When Imitation Learning Outperforms Reinforcement Learning in Surgical Action Planning

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Surgical action triplet prediction has primarily focused on recognition tasks for activity analysis. However, real-time surgical assistance requires next action prediction for planning and control applications. Teleoperated robotic surgery provides a natural interface to acquire expert demonstrations for imitation learning (IL), while reinforcement learning (RL) could in principle discover new strategies and achieve beyond expert-level performance. The question of when to use IL versus RL in surgical domains remains largely unexplored. We conducted a comprehensive comparison of IL versus RL approaches for surgical action planning on the CholecT50 dataset, evaluating both recognition accuracy and planning capability. Our baseline uses supervised learning on expert demonstration videos to learn surgical behavior through direct observation. 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 smooth planning degradation (33.6% at 1s to 29.2% at 10s). Surprisingly, sophisticated RL approaches failed to improve upon this baseline. 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 · Surgical Action Prediction · Surgical AI

1 Introduction

Surgical action planning in teleoperated robotic surgery represents one of the most challenging applications of artificial intelligence in healthcare [**?**,**?**]. Recent advances in surgical action triplet recognition [**?**] and surgical gesture prediction [**?**,**?**] have demonstrated the potential for AI-driven surgical assistance. Teleoperated robotic surgery provides a natural interface for acquiring expert demonstrations for imitation learning. 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. Recent work in surgical automation [?] has shown promise for integrating RL with imitation learning, while vision transformers have emerged as powerful tools for surgical video analysis [?,?,?]. Additionally, advances in long-term surgical workflow prediction [?] demonstrate the potential for anticipatory surgical planning systems. Expert surgical demonstrations represent decades 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.

Contributions: (1) *First systematic IL vs RL comparison for surgical action planning*: We provide the first comprehensive evaluation comparing imitation learning and reinforcement learning approaches for surgical action planning, addressing a fundamental methodological question in surgical AI. (2) *Surprising negative results with clear performance gaps*: We demonstrate that sophisticated RL methods consistently underperform our IL baseline—world model RL drops to 3.1% mAP at 10s vs 29.2% for DARIL, while direct video RL achieves only 15.9% vs 29.2%, challenging assumptions about RL superiority in sequential decision making. (3) *Novel DARIL architecture achieving strong temporal consistency*: Our dual-task autoregressive imitation learning approach maintains robust performance across planning horizons (34.6% current action mAP degrading smoothly to 29.2% at 10-second planning), establishing a new baseline for surgical action planning. (4) *Insights into evaluation bias and domain characteristics*: We identify how expert demonstration alignment with test distributions systematically favors IL over potentially valid RL policies, providing crucial guidance for evaluation design and method selection in expert domains like surgery.

2 Methods

2.1 Dataset

We evaluate our approaches on the CholecT50 dataset [?], which contains 50 laparoscopic cholecystectomy videos with expert-level surgical demonstrations. The dataset provides frame-level annotations for surgical action triplets, consisting of 100 distinct triplet classes combining 6 instruments, 10 verbs, and 15 targets. Videos are sampled at 1 frame per second, with the training set (40 videos) containing 78,968 frames and the test set (10 videos) containing 21,895 frames, totaling approximately 100,863 annotated frames. The dataset represents high-quality expert demonstrations from experienced surgeons performing full procedures.

2.2 Problem Formulation

We formulate surgical action triplet prediction as a sequential decision making problem. Given a sequence of surgical video frames $\{f_1, f_2, ..., f_t\}$, the task is to predict future action triplets $\{a_{t+1}, a_{t+2}, ..., a_{t+H}\}$ where H represents the prediction horizon.

Each action triplet $a_i = (I_i, V_i, T_i)$ consists of an instrument $I_i \in \mathcal{I}$, verb $V_i \in \mathcal{V}$, and target $T_i \in \mathcal{T}$ from predefined vocabularies. We evaluate both single-step prediction ($H = 1$) for recognition comparison and multi-step prediction ($H > 1$) for planning assessment.

