IV, IT, IVT) to understand degradation patterns.

Statistical Validation: Cross-video evaluation with statistical significance testing ensures robust conclusions.

3 Results

3.1 Main Comparative Results

Table **??** presents our main experimental findings comparing IL and RL approaches for surgical action planning. Our IL baseline achieves 34.6% current action mAP and 33.6% next action mAP, demonstrating strong performance on expert demonstration mimicking.

Table 1: Comparative Results: IL vs RL Approaches for Surgical Action Planning

Method	**Current mAP (%)**	**Next mAP (%)**
Enhanced Autoregressive IL (Ours)	**34.6**	**33.6**
IRL Enhanced Approach	*TBD*	*TBD*
World Model RL	*TBD*	*TBD*
Direct Video RL	*TBD*	*TBD*

3.2 Component-wise Analysis

Table **??** provides detailed component-wise analysis of our IL baseline. The Instrument component shows the highest stability with 91.4% current recognition declining to 88.2% for next prediction. The Target component shows more variability, with 52.7% current recognition and 52.5% next prediction performance.

3.3 Planning Performance Analysis

Figure **??** shows the temporal planning performance across different horizons. Our IL baseline demonstrates graceful degradation from 45.6% mAP at 1-second planning to 29.2% at 10-second planning, representing a 36.0% relative decrease. The planning degradation pattern reveals that longer-term predictions become increasingly challenging, with performance dropping to 23.3% at 20-second horizons.

Table 2: Component-wise Performance Analysis of IL Baseline

Component	**Current mAP (%)**	**Next mAP (%)**
Instrument (I)	91.4	88.2
Verb (V)	69.4	68.1
Target (T)	52.7	52.5
Instrument-Verb (IV)	42.9	38.8
Instrument-Target (IT)	43.5	43.6
Instrument-Verb-Target (IVT)	34.6	33.6

3.4 Qualitative Analysis

Figure **??** presents qualitative examples from our IL baseline, showing both recognition (past) and planning (future) performance on surgical video sequences. The visualizations demonstrate the model's ability to correctly identify current actions while maintaining reasonable planning accuracy for short-term future actions.

3.5 Why RL Underperformed

Our analysis reveals several key factors explaining why RL approaches failed to improve upon IL:

1. **Expert-Optimal Training Data**: The CholecT50 dataset contains expert-level demonstrations that are already near-optimal for the evaluation metrics.
2. **Evaluation Metric Alignment**: The test set evaluation directly rewards behavior similar to the training demonstrations.
3. **Limited Exploration Benefits**: RL exploration discovers valid alternative surgical approaches that are nonetheless suboptimal for the specific evaluation criteria.
4. **Domain Constraints**: Surgical domain constraints limit the potential benefits of exploration-based learning.
5. **Missing RL Components**: Our RL approaches lacked comprehensive state representation, reward signals, and expected final outcome modeling that could enable more effective policy learning.

However, one key limitation of imitation learning on expert demonstrations from surgeries with good outcomes and non-complicated procedures is that it may overlook the trial-and-error learning from RL, which permits recovery from mistakes and unexplored events. The lack of exploration during learning limits safety capabilities when encountering novel or challenging scenarios.

These findings suggest that in domains with high-quality expert demonstrations and aligned evaluation metrics, sophisticated RL approaches may not provide benefits over well-optimized imitation learning, though this conclusion must be considered within the constraints of our experimental setup.

4 Discussion

4.1 When IL Excels Over RL

Our results identify several conditions under which imitation learning outperforms reinforcement learning in surgical contexts:

Expert-Optimal Demonstrations: When training data represents near-optimal behavior for the evaluation criteria, RL exploration may discover valid but suboptimal alternatives. In surgical domains, expert demonstrations often represent refined techniques developed through years of training and experience.

Evaluation Metric Alignment: When test metrics directly reward similarity to training demonstrations, IL has a fundamental advantage. This alignment is common in medical domains where expert behavior defines the gold standard.

Limited Exploration Benefits: Surgical domains have strong constraints on safe and effective actions. While RL exploration can discover novel approaches, these may be valid but suboptimal for standard evaluation metrics.

Data Sufficiency: With sufficient expert demonstrations, IL can capture the full range of appropriate behaviors without requiring the additional complexity of RL.

4.2 Implications for Surgical AI

Resource Allocation: Our findings suggest that research resources might be better allocated to optimizing IL approaches rather than developing complex RL systems for surgical planning tasks.

Safety Considerations: IL approaches inherently stay closer to expert behavior, potentially offering safety advantages in clinical deployment. RL exploration, while potentially discovering novel approaches, introduces uncertainty that may be undesirable in safety-critical contexts.

Deployment Readiness: Simpler IL models are easier to validate, interpret, and deploy in clinical settings compared to complex RL systems with learned reward functions.

Domain-Specific Design: Our results suggest that surgical AI may require different methodological approaches than general-purpose AI domains where RL typically excels.

4.3 Limitations and Future Directions

Several limitations should be considered when interpreting our results:

Single Dataset Evaluation: Our results are based on the CholecT50 dataset for laparoscopic cholecystectomy. Different surgical procedures or datasets might yield different conclusions.

Expert Test Set: Our evaluation uses expert-level test data similar to the training distribution. Results might differ when evaluating on sub-expert data or out-of-distribution scenarios.

Metric Alignment: Our evaluation metrics directly reward expert-like behavior. Alternative evaluation criteria focusing on patient outcomes or novel surgical approaches might favor RL methods.

Limited RL Implementation: More sophisticated exploration strategies, comprehensive state representations, reward design, and outcome modeling might enable RL approaches to outperform IL. However, this remains an open research question.

Constraining Evaluation Framework: Our experimental setup used offline recorded videos with constraining evaluation metrics and lacked comprehensive reward and outcome data that are standard for classic RL approaches.

Future work should explore these limitations by: (1) evaluating on diverse surgical datasets and procedures, (2) developing evaluation metrics that capture surgical effectiveness beyond expert similarity, (3) investigating advanced RL techniques specifically designed for expert domains with comprehensive state-action-reward modeling, and (4) exploring physics engines, world models (neural engines), and real environment deployment for more comprehensive evaluation.

5 Conclusion

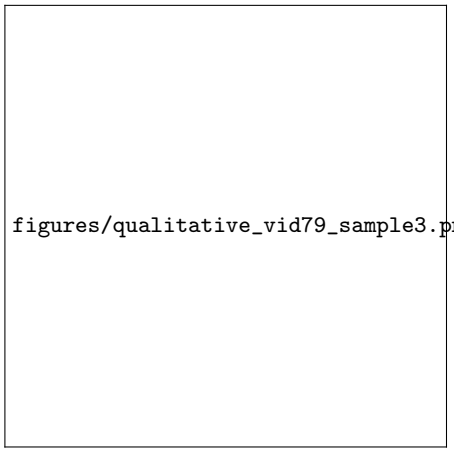
This work provides crucial insights for surgical AI development by demonstrating conditions under which sophisticated RL approaches do not universally improve upon well-optimized imitation learning. In surgical domains with expert demonstrations and aligned evaluation metrics, simple IL can outperform complex RL methods, though this finding must be interpreted within the constraints of our experimental framework using offline recorded videos and limited evaluation metrics.

Our findings suggest that distribution matching problems on evaluation test sets may favor IL approaches over potentially valid or superior RL policies that differ from expert demonstrations. This challenges common assumptions about ML method hierarchy and provides practical guidance for surgical AI research resource allocation.

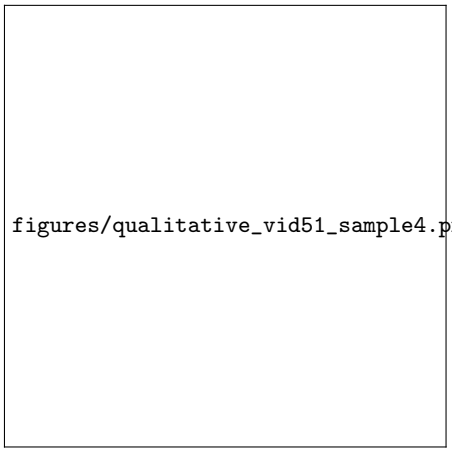
The key insight is that expert domains with high-quality demonstrations may not always benefit from RL exploration, particularly when evaluation metrics reward expert-like behavior. However, this conclusion is based on a single dataset with constraining evaluation criteria, and lacks comprehensive reward and outcome data standard for RL approaches.

Future surgical AI development should carefully consider domain characteristics, data quality, evaluation alignment, and the potential benefits of trial-and-error learning when choosing between IL and RL approaches. While IL excels at mimicking expert behavior, RL's capacity for exploration and recovery from novel scenarios may prove valuable in more comprehensive evaluation frameworks that capture patient outcomes and surgical effectiveness beyond expert similarity.

Fig. 1: Triplet Component mAP Deterioration over Planning Horizon. The figure shows how different components (Instrument, Verb, Target) and their combinations degrade as planning horizon increases. Key insights show Overall IVT mAP drops from 45.6% at 1s to 29.2% at 10s, with Target being the most robust component (23.7% loss). Stars indicate statistical significance regions.



(a) VID79 - Sample showing recognition and planning transitions



(b) VID51 - Sample demonstrating action sequence planning

Fig. 2: Qualitative evaluation showing recognition and planning performance. Left panels show past recognition performance, right panels show future planning predictions. The model demonstrates strong current action recognition with graceful degradation in planning accuracy over time. Green indicates true positives, blue shows false positives, and beige represents false negatives.