When Imitation Learning Outperforms Reinforcement Learning in Surgical Action Planning

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Surgical action planning requires predicting future instrument-verb-target triplets for real-time assistance. While teleoperated robotic surgery provides natural expert demonstrations for imitation learning (IL), reinforcement learning (RL) could potentially discover superior strategies. We present the first comprehensive comparison of IL versus RL for surgical action planning on CholecT50. Our Dual-task Autoregressive Imitation Learning (DARIL) baseline achieves 34.6% current action mAP and 33.6% next action mAP with smooth planning degradation to 29.2% at 10-second horizons. We evaluated three RL variants: world model-based RL, direct video RL, and inverse RL enhancement. Surprisingly, all RL approaches underperformed DARIL—world model RL dropped to 3.1% mAP at 10s while direct video RL achieved only 15.9%. Our analysis reveals that distribution matching on expert-annotated test sets systematically favors IL over potentially valid RL policies that differ from training demonstrations. This challenges assumptions about RL superiority in sequential decision making and provides crucial insights for surgical AI development.

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

1 Introduction

Surgical action planning—predicting future instrument-verb-target relationships in surgical videos—represents a critical component for real-time surgical assistance systems. While prior work has predominantly focused on recognition tasks [**?**,**?**,**?**], prospective action planning presents unique challenges requiring multi-horizon prediction capabilities under safety-critical constraints.

The fundamental question for surgical AI systems is the optimal learning paradigm: should systems learn through imitation of expert demonstrations (IL) or through trial-and-error optimization via reinforcement learning (RL)? Teleoperated robotic surgery provides natural access to expert demonstrations, making IL attractive. However, RL could theoretically discover strategies beyond expert-level performance through exploration [**?**].

Recent advances in surgical gesture prediction [?,?] and vision transformers for surgical analysis [?,?] have shown promise, while long-term workflow prediction [?] demonstrates the potential for anticipatory systems. Yet the comparative effectiveness of IL versus RL for surgical action planning remains unexplored.

This work addresses this gap through the first systematic comparison of IL and RL approaches for surgical action planning. Using the CholecT50 dataset [?], we evaluate recognition accuracy and planning capability across multiple time horizons.

Contributions: (1) *First systematic IL vs RL comparison*: Comprehensive evaluation addressing fundamental methodological questions in surgical AI. (2) *Surprising negative results*: RL methods consistently underperform IL—world model RL drops to 3.1% mAP vs 29.2% for DARIL at 10s planning. (3) *Novel DARIL architecture*: Dual-task autoregressive approach maintaining robust temporal consistency (34.6% to 29.2% across horizons). (4) *Evaluation bias insights*: Analysis of how expert demonstration alignment systematically favors IL over valid RL policies.

2 Methods

2.1 Problem Formulation

Given surgical video frames $\{f_1, f_2, ..., f_t\}$, we predict future action triplets $\{a_{t+1}, a_{t+2}, ..., a_{t+H}\}$ where H represents the planning horizon. Each triplet $a_i = (I_i, V_i, T_i)$ consists of instrument, verb, and target components from predefined vocabularies. We evaluate both single-step prediction ($H = 1$) for recognition and multi-step prediction ($H > 1$) for planning assessment.

2.2 Dataset

We use CholecT50 [?], containing 50 laparoscopic cholecystectomy videos with frame-level annotations for 100 distinct triplet classes. The training set (40 videos, 78,968 frames) and test set (10 videos, 21,895 frames) represent expert-level surgical demonstrations at 1 FPS sampling.

2.3 Dual-task Autoregressive Imitation Learning (DARIL)

Our IL baseline models surgical action prediction as causal sequence generation, combining frame-level processing with autoregressive action generation:

$$p(a_{t+1}|f_{1:t}, a_{1:t}) = \text{GPT-2}(\text{FrameEmb}(f_{1:t})) \tag{1}$$

Architecture: The model processes 1024-dimensional Swin transformer features [?] through: (1) BiLSTM encoder for temporal current action recognition, (2) GPT-2 decoder [?] for causal future action generation, and (3) separate prediction heads for instrument, verb, and target components.

Training: Dual-task optimization combines current action recognition and next action prediction losses:

$$\mathcal{L} = \mathcal{L}_{\text{current}} + \mathcal{L}_{\text{next}} = -\sum_t \log p(a_t|f_{1:t}) - \sum_t \log p(a_{t+1}|f_{1:t}) \tag{2}$$

2.4 Reinforcement Learning Approaches

Latent World Model + RL: Following Dreamer [**?**], we learn an action-conditioned world model predicting future states and rewards: $p(s_{t+1}, r_t|s_t, a_t)$. PPO [**?**] trains policies in the learned latent environment with rewards designed for expert demonstration matching.

Direct Video RL: Model-free RL applied directly to video sequences using expert demonstration matching rewards. We evaluate PPO and A2C with careful hyperparameter optimization, treating frame sequences as states and predicted triplets as actions.