2.3 Autoregressive Imitation Learning Baseline

Our IL approach models surgical action prediction as a causal sequence generation problem. The architecture combines frame-level feature extraction with autoregressive action generation:

$$p(a_{t+1}|f_1, ..., f_t, a_1, ..., a_t) = \text{GPT-2}(\text{FrameEmb}(f_1, ..., f_t), a_1, ..., a_t) \quad (1)$$

The model consists of three components:

1. **Frame Processing**: Pre-trained visual features are processed through learned embeddings to create temporal representations
2. **GPT-2 Backbone**: A transformer decoder models causal dependencies between frames and actions
3. **Action Prediction**: Separate heads predict instrument, verb, and target components with IVT-based optimization

Training optimizes the standard imitation learning objective:

$$\mathcal{L}_{IL} = -\sum_{t=1}^{T} \log p(a_t|f_{1:t}, a_{1:t-1}; \theta) \quad (2)$$

2.4 Reinforcement Learning Approaches

We evaluate three RL variants for surgical action prediction:

World Model-Based RL This approach learns a conditional world model predicting future states given current state and action:

$$p(s_{t+1}, r_t|s_t, a_t) = \text{WorldModel}(s_t, a_t; \phi) \quad (3)$$

The world model enables planning through rollouts, with policy optimization using PPO on the learned dynamics. Rewards are designed to match expert demonstrations and encourage task-relevant behavior.

Direct Video RL This model-free approach learns policies directly from video observations without explicit world models. The environment provides frame sequences as states, with actions representing predicted triplets. Rewards incorporate both demonstration matching and task-specific objectives.

Inverse RL Enhancement Building on the IL baseline, this approach learns reward functions from expert trajectories using Maximum Entropy IRL, then applies lightweight policy improvement through GAIL. This enables scenario-specific enhancements while maintaining IL performance for routine cases.

2.5 Evaluation Framework

We introduce a dual evaluation protocol comparing recognition accuracy and planning capability:

Recognition Evaluation: Standard mAP computation on single-step predictions, comparing with existing CholecT50 benchmarks.

Planning Evaluation: Multi-horizon prediction assessment measuring:

- Temporal consistency across prediction horizons (1, 3, 5, 10 steps)
- Trajectory-level planning accuracy using mAP degradation analysis
- Scenario-specific performance on complex vs. routine surgical phases

2.6 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. We refer to this approach as Dual-task Autoregressive Imitation Learning (DARIL).

Architecture: The model processes 1024-dimensional Swin transformer features [**?**] extracted from surgical video frames using distilled feature extraction [**?**]. 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.

2.7 RL Approaches Evaluated

Latent World Model + RL: We develop an action-conditioned world model that predicts future states and rewards given current states and actions, following the paradigm established by Dreamer [**?**]. The goal is to create a latent simulator that enables policy learning in the learned environment. PPO [**?**] is trained in this simulated latent environment for action planning.

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

Inverse RL with Learned Rewards: We implement Maximum Entropy IRL [**?**] with sophisticated negative generation to learn reward functions from expert demonstrations. Negative examples are generated by sampling actions that deviate from expert demonstrations, and the learned reward function is used to guide policy optimization during training by providing dense feedback signals that encourage expert-like behavior while penalizing deviations.

2.8 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. We perform bootstrap sampling across video splits and apply paired t-tests to assess statistical significance of performance differences between methods.

3 Results

3.1 Main Comparative Results

Table **??** presents our main experimental findings comparing IL and RL approaches for surgical action planning. Our DARIL baseline achieves 34.6% current action mAP and 33.6% next action mAP, demonstrating strong performance on expert demonstration learning. Figure **??** provides a comprehensive view of how each method performs across different planning horizons, highlighting the consistent superiority of our DARIL approach.

Table 1. Comparative Results: IL vs RL Approaches for Surgical Action Planning. All values are IVT mAP scores in %. $\text{mAP}_{t=0}$ refers to current action recognition, mAP_{t+1s} to 1-second prediction, etc.

Method	$\mathbf{mAP}_{t=0}$	$\mathbf{mAP}_{t+1s}$	$\mathbf{mAP}_{t+5s}$	$\mathbf{mAP}_{t+10s}$
DARIL (Ours)	**34.6**	**33.6**	**31.2**	**29.2**
DARIL + Direct Video RL	33.2	22.6	19.3	15.9
DARIL + IRL	33.1	32.1	29.6	28.1
Latent World Model + RL	33.1	14.0	9.1	3.1

3.2 Component-wise Analysis

Table **??** provides detailed component-wise analysis of our DARIL 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.

Table 2. Component-wise Performance Analysis of DARIL Baseline. All values are mAP scores in % for each component and combination.

Component	$\mathbf{mAP}_{t=0}$	$\mathbf{mAP}_{t+1s}$
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

3.3 Planning Performance Analysis

Figure **??** shows the temporal planning performance across different horizons. Our DARIL baseline demonstrates smooth degradation from 33.6% mAP at 1-second planning to 29.2% at 10-second planning, representing a 13.1% relative decrease. The planning degradation pattern reveals that longer-term predictions become increasingly challenging, with performance dropping to 23.3% at 20-second horizons.

3.4 Qualitative Analysis

Figure **??** presents qualitative examples from our DARIL 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. This aligns with recent observations that expert demonstrations can significantly improve RL learning efficiency in surgical domains [**?**].

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.

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. Liu, Y., et al.: Lovit: Long video transformer for surgical phase recognition. Medical Image Analysis **99**, 103366 (2025)
9. 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)
10. Liu, J., et al.: Surgical task automation using actor-critic frameworks and self-supervised imitation learning. arXiv preprint arXiv:2409.02724 (2024)
11. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)

12. Levine, S., et al.: End-to-end training of deep visuomotor policies. Journal of Machine Learning Research **17**(1), 1334–1373 (2016)
13. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
14. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
15. Yamlahi, A., et al.: Self-distillation for surgical action recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 637–646. Springer (2023)
16. Hafner, D., et al.: Dream to control: Learning behaviors by latent imagination. In: ICLR (2020)
17. Ziebart, B.D., et al.: Maximum entropy inverse reinforcement learning. In: AAAI, pp. 1433–1438 (2008)
18. Schulman, J., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
19. Nwoye, C.I., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)

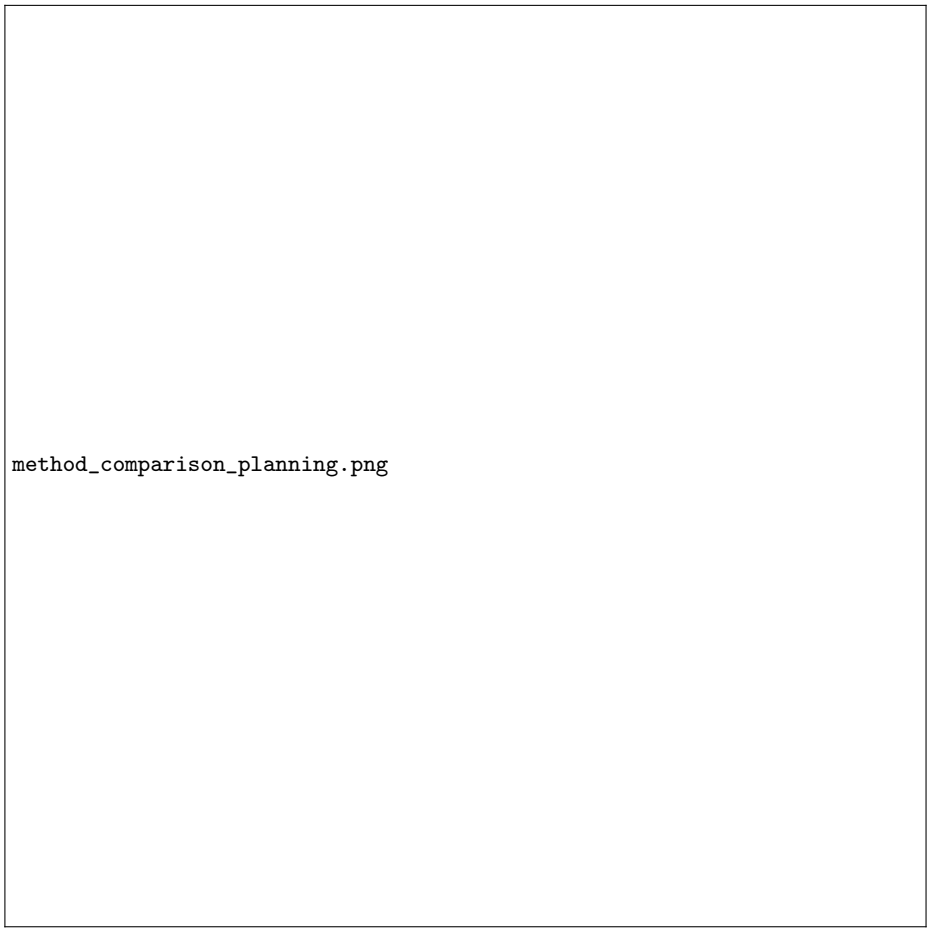

Fig. 1. Planning Performance Comparison Across Methods and Time Horizons. The figure shows mAP performance for each method (DARIL, DARIL+IRL, Latent World Model+RL, Direct Video RL) across different planning horizons (1s, 2s, 3s, 5s, 10s, 20s). DARIL demonstrates superior performance across all time horizons, with RL approaches showing varying degrees of performance degradation. Error bars indicate 95% confidence intervals across video splits.

Fig. 2. 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 33.6% at 1s to 29.2% at 10s, with Target being the most robust component (23.7% loss). Stars indicate statistical significance regions.

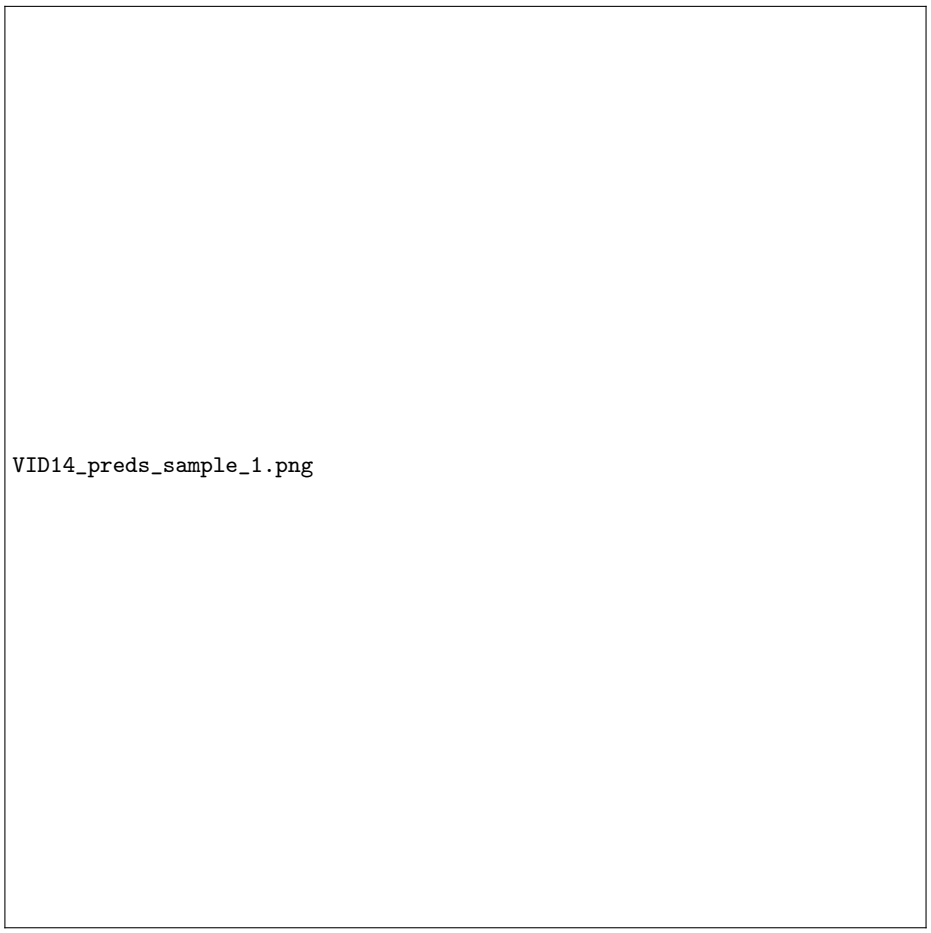

Fig. 3. Qualitative evaluation showing recognition and planning performance from a representative video sequence. VID51 sample demonstrating action sequence planning. Left panels show past recognition performance, right panels show future planning predictions. The model demonstrates strong current action recognition with smooth degradation in planning accuracy over time. Green indicates true positives, blue shows false positives, and beige represents false negatives.