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. Surgical action planning requires learning from expert demonstrations while ensuring safe and effective decision-making. While reinforcement learning (RL) has shown promise in various domains, its effectiveness compared to imitation learning (IL) in surgical contexts remains unclear. We conducted a comprehensive comparison of IL versus RL approaches for surgical action planning on the CholecT50 dataset. Our baseline autoregressive transformer achieves strong performance through expert demonstration learning. 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 45.6% current action mAP and 44.9% next action mAP with graceful planning degradation (47.1% at 1s to 29.1% at 10s). Surprisingly, sophisticated RL approaches failed to improve upon this baseline, achieving comparable or slightly worse performance. In surgical domains with expert demonstrations, well-optimized imitation learning can outperform complex RL approaches. This challenges the assumption that RL universally improves upon IL and provides crucial insights for surgical AI development.

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

1 Introduction

Surgical action planning represents one of the most challenging applications of artificial intelligence in healthcare, requiring models to learn from expert demonstrations while ensuring safe and effective decision-making. 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 key finding challenges conventional wisdom: sophisticated RL approaches fail to improve upon a well-optimized IL baseline, achieving comparable or worse performance across multiple 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 45.6% current action mAP and 44.9% next action mAP, outperforming all RL variants tested.

Notably, the IRL enhanced approach achieves 44.2% current mAP and 43.8% next mAP, representing a 1.4% and 1.1% decrease from the IL baseline, respectively. Similarly, world model RL achieves 42.1% and 41.6% for current and next action prediction, while direct video RL achieves 43.9% and 43.1%.

3.2 Planning Performance Analysis

Figure **??** shows the temporal planning performance across different horizons. Our IL baseline demonstrates graceful degradation from 47.1% mAP at 1-second planning to 29.1% at 10-second planning, representing a 38.2% relative decrease.

The planning degradation pattern is consistent across all methods, suggesting that this limitation is fundamental to the temporal planning task rather than method-specific. However, the IL baseline consistently maintains higher absolute performance at all horizons.

3.3 Component-wise Analysis

Table **??** provides detailed component-wise analysis of our IL baseline. The Instrument component shows the highest stability with 90.3% current recognition declining to 87.9% for next prediction. The Target component shows more variability, with 57.1% current recognition and 56.1% next prediction performance.

Combination components (IV: 45.7%, IT: 53.2%) show the expected multiplicative effects of their constituent components, with performance levels that reflect the interaction complexity.

3.4 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.

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.

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.

Exploration Strategies: More sophisticated exploration strategies or reward design might enable RL approaches to outperform IL. However, this remains an open research question.

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, and (3) investigating advanced RL techniques specifically designed for expert domains.

5 Conclusion

This work provides crucial insights for surgical AI development by demonstrating that 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.

Our findings challenge common assumptions about ML method hierarchy and provide practical guidance for surgical AI research resource allocation. The key insight is that expert domains with high-quality demonstrations may not benefit from RL exploration, particularly when evaluation metrics reward expert-like behavior.

Future surgical AI development should carefully consider domain characteristics, data quality, and evaluation alignment when choosing between IL and RL approaches. In many cases, focusing optimization efforts on IL rather than complex RL systems may yield better results with lower complexity and risk.