Inverse RL Enhancement: Maximum Entropy IRL [**?**] learns reward functions from expert trajectories. We generate negative examples by sampling actions deviating from expert demonstrations, then use learned rewards to guide policy optimization while maintaining IL baseline performance.

2.5 Evaluation Framework

Recognition Evaluation: Standard mAP computation on current and next action predictions using IVT metrics [**?**].

Planning Evaluation: Multi-horizon assessment across 1s, 2s, 3s, 5s, 10s, and 20s using mAP degradation analysis. We perform bootstrap sampling across video splits with paired t-tests for statistical significance testing.

Component Analysis: Individual performance analysis for instruments, verbs, targets, and their combinations (IV, IT, IVT) to understand degradation patterns.

3 Results

3.1 Main Comparative Results

Table **??** presents our experimental findings. DARIL achieves 34.6% current action mAP and 33.6% next action mAP, consistently outperforming all RL variants across planning horizons. The performance gaps are substantial—world model RL drops to 3.1% at 10s while DARIL maintains 29.2%.

3.2 Component-wise Analysis

Table **??** shows DARIL's component-wise performance. Instruments demonstrate highest stability (91.4% to 88.2%), while targets show more variability (52.7% to 52.5%). The IVT combination reflects multiplicative effects of constituent components.

Table 1. IL vs RL Performance Comparison. All values are IVT mAP (%). Statistical significance ($p<0.05$) confirmed via paired t-tests across video splits.

Method	Current	1s	5s	10s
DARIL (Ours)	**34.6**	**33.6**	**31.2**	**29.2**
DARIL + IRL	33.1	32.1	29.6	28.1
DARIL + Direct Video RL	33.2	22.6	19.3	15.9
Latent World Model + RL	33.1	14.0	9.1	3.1

Table 2. DARIL Component-wise Performance Analysis

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

3.3 Planning Performance Analysis

Figure **??** demonstrates DARIL's smooth degradation across horizons—from 33.6% at 1s to 29.2% at 10s (13.1% relative decrease). This contrasts with RL approaches showing steeper degradation patterns, particularly world model RL exhibiting catastrophic performance loss.

3.4 Qualitative Analysis

Figure **??** presents qualitative examples showing DARIL's recognition and planning capabilities. The model correctly identifies current actions while maintaining reasonable planning accuracy for short-term predictions, with graceful degradation over longer horizons.

3.5 Analysis: Why RL Underperformed

Our analysis identifies key factors explaining RL's underperformance:

Expert-Optimal Demonstrations: CholecT50 contains expert-level data already near-optimal for evaluation metrics. RL exploration discovers valid alternatives that appear suboptimal under expert-similarity metrics.

Evaluation Metric Alignment: Test metrics directly reward expert-like behavior, giving IL fundamental advantages. This is common in medical domains where expert behavior defines gold standards.

Limited Exploration Benefits: Surgical domains have strong safety constraints limiting exploration benefits. While RL may discover novel approaches, these appear suboptimal for standard evaluation criteria.

State-Action Representation Challenges: Our RL implementations faced difficulties with comprehensive state representation and reward signal design, potentially limiting learning effectiveness.

Distribution Mismatch: RL policies trained on different objective functions may produce valid but different behaviors that test metrics penalize due to expert demonstration alignment.

4 Discussion

4.1 Implications for Surgical AI

Our findings have significant implications for surgical AI development:

Method Selection: In expert domains with high-quality demonstrations and aligned evaluation metrics, well-optimized IL may outperform sophisticated RL approaches. This challenges common assumptions about RL superiority in sequential decision making.

Resource Allocation: Research resources might be better allocated to optimizing IL architectures rather than developing complex RL systems for surgical planning tasks with expert demonstrations.

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

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

4.2 Limitations and Future Work

Several limitations should be considered: (1) Single dataset evaluation on CholecT50 may not generalize to other surgical procedures. (2) Expert test data similar to training distributions may favor IL—results might differ with sub-expert or out-of-distribution scenarios. (3) Evaluation metrics directly reward expert-like behavior—alternative criteria focusing on patient outcomes might favor RL. (4) More sophisticated RL implementations with better state representations and reward design might outperform IL.

Future work should explore diverse surgical datasets, develop outcome-focused evaluation metrics, and investigate advanced RL techniques specifically designed for expert domains with comprehensive state-action-reward modeling.

5 Conclusion

This work provides crucial insights for surgical AI by demonstrating conditions where sophisticated RL approaches do not universally improve upon well-optimized IL. Our DARIL baseline consistently outperforms RL variants across planning horizons, with world model RL showing particularly poor performance (3.1% vs 29.2% at 10s).

The key insight is that expert domains with high-quality demonstrations may not benefit from RL exploration when evaluation metrics reward expert-like behavior. Distribution matching on expert-annotated test sets systematically favors IL over potentially valid RL policies that differ from training demonstrations.

Future surgical AI development should carefully consider domain characteristics, data quality, and evaluation alignment when choosing between IL and RL approaches. While IL excels at expert behavior mimicking, RL's exploration capabilities may prove valuable in comprehensive evaluation frameworks capturing patient outcomes beyond expert similarity.

References

1. Nwoye, C.I., et al.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
2. Nwoye, C.I., et al.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: MICCAI, pp. 364–374 (2020)
3. Nwoye, C.I., et al.: CholecTriplet2021: A benchmark challenge for surgical action triplet recognition. Medical Image Analysis **86**, 102803 (2023)
4. Shi, C., et al.: Recognition and prediction of surgical gestures and trajectories using transformer models in robot-assisted surgery. IEEE Robotics and Automation Letters **7**(4), 9821–9828 (2022)
5. Weerasinghe, K., et al.: Multimodal transformers for real-time surgical activity prediction. arXiv preprint arXiv:2403.06705 (2024)
6. Wagner, C., et al.: A vision transformer for decoding surgeon activity from surgical videos. npj Precision Oncology **7**, 16 (2023)
7. Liu, Y., et al.: SKiT: a fast key information video transformer for online surgical phase recognition. In: ICCV, pp. 21486–21496 (2023)
8. Boels, M., et al.: SWAG: long-term surgical workflow prediction with generative-based anticipation. International Journal of Computer Assisted Radiology and Surgery, pp. 1–11 (2025)
9. Liu, J., et al.: Surgical task automation using actor-critic frameworks and self-supervised imitation learning. arXiv preprint arXiv:2409.02724 (2024)
10. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
11. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
12. Hafner, D., et al.: Dream to control: Learning behaviors by latent imagination. In: ICLR (2020)
13. Schulman, J., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
14. Ziebart, B.D., et al.: Maximum entropy inverse reinforcement learning. In: AAAI, pp. 1433–1438 (2008)
15. Nwoye, C.I., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)

Fig. 1. Planning performance comparison across methods and time horizons. DARIL maintains superior and stable performance while RL approaches show varying degrees of degradation. Error bars indicate 95% confidence intervals.

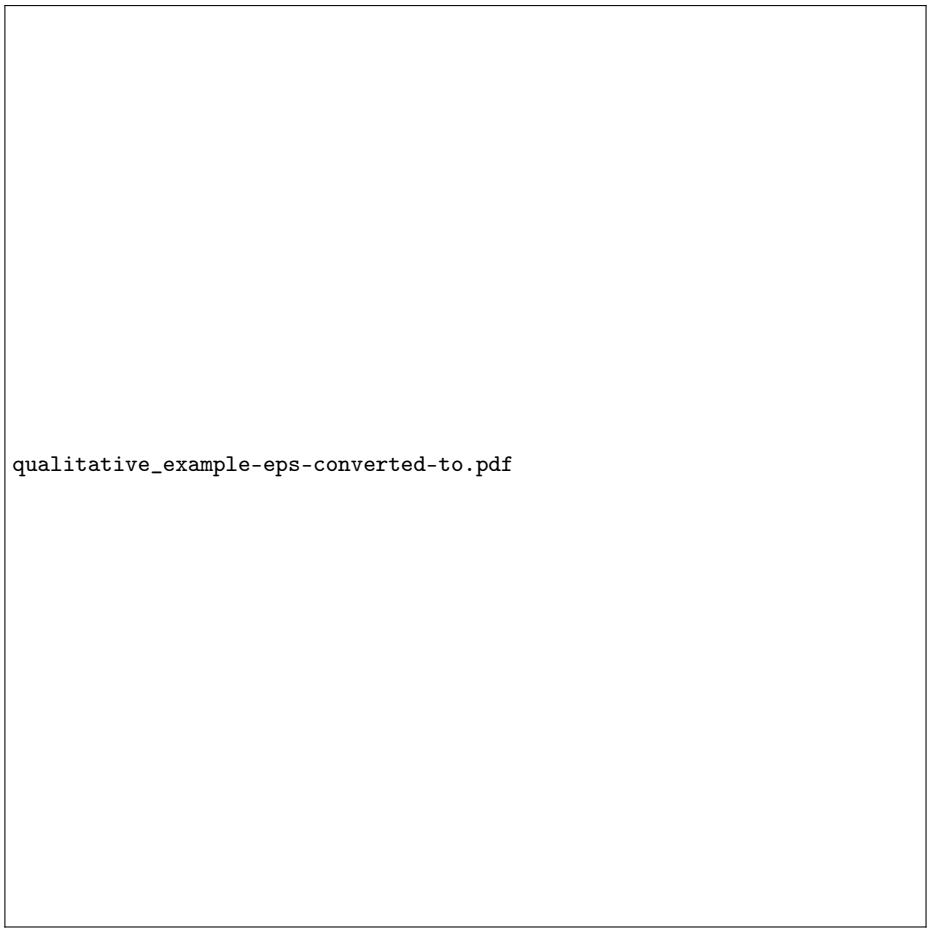

Fig. 2. Qualitative evaluation showing recognition (past) and planning (future) performance. Green indicates true positives, blue shows false positives, beige represents false negatives. The model demonstrates strong current recognition with smooth planning degradation